On Bayesian index policies for sequential resource allocation

Emilie Kaufmann, CNRS & CRIStAL (Université de Lille)

7 janvier 2016

Résumé

This paper is about index policies for minimizing (frequentist) regret in a stochastic multi-armed bandit model, that are inspired by a Bayesian view on the problem. Our main contribution is to prove the asymptotic optimality of Bayes-UCB, an algorithm based on quantiles of posterior distributions, when the rewards distributions belong to a one-dimensional exponential family, for a large class of prior distributions. We also show that the Bayesian literature gives new insight on what kind of exploration rates could be used in frequentist, UCB-type algorithms. Indeed, approximations of the Bayesian optimal solution or the Finite Horizon Gittins indices suggest the introduction of two algorithms, KL-UCB$^+$ and KL-UCB-H$^+$, whose asymptotic optimality is also established.

1 Introduction

This paper present new analyses of Bayesian-flavored strategies for sequential resource allocation in an unknown, stochastic environment modeled as a multi-armed bandit. A *stochastic multi-armed bandit model* is a set of K probability distributions, $\nu_1, \dots, \nu_K$, called arms, with which an agent interacts in a sequential way. At round t the agent, who does not know the arms' distributions, chooses an arm A_t. The draw of this arm produces an independent sample X_t from the associated probability distribution ν_{A_t}, often interpreted as a reward. The arms can indeed be viewed as those of different slot machines, also called *one-armed bandit*, generating rewards according to some underlying probability distribution.

In several applications, that range from the motivating example of clinical trials [31] to the more modern motivation of online advertisement (e.g., [14]), the goal of the agent is to adjust his strategy $\mathcal{A} = (A_t)_{t\in\mathbb{N}}$ in order to maximize the rewards accumulated during his interaction with the bandit model. Note that the strategy of the agent, also called *bandit algorithm*, is sequential in the sense that the arm A_t is chosen based on the previous observations $A_1, X_1, \dots, A_{t-1}, X_{t-1}$. More precisely, the goal is to design a sequential strategy maximizing the expectation of the sum of rewards up to some horizon T. If $\mu_1, \dots, \mu_K$ denote the means of the arms, and $\mu^* = \max_a \mu_a$, this is equivalent to minimize the *regret*, defined as the expected difference between the reward accumulated up to time T by the strategy that knows the arm with highest mean and always plays it, and the reward accumulated by a strategy $\mathcal{A}$:

$$\mathrm{R}(T, \mathcal{A}) := \mathbb{E}\left[T\mu^* - \sum_{t=1}^{T} X_t\right] = \mathbb{E}\left[\sum_{t=1}^{T}(\mu^* - \mu_{A_t})\right]. \tag{1}$$

This paper focuses on good strategies in *parametric* bandit models, in which the distribution of arm a depends on some parameter θ_a : he write $\nu_a = \nu_{\theta_a}$. Just like in every parametric model, two different views can be adopted. In the frequentist view, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ is an unknown parameter whereas in the Bayesian view $\boldsymbol{\theta}$ is a random variable, drawn from a prior distribution Π. The expectation in (1) can thus

be taken under any of this two probabilistic models, leading in the first setting to the notion of frequentist regret, that depends on the parameter $\boldsymbol{\theta}$:

$$\mathrm{R}_{\boldsymbol{\theta}}(T, \mathcal{A}) := \mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{t=1}^{T}(\mu^* - \mu_{A_t})\right] = \sum_{a=1}^{K}(\mu^* - \mu_a)\mathbb{E}_{\boldsymbol{\theta}}[N_a(T)], \tag{2}$$

where $N_a(T) = \sum_{t=1}^{T} \mathbb{1}_{(A_t=a)}$ is the number of times arm a has been drawn up to time T, and in the second case to the notion of Bayesian regret, that depends on the prior distribution Π :

$$\mathcal{R}_{\Pi}(T, \mathcal{A}) := \mathbb{E}^{\Pi}\left[\sum_{t=1}^{T}(\mu^* - \mu_{A_t})\right] = \int \mathrm{R}_{\boldsymbol{\theta}}(T, \mathcal{A}) d\Pi(\boldsymbol{\theta}), \tag{3}$$

where $\mathbb{E}^{\Pi}$ includes an average over $\boldsymbol{\theta} \sim \Pi$. Bayesian regret is sometimes referred to as *Bayes risk* in the literature, and we will use this terminology in the rest of the paper.

The first bandit strategy was introduced by Thompson in 1933 [31] in a Bayesian framework, and a large part of the early works on bandit models was adopting the same perspective [9, 7, 18, 8]. Indeed, as Bayes risk minimization has an *exact*—yet often intractable—solution, finding ways to efficiently compute this solution was an important line of research. This will be explained in more details in Section 3. Frequentist regret minimization has no exact solution, however since 1985 and the seminal work of Lai and Robbins [24], there is a precise characterization of good bandit algorithms in a frequentist sense. They show that for any *uniformly consistent policy* $\mathcal{A}$ (i.e. such that for all $\boldsymbol{\theta}$, $\mathcal{R}_{\boldsymbol{\theta}}(\mathcal{A}, T) = o(T^{\alpha})$ for all $\alpha \in]0, 1]$), the number of draws of any sub-optimal arm a ($\mu_a < \mu^*$) is asymptotically lower bounded as follows :

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[N_a(T)]}{\log T} \geq \frac{1}{\mathrm{KL}(\nu_{\theta_a}, \nu_{\theta^*})}, \tag{4}$$

where $\mathrm{KL}(\nu, \nu')$ denotes the Kullback-Leibler divergence between the distributions ν and ν'. From (2), this yield a lower bound on the regret.

This result holds for simple parametric bandit models, including exponential family bandit models presented in Section 2, that will be our main focus in this paper. It paved the way to a new line of research, trying to build *asymptotically optimal* strategies, that is strategies matching the lower bound (4), for some classes of distributions. Most of the algorithms proposed since then belong to the family of *index policies*, that compute at each round one index per arm, depending on the history of rewards observed from this arm only, and select the arm with largest index. More precisely, they are UCB-type algorithms, building confidence intervals for the means of the arms and choosing as an index for each arm the associated Upper Confidence Bound. The design of the confidence intervals has been successively improved [1, 6, 5, 4, 19, 12] so as to obtain simple index policies for which non-asymptotic upper bound on the regret can be given. Among them, the KL-UCB algorithm [12] matches the lower bound (4). As they use confidence intervals on unknown parameters, all these index policies are based on *frequentist tools*. However, it is interesting to note that the first index policy was introduced by Gittins in 1979 [18] to solve a Bayesian multi-armed bandit problem and is based on *Bayesian tools*, i.e. on exploiting the posterior distribution on the parameter of each arm.

Note that tools and objectives should be separated : one can compute the Bayes risk of an algorithm based on frequentist tools or the (frequentist) regret of an algorithm based on Bayesian tools. In this paper, we focus on the latter and advocate the use of index policies inspired by Bayesian tools for minimizing regret. Our main contribution is to prove the asymptotic optimality of the Bayes-UCB algorithm introduced by [20] for any exponential bandit models and for a large class of prior distributions. Our

analysis relies on two ingredients : tight bounds on the tail of posterior distributions, and a deviation inequality involving alternative exploration rate. This last tool also allow us to prove the asymptotic optimality of two variants of KL-UCB, called KL-UCB$^+$ and KL-UCB-H$^+$, that display improved empirical performance. Interestingly, the alternative exploration rate used by these two algorithms is already suggested by asymptotic approximations of the Bayesian exact solution or the Finite-Horizon Gittins indices.

Over the past few years, another Bayesian algorithm, Thompson Sampling, has become increasingly popular for its good empirical performance. This randomized algorithm, that draws each arm according to its posterior probability of being optimal, was introduced in 1933 as the very first bandit algorithm [31] but the first logarithmic upper bound on its regret dates back to 2012 [2]. Now, this strategy is know to be asymptotically optimal in exponential family bandit models, for specific choices of prior distributions [21, 3, 22]. Our experiments of Section 6 highlight that the index policy presented in this paper are also competitive with Thompson Sampling.

The paper is structured as follow. In Section 2, we introduce the exponential family bandit models that we consider in the rest of the paper, and the associated Bayesian tools. After recalling the notion of Bayesian optimal solution and Gittins indices in Section 3, we explain in Section 4 how they suggest a modification in the KL-UCB algorithm, leading to the KL-UCB$^+$ and the KL-UCB-H$^+$ algorithms. Section 5 is dedicated to our analysis of the Bayes-UCB algorithm, from which the asymptotic optimality of KL-UCB$^+$ and KL-UCB-H$^+$ also follows. In Section 6, we investigate numerically the performance of these three asymptotically optimal, Bayesian-flavored index policies.

Notation We denote by $(Y_{a,s})$ the sequence of successive rewards generated by arm a in the bandit model. Given a bandit algorithm, $N_a(t) = \sum_{k=1}^{t} \mathbb{1}_{(A_k=a)}$ is the number of draws of arm a up to round t. In particular, when arm A_t is chosen at round t, the observed reward X_t satisfies $X_t = Y_{a,N_a(t)}$. Letting $\hat{\mu}_{a,s} = \frac{1}{s}\sum_{k=1}^{s} Y_{a,k}$ be the empirical mean of the first s rewards from a, the empirical mean of arm a after t rounds of the bandit algorithm, $\hat{\mu}_a(t)$, satisfies

$$\hat{\mu}_a(t) = \hat{\mu}_{a,N_a(t)} = \frac{1}{N_a(t)} \sum_{s=1}^{N_a(t)} Y_{a,s}.$$

2 (Bayesian) exponential family bandit models

In the rest of the paper, we consider the important class of *exponential family bandit models*, in which the arms belong to a one-parameter canonical exponential family.

2.1 Exponential families

A one-parameter canonical exponential family is a set $\mathcal{P}$ of probability distributions, indexed by $\theta \in \mathbb{R}$ called the natural parameter, that is defined by

$$\mathcal{P} = \{\nu_\theta, \theta \in \Theta : \nu_\theta \text{ has a density } f_\theta(x) = \exp(\theta x - b(\theta)) \text{ w.r.t } \xi\},$$

where $\Theta =]\theta^-, \theta^+[$ is an open interval, b a twice-differentiable and convex function (called the log-partition function) and ξ a reference measure. Examples of such distributions are given in Table 1 below.

If $X \sim \nu_\theta$, it can be shown that $\mathbb{E}[X] = \dot{b}(\theta)$ and $\mathrm{Var}[X] = \ddot{b}(\theta) > 0$. Thus there is a one-to-one mapping between the natural parameter θ and the mean $\mu = \dot{b}(\theta)$, and distributions in an exponential

Distribution	Density	Mean μ	Parameter θ	$b(\theta)$
Bernoulli $\mathcal{B}(\lambda)$	$\lambda^x(1-\lambda)^{1-x}\mathbb{1}_{\{0,1\}}(x)$	λ	$\log\frac{\lambda}{1-\lambda}$	$\log(1+e^\theta)$
Poisson $\mathcal{P}(\lambda)$	$\frac{\lambda^x}{x!}e^{-\lambda}\mathbb{1}_{\mathbb{N}^*}(x)$	λ	$\log(\lambda)$	e^θ
Gaussian $\mathcal{N}(\lambda,\sigma^2)$ (σ^2 known)	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\lambda)^2}{2\sigma^2}}$	λ	$\frac{\lambda}{\sigma^2}$	$\frac{\sigma^2\theta^2}{2}$
Gamma $\Gamma(k,\lambda)$ (k known)	$\frac{\lambda^k}{\Gamma(k)}x^{k-1}e^{-\lambda x}\mathbb{1}_{\mathbb{R}^+}(x)$	k/λ	$-\lambda$	$-k\log(-\theta)$

TABLE 1 – Examples of exponential families and associated divergence.

family can be alternatively parametrized by their mean. Letting $\mathrm{J} := \dot{b}(\Theta)$, for $\mu \in \mathrm{J}$ we denote by ν^μ the distribution in $\mathcal{P}$ that has mean μ : $\nu^\mu = \nu_{\dot{b}^{-1}(\mu)}$. The variance $V(\mu)$ of the distribution ν^μ is related to its mean in the following way :

$$\mathrm{V}(\mu) = \ddot{b}(\dot{b}^{-1}(\mu)).$$

The Kullback-Leibler divergence between distributions in an exponential family has a closed form expression as a function of the natural parameters. Letting $\mathrm{K}(\theta,\lambda)$ be the Kullback-Leibler divergence between the distributions parameterized by θ and λ, one has

$$\mathrm{K}(\theta,\lambda) := \mathrm{KL}(\nu_\theta,\nu_\lambda) = \dot{b}(\theta)(\theta-\lambda) - b(\theta) + b(\lambda). \tag{5}$$

We also let $\mathrm{d}(\mu,\mu')$ be the KL-divergence between the distributions of means μ and μ' :

$$\mathrm{d}(\mu,\mu') := \mathrm{KL}(\nu^\mu,\nu^{\mu'}) = \mathrm{K}(\dot{b}^{-1}(\mu),\dot{b}^{-1}(\mu')).$$

Closed-form for d in the examples of exponential families given in Table 1 are available (see [12]), which allows to define the associated KL-UCB index :

$$u_a(t) = \sup\left\{q \in \mathrm{J} : N_a(t)d(\hat{\mu}_a(t),q) \le \log(t\log^c(t))\right\}.$$

The *exploration rate*, which is here $\log(t\log^c(t))$, controls the probability with which $u_a(t)$ is an upper bound on μ_a. Using that $y \mapsto d(x,y)$ is convex and non-decreasing when $y > x$, $u_a(t)$ can be easily (approximately) computed in practice, using dichotomic search for example.

2.2 Posterior distributions in exponential families

Assume that θ is drawn from a distribution on Θ that has density h with respect to the Lebesgue measure, and let (Y_s) be a sequence of observations i.i.d. conditionally to θ, with distribution ν_θ. The posterior distribution of θ given the first n observations has density

$$p(\theta|Y_1,\dots,Y_n) \propto \exp\left(n\left(\theta\hat{\mu}_n - b(\theta)\right)\right)h(\theta), \quad \text{with} \quad \hat{\mu}_n = \frac{1}{n}\sum_{s=1}^n Y_s.$$

This distribution depends on two sufficient statistics : the number of observations, n, and the empirical mean of observations, $\hat{\mu}_n$. Indeed, the posterior distribution of θ after n observations is $p_{n,\hat{\mu}_n}$, using the notation

$$p_{n,x}(\mathcal{I}) := \frac{\int_{\mathcal{I}}\exp(n(\theta x - b(\theta))h(\theta)d\theta}{\int_\Theta \exp(n(\theta x - b(\theta))h(\theta)d\theta} \quad \text{for} \quad \mathcal{I} \subset \Theta.$$

A prior distribution on θ with density h defined on Θ is equivalent to a prior distribution on the mean $\mu = \dot{b}(\theta)$ with density f defined on $\mathrm{J} = \dot{b}(\Theta)$, where f and h are related by

$$\forall \theta \in \Theta,\ h(\theta) = f(\dot{b}(\theta))\ddot{b}(\theta) \quad \Leftrightarrow \quad \forall u \in \mathrm{J},\ f(u)V(u) = h(\dot{b}^{-1}(u)).$$

The posterior distribution of μ given n observations is also a function of n and $\hat{\mu}_n$, and we denote by $\pi_{n,x}$ the posterior distribution on μ after n observations with empirical mean x :

$$\pi_{n,x}(\mathcal{J}) := \frac{\int_{\mathcal{J}} \exp(n(\dot{b}^{-1}(u)x - b(\dot{b}^{-1}(u))))f(u)du}{\int_{\mathrm{J}} \exp(n(\dot{b}^{-1}(u)x - b(\dot{b}^{-1}(u))))f(u)du} \quad \text{for} \quad \mathcal{J} \subset \mathrm{J}.$$

Putting a prior distribution on the mean is often more natural and there exists families of conjugated prior on the mean, given in Table 2.

Distribution	Prior distribution on μ	Posterior distribution on μ after n observations with empirical mean x
$\mathcal{B}(\mu)$	$\mathrm{Beta}(a,b)$	$\mathrm{Beta}(a+nx, b+n(1-x))$
$\mathcal{P}(\mu)$	$\Gamma(c,d)$	$\Gamma(c+nx, d+n)$
$\mathcal{N}(\mu,\sigma^2)$	$\mathcal{N}(\mu_0, m_0^{-1})$	$\mathcal{N}\left(\frac{m_0\mu_0+nx\sigma^{-2}}{m_0+n\sigma^{-2}}, (m_0+n\sigma^{-2})^{-1}\right)$
$\Gamma(k,k/\mu)$	$\mathrm{Inv}\Gamma(c,d)$	$\mathrm{Inv}\Gamma(c+kn, d+knx)$

TABLE 2 – Conjugate prior on the mean and associated posterior distributions.

2.3 Exponential family bandit model

In the sequel, we fix an exponential family $\mathcal{P}$ and consider a bandit model $\nu^{\boldsymbol{\mu}} = (\nu^{\mu_1}, \dots, \nu^{\mu_K})$, where ν^{μ_a} belongs to $\mathcal{P}$ and has mean μ_a. We restrict our attention to Bayesian bandit models with a product prior on $\boldsymbol{\mu}$, such that $\mu_1, \dots, \mu_K$ are independent, and μ_a is drawn from a prior distribution on $\mathrm{J} = \dot{b}(\Theta)$ that has density f_a with respect to the Lebesgue measure. Inspired by the notation introduced in Section 2.2, we let $\pi_{a,n,x}$ be the posterior distribution on μ_a after n observations from this arm that have yield an empirical mean x. The posterior distribution on μ_a at the end of round t is therefore

$$\pi_a^t := \pi_{a,N_a(t),\hat{\mu}_a(t)}.$$

The following rewriting of $\pi_{a,n,x}$ will be very useful in the sequel to obtain tight bounds on the tail of the posterior distribution.

Lemma 1.

$$\pi_{a,n,x}(\mathcal{J}) = \frac{\int_{\mathcal{J}} \exp(-nd(x,u))f_a(u)du}{\int_{\mathrm{J}} \exp(-nd(x,u))f_a(u)du}, \ \textit{for all} \ \mathcal{J} \subset \mathrm{J}.$$

Proof Let $\mathcal{J} \subset \mathrm{J}$. One has

$$\begin{aligned}
\pi_{a,n,x}(\mathcal{J}) &= \frac{\int_{\mathcal{J}} \exp(n(\dot{b}^{-1}(u)x - b(\dot{b}^{-1}(u))))f_a(u)du}{\int_{\mathrm{J}} \exp(n(\dot{b}^{-1}(u)x - b(\dot{b}^{-1}(u))))f_a(u)du} \times \frac{e^{-n(x\dot{b}^{-1}(x)-b(\dot{b}^{-1}(x)))}}{e^{-n(x\dot{b}^{-1}(x)-b(\dot{b}^{-1}(x)))}} \\
&= \frac{\int_{\mathcal{J}} \exp(-n(x(\dot{b}^{-1}(x) - \dot{b}^{-1}(u)) - b(\dot{b}^{-1}(x)) + b(\dot{b}^{-1}(u)))f_a(u)du}{\int_{\mathrm{J}} \exp(-n(x(\dot{b}^{-1}(x) - \dot{b}^{-1}(u)) - b(\dot{b}^{-1}(x)) + b(\dot{b}^{-1}(u)))f_a(u)du} \\
&= \frac{\int_{\mathcal{J}} \exp(-nd(x,u))f_a(u)du}{\int_{\mathrm{J}} \exp(-nd(x,u))f_a(u)du},
\end{aligned}$$

using the closed form expression (5) and the fact that $\theta = \dot{b}^{-1}(\mu)$.

3 Bayesian optimal solution and Gittins indices

In a Bayesian framework, the interaction of an agent with a multi-armed bandit can be modeled by a Markov Decision Process (MDP), in which the state Π_t is the current posterior distribution over the parameter of the arms. In exponential bandit models, the posterior over $\boldsymbol{\mu}$ is $\Pi_t = \bigotimes \pi_a^t$. There are K possible actions and when action A_t is chosen in state Π_t, the observed reward X_t is a sample from arm A_t, that satisfies, conditionally to the past,

$$\left\{ \begin{array}{rcl} X_t & \sim & \nu^{\mu} \\ \mu & \sim & \Pi_t(A_t). \end{array} \right.$$

The new state is $\Pi^{t+1} = \bigotimes \pi_a^{t+1}$ with $\pi_a^{t+1} = \pi_a^t$ for all $a \neq A_t$ and the density of $\pi_{A_t}^{t+1}$ gets updated according to

$$d\pi_{A_t}^{t+1}(u) \propto \exp(-(\dot{b}^{-1}(u)X_t - b(\dot{b}^{-1}(u))))d\pi_{A_t}^{t}(u).$$

Bayes risk minimization, or reward maximization under the Bayesian probabilistic model, is equivalent to solving this MDP for the finite-horizon criterion, which boils down to finding a strategy of the form $A_t = g(\Pi_t)$ for some deterministic function g, that maximizes

$$\mathbb{E}^{\Pi}\left[\sum_{t=1}^{T} X_t^g\right], \tag{6}$$

where $(X_t^g)_t$ is the sequence of rewards obtained under policy g. From the theory of MDPs (see e.g., [28]), the optimal policy is solution of dynamic programming equations and can be computed by induction. However, due to the very large, if not infinite, state space (the set of possible posterior distributions over $\boldsymbol{\mu}$), the computation is often intractable. In a slightly different setting, Gittins proved in 1979 [18] that the apparently intractable optimal policy reduces to an index policy, with corresponding indices later called the *Gittins indices*. He considers the discounted Bayesian multi-armed bandit problem, in which the goal is to find a policy g that minimizes

$$\mathbb{E}^{\Pi}\left[\sum_{t=1}^{\infty} \alpha^{t-1} X_t^g\right],$$

for some discount parameter $\alpha \in]0,1[$. Interestingly, it was proved in [8] that the discount is necessary for this reduction to hold : in particular, the policy minimizing (6) is *not* an index policy.

However, the notion of Gittins indices is a powerful concept, that can also be defined in a finite horizon multi-armed bandit. The Finite-Horizon Gittins index of an arm depends on the current posterior distribution on its mean ($\pi = \pi_a^t$) and on the remaining time to play ($r = T - t$). It can be interpreted as the price worth paying for playing an arm with posterior π at most r times. Indeed, for $\lambda > 0$ consider the following game, called $\mathcal{C}_\lambda$, in which a player can either pay λ and draw the arm to receive a sample Y_t, which results in a reward $Y_t - \lambda$, or stop playing, which yields no reward. As precisely defined below, the Gittins index is the critical value of λ for which the optimal policy in $\mathcal{C}_\lambda$ is to stop playing the arm from the beginning.

Definition 2. *The Finite-Horizon Gittins index for a current posterior π and remaining time r is $G(\pi, r) = \inf\{\lambda \in \mathbb{R} : V_\lambda^*(\pi, r) = 0\}$, with*

$$V_\lambda^*(\pi, r) = \sup_{0 \leq \tau \leq r} \mathbb{E}_{\substack{Y_t \overset{i.i.d}{\sim} \nu^\mu \\ \mu \sim \pi}} \left[\sum_{t=1}^{\tau} (Y_t - \lambda) \right],$$

where the supremum is taken over all stopping time τ smaller than r a.s., with the convention $\sum_{t=1}^{0} \cdot = 0$.

Computing the FH-Gittins indices requires to compute $V_\lambda^*(\pi, r)$ for several values of λ in order to find the critical value (dichotomic search can be used). Each computation boils down to solving a MDP, but on a smaller state space : the possible posterior distributions on the mean of a single arm. Even if it is much easier than the computation of the optimal policy, computing the FH-Gittins indices can still be costly, and finding efficient methods to do it is still an area of investigation (see [27]). In the particular case of Bernoulli bandit models with a uniform prior over the mean, the set of (Beta) posterior is parametrized by two integers (number of zeros and ones observed so far), and for small horizons (up to $T = 1000$), it is possible to implement the index policy associated to the Finite-Horizon Gittins indices :

$$A_{t+1} = \operatorname*{argmax}_{a=1,\dots,K} G(\pi_a^t, T - t).$$

We refer to this index policy as the Finite-Horizon Gittins algorithm. Although this algorithm does not coincide with the Bayesian optimal solution, we believe it is a good approximation. This is supported by simulations performed in a two-armed Bernoulli bandit problem, for which we compute the Bayes risk of the optimal strategy and that of the FH-Gittins algorithm up to horizon $T = 70$, as presented in Figure 1. For small horizons, [17] propose a comparison of different algorithms with the Bayesian optimal solution and similarly notice that the Bayes-risk of FH-Gittins (called Λ-strategy) is very close to the optimal value, for various choices of prior and horizons.

More interestingly, the FH-Gittins algorithm also appears to perform well when evaluated with respect to the frequentist regret, as illustrated in experiments reported in Section 6 and we believe it is a good idea to employ this algorithm when it is efficiently implementable, that is for small horizons. The first theoretical elements to support this claim were recently obtained by [25], who present the first logarithmic upper bound on the regret of the Finite-Horizon Gittins algorithm, in the particular case of Gaussian bandit models. In this paper, we provide new theoretical guarantees for another Bayesian index policy, much easier to implement, Bayes-UCB. Before introducing this algorithm in Section 5, we present in the next section other index policies that can be related to the Finite-Horizon Gittins indices.

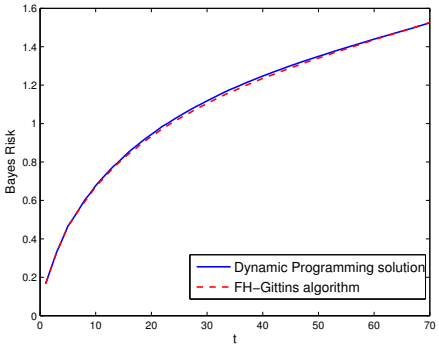

FIGURE 1 – Bayes risk of the optimal strategy (blue) and FH-Gittins (dashed red) estimated using $N = 10^6$ replications of a bandit game, for which the means are drawn from $\mathcal{U}([0,1])$

4 Algorithms inspired by the Bayesian optimal solution

4.1 An asymptotic lower bound on the Bayes risk

The Bayesian optimal solution is introduced in the previous section, but the order of magnitude of its Bayes risk is not given. In the paper [23], Lai shows that, in exponential family bandit models, the Bayes risk of *any* strategy is asymptotically lower bounded by $C_0(\pi)\log^2(T)$, when $C_0(\pi)$ is a prior-dependent constant. He also provides matching strategies, implying in particular that the Bayes risk of the Bayesian solution is of order $\log^2(T)$.

For product prior distributions, which is the particular case studied in this paper, Theorem 3 below provides a lower bound on the Bayes risk that is slightly more general than Lai's result in the sense that it does not require the prior distribution on the natural parameter of each arm to have a compact support. The proof of this result, provided in Appendix B, follows however closely that of [23]. Before stating Theorem 3, we introduce some useful notation. For $a = 1,\dots,K$, we let $\boldsymbol{\theta}_{-a} = (\theta_1,\dots,\theta_{a-1},\theta_{a+1},\dots,\theta_K)$ be the vector of Θ^{K-1} that consists of all components of $\boldsymbol{\theta}$ except component number a. We let $\theta_a^* = \max_{i\neq a}\theta_i$, so that θ_a^* only depends on $\boldsymbol{\theta}_{-a}$.

Theorem 3. *Let H be a prior distribution on Θ^K that has a product form, such that each marginal has a density h_a with respect to the Lebesgue measure λ that satisfies $h_a(\theta) > 0$ for all $\theta \in \Theta$. Letting H_{-a} the marginal distribution of $\boldsymbol{\theta}_{-a}$, that has density $\prod_{i\neq a} h_i(\theta_i)$ with respect to $\lambda^{\otimes K-1}$, one assume that*

$$\forall a = 1,\dots,K, \qquad \int_{\Theta^{K-1}} h_a(\theta_a^*)dH_{-a}(\boldsymbol{\theta}_{-a}) < \infty.$$

Under the prior distribution H, the Bayes risk of any strategy $\mathcal{A}$ satisfies

$$\liminf_{T\to\infty} \frac{\mathcal{R}^H(T,\mathcal{A})}{\log^2(T)} \geq \frac{1}{2}\sum_{a=1}^{K}\int_{\Theta^{K-1}} h_a(\theta_a^*)dH_{-a}(\boldsymbol{\theta}_{-a}).$$

For exponential family bandit models with a product prior, Lai provides the first (asymptotic) prior-dependent Bayes risk upper bounds, when Θ is compact. Letting $[\mu^-,\mu^+] = \dot{b}(\Theta)$, he shows in particular that the index policy associated to

$$I_a(t) = \sup\left\{q \in [\mu^-,\mu^+] : N_a(t)\overline{d}(\hat{\mu}_a(t),q) \leq \log\left(\frac{T}{N_a(t)}\right)\right\}, \tag{7}$$

where $\overline{d}(x,y) = d\left(\max(\mu^-, \min(\mu^+, x)), y\right)$, has a Bayes risk that asymptotically matches the lower bound of Theorem 3. This index policy is reminiscent of KL-UCB. Beyond the fact that a regularized version of the divergence function d is used, the main difference with KL-UCB is the use of $\log(T/N_a(t))$ as an exploration rate in place of $\log t$.

While a recent line of research on Bayesian randomized algorithms (e.g. Thompson Sampling) has provided Bayes risk upper bound in quite general settings ([30, 29]), to the best of our knowledge, no upper bound scaling in $\log^2(T)$ has been obtained for exponential family bandit models since the work of Lai. [10, 26] provide the first prior-dependent upper bounds on the Bayes-risk of Thompson Sampling, in a particular case quite different from our setting : a two-armed bandit model in which the means of the arms are known up to a permutation and the joint prior distribution is thus supported on (μ_1, μ_2) and (μ_2, μ_1). In Section 6.2, we investigate numerically the optimality of the Bayesian index policies discussed in the rest of the paper with respect to the lower bound of Theorem 3.

4.2 Approximating the Gittins indices

As discussed in the previous section, the FH-Gittins strategy, that is the index policy associated to

$$J_a(t) = G(\pi_a^t, T-t),$$

is conjectured to be a good approximation of the Bayesian optimal policy, yet the above indices remain difficult to compute. Building on approximations of the Finite-Horizon Gittins indices that can be extracted from the literature permits to obtain a related *efficient* index policy.

Recall from Definition 2 that the Finite-Horizon Gittins index takes the form

$$G(\pi, r) = \inf\left\{\lambda \in \mathbb{R} : V_\lambda^*(\pi, r) = 0\right\},$$

where $V_\lambda^*(\pi, r)$ correspond to the optimal value function associated a calibration game $\mathcal{C}_\lambda$. In the paper [11], Burnetas and Katehakis propose tight bounds on the value function $V_\lambda^*(\pi_{a,n,x}, r)$ for exponential family bandits. These bounds permit to derive asymptotic approximations of the FH-Gittins indices, when r is large, and to show that, for large values of the remaining time $T-t$,

$$J_a(t) \simeq \sup\left\{q \in [\mu^-, \mu^+] : N_a(t)\tilde{d}(\hat{\mu}_a(t), q) \leq \log\left(\frac{T-t}{N_a(t)}\right)\right\}. \tag{8}$$

This approximation is valid under the assumption that Θ is compact : $[\mu^-, \mu^+] = \dot{b}(\Theta)$ and $\tilde{d}$ is another regularization of the divergence function d, such that, for any y, $\tilde{d}(x,y) = d(x,y)$ for $x > \mu^-$ and for $x \leq \mu^-$,

$$\tilde{d}(x,y) = d(\mu^-, y) + (\dot{b}^{-1}(y) - \dot{b}^{-1}(\mu^-))(\mu^- - x).$$

In the particular case of Gaussian bandit models, the work of Chang and Lai ([13]) on the approximation of discounted Gittins indices can also be adapted to obtain approximations of the Finite-Horizon Gittins indices, showing the same tendency as in (8) : compared to the corresponding KL-UCB index, here the $\log t$ is replaced by $\log((T-t)/N_a(t))$. This alternative exploration rate also appears in the non-asymptotic lower bound on the Gaussian Gittins index obtained by [25].

4.3 Alternative exploration rates

The indices (7) and (8) obtained above reveal that asymptotic approximations of the Bayesian optimal policy or the Finite-Horizon Gittins indices both suggest the use of an alternative exploration rate in the KL-UCB algorithm, in which the $\log(t)$ is replaced by a quantity that decreases when then number of draws of the arm $N_a(t)$ increases.

Interestingly, the use of alternative exploration rates in UCB-type algorithms has appeared before in the bandit literature. For example the MOSS algorithm [4], associated to the index

$$\hat{\mu}_a(t) + \sqrt{\frac{\log\left(T/(KN_a(t))\right)}{N_a(t)}}$$

is designed to be optimal in a minimax sense for bandit models with sub-gaussian rewards : the algorithm achieves a $O(\sqrt{KT})$ distribution-independent upper bound on the regret. Besides, it was noticed by [16] that the use of the exploration rate $\log(t/N_a(t))$ in place of $\log(t)$ in the KL-UCB algorithm leads to better empirical performance.

We formally define below two index policies that are variants of KL-UCB. KL-UCB-H$^+$ requires the knowledge of the horizon T, and is suggested by approximations of the Bayesian optimal strategy or the Gittins indices, whereas KL-UCB$^+$ is an anytime version of this strategy, that was already shown to perform well in practice. In the next section we will see that the analysis of the Bayes-UCB algorithm also give us the asymptotic optimality of these two algorithms, and their practical improvement over KL-UCB will be illustrated in Section 6.

Definition 4. *Let $c \geq 0$. We define KL-UCB$^+$ and KL-UCB-H$^+$ with parameter c as the index policies respectively associated to the indices*

$$u_a^+(t) = \sup\left\{q \geq \hat{\mu}_a(t) : N_a(t)d(\hat{\mu}_a(t), q) \leq \log\left(\frac{t\log^c t}{N_a(t)}\right)\right\}, \tag{9}$$

$$u_a^{H,+}(t) = \sup\left\{q \geq \hat{\mu}_a(t) : N_a(t)d(\hat{\mu}_a(t), q) \leq \log\left(\frac{T\log^c T}{N_a(t)}\right)\right\}. \tag{10}$$

5 Bayes-UCB : a simple and optimal Bayesian index policy

5.1 Algorithm and main result

The Bayes-UCB algorithm is an index policy that was introduced by [20] in the context of parametric bandit models. Given a prior distribution on the parameters of the arms, the index used for each arm is a well-chosen quantile of the (marginal) posterior distributions of its mean. In the particular case of exponential family bandit models, given a product prior distribution on the means $\pi^0 = \pi_1^0 \otimes \cdots \otimes \pi_K^0$, the Bayes-UCB index is

$$q_a(t) = Q\left(1 - \frac{1}{t(\log t)^c}; \pi_a^t\right) = Q\left(1 - \frac{1}{t(\log t)^c}; \pi_{N_a(t),\hat{\mu}_a(t)}\right),$$

where $Q(\alpha;\pi)$ is the quantile of order α of the distribution π (that is, $\mathbb{P}_{X\sim\pi}(X \leq Q(\alpha;\pi)) = \alpha$) and c is a real parameter.

While the efficiency of Bayes-UCB has been demonstrated even beyond bandit models with independent arms, the only available regret bound holds for Bernoulli bandit models when a uniform prior

distribution on the mean of each arm is used. In this section, we provide new regret bounds for general exponential family bandit models, showing that a slight variant of Bayes-UCB is asymptotically optimal for a large class of prior distributions.

We fix an exponential family, characterized by its log-partition function b and the interval $\Theta =]\theta^-, \theta^+[$ of possible natural parameters. We analyze Bayes-UCB for exponential bandit models satisfying the following.

Assumption 5. *There exists $\mu_0^- > \dot{b}(\theta^-)$ and $\mu_0^+ < \dot{b}(\theta^+)$ such that*

$$\forall a \in \{1, \dots, K\}, \quad \mu_0^- \leq \mu_a \leq \mu_0^+.$$

For the exponential families of Table 1, this assumption requires that the means of all arms are inside a compact interval that does not contain zero, and neither zero nor one in the Bernoulli case (i.e. there exists $\alpha > 0$ such that $\mu_a \in [\alpha, 1-\alpha]$ for all a). We now introduce a regularized version of the Bayes-UCB index, that relies on the knowledge of μ_0^- and μ_0^+, as

$$\overline{q}_a(t) = Q\left(1 - \frac{1}{t(\log t)^c}; p_{N_a(t), \bar{\mu}_a(t)}\right), \tag{11}$$

where $\bar{\mu}_a(t) = \min\left(\max(\hat{\mu}_a(t), \mu_0^-), \mu_0^+\right)$.

Theorem 6. *Let ν^{μ} be an exponential bandit model satisfying Assumption 5. Assume that for all a, π_a^0 has a density f_a with respect to the Lebesgue measure such that $f_a(u) > 0$ for all $u \in \mathrm{J} = \dot{b}(\Theta)$. Let $\varepsilon > 0$. The algorithm that draws each arm once and for $t \geq K$ selects at time $t+1$*

$$A_{t+1} = \operatorname*{argmax}_a \overline{q}_a(t),$$

with $\overline{q}_a(t)$ defined in (11), satisfies

$$\forall a \neq a^*, \quad \mathbb{E}[N_a(T)] \leq \frac{1+\varepsilon}{d(\mu_a, \mu^*)} \log(T) + o_\varepsilon\left(\log(T)\right).$$

5.2 Tail bounds for posterior distributions

Just like the analysis of [20], the analysis of Bayes-UCB that we give in the next section relies on tight bounds on the tails of posterior distributions, that permit to control quantiles. These bounds are expressed with the Kullback-Leibler divergence function d. Therefore, an additional tool in the proof is the control of the deviations of the empirical mean rewards from the true mean reward, measured with this divergence function, which follows from the work of [12].

In the particular case of Bernoulli bandit models, Bayes-UCB uses quantiles of Beta posterior distributions, and a specific argument, namely the fact that $\mathrm{Beta}(a, b)$ is the distribution of the a-th order statistic among $a + b - 1$ uniform random variables, permits to relate a Beta distribution (and its tails) to a Binomial distribution (and its tails). This 'Beta-Binomial trick' is also used extensively in the analysis of Thompson Sampling for Bernoulli bandits proposed by [2, 21, 3]. Note that this argument can only be used for Beta distribution with integer parameters, which rules out many possible prior distributions. Using specific arguments, it would also be possible to derive posterior bounds for Gaussian bandit models, using known tails bounds for the Gaussian posterior distribution (see Theorem 1.2.3. of [15]). For

exponential family bandit models, an upper bound on the tail of the posterior distribution was obtained by [22] when the Jeffrey's prior is used.

Lemma 7 below present more general results, that hold for any class of exponential family bandit models and any prior distribution with a density that is positive on $\mathrm{J} = \dot{b}(\Theta)$. For such (proper) prior distributions, we give deterministic upper and lower bounds on the corresponding posterior probabilities $\pi_{a,n,x}([v,\mu^+[)$. Compared to the result of [22], which is not presented in this deterministic way, Lemma 7 is based on a different rewriting of the posterior distribution, given in Lemma 1.

Lemma 7. *Let μ_0^-, μ_0^+ be such that $\dot{b}(\mu_0^-) > \theta^-$ and $\dot{b}(\mu_0^+) < \theta^+$.*

1. *There exists two positive constants A and B such that for all x, v that satisfy $\mu_0^- < x < v < \mu_0^+$, for all $n \geq 1$, for all $a \in \{1, \dots, K\}$,*

$$An^{-1}e^{-nd(x,v)} \leq \pi_{a,n,x}([v,\mu^+[) \leq B\sqrt{n}e^{-nd(x,v)}.$$

2. *There exists a constant C such that for all x, v that satisfy $\mu_0^- < v \leq x < \mu_0^+$, for all $n \geq 1$, for all $a \in \{1, \dots, K\}$,*

$$\pi_{a,n,x}([v,\mu^+[) \geq \frac{1}{C\sqrt{n}+1}.$$

The constants A, B, C depend on μ_0^-, μ_0^+, b and the prior densities.

5.3 A finite-time analysis

We give here the proof of Theorem 6. To ease the notation, assume that arm 1 is an optimal arm, and let a be a suboptimal arm.

$$\mathbb{E}[N_a(T)] = \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbb{1}_{(A_{t+1}=a)}\right] = 1 + \mathbb{E}\left[\sum_{t=K}^{T-1} \mathbb{1}_{(A_{t+1}=a)}\right].$$

We introduce a truncated version of the KL-divergence, $d^+(x,y) := d(x,y)\mathbb{1}_{(x<y)}$ and let g_t be a decreasing sequence, that will be specified later.

Using that, by definition of the algorithm, if a is played at round $t+1$, it holds in particular that $\overline{q}_a(t) \geq \overline{q}_1(t)$, one has

$$\begin{aligned}(A_{t+1} = a) &\subseteq (\mu_1 - g_t \geq \overline{q}_1(t)) \bigcup (\mu_1 - g_t \leq \bar{q}_1(t), A_{t+1} = a) \\ &\subseteq (\mu_1 - g_t \geq \overline{q}_1(t)) \bigcup (\mu_1 - g_t \leq \bar{q}_a(t), A_{t+1} = a)\,.\end{aligned}$$

This yields

$$\mathbb{E}[N_a(T)] \leq 1 + \sum_{t=K}^{T-1} \mathbb{P}\left(\mu_1 - g_t \geq \bar{q}_1(t)\right) + \sum_{t=K}^{T-1} \mathbb{P}\left(\mu_1 - g_t \leq \bar{q}_a(t), A_{t+1} = a\right).$$

The posterior bounds established in Lemma 7 permit to further upper bound the two sums in the right-hand side of the above inequality. With C defined in Lemma 7, we introduce t_0, defined by

$$t \geq t_0 \;\Rightarrow\; \left(\mu_1 - g_t \geq \mu_0^- \text{ and } (t\log(t)^c - 1)^2 \geq C^2 t\right).$$

On the one hand, for $t \geq t_0$,

$$\begin{aligned}(\mu_1 - g_t \geq \bar{q}_1(t)) &= \left(\pi_{1,N_1(t),\bar{\mu}_1(t)}([\mu_1 - g_t, \mu^+[) \leq \frac{1}{t\log^c t}\right) \\ &= \left(\pi_{1,N_1(t),\bar{\mu}_1(t)}([\mu_1 - g_t, \mu^+[) \leq \frac{1}{t\log^c t}, \bar{\mu}_1(t) \leq \mu_1 - g_t\right),\end{aligned}$$

since by the lower bound in the second statement of Lemma 7,

$$\begin{aligned}&\left(\pi_{1,N_1(t),\bar{\mu}_1(t)}([\mu_1 - g_t, \mu^+[) \leq \frac{1}{t\log^c t}, \bar{\mu}_1(t) \geq \mu_1 - g_t\right) \\ &\quad\subset \left(\frac{1}{C\sqrt{N_1(t)}+1} \leq \frac{1}{t\log^c t}\right) \subset \left(N_1(t) \geq \left(\frac{t\log^c t - 1}{C}\right)^2\right) \\ &\quad\subset (N_1(t) \geq t) = \varnothing.\end{aligned}$$

Now using the lower bound in the first statement of Lemma 7,

$$\begin{aligned}(\mu_1 - g_t \geq \bar{q}_1(t)) &\subseteq \left(\frac{Ae^{-N_1(t)d(\bar{\mu}_1(t),\mu_1-g_t)}}{N_1(t)} \leq \frac{1}{t\log^c t}, \bar{\mu}_1(t) \leq \mu_1 - g_t\right) \\ &\subset \left(N_1(t)d^+(\hat{\mu}_1(t), \mu_1 - g_t) \geq \log\left(\frac{At\log^c t}{N_1(t)}\right)\right).\end{aligned}$$

On the other hand,

$$\begin{aligned}&\sum_{t=K}^{T-1} \mathbb{P}\left(\mu_1 - g_t \leq \bar{q}_a(t), A_{t+1} = a\right) \\ &= \sum_{t=K}^{T-1} \mathbb{P}\left(\pi_{a,N_a(t),\bar{\mu}_a(t)}([\mu_1 - g_t, \mu^+[) \geq \frac{1}{t\log^c t}, A_{t+1} = a\right) \\ &\leq \sum_{t=K}^{T-1} \mathbb{P}\left(\bar{\mu}_a(t) < \mu_1 - g_t, \pi_{a,N_a(t),\bar{\mu}_a(t)}([\mu_1 - g_t, \mu^+[) \geq \frac{1}{t\log^c t}, A_{t+1} = a\right) \\ &\qquad + \sum_{t=K}^{T-1} \mathbb{P}\left(\bar{\mu}_a(t) \geq \mu_1 - g_t, A_{t+1} = a\right).\end{aligned} \tag{12}$$

Using Lemma 7, the first sum in (12) is upper bounded by

$$\begin{aligned}&\sum_{t=K}^{T-1} \mathbb{P}\left(B\sqrt{N_a(t)}e^{-N_a(t)d^+(\bar{\mu}_a(t),\mu_1-g_t)} \geq \frac{1}{t\log^c t}, A_{t+1} = a\right) \\ &\leq \sum_{t=K}^{T-1}\sum_{s=1}^{t} \mathbb{P}\left(B\sqrt{s}e^{-sd^+(\bar{\mu}_{a,s},\mu_1-g_t)} \geq \frac{1}{t\log^c t}, N_a(t) = s, A_{t+1} = a\right) \\ &\leq \sum_{t=K}^{T-1}\sum_{s=1}^{t} \mathbb{P}\left(sd^+(\bar{\mu}_{a,s}, \mu_1 - g_s) \leq \log(T\log^c T) + \log(B) + \frac{1}{2}\log s, N_a(t) = s, A_{t+1} = a\right) \\ &\leq \sum_{s=1}^{T} \mathbb{P}\left(sd^+(\bar{\mu}_{a,s}, \mu_1 - g_s) \leq \log T + c\log\log T + \log(B) + \frac{1}{2}\log s\right) \\ &\leq \sum_{s=1}^{T} \mathbb{P}\left(sd^+(\hat{\mu}_{a,s}, \mu_1 - g_s) \leq \log T + c\log\log T + \log(B) + \frac{1}{2}\log s\right) + \sum_{s=1}^{T} \mathbb{P}(\hat{\mu}_{a,s} < \mu_0^-).\end{aligned}$$

And by Chernoff inequality,

$$\sum_{s=1}^{T} \mathbb{P}(\hat{\mu}_{a,s} < \mu_0^-) \le \sum_{s=1}^{\infty} \exp(-sd(\mu_0^-, \mu_a)) = \frac{1}{1 - e^{-d(\mu_0^-, \mu_a)}}.$$

Still using Chernoff inequality, the second sum in (12) is upper bounded by

$$\begin{aligned}
&\sum_{t=K}^{T-1} \mathbb{P}\left(\hat{\mu}_a(t) \ge \mu_1 - g_t, A_{t+1} = a\right) \le \sum_{t=K}^{T-1} \mathbb{P}\left(\hat{\mu}_a(t) \ge \mu_1 - g_{N_a(t)}, A_{t+1} = a\right) \\
&\le \sum_{t=K}^{T-1} \sum_{s=1}^{t} \mathbb{P}\left(\hat{\mu}_{a,s} \ge \mu_1 - g_s, N_a(t) = s, A_{t+1} = a\right) \\
&\le \sum_{s=1}^{T} \mathbb{P}\left(\hat{\mu}_{a,s} \ge \mu_1 - g_s\right) \le \sum_{s=1}^{\infty} \exp(-sd(\mu_1 - g_s, \mu_a)) := N_0 < +\infty.
\end{aligned}$$

Putting things together, we proved that, if $N := \max(t_0, N_0 + (1 - e^{-d(\mu_0^-, \mu_a)})^{-1}) + 1$, one has

$$\begin{aligned}
\mathbb{E}[N_a(T)] \le N + \underbrace{\sum_{t=K}^{T-1} \mathbb{P}\left(N_1(t) d^+(\hat{\mu}_1(t), \mu_1 - g_t) \ge \log\left(\frac{At\log^c t}{N_1(t)}\right)\right)}_{T_1} \\
+ \underbrace{\sum_{s=1}^{T} \mathbb{P}\left(sd^+(\hat{\mu}_{a,s}, \mu_1 - g_s) \le \log T + c\log\log T + \log(B) + \frac{1}{2}\log s\right)}_{T_2}
\end{aligned}$$

Term T1 is shown below to be of order $o(\log(T))$, as $\hat{\mu}_1(t)$ cannot be too far from $\mu_1 - g_t$. Note however that the deviation is expressed with $\log(t/N_1(t))$ in place of the traditional $\log(t)$, which makes the proof of Lemma 8 more intricate. In particular, Lemma 8 applies to a specific sequence (g_t) defined therein, and a similar result could not be obtained for the choice $g_t = 0$, unlike Lemma 9 below.

Lemma 8. *Let g_t be such that $d(\mu_1 - g_t, \mu_1) = \frac{1}{\log(t)}$. If $c \ge 7$, for all A, if $t \ge \exp(\max(\sqrt{3}, A^{-1/7}))$,*

$$\mathbb{P}\left(N_1(t) d^+(\hat{\mu}_1(t), \mu_1 - g_t) \ge \log\frac{At\log^c t}{N_1(t)}\right) \le e\left(\frac{1}{At\log t} + \frac{3\log\log t + \log A}{At\log^2 t} + \frac{1}{At\log^3 t}\right) + \frac{1}{t^2}.$$

From Lemma 8, one has

$$\begin{aligned}
(T_1) &\le e\sum_{t=K}^{T-1} \frac{\log^2 t + 3(\log t)\log\log(t) + \log A\log t + 1}{At(\log^3 t)} + \sum_{t=K}^{T-1} \frac{1}{t^2} \\
&\le \frac{e}{A}\left(2 + \frac{3}{e} + \frac{\log A}{\log K}\right)\sum_{t=K}^{T-1} \frac{1}{t\log(t)} + \frac{\pi^2}{6} \\
&\le \frac{e}{A}\left(2 + \frac{3}{e} + \frac{\log A}{\log K}\right)\log\log T + \frac{\pi^2}{6}.
\end{aligned}$$

The following lemma permits to give an upper bound on Term T2.

Lemma 9. *Let f, g, h be three functions such that*

$$f(s) \underset{s\to\infty}{\longrightarrow} \infty, \quad g(s) \underset{s\to\infty}{\longrightarrow} 0 \quad \text{and} \quad \frac{h(s)}{s} \underset{s\to\infty}{\longrightarrow} 0,$$

with g and $s \mapsto h(s)/s$ non-increasing for s large enough. For all $\varepsilon > 0$ there exists a (problem-dependent) constant $N_a(\varepsilon)$ such that for all $T \geq N_a(\varepsilon)$,

$$\sum_{s=1}^{T} \mathbb{P}\left(sd^+(\hat{\mu}_{a,s}, \mu_1 - g(s)) \leq f(T) + h(s)\right) \leq \frac{1+\varepsilon}{d(\mu_a,\mu_1)} f(T) + \sqrt{f(T)}\sqrt{\frac{8V_a^2\pi(1+\varepsilon)^3 d'(\mu_a,\mu_1)^2}{d(\mu_a,\mu_1)^3}} + 8(1+\varepsilon)^2 V_a^2 \left(\frac{d'(\mu_a,\mu_1)}{d(\mu_a,\mu_1)}\right)^2 \frac{1}{1-e^{-d(\mu_0^-,\mu_a)}} + 1,$$

with $V_a = \sup_{\mu\in[\mu_a,\mu_1]} V(\mu)$.

Let $\varepsilon > 0$. Using Lemma 9, with $f(s) = \log(s) + c\log\log(s) + \log(B)$, $g(s) = g_s$ defined in Lemma 8 and $h(s) = \frac{1}{2}\log(s)$, there exists problem dependent constants C and $D(\varepsilon)$ such that

$$(T_2) \leq \frac{1+\varepsilon}{d(\mu_a,\mu_1)}(\log T + c\log\log T) + C\sqrt{\log T + c\log\log T} + D(\varepsilon).$$

Putting together the upper bounds on (T1) and (T2) yields, for all $\varepsilon > 0$,

$$\mathbb{E}[N_a(T)] \leq \frac{1+\varepsilon}{d(\mu_a,\mu^*)}\log(T) + O_\varepsilon(\sqrt{\log(T)}),$$

which concludes the proof.

□

5.4 Asymptotic optimality of KL-UCB$^+$ and KL-UCB-H$^+$

A key step in the analysis of Bayes-UCB is the control of the probability of the event

$$\left(N_1(t)d^+(\hat{\mu}_1(t), \mu_1 - g_t) \geq \log\frac{At\log^c t}{N_1(t)}\right),$$

in which the exploration rate $\log(t/N_1(t))$ appears. This control is obtained in Lemma 8 which can also be used to analyze the KL-UCB$^+$ and KL-UCB-H$^+$ algorithms, introduced in Definition 4, that are based on such alternative exploration rates. The following theorem proves the asymptotic optimality of these two index policies, that are respectively associated to the indices $u_a^+(t)$ and $u_a^{H,+}(t)$ defined in (9) and (10), that depend on a parameter c.

Theorem 10. *Let $c \geq 7$. Each of the index policy associated to the indices defined by (9) and (10) satisfies, for all $\varepsilon > 0$,*

$$\mathbb{E}[N_a(T)] \leq \frac{1+\varepsilon}{d(\mu_a,\mu^*)}\log(T) + O_\varepsilon(\sqrt{\log(T)}).$$

Proof of Theorem 10 We first give an analysis of the index policy associated to $u_a^+(t)$. Introducing g_t defined by $d(\mu_1 - g_t, \mu_1) = \frac{1}{\log(t)}$, one can write a decomposition similar to that used in the proof of Theorem 6 :

$$\begin{aligned}
\mathbb{E}[N_a(T)] &\leq 1 + \sum_{t=K}^{T-1} \mathbb{P}\left(\mu_1 - g_t \geq u_1^+(t)\right) \\
&\quad + \sum_{t=K}^{T-1} \mathbb{P}(\mu_1 - g_t \leq u_a^+(t), A_{t+1} = a) \\
&\leq 1 + \sum_{t=K}^{T-1} \mathbb{P}\left(N_1(t) d^+(\hat{\mu}_1(t), \mu_1 - g_t) \geq \log\left(\frac{t\log^c(t)}{N_a(t)}\right)\right) \qquad (13) \\
&\quad + \sum_{t=K}^{T-1} \mathbb{P}\left(N_a(t) d^+(\hat{\mu}_a(t), \mu_1 - g_t) \leq \log\left(T\log^c(T)\right), A_{t+1} = a\right), \qquad (14)
\end{aligned}$$

using the definition of $u_a^+(t)$ and the fact that $t\log^c t / N_a(t) \leq T\log^c T$. Lemma 8 can be applied (with $A = 1$) to show that the sum in (13) is of order $o(\log(T))$, while the sum in (14) can be rewritten and upper bounded using Lemma 9 : for all $\varepsilon > 0$, the result follows from

$$\begin{aligned}
&\mathbb{E}\sum_{t=K}^{T-1}\sum_{s=1}^{t} \mathbb{1}_{(sd^+(\hat{\mu}_{a,s}, \mu_1 - g_t) \leq \log(T\log^c T))} \mathbb{1}_{(A_{t+1}=a, N_a(t)=s)} \\
&\leq \sum_{s=1}^{T-1} \mathbb{P}\left(sd^+(\hat{\mu}_{a,s}, \mu_1 - g_t) \leq \log\left(T\log^c(T)\right)\right) \leq \frac{1+\varepsilon}{d(\mu_a, \mu_1)} \log(T\log^c(T)) + o_\varepsilon(\log(T)).
\end{aligned}$$

For the index policy associated to $u_a^{H,+}(t)$, using a similar decomposition,

$$\begin{aligned}
\mathbb{E}[N_a(T)] &\leq 1 + \sum_{t=K}^{T-1} \mathbb{P}\left(\mu_1 - g_t \geq u_1^{H,+}(t)\right) \\
&\quad + \sum_{t=K}^{T-1} \mathbb{P}(\mu_1 - g_t \leq u_a^{H,+}(t), A_{t+1} = a) \\
&\leq 1 + \sum_{t=K}^{T-1} \mathbb{P}\left(N_1(t) d^+(\hat{\mu}_1(t), \mu_1 - g_t) \geq \log\left(\frac{T\log^c(T)}{N_a(t)}\right)\right) \qquad (15) \\
&\quad + \sum_{t=K}^{T-1} \mathbb{P}\left(N_a(t) d^+(\hat{\mu}_a(t), \mu_1 - g_t) \leq \log\left(T\log^c(T)\right), A_{t+1} = a\right). \qquad (16)
\end{aligned}$$

The sums in (16) and (14) are the same, and lower bounding $T\log^c T$ by $t\log^c t$ in each term of the sum in (15) shows that it is upper bounded by (13). Thus, this index policy is also asymptotically optimal.

6 Numerical experiments

6.1 Regret minimization

We first perform experiments with a moderate horizon $T = 1000$, which permits to include the Finite-Horizon Gittins algorithm discussed in Section 3. Figure 2 displays the regret of KL-UCB, Thompson Sampling and the four Bayesian index policies discussed in this paper, in two instances of two-armed

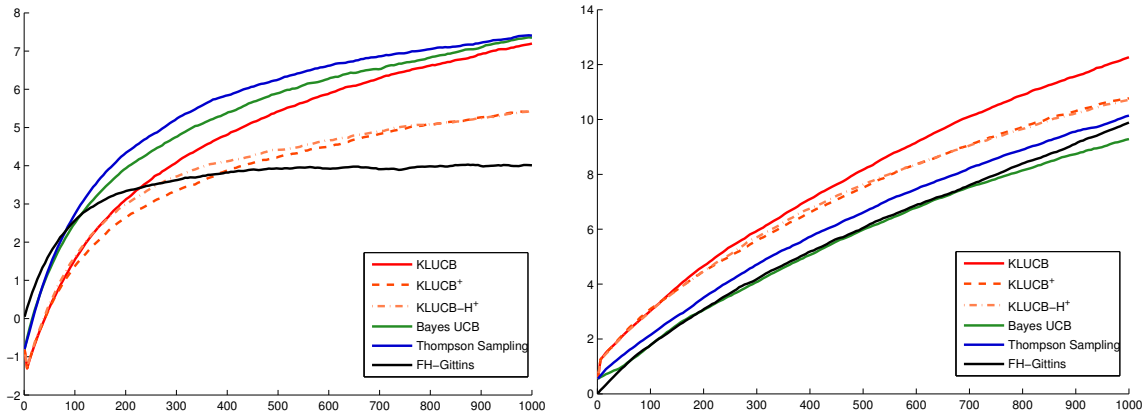

FIGURE 2 – Regret on two-armed Bernoulli bandits ($\boldsymbol{\mu} = [0.05\ 0.15]$ (left) $\boldsymbol{\mu} = [0.75\ 0.8]$ (right)) up to horizon $T = 1000$, averaged over $N = 10000$ simulations

Bernoulli bandit problems. The Bayesian index policies display comparable, if not better, performance than KL-UCB and Thompson Sampling. In particular, FH-Gittins appears to be significantly better than the other algorithms on the instance with small rewards. For a larger horizon $T = 20000$, we then run experiments on a bandit model with continuous rewards that follow an exponential distribution (which is a particular case of Gamma distribution, with parameter $k = 1$, see Table 1). In this setting, Bayes-UCB and Thompson Sampling are implemented using a $\mathrm{InvGamma}(1, 1)$ prior on the means. In this setting, Bayes-UCB, KL-UCB$^+$ and KL-UCB-H$^+$ improve over KL-UCB, and are also competitive with Thompson Sampling. As already noted in several works (e.g. [12]), the Lai and Robbins lower bound, that is asymptotic, is quite pessimistic for finite (even large) horizons.

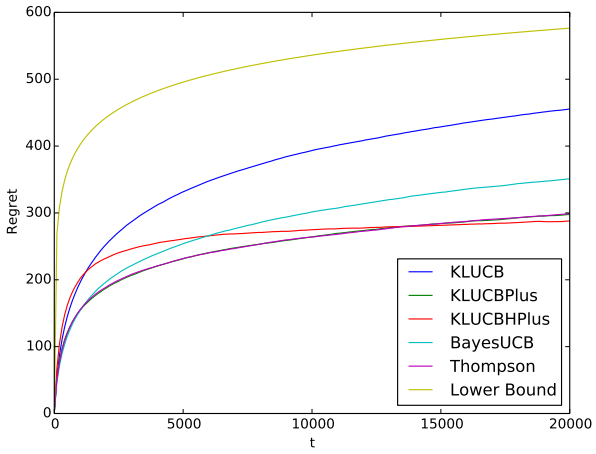

FIGURE 3 – Regret on a five-armed bandit with Exponential distributions with means $\boldsymbol{\mu} = [1\ 1.5\ 2\ 2.5\ 3]$ up to horizon $T = 20000$, averaged over $N = 50000$ simulations

6.2 Bayes risk minimization

In this paper, Bayes risk minimization and its exact solution is mostly presented as an inspiration to come up with better algorithms for regret minimization. However, it is also interesting to understand whether the proposed algorithm are good approximation of the Bayesian solution, i.e. whether they match the asymptotic lower bound of Theorem 3.

We report here results of experiments in Bernoulli bandit models with a uniform prior of the means. In this setting, some computations (see Section B.4) show that the lower bound rewrites

$$\liminf_{T\to\infty} \frac{\mathcal{R}(T,\mathcal{A})}{\log^2(T)} \geq \frac{K-1}{K+1}.$$

In particular, we see that the asymptotic rate of the Bayesian regret is (almost) independent of the number of arms. For several values of K, we display on Figure 4 the Bayes risk $\mathcal{R}_T(\mathcal{A}_{(T)})$ of several algorithms, together with the theoretical lower bound, as a function of $\log^2(T)$.

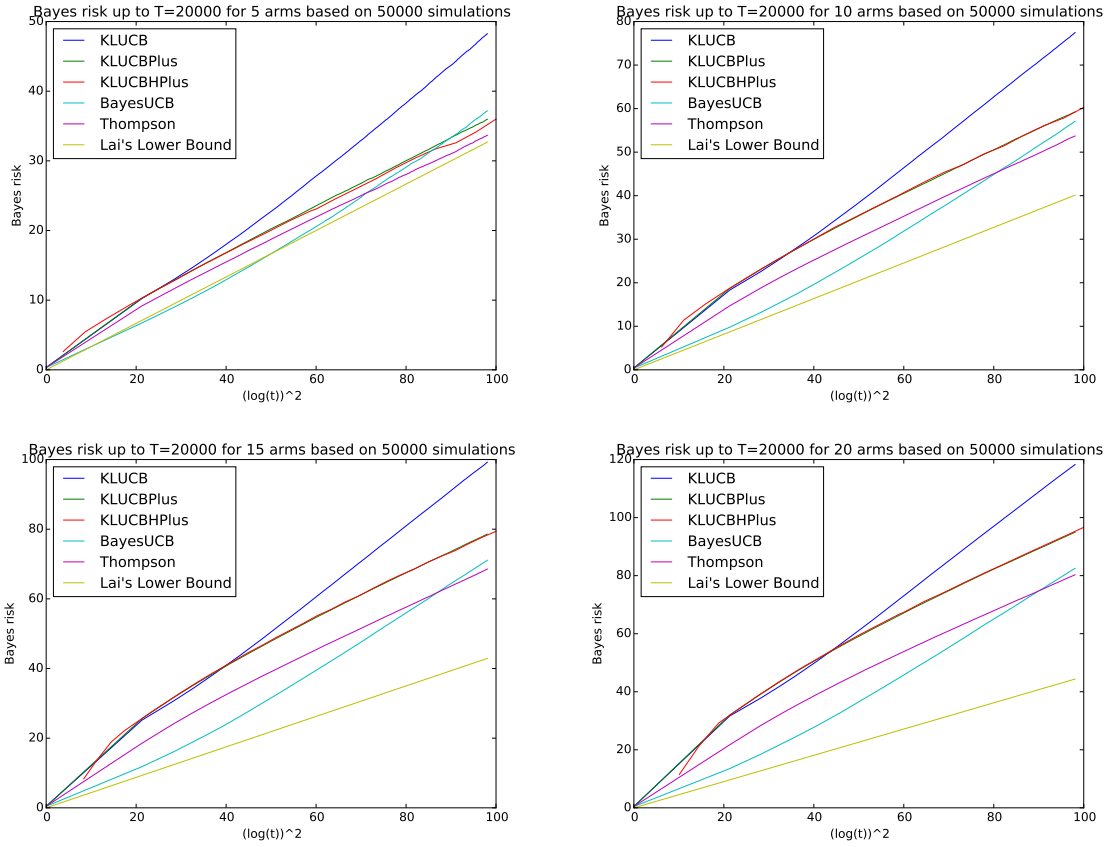

FIGURE 4 – Bayes risk up to $T = 20000$ on a Bernoulli bandit model with a uniform prior on the K arms, for $K = 5, 10, 15, 20$, averaged over $N = 50000$ simulations.

For each value of K, we observe that all the algorithms have a Bayes risk that seems to be affine in $\log^2(T)$. For Thompson Sampling, KL-UCB$^+$ and KL-UCB-H$^+$ the slope is close to $(K-1)/(K+1)$, whereas for KL-UCB and Bayes-UCB it is strictly larger. This leads to the conjecture that the first three

algorithms are asymptotically optimal in a Bayesian sense. It is to be noted that, while the Bayes risk of these algorithms seems to be of order $(K-1)/(K+1)\log^2(T) + C(K)$ for large values of T, the second order term $C(K)$ appears to be increasing significantly with the number of arms. Compared to Lai and Robbins lower bound on the regret, this lower bound does not appear to be over-pessimistic in finite time.

7 Conclusion

In the context of exponential family bandit models, this paper provides the first analysis of a Bayesian algorithm that holds for a wide class of prior distributions, namely all distributions that have positive density with respect to the Lebesgue measure. It also provides theoretical justifications for the use of the KL-UCB$^+$ algorithm together with a new insight on the alternative exploration rate used by this algorithm. An interesting future direction of research would be to better understand the Finite-Horizon Gittins strategy, which performs well in practice, but whose asymptotic optimality is still to be established.

Références

[1] R. Agrawal. Sample mean based index policies with O(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4) :1054–1078, 1995.

[2] S. Agrawal and N. Goyal. Analysis of Thompson Sampling for the multi-armed bandit problem. In *Proceedings of the 25th Conference On Learning Theory*, 2012.

[3] S. Agrawal and N. Goyal. Further Optimal Regret Bounds for Thompson Sampling. In *Proceedings of the 16th Conference on Artificial Intelligence and Statistics*, 2013.

[4] J-Y. Audibert and S. Bubeck. Regret Bounds and Minimax Policies under Partial Monitoring. *Journal of Machine Learning Research*, 2010.

[5] J-Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19), 2009.

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

[7] R. Bellman. A problem in the sequential design of experiments. *The indian journal of statistics*, 16(3/4) :221–229, 1956.

[8] D.A. Berry and B. Fristedt. *Bandit Problems. Sequential allocation of experiments*. Chapman and Hall, 1985.

[9] R.N Bradt, S.M. Johnson, and S. Karlin. On sequential designs for maximizing the sum of n observations. *Annals of Mathematical Statistics*, 27(4) :1060–1074, 1956.

[10] S. Bubeck and C.-Y. Liu. Prior-free and prior-dependent regret bounds for Thompson Sampling. In *Advances in Neural Information Processing Systems*, 2013.

[11] A. Burnetas and M. Katehakis. Asymptotic Bayes Analysis for the finite horizon one armed bandit problem. *Probability in the Engineering and Informational Sciences*, 17 :53–82, 2003.

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

[13] F. Chang and T. Lai. Optimal stopping and dynamic allocation. *Advances in Applied Probability*, 19 :829–853, 1987.

[14] O. Chapelle and L. Li. An empirical evaluation of Thompson Sampling. In *Advances in Neural Information Processing Systems*, 2011.

[15] R. Durrett. *Probability : Theory and Examples*. Cambridge University Press, 2010.

[16] A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Conference on Learning Theory*, 2011.

[17] J. Ginebra and M.K. Clayton. Small-sample performance of Bernoulli two-armed bandit Bayesian strategies. *Journal of Statistical Planning and Inference*, 79(1) :107–122, 1999.

[18] J.C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, 41(2) :148–177, 1979.

[19] J. Honda and A. Takemura. An Asymptotically Optimal Bandit Algorithm for Bounded Support Models. In *Proceedings of the 23rd Conference on Learning Theory*, 2010.

[20] E. Kaufmann, O. Cappé, and A. Garivier. On Bayesian Upper-Confidence Bounds for Bandit Problems. In *Proceedings of the 15th conference on Artificial Intelligence and Statistics*, 2012.

[21] E. Kaufmann, N. Korda, and R. Munos. Thompson Sampling : an Asymptotically Optimal Finite-Time Analysis. In *Proceedings of the 23rd conference on Algorithmic Learning Theory*, 2012.

[22] N. Korda, E. Kaufmann, and R. Munos. Thompson Sampling for 1-dimensional Exponential family bandits. In *Advances in Neural Information Processing Systems*, 2013.

[23] T.L. Lai. Adaptive treatment allocation and the multi-armed bandit problem. *Annals of Statistics*, 15(3) :1091–1114, 1987.

[24] T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1) :4–22, 1985.

[25] T. Lattimore. Regret analysis of the finite-horizon gittins index strategy for multi-armed bandits. *arXiv :1511.06014*, 2015.

[26] C.-Y. Liu and L. Li. On the prior sensitivity of thompson sampling. *arXiv :1506.03378*, 2015.

[27] J. Nino-Mora. Computing a Classic Index for Finite-Horizon Bandits. *INFORMS Journal of Computing*, 23(2) :254–267, 2011.

[28] M.L. Puterman. *Markov Decision Processes. Discrete Stochastic. Dynamic Programming.* Wiley, 1994.

[29] D. Russo and B. Van Roy. Learning to optimize via information direct sampling. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[30] D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research (to appear)*, 2014.

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

The supplementary material is structured as follows. Appendix A presents Pinsker-like inequalities, that is quadratic approximations of the Kullback-Leibler divergence functions, when the natural parameters of the distributions belong to some compact interval. These inequalities are used throughout Appendix B and Appendix C, respectively dedicated to the proof of the asymptotic lower bound on the Bayes risk stated in Theorem 3 and to the proof of the posterior tail bounds given in Lemma 7. Appendix D gathers the proofs of the lemmas introduced in the finite-time analysis of Bayes-UCB.

A Pinsker-like inequalities

For on any compact $\mathcal{C} \subset \Theta$, one can obtain quadratic approximations of the KL-divergence as a function of either the natural parameters or the means. These useful inequalities are stated in Proposition 11

Proposition 11. *Let $\mathcal{C}$ be a compact subset of Θ. Introducing*

$$c_1 := \inf_{\theta \in \mathcal{C}} \ddot{b}(\theta) > 0 \quad \textit{and} \quad c_2 := \sup_{\theta \in \mathcal{C}} \ddot{b}(\theta) < \infty, \tag{17}$$

one has

$$\forall (\theta, \theta') \in (\mathcal{C})^2, \quad \frac{c_1}{2}(\theta - \theta')^2 \leq \quad K(\theta, \theta') \leq \quad \frac{c_2}{2}(\theta - \theta')^2, \tag{18}$$

$$\forall (x, v) \in (\dot{b}(\mathcal{C}))^2, \quad \frac{1}{2c_2}(x - v)^2 \leq \quad d(x, v) \leq \quad \frac{1}{2c_1}(x - v)^2. \tag{19}$$

If $(x, v) \in (\dot{b}(\mathcal{C}))^2$ are such that $x < v$, one has

$$\dot{b}^{-1}(v) - \dot{b}^{-1}(x) \leq \frac{1}{c_1}(v - x). \tag{20}$$

Proof These three statements follow from Lagrange formulas. For example to derive (19), given that $d(x, y) = \mathrm{K}(\dot{b}^{-1}(x), \dot{b}^{-1}(y))$, it can be shown, using the close form expression (5), that

$$\frac{d}{dx} d(x, v) = \dot{b}^{-1}(x) - \dot{b}^{-1}(v) \quad \text{and} \quad \frac{d^2}{d^2 x} d(x, v) = \frac{1}{\ddot{b}(\dot{b}^{-1}(x))}.$$

From the second-order Lagrange formula applied to $x \mapsto d(x, v)$, there exists $c \in]x, v[$ (or $]v, x[$) such that

$$d(x, v) = \frac{1}{2} \frac{1}{\ddot{b}(\dot{b}^{-1}(c))} (x - v)^2 \leq \frac{1}{2c_1}(x - v)^2.$$

The other inequalities are obtained using similar arguments.

B Lower bound on the Bayesian regret

B.1 Proof of Theorem 3

Let $\mathcal{A}$ be a bandit algorithm. Introducing

$$C_{\mathrm{opt}} = \frac{1}{2} \sum_{a=1}^{K} \int_{\Theta^{K-1}} h_a(\theta_a^*) dH_{-a}(\boldsymbol{\theta}_{-a}),$$

we assume that $\mathcal{A}$ satisfies the following : there exists constants $C > C_{\text{opt}}$ and $T_0 > 0$ such that

$$\forall T \geq T_0,\ \mathcal{R}^H(T,\mathcal{A}) \leq C(\log T)^2. \tag{21}$$

Note that if $\mathcal{A}$ does not satisfy the above assumption, the desired conclusion follows directly :

$$\liminf_{T\to\infty} \frac{\mathcal{R}^H(T,\mathcal{A})}{\log^2(T)} \geq C_{\text{opt}}.$$

The Bayes risk of $\mathcal{A}$ rewrites

$$\begin{aligned}\mathcal{R}^H(T,\mathcal{A}) &= \mathbb{E}[R_{\boldsymbol{\theta}}(T,\mathcal{A})] = \mathbb{E}\left[\sum_{a=1}^{K}(\dot{b}(\theta^*)-\dot{b}(\theta_a))\mathbb{E}_{\boldsymbol{\theta}}[N_a(T)]\right]\\ &= \sum_{a=1}^{K}\int_{\{\boldsymbol{\theta}\in\Theta^K:\theta_a<\theta_a^*\}}(\dot{b}(\theta_a^*)-\dot{b}(\theta_a))\mathbb{E}_{\boldsymbol{\theta}}[N_a(T)]dH(\boldsymbol{\theta})\end{aligned}$$

Letting $\mathcal{T}_a$ be the a-th term in this last sum, one has

$$\begin{aligned}\mathcal{T}_a &= \int_{\Theta^{K-1}}\int_{\{\theta_a\in\Theta:\theta_a<\theta_a^*\}}(\dot{b}(\theta_a^*)-\dot{b}(\theta_a))\mathbb{E}_{\boldsymbol{\theta}}[N_a(T)]h_a(\theta_a)d\theta_a\ dH_{-a}(\boldsymbol{\theta}_{-a})\\ &= \int_{\Theta^{K-1}}\int_0^{\theta_a^*-\theta^-}(\dot{b}(\theta_a^*)-\dot{b}(\theta_a^*-t))\mathbb{E}_{\boldsymbol{\theta}_{a,t}}[N_a(T)]h_a(\theta_a^*-t)dt\ dH_{-a}(\boldsymbol{\theta}_{-a}),\end{aligned}$$

where $\boldsymbol{\theta}_{a,t} := (\theta_1,\dots,\theta_{a-1},\theta_a^*-t,\theta_{a+1},\dots,\theta_K)$ and θ^- denotes the lower bound of Θ : we let $\Theta =]\theta^-,\theta^+[$.

Let $\gamma\in]0,1[$ and let $B=[b^-,b^+]$ be a compact subset of Θ. For T large enough, more precisely such that

$$1/(b^- - \theta^-) < \log T < T^{\frac{1-\gamma}{2}},$$

reducing the integration domain by first letting $\boldsymbol{\theta}_{-a}\in B^{K-1}$ and then $t\in[T^{-(1-\gamma)/2},(\log T)^{-1}]$, one can write

$$\begin{aligned}\mathcal{T}_a &\geq \int_{B^{K-1}}\int_{T^{-(1-\gamma)/2}}^{(\log T)^{-1}}(\dot{b}(\theta_a^*)-\dot{b}(\theta_a^*-t))\mathbb{E}_{\boldsymbol{\theta}_{a,t}}[N_a(T)]h_a(\theta_a^*-t)dtdH_{-a}(\boldsymbol{\theta}_{-a})\\ &\geq (1-\gamma)\int_{B^{K-1}}\int_{T^{-(1-\gamma)/2}}^{(\log T)^{-1}}\frac{2h_a(\theta_a^*)\mathrm{K}(\theta_a^*-t,\theta_a^*+\zeta t)\mathbb{E}_{\boldsymbol{\theta}_{a,t}}[N_a(T)]}{t}dtdH_{-a}(\boldsymbol{\theta}_{-a}).\end{aligned}$$

The last inequality follows from the technical lemma stated below, in which the constant ζ is defined.

Lemma 12. *Let $\gamma>0$. There exists $\zeta\in]0,1[$ and $t_0>0$ such that for all $\boldsymbol{\theta}_{-a}\in B^{K-1}$ and $0\leq t\leq t_0$,*

$$\forall\theta\in B,\quad \frac{(\dot{b}(\theta)-\dot{b}(\theta-t))h_a(\theta-t)}{\mathrm{K}(\theta-t,\theta+\zeta t)} \geq (1-\gamma)\frac{2h_a(\theta)}{t}.$$

Now we need to give a lower bound on $\mathbb{E}_{\boldsymbol{\theta}_{a,t}}[N_a(T)]$, that will subsequently be integrated over $B^{K-1}\times[T^{-(1-\gamma)/2},(\log T)^{-1}]$. Lai and Robbins provide such a lower bound in [24], but under the assumption (not satisfied here) that, for all $\alpha\in]0,1[$, $\mathcal{A}$ has a $o(T^\alpha)$ regret on every bandit model. Moreover their lower bound is asymptotic, which makes it more complicated to integrate. Lemma 13 below provides a non-asymptotic lower bound on $\mathbb{E}_{\boldsymbol{\theta}_{a,t}}[N_a(T)]$, that also follows from a change of distribution argument.

Lemma 13. *Let $\zeta \in]0,1[$ and $B = [b^-, b^+] \subset \Theta$. Introducing*

$$e_{T,t}(\boldsymbol{\theta}_{-a}) = \inf\left\{\mathbb{E}_{\boldsymbol{\theta}}[T - N_a(T)] : \theta_a \in \Theta, \theta_a^* + \zeta t/2 \le \theta_a \le \theta_a^* + \zeta t\right\},$$

for every $\gamma \in]0,1[$ there exists positive constants C_1, t_1 and T_1 (that depend on B, γ and ζ) such that if $t \le t_1$ and $Tt^2 > T_1$, $\forall \boldsymbol{\theta}_{-a} \in B^{K-1}$,

$$\mathbb{E}_{\boldsymbol{\theta}_{a,t}}[N_a(T)] \ge \frac{(1-\gamma)\log(Tt^2)}{\mathrm{K}(\theta_a^* - t, \theta_a^* + \zeta t)}\left(1 - e^{-C_1\log(Tt^2)} - \frac{2t^2 e_{T,t}(\boldsymbol{\theta}_{-a})}{(Tt^2)^{\frac{\gamma}{2}}}\right).$$

Using Lemma 13, if T satisfies moreover $\log(T) \ge 1/\min(t_0, t_1, \gamma/\log(T_1))$,

$$\mathcal{T}_a \ge 2(1-\gamma)^2(\mathcal{I}_1(T) - \mathcal{I}_2(T) - 2\mathcal{I}_3(T)),$$

where

$$\begin{aligned}
\mathcal{I}_1(T) &:= \int_{B^{K-1}} h_a(\theta_a^*) \int_{T^{-(1-\gamma)/2}}^{(\log T)^{-1}} \frac{\log(Tt^2)}{t} dt\, dH_{-a}(\boldsymbol{\theta}_{-a}),\\
\mathcal{I}_2(T) &:= \int_{B^{K-1}} h_a(\theta_a^*) \int_{T^{-(1-\gamma)/2}}^{(\log T)^{-1}} \frac{\log(Tt^2)}{t} \frac{1}{(Tt^2)^{C_1}} dt\, dH_{-a}(\boldsymbol{\theta}_{-a}),\\
\mathcal{I}_3(T) &:= \int_{B^{K-1}} h_a(\theta_a^*) \int_{T^{-(1-\gamma)/2}}^{(\log T)^{-1}} \frac{\log(Tt^2)}{t} t^2 (Tt^2)^{-\frac{\gamma}{2}} e_{T,t}(\boldsymbol{\theta}_{-a}) dt\, dH_{-a}(\boldsymbol{\theta}_{-a}).
\end{aligned}$$

First, an explicit calculation yields

$$\mathcal{I}_1(T) = \frac{1}{4}\left(\left(1 - \frac{2\log\log(T)}{\log(T)}\right)^2 - \gamma^2\right)\log^2(T) \int_{B^{K-1}} h_a(\theta_a^*) dH_{-a}(\boldsymbol{\theta}_{-a}),$$

which shows that

$$\mathcal{I}_1(T) \underset{T\to\infty}{\sim} \frac{1}{4}(1-\gamma^2)\left(\int_{B^{K-1}} h_a(\theta_a^*) dH_{-a}(\boldsymbol{\theta}_{-a})\right)\log^2(T).$$

Then, for every $\varepsilon > 0$, there exists $T_2(\varepsilon)$ such that for all $T \ge T_2(\varepsilon)$, for all $t \ge T^{-(1-\gamma)/2}$, $1/(Tt^2)^{C_1} \le \varepsilon$. Hence, for $T \ge T_2(\varepsilon)$,

$$\mathcal{I}_2(T) \le \varepsilon \mathcal{I}_1(T).$$

This proves that $\mathcal{I}_2(T) = \underset{T\to\infty}{o}(\log^2(T))$.

Finally, to prove that $\mathcal{I}_3(T) = \underset{T\to\infty}{o}(\log^2(T))$, we start by writing

$$\mathcal{I}_3(T) = \int_{T^{-(1-\gamma)/2}}^{(\log T)^{-1}} \frac{\log(Tt^2)}{t} (Tt^2)^{-\frac{\gamma}{2}} t^2 \left(\int_{B^{K-1}} e_{T,t}(\boldsymbol{\theta}_{-a}) h_a(\theta_a^*) dH_{-a}(\boldsymbol{\theta}_{-a})\right) dt.$$

and we provide an upper bound on the inner integral. First note that if $\boldsymbol{\theta}$ is such that $\theta_a > \theta_a^*$, one has

$$R_{\boldsymbol{\theta}}(T, \mathcal{A}) \ge (\dot{b}(\theta_a) - \dot{b}(\theta_a^*))\mathbb{E}_{\boldsymbol{\theta}}[T - N_a(T)].$$

Using (21) together with this last inequality, one obtains, for every t,

$$\begin{aligned} C\log^2(T) &\geq \int_{\{\boldsymbol{\theta}\in B^K:\theta_a^*+\zeta t/2<\theta_a<\theta_a^*+\zeta t\}} R_T(\mathcal{A},\boldsymbol{\theta})dH(\boldsymbol{\theta}) \\ &\geq \int_{B^{K-1}}\int_{\theta_a^*+\zeta t/2}^{\theta_a^*+\zeta t}(\dot{b}(\theta_a)-\dot{b}(\theta_a^*))\mathbb{E}_{\boldsymbol{\theta}}[T-N_a(T)]h_a(\theta_a)d\theta_a\, dH_{-a}(\boldsymbol{\theta}_{-a}) \\ &\geq \int_{B^{K-1}} e_{T,t}(\boldsymbol{\theta}_{-a})\int_{\theta_a^*+\zeta t/2}^{\theta_a^*+\zeta t}(\dot{b}(\theta_a)-\dot{b}(\theta_a^*))h_a(\theta_a)d\theta_a\, dH_{-a}(\boldsymbol{\theta}_{-a}). \end{aligned}$$

With $B=[b^-,b^+]$, let t_2 be such that the compact $B'=[b^-+\zeta t_2/2, b^++\zeta t_2]$ is included in Θ. As h_a is uniformly continuous and bounded on B', there exists t_2 such that and for all $\theta_a^*\in B$, for all $t\leq t_2$,

$$\inf_{[\theta_a^*+\zeta t/2,\theta_a^*+\zeta t]} h_a(\theta)\geq\frac{2}{3}h_a(\theta_a^*).$$

Let $t\leq t_2$. Introducing $c_1=\inf_{\theta\in B'}\ddot{b}(\theta)>0$, using the Lagrange formula,

$$\begin{aligned} C\log^2(T) &\geq \frac{2c_1}{3}\int_{B^{K-1}} e_{T,t}(\boldsymbol{\theta}_{-a})\int_{\theta_a^*+\zeta t/2}^{\theta_a^*+\zeta t}(\theta_a-\theta_a^*)h_a(\theta_a^*)d\theta_a\, dH_{-a}(\boldsymbol{\theta}_{-a}) \\ &= \frac{c_1}{4}\zeta^2t^2\int_{B^{K-1}} e_{T,t}(\boldsymbol{\theta}_{-a})h_a(\theta_a^*)dH_{-a}(\boldsymbol{\theta}_{-a}). \end{aligned}$$

Finally, if $T^{-\frac{1-\gamma}{2}}\leq t\leq t_2$,

$$\int_{B^{K-1}} e_{T,t}(\boldsymbol{\theta}_{-a})h_a(\theta_a^*)dH_{-a}(\boldsymbol{\theta}_{-a})\leq\frac{4C}{c_1\zeta^2}\frac{\log^2(T)}{t^2}\leq\frac{4C}{c_1\zeta^2\gamma^2}\frac{(\log(Tt^2))^2}{t^2}.$$

For T satisfying $\log(T)^{-1}\leq t_2$, $[T^{-(1-\gamma)/2},(\log T)^{-1}]\subseteq[T^{-\frac{1-\gamma}{2}},t_2]$ and

$$\begin{aligned} \mathcal{I}_3(T) &\leq \int_{T^{-(1-\gamma)/2}}^{(\log T)^{-1}}\frac{\log(Tt^2)}{t}(Tt^2)^{-\frac{\gamma}{2}}t^2\left(\frac{4C}{c_1\zeta^2\gamma^2}\frac{(\log(Tt^2))^2}{t^2}\right)dt \\ &= \frac{4C}{c_1\zeta^2\gamma^2}\int_{T^{-(1-\gamma)/2}}^{(\log T)^{-1}}\frac{\log(Tt^2)}{t}\left(\frac{\left(\log(Tt^2)\right)^2}{(Tt^2)^{\frac{\gamma}{2}}}\right)dt. \end{aligned}$$

Let $\varepsilon>0$. As $u\mapsto\log^2(u)/(u^{\gamma/2})$ tends to zero when u tends to infinity, and $Tt^2\geq T^\gamma$ for $t\geq T^{-(1-\gamma)/2}$, there exists a constant $T_3(\varepsilon)$ such that

$$\text{for } T\geq T_3(\varepsilon), \text{for } t\geq T^{-(1-\gamma)/2},\quad \frac{\left(\log(Tt^2)\right)^2}{(Tt^2)^{\frac{\gamma}{2}}}\leq\varepsilon.$$

Hence, for $T\geq T_3(\varepsilon)$,

$$\mathcal{I}_3(T)\leq\varepsilon\frac{4C}{c_1\zeta^2\gamma^2}\int_{T^{-(1-\gamma)/2}}^{(\log T)^{-1}}\frac{\log(Tt^2)}{t}dt=\varepsilon\frac{C}{c_1\zeta^2\gamma^2}\left(\left(1-\frac{2\log\log(T)}{\log(T)}\right)^2-\gamma^2\right)\log^2(T),$$

which proves that $\mathcal{I}_3(T)=o\left(\log^2(T)\right)$.

Putting everything together, we proved that, for every algorithm $\mathcal{A}$, for every $\gamma > 0$, for every compact $B \subset \Theta$,

$$\liminf_{T\to\infty} \frac{\mathcal{R}_T(\mathcal{A}, H)}{\log^2(T)} \geq (1-\gamma)^2(1-\gamma^2)\frac{1}{2}\sum_{a=1}^{K}\int_{B^{K-1}} h_a(\theta_a^*)dH_{-a}(\boldsymbol{\theta}_{-a}).$$

Taking the supremum over all compact set B yields, for every $\gamma > 0$,

$$\liminf_{T\to\infty} \frac{\mathcal{R}_T(\mathcal{A}, H)}{\log^2(T)} \geq (1-\gamma)^2(1-\gamma^2)\frac{1}{2}\sum_{a=1}^{K}\int_{\Theta^{K-1}} h_a(\theta_a^*)dH_{-a}(\boldsymbol{\theta}_{-a}),$$

provided the integral in the right-hand side is finite. Letting γ go to zero concludes the proof.

B.2 Proof of Lemma 12

Let $\zeta \in]0,1[$ be fixed . As $B = [b^-, b^+]$ is strictly included in Θ, there exists t_1 such that $\mathcal{C} := [b^- - t_1, b^+ + \zeta t_1]$ in included in Θ. For $(\theta, t) \in B \times [0, t_1]$ we define

$$f(\theta, t) = \frac{(1+\zeta)^2 t}{2}\frac{(\dot{b}(\theta) - \dot{b}(\theta - t))h_a(\theta - t)}{K(\theta - t, \theta + \zeta t)h_a(\theta)}.$$

f is continuous on $B \times [0, t_1]$ and it can be checked that

$$\lim_{(\theta,t)\to(\theta_0,0)} f(\theta, t) = 1.$$

As f is uniformly continuous, there exists $t_0 \leq t_1$, such that for all $t \leq t_0$, for all $\theta \in B$,

$$|f(\theta, t) - 1| \leq \frac{\gamma}{2},$$

which rewrites

$$\left|\frac{(\dot{b}(\theta) - \dot{b}(\theta - t))h_a(\theta - t)}{K(\theta - t, \theta + \zeta t)} - \frac{2h_a(\theta)}{(1+\zeta)^2 t}\right| \leq \frac{\gamma}{2}\frac{2h_a(\theta)}{(1+\zeta)^2 t}$$

hence, for $t \leq t_0$, one has

$$\frac{(\dot{b}(\theta) - \dot{b}(\theta - t))h_a(\theta - t)}{K(\theta - t, \theta + \zeta t)} \geq \frac{1 - \frac{\gamma}{2}}{(1+\zeta)^2}\frac{2h_a(\theta)}{t}.$$

Applying this to ζ such that $1 + \zeta = \sqrt{\frac{1-\frac{\gamma}{2}}{1-\gamma}}$ concludes the proof.

B.3 Proof of Lemma 13

Let $\zeta \in]0,1[$ be fixed and define t_1 and $\mathcal{C} = [b^- - t_1, b^+ + \zeta t_1] \subset \Theta$ as in the proof of Lemma 12. Let $t \leq t_1$ and fix $\boldsymbol{\theta}_{-a} \in B^{K-1}$. First, using Markov inequality,

$$\mathbb{E}_{\boldsymbol{\theta}_{a,t}}[N_a(T)] \geq \frac{(1-\gamma)\log(Tt^2)}{K(\theta_a^* - t, \theta_a^* + \zeta t)}\mathbb{P}_{\boldsymbol{\theta}_{a,t}}\left(N_a(T) \geq \frac{(1-\gamma)\log(Tt^2)}{K(\theta_a^* - t, \theta_a^* + \zeta t)}\right).$$

Thus it is sufficient to prove that

$$\mathbb{P}_{\boldsymbol{\theta}_{a,t}}\left(N_a(T) \leq \frac{(1-\gamma)\log(Tt^2)}{K(\theta_a^* - t, \theta_a^* + \zeta t)}\right) \leq e^{-C_1\log(Tt^2)} + \frac{2t^2 e_{T,t}(\boldsymbol{\theta}_{-a})}{(Tt^2)^{\frac{\gamma}{2}}}. \tag{22}$$

As $t \le t_1$, the set $\{\theta_a : \theta_a^* + \zeta t/2 \le \theta_a \le \theta_a^* + \zeta t\}$ is a compact set included in $\mathcal{C}$, therefore there exists $\boldsymbol{\lambda} \in B^{K-1} \times \mathcal{C}$ that attains the infimum in the definition of $e_{T,t}(\boldsymbol{\theta}_{-a})$:

$$e_{T,t}(\boldsymbol{\theta}_{-a}) = \mathbb{E}_{\boldsymbol{\lambda}}[T - N_a(T)],$$

with $\boldsymbol{\lambda}_{-a} = \boldsymbol{\theta}_{-a}$ and $\lambda_a = \theta_a^* + \varepsilon t$, for some $\varepsilon \in [\zeta/2, \zeta]$. Using Markov inequality,

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\lambda}}\left(N_a(T) \le \frac{(1-\gamma)\log(Tt^2)}{K(\theta_a^* - t, \theta_a^* + \zeta t)}\right) &\le \frac{\mathbb{E}_{\boldsymbol{\lambda}}[T - N_a(T)]}{T - \frac{(1-\gamma)\log(Tt^2)}{K(\theta_a^* - t, \theta_a^* + \zeta t)}} \\ &= \frac{e_{T,t}(\boldsymbol{\theta}_{-a})}{T\left(1 - \frac{(1-\gamma)\log(Tt^2)}{K(\theta_a^* - t, \theta_a^* + \zeta t)T}\right)}.\end{aligned}$$

Introducing $c_1 = \inf_{\theta \in \mathcal{C}} \ddot{b}(\theta)$, using (18) in Proposition 11, for $t \le t_1$,

$$\frac{(1-\gamma)\log(Tt^2)}{K(\theta_a^* - t, \theta_a^* + \zeta t)T} \le \frac{2(1-\gamma)\log(Tt^2)}{c_1(1+\zeta)^2(Tt^2)} \le \frac{1}{2},$$

where the last inequality holds to Tt^2 large enough. Thus there exists $T_1 > 0$ such that for $t \ge t_1$ and $Tt^2 \ge T_1$,

$$\mathbb{P}_{\boldsymbol{\lambda}}\left(N_a(T) \le \frac{(1-\gamma)\log(Tt^2)}{K(\theta_a^* - t, \theta_a^* + \zeta t)}\right) \le \frac{2e_{T,t}(\boldsymbol{\theta}_{-a})}{T}. \tag{23}$$

Introducing the log likelihood ratio $L_n = \sum_{s=1}^n \log \frac{f_{\theta_a^* - t}(Y_{a,s})}{f_{\lambda_a}(Y_{a,s})}$, where $Y_{a,s}$ are i.i.d. samples of the distribution of arm a, one can write

$$\begin{aligned}&\mathbb{P}_{\boldsymbol{\theta}_{a,t}}\left(N_a(T) \le \frac{(1-\gamma)\log(Tt^2)}{K(\theta_a^* - t, \theta_a^* + \zeta t)}\right) \\ &\le \mathbb{P}_{\boldsymbol{\theta}_{a,t}}\left(N_a(T) \le \frac{(1-\gamma)\log(Tt^2)}{K(\theta_a^* - t, \theta_a^* + \zeta t)}, L_{N_a(T)} \le \left(1 - \frac{\gamma}{2}\right)\log(Tt^2)\right) \qquad (24) \\ &\quad + \mathbb{P}_{\boldsymbol{\theta}_{a,t}}\left(\max_{n \le \frac{(1-\gamma)\log(Tt^2)}{K(\theta_a^* - t, \theta_a^* + \zeta t)}} L_n \ge \left(1 - \frac{\gamma}{2}\right)\log(Tt^2)\right) \qquad (25)\end{aligned}$$

An upper bound on Term (24) follows from a change of distribution argument. Let $\mathcal{E}$ be the event

$$\mathcal{E} := \left\{N_a(T) \le \frac{(1-\gamma)\log(Tt^2)}{K(\theta_a^* - t, \theta_a^* + \zeta t)}, L_{N_a(T)} \le \left(1 - \frac{\gamma}{2}\right)\log(Tt^2)\right\}$$

As $\mathcal{E} \in \mathcal{F}_{N_a(T)}$, one has

$$\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}) = \mathbb{E}_{\boldsymbol{\theta}_{a,t}}\left[\mathbb{1}_{\mathcal{E}} \exp\left(-L_{N_a(T)}\right)\right] \ge \exp\left(-\left(1 - \frac{\gamma}{2}\right)\log(Tt^2)\right)\mathbb{P}_{\boldsymbol{\theta}_{a,t}}(\mathcal{E}).$$

Thus, using moreover (23),

$$\begin{aligned}(24) &\le (Tt^2)^{1-\frac{\gamma}{2}}\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}) \le (Tt^2)^{1-\frac{\gamma}{2}}\mathbb{P}_{\boldsymbol{\lambda}}\left(N_a(T) \le \frac{(1-\gamma)\log(Tt^2)}{K(\theta_a^* - t, \theta_a^* + \zeta t)}\right) \\ &\le 2t^2\left(Tt^2\right)^{-\frac{\gamma}{2}} e_{N_t}(\boldsymbol{\theta}_{-a}).\end{aligned}$$

An upper bound of Term (25) follows from a concentration inequality specific to exponential families, stated as Lemma 14, whose proof is given below, for the sake of completeness.

Lemma 14. *Let the (Y_i) be i.i.d with distribution ν_θ and mean $\mu = \dot{b}(\theta)$.*

$$\mathbb{P}\left(\max_{n\leq N}\sum_{s=1}^{n}(\mu - Y_i) \geq x\right) \leq \exp\left(-Nd\left(\mu - \frac{x}{N}, \mu\right)\right)$$

Introducing the notation $\overline{\theta}_a = \theta_a^* - t$ and

$$K_T = \frac{(1-\gamma)\log(Tt^2)}{K(\theta_a^* - t, \theta_a^* + \zeta t)} = \frac{(1-\gamma)\log(Tt^2)}{K(\overline{\theta}_a, \theta_a^* + \zeta t)},$$

the log likelihood ratio can be made explicit, and satisfies, for $n \leq K_T$,

$$\begin{aligned} L_n &= \sum_{s=1}^{n}(\overline{\theta}_a - \lambda_a)Y_{a,s} - b(\overline{\theta}_a) + b(\lambda_a) \\ &= (\overline{\theta}_a - \lambda_a)\sum_{s=1}^{n}(Y_{a,s} - \dot{b}(\overline{\theta}_a)) + nK(\overline{\theta}_a, \lambda_a). \\ &\leq (\lambda_a - \overline{\theta}_a)\sum_{s=1}^{n}(\dot{b}(\overline{\theta}_a) - Y_{a,s}) + (1-\gamma)\log(Tt^2). \end{aligned}$$

Term (25) is upper bounded by

$$\begin{aligned} &\mathbb{P}_{\boldsymbol{\theta}_{a,t}}\left(\max_{n\leq K_T}\left[(\lambda_a - \overline{\theta}_a)\sum_{s=1}^{n}(\dot{b}(\overline{\theta}_a) - Y_{a,s}) + (1-\gamma)\log(Tt^2)\right] \geq \left(1 - \frac{\gamma}{2}\right)\log(Tt^2)\right) \\ &\leq \mathbb{P}_{\boldsymbol{\theta}_{a,t}}\left(\max_{n\leq K_T}\sum_{s=1}^{n}(\dot{b}(\overline{\theta}_a) - Y_{a,s}) \geq \frac{\gamma}{2}\frac{\log(Tt^2)}{\lambda_a - \overline{\theta}_a}\right). \end{aligned}$$

Under $\boldsymbol{\theta}_{a,t}$, the sequence $Y_{a,s}$ is i.i.d with distribution $\nu_{\overline{\theta}_a}$. Therefore, using Lemma 14 one obtains, with the notation $\overline{\mu}_a = \dot{b}(\overline{\theta}_a)$,

$$(25) \leq \exp\left(-K_T d\left(\overline{\mu}_a - \frac{\gamma K(\theta_a^* - t, \theta_a^* + \zeta t)}{2(1-\gamma)(\varepsilon+1)t}, \overline{\mu}_a\right)\right).$$

Letting $c_1 = \inf_{\theta\in\mathcal{C}}\ddot{b}(\theta)$ and $c_2 = \inf_{\theta\in\mathcal{C}}\ddot{b}(\theta)$, from (18) in Proposition 11,

$$\frac{\gamma K(\theta_a^* - t, \theta_a^* + \zeta t)}{2(1-\gamma)(\varepsilon+1)t} \in \left[\frac{\gamma}{2(1-\gamma)}\frac{c_1}{2}(\zeta+1)^2t^2; \frac{\gamma}{2(1-\gamma)}\frac{c_2}{2}(\zeta+1)^2t^2\right].$$

Thus, for t small enough, $\overline{\mu}_a$ and $\overline{\mu}_a - \frac{\gamma K(\theta_a^* - t, \theta_a^* + \zeta t)}{2(1-\gamma)(\varepsilon+1)t}$ belong to a compact $\mathcal{C}'$ satisfying $\mathcal{C} \subseteq \mathcal{C}' \subseteq \Theta$. Letting $c_2' = \sup_{\theta\in\mathcal{C}'}\ddot{b}(\theta)$, using (19),

$$\begin{aligned} (25) &\leq \exp\left(-\frac{K_T}{2c_2'}\left(\frac{\gamma K(\theta_a^* - t, \theta_a^* + \zeta t)}{2(1-\gamma)(\varepsilon+1)t}\right)^2\right) \\ &= \exp\left(-\log(Tt^2)\frac{\gamma^2}{8(1-\gamma)c_2'}\frac{K(\theta_a^* - t, \theta_a^* + \zeta t)}{(1+\varepsilon)^2t^2}\right) \\ &\leq \exp\left(-\log(Tt^2)\frac{\gamma^2 c_1}{8(1-\gamma)c_2'}\frac{c_1(1+\zeta)^2}{(1+\varepsilon)^2}\right). \end{aligned}$$

Letting $C_1 = \frac{\gamma^2 c_1}{8(1-\gamma)c_2'} \frac{c_1(1+\zeta)^2}{(1+\varepsilon)^2}$, from the upper bounds obtained on (24) and (25), it follows that

$$\mathbb{P}_{\boldsymbol{\theta}_{a,t}}\left(N_a(T) \leq \frac{(1-\gamma)\log(Tt^2)}{K(\theta_a^* - t, \theta_a^* + \zeta t)}\right) \leq 2t^2\left(Tt^2\right)^{-\frac{\gamma}{2}} e_{N_t}(\boldsymbol{\theta}_{-a}) + e^{-C_1\log(Tt^2)},$$

provided that $t \leq t_1$ and $Tt^2 \geq T_1$, which concludes the proof.

□

Proof of Lemma 14 The proof follows from the Chernoff technique and a maximal inequality.

Let $S_n = \sum_{s=1}^n (\mu - Y_i)$. For every $\lambda > 0$,

$$\mathbb{P}\left(\max_{n\leq N} S_n \geq x\right) = \mathbb{P}\left(\max_{n\leq N} e^{\lambda S_n} \geq e^{\lambda x}\right) \leq e^{-\lambda x}\mathbb{E}\left[e^{\lambda S_N}\right], \tag{26}$$

where the last inequality is a consequence of Doob's maximal inequality applied to $M_n = e^{\lambda S_n}$, which is a sub-martingale with respect to the filtration generated by the (Y_i). Indeed, using the convexity of the mapping $x \mapsto e^{\lambda x}$,

$$\begin{aligned}\mathbb{E}\left[M_n - M_{n-1}|\mathcal{F}_{n-1}\right] &= e^{\lambda S_n}\mathbb{E}\left[e^{\lambda(S_n - S_{n-1})} - 1|\mathcal{F}_{n-1}\right] \\ &\geq e^{\lambda S_n}\lambda\mathbb{E}\left[S_n - S_{n-1}|\mathcal{F}_{n-1}\right] = 0.\end{aligned}$$

Using the independence of the Y_i and $\mathbb{E}[e^{\lambda Y_i}] = \exp(b(\theta+\lambda) - b(\theta)$ for any $\lambda \in \mathbb{R}$, it can be show that

$$e^{-\lambda x}\mathbb{E}\left[e^{\lambda S_N}\right] = \exp\left(-N\left[\lambda\left(\frac{x}{N} - \dot{b}(\theta)\right) + b(\theta) - b(\theta-\lambda)\right]\right).$$

The exponent is minimized of λ^* satisfying $\dot{b}(\theta - \lambda^*) = \dot{b}(\theta) - x/N$ and

$$\begin{aligned}e^{-\lambda^* x}\mathbb{E}\left[e^{\lambda^* S_N}\right] &= \exp\left(-N\left[\dot{b}(\theta-\lambda^*)(-\lambda^*) - \dot{b}(\theta-\lambda^*) + b(\theta)\right]\right) \\ &= \exp(-NK(\theta-\lambda^*,\theta)) = \exp\left(-Nd\left(\mu - \frac{x}{N}, \mu\right)\right).\end{aligned}$$

The conclusion follows by plugging λ^* in (26).

B.4 The lower bound for Bernoulli bandits

As pointed out by [23], in the particular case in which $h_a(\theta) = q(\theta)$ for all $a = 1, \ldots, K$, using the fact that the distribution of $\max_{a\in\mathcal{S}} \theta_a$ has density $kq(\theta)Q^{k-1}(\theta)$ where Q is the c.d.f. of the distribution with density q and $k = |\mathcal{S}|$, the constant in the lower bound can be expressed

$$\frac{1}{2}\sum_{a=1}^K \int_{\Theta^{K-1}} h_a(\theta_a^*) dH_{-a}(\boldsymbol{\theta}_{-a}) = \frac{K(K-1)}{2}\int_\Theta q^2(\theta)Q^{K-2}(\theta)d\theta. \tag{27}$$

Now consider a Bernoulli bandit model with K arms, with a uniform prior distribution on the mean of each arm. The set of Bernoulli distribution of means $\mu \in [0,1]$ form an exponential family when each distribution is parametrized by the natural parameter $\theta = \log(\mu/(1-\mu))$. This exponential family is characterized by

$$\Theta = \mathbb{R}, \quad b(\theta) = \log(1 + e^\theta),$$

and the reference measure is the Lebesgue measure. As each mean μ_a is drawn from a uniform distribution on $[0,1]$, the associated natural parameter θ_a is drawn from a distribution on $\mathbb{R}$ having respectively density and c.d.f.

$$q(\theta) = \frac{e^\theta}{(1+e^\theta)^2} \quad \text{and} \quad Q(\theta) = \frac{e^\theta}{1+e^\theta}.$$

Using the formula (27), the constant of the lower bound is

$$\begin{aligned}\frac{K(K-1)}{2}\int_{-\infty}^{+\infty}\frac{e^{K\theta}}{(1+e^\theta)^{K+2}}d\theta &= \frac{K(K-1)}{2}\int_0^\infty \frac{x^{K-1}}{(1+x)^{K+2}}dx \\ &= \frac{K(K-1)}{2}\frac{1}{K(K+1)},\end{aligned}$$

where the integral is computed using by inducting, using a by part integration. Finally, the asymptotic rate of the Bayes risk for a Bernoulli bandit model with K arms and a uniform prior on their means is

$$\frac{1}{2}\frac{K-1}{K+1}\log^2(T).$$

C Posterior tail bounds

Let μ^-, μ^+ be such that $\mathrm{J} := \dot{b}(\Theta) = (\mu^-, \mu^+)$. We give here the proof of Lemma 7, that follows directly from bounds on

$$\pi_{n,x}([v,\mu^+[) := \frac{\int_v^{\mu^+}\exp(-nd(x,u))f_0(u)du}{\int_{\mathrm{J}}\exp(-nd(x,u))f_0(u)du}$$

for a function f_0 satisfying $f_0(u) > 0$ for all $u \in \mathrm{J}$. We fix μ_0^-, μ_0^+ such that $\dot{b}(\theta^-) < \mu_0^- < \mu_0^+ < \dot{b}(\theta^+)$.

First, we fix a compact $\mathcal{C}$ included in Θ such that $[\mu_1^-, \mu_1^+] := \dot{b}(\mathcal{C})$ satisfy

$$\dot{b}(\theta^-) < \mu_1^- < \mu_0^- < \mu_0^+ < \mu_1^+ < \dot{b}(\theta^+).$$

We let $\mathrm{J}_\mathcal{C} = [\mu_1^-, \mu_1^+]$ and c_1 and c_2 be the upper and lower bounds on $\ddot{b}$ on $\mathcal{C}$, defined as (17). We will often use the quadratic bounds on the Kullback-Leibler divergence on this compact, that are stated in Proposition 11.

Let x, v such that $\mu_0^- < x < v < \mu_0^+$. One has

$$\pi_{n,x}([v,\mu^+[) = \frac{\int_v^{\mu^+} e^{-nd(x,u)}f_0(u)du}{\int_{\mu^-}^{\mu^+} e^{-nd(x,u)}f_0(u)du}. \tag{28}$$

For any $V_{n,x} \subset \mathrm{J}$,

$$\pi_{n,x}([v,\mu^+[) \leq \frac{e^{-nd(x,v)}\int_v^{\mu^+}f_0(u)du}{\int_{V_{n,x}} e^{-nd(x,u)}f_0(u)du} \leq \frac{e^{-nd(x,v)}}{\int_{V_{n,x}} e^{-nd(x,u)}f_0(u)du}.$$

We now choose $V_{n,x} = \{u \in \mathrm{J}_\mathcal{C} : \frac{n(x-u)^2}{2c_1} \leq 1\}$. From (19), $nd(x,u) \leq 1$ on $V_{n,x}$. Hence

$$\int_{V_{n,x}} e^{-nd(x,u)}f_0(u)du \geq e^{-1}\inf_{u\in\mathrm{J}_\mathcal{C}} f_0(u)\int_{V_{n,x}} 1du$$

and

$$\begin{aligned}\int_{V_{n,x}} 1du &= \lambda\left([\mu_1^-,\mu_1^+]\cap\left[x-\sqrt{\frac{2c_1}{n}},x+\sqrt{\frac{2c_1}{n}}\right]\right)\\ &\geq \min\left(\sqrt{\frac{2c_1}{n}},\mu_1^+-\mu_1^-\right).\end{aligned}$$

The following inequality yields the upper bound in statement 1 :

$$\pi_{n,x}([v,\mu^+[) \leq \frac{e}{\sqrt{2c_1}\inf_{u\in J_{\mathcal{C}}} f_0(u)}\max\left(\sqrt{n},\frac{\sqrt{2c_1}}{(\mu_1^+-\mu_1^-)}\right)e^{-nd(x,v)}.$$

As $e^{-nd(x,u)}\leq 1$, the denominator in (28) is upper bounded by 1, thus

$$\pi_{n,x}([v,\mu^+[) \geq \int_v^{\mu^+} e^{-nd(x,u)}f_0(u)du \geq \int_v^{\mu_1^+} e^{-nd(x,u)}f_0(u)du.$$

This last integral can be lower bounded in the following way :

$$\begin{aligned}\int_v^{\mu_1^+} e^{-nd(x,u)}f_0(u)du &= e^{-nd(x,v)}\int_v^{\mu_1^+} e^{-n[d(x,u)-d(x,v)]}f_0(u)du\\ &= e^{-nd(x,v)}\int_v^{\mu_1^+} e^{-n\left[d(v,u)+(\dot b^{-1}(u)-\dot b^{-1}(v))(v-x)\right]}f_0(u)du\\ &\geq e^{-nd(x,v)}\int_v^{\mu_1^+} e^{-n\left[\frac{1}{2c_1}(u-v)^2+\frac{1}{c_1}(u-v)(v-x)\right]}f_0(u)du, \qquad (29)\end{aligned}$$

where the last inequality follows from (19) and (20). We let

$$\phi(u) = \frac{1}{2c_1}\left[(u-v)^2+2(u-v)(v-x)\right].$$

One has $\phi'(u)=(u-x)/c_1$, thus ϕ is strictly increasing on $[v,\mu_1^+]$ and it can be checked that $\phi^{-1}(y)=x+\sqrt{(v-x)^2+2c_1y}$. Thus letting $y=n\phi(u)$, one has

$$du = \frac{c_1}{n\sqrt{(v-x)^2+2c_1y/n}}dy,$$

and

$$\begin{aligned}(29) &= \frac{c_1e^{-nd(x,v)}}{n}\int_0^{\frac{n}{2c_1}\left[(\mu_1^+-v)^2+2(\mu_1^+-v)(v-x)\right]}\frac{e^{-y}}{\sqrt{(v-x)^2+2c_1y/n}}f_0(y)dy\\ &\geq \frac{c_1e^{-nd(x,v)}\min_{J_{\mathcal{C}}}f_0}{n}\int_0^{\frac{n}{2c_1}\left[(\mu_1^+-v)^2+2(\mu_1^+-v)(v-x)\right]}\frac{e^{-y}dy}{\sqrt{(v-x)^2+(\mu_1^+-v)^2+2(\mu_1^+-v)(v-x)}}\\ &\geq \frac{c_1e^{-nd(x,v)}\min_{J_{\mathcal{C}}}f_0}{n(\mu_1^+-x)}\left(1-e^{-\frac{n}{2c_1}(\mu_1^+-v)^2}\right).\end{aligned}$$

Finally, using that $\mu_0^-<x$ and $v\leq\mu_0^+$, one obtains

$$\pi_{n,x}([v,\mu^+[) \geq \left(\frac{c_1(1-e^{-\frac{(\mu_1^+-\mu_0^+)^2}{2c_1}})\min_{J_{\mathcal{C}}}f_0}{(\mu_1^+-\mu_0^-)}\right)\frac{1}{n}e^{-nd(x,v)},$$

which yields the lower bound in statement 1.

We now prove statement 2. Let x, v such that $\mu_0^- < v \le x < \mu_0^+$. As $[x, \mu_1^+] \subset [v, \mu^+[$, one has

$$\begin{aligned}\pi_{n,x}([v,\mu^+[) &\ge \frac{\int_x^{\mu_1^+} e^{-nd(x,u)} f_0(u)du}{\int_{[\mu^-,x[\cup]\mu_1^+,\mu^+[} e^{-nd(x,u)} f_0(u)du + \int_x^{\mu_1^+} e^{-nd(x,u)} f_0(u)du} \\ &\ge \frac{\int_x^{\mu_1^+} e^{-nd(x,u)} f_0(u)du}{1 + \int_x^{\mu_1^+} e^{-nd(x,u)} f_0(u)du}\end{aligned}$$

Given that $x \mapsto x/(1+x)$ is non-decreasing, we now provide a lower bound on $\int_x^{\mu_1^+} e^{-nd(x,u)} f_0(u)du$.

$F_x : u \mapsto \sqrt{d(x,u)}$ is a one-to-one mapping between $[x, \mu_1^+]$ and $[0, \sqrt{d(x,\mu_1^+)}]$. Moreover, letting $d'(x,u) = \frac{d}{du} d(x,u) = \frac{u-x}{\frac{1}{2}b(b^{-1}(u))} = \frac{u-x}{V(u)}$,

$$F_x'(u) = \frac{d'(x,u)}{2\sqrt{d(x,u)}} \underset{u\to x}{\sim} \frac{(u-x)/V(u)}{2\sqrt{\frac{1}{2}(x-u)^2/V(x)}} \underset{u\to x}{\longrightarrow} \sqrt{\frac{V(x)}{2}}.$$

F_x' is continuous on $[x, \mu_1^+]$ and strictly positive, thus the inverse mapping $\phi_x : [0, \sqrt{d(x,\mu_1^+)}] \to [x, \mu_1^+]$ is well defined and differentiable. Letting $u = \phi_x(y)$, one has

$$\begin{aligned}\int_x^{\mu_1^+} e^{-nd(x,u)} f_0(u)du &= \int_0^{\sqrt{d(x,\mu_1^+)}} e^{-ny^2} f_0(\phi_x(y)) \frac{2\sqrt{d(x,\phi_x(y))}}{d'(x,\phi_x(y))} dy \\ &\ge \inf_{u\in \mathrm{J}_C} f_0(u) \frac{2}{\sqrt{n}} \int_0^{\sqrt{nd(x,\mu_1^+)}} e^{-y^2} \frac{\sqrt{d(x,\phi_x\left(\frac{y}{\sqrt{n}}\right))}}{d'(x,\phi_x\left(\frac{y}{\sqrt{n}}\right))} dy.\end{aligned}$$

The mapping $(x,u) \mapsto \sqrt{d(x,u)}/d'(x,u)$ is continuous and strictly positive on the compact set $\mathcal{S} = \{(x,u) \in [\mu_0^-, \mu_0^+] \times [\mu_1^-, \mu_1^+] : x \le u \le \mu_1^+\}$ therefore, one can define

$$c = \inf_{(x,u)\in\mathcal{S}} \frac{\sqrt{d(x,u)}}{d'(x,u)} > 0.$$

For $n \ge 1$, one has

$$\int_x^{\mu_1^+} e^{-nd(x,u)} f_0(u)du \ge \frac{1}{\sqrt{n}} \left(2c \inf_{u\in \mathrm{J}_C} f_0(u) \int_0^{\sqrt{d(\mu_0^+,\mu_1^+)}} e^{-y^2} dy \right).$$

Thus there exists a constant $C' = C'(\mu_1^-, \mu_1^+, f_0) > 0$ such that

$$\pi_{n,x}([v,\mu^+[) \ge \frac{\frac{C'}{\sqrt{n}}}{1 + \frac{C'}{\sqrt{n}}} = \frac{1}{(1/C')\sqrt{n} + 1},$$

which concludes the proof.

D Finite-time analysis

D.1 Proof of Lemma 8

To upper bound

$$(A) := \mathbb{P}\left(N_1(t)d^+(\hat{\mu}_1(t), \mu_1 - g_t) \geq \log \frac{At\log^c t}{N_1(t)}\right),$$

we consider two cases in which arm 1 has or not been drawn a lot.

$$(A) \leq \underbrace{\mathbb{P}\left(N_1(t)d^+(\hat{\mu}_1(t), \mu_1 - g_t) \geq \log\left(\frac{At\log^c t}{N_1(t)}\right), N_1(t) \leq \log^4(t)\right)}_{A_1}$$
$$+ \underbrace{\mathbb{P}\left(N_1(t)d^+(\hat{\mu}_1(t), \mu_1 - g_t) \geq \log\left(\frac{At\log^c t}{N_1(t)}\right), N_1(t) > \log^4(t)\right)}_{A_2}$$

To upper bound term A_1, we write

$$\begin{aligned}
&\left(N_1(t)d^+(\hat{\mu}_1(t), \mu_1 - g_t) \geq \log\left(\frac{At\log^c t}{N_1(t)}\right), N_1(t) \leq \log^4(t)\right) \\
&\subseteq \left(N_1(t)d^+\left(\hat{\mu}_1(t), \mu_1\right) \geq \log(At) + c\log\log(t) - 4\log\log(t), N_1(t) \leq \log^4(t)\right) \\
&\subseteq \left(N_1(t)d^+\left(\hat{\mu}_1(t), \mu_1\right) \geq \log(At) + 3\log\log t\right),
\end{aligned}$$

using that $c \geq 7$. The self-normalized concentration inequality proved in [12] and stated in Lemma 15 permits to further upper bound A_1 :

$$(A_1) \leq e\frac{\log^2 t + 3(\log t)\log\log(t) + \log(A)\log t + 1}{At\log^3 t}.$$

Lemma 15.

$$\mathbb{P}\left(\exists s \in \{1, \ldots, t\} : sd^+\left(\mu_{1,s}, \mu_1\right) \geq \delta\right) \leq (\delta\log(t) + 1)\exp(-\delta + 1).$$

To upper bound term A_2, if t is such that $\log^7 t \geq A^{-1}$, we write

$$\begin{aligned}
&\left(N_1(t)d^+(\hat{\mu}_1(t), \mu_1 - g_t) \geq \log\left(\frac{At\log^c t}{N_1(t)}\right), N_1(t) \geq \log^4(t)\right) \\
&\subseteq \left(N_1(t)d^+\left(\hat{\mu}_1(t), \mu_1 - g_t\right) \geq 0, N_1(t) \geq \log^4(t)\right) \\
&\subseteq \left(\hat{\mu}_1(t) \leq \mu_1 - g_t, N_1(t) \geq \log^4(t)\right),
\end{aligned}$$

Thus, if t is such that $t \geq \exp(\sqrt{3})$ (which implies $\log^3 t \geq 3\log t$),

$$\begin{aligned}
(A_2) &\leq \mathbb{P}\left(\hat{\mu}_1(t) \leq \mu_1 - g_t, N_1(t) \geq \log^4(t)\right) \leq \mathbb{P}\left(\exists s \in [\lceil\log(t)^4\rceil; t] \ : \ \hat{\mu}_{1,s} \leq \mu_1 - g_t\right) \\
&\leq \sum_{s=\lceil\log(t)^4\rceil}^{t} \mathbb{P}\left(\hat{\mu}_{1,s} \leq \mu_1 - g_t\right) \leq \sum_{s=\lceil\log(t)^4\rceil}^{t} e^{-sd(\mu_1 - g_t, \mu_1)} \\
&\leq te^{-(\log t)^4 d(\mu_1 - g_t, \mu_1)} = te^{-(\log t)^3} \leq te^{-3\log t} = \frac{1}{t^2}.
\end{aligned}$$

Combining this with the upper bound on A_1 yields the result.

D.2 Proof of Lemma 9

The quantity to be upper bounded is

$$(B) := \sum_{s=1}^{T} \mathbb{P}\left(s d^+\left(\hat{\mu}_{a,s}, \mu_1 - g(s)\right) \leq f(T) + h(s)\right).$$

The function $w(q) = d^+(\hat{\mu}_{a,s}, q)$ is convex and differentiable and $w'(q) = \frac{q-\hat{\mu}_{a,s}}{V(q)} \mathbb{1}_{(\hat{\mu}_{a,s} \leq q)}$, thus, if $\hat{\mu}_{a,s}$ is larger than μ_0^-,

$$d^+(\hat{\mu}_{a,s}, \mu_1 - g(s)) \geq d^+(\hat{\mu}_{a,s}, \mu_1) - g(s)\frac{\mu_1 - \hat{\mu}_{a,s}}{V(\mu_1)} \geq d^+(\hat{\mu}_{a,s}, \mu_1) - g(s)\frac{\mu_1 - \mu_0^-}{V(\mu_1)}\,.$$

Therefore

$$\begin{aligned}(B) &\leq \sum_{s=1}^{T} \mathbb{P}\left(d^+(\hat{\mu}_{a,s}, \mu_1) \leq \frac{f(T)}{s} + g(s)\frac{\mu_1 - \mu_0^-}{V(\mu_1)} + \frac{h(s)}{s}\right) + \sum_{s=1}^{T} \mathbb{P}(\hat{\mu}_{a,s} < \mu_0^-) \\ &\leq \sum_{s=1}^{T} \mathbb{P}\left(d^+(\hat{\mu}_{a,s}, \mu_1) \leq \frac{f(T)}{s} + r(s)\right) + \frac{1}{1 - e^{-d(\mu_0^-, \mu_a)}},\end{aligned}$$

using Chernoff inequality and introducing

$$r(s) := g(s)\frac{\mu_1 - \mu_0^-}{V(\mu_1)} + \frac{h(s)}{s}.$$

Let $\varepsilon > 0$. One also introduce

$$K_T(\varepsilon) := \left\lceil \frac{(1+\varepsilon) f(T)}{d(\mu_a, \mu_1)} \right\rceil.$$

From the assumptions on f, g and h, there exists s_0 such that r is non-increasing for $s \geq s_0$ and one has

$$K_T(\varepsilon) \underset{T \to \infty}{\longrightarrow} \infty \quad \text{and} \quad r(s) \underset{s \to \infty}{\longrightarrow} 0.$$

For T such that $K_T \geq s_0$,

$$(B) \leq K_T + \sum_{s=K_T+1}^{T} \mathbb{P}\left(d^+(\hat{\mu}_{a,s}, \mu_1) \leq \frac{f(T)}{s} + r(K_T)\right) + C_a,$$

with $C_a = 1/\left(1 - e^{-d(\mu_0^-, \mu_a)}\right)$. As $r(K_T) \to 0$, there exists $N_a(\varepsilon)$ such that

$$T \geq N_a(\varepsilon) \;\Rightarrow\; r(K_T) \leq d(\mu_a, \mu_1)\frac{\varepsilon}{1+\varepsilon}.$$

Then, if $T \geq N_a(\varepsilon)$, one has, for all $s \geq K_T + 1$,

$$\frac{f(T)}{s} + r(K_T) \leq d(\mu_a, \mu_1)$$

and there exists $\mu^*(s) \in]\mu_a; \mu_1[$ such that $d(\mu^*(s), \mu_1) = \frac{f(T)}{s} + r(K_T)$. Then, using Chernoff inequality and the inequality

$$\forall \mu > \mu', \;\; d(\mu, \mu') \geq \frac{1}{2\sup_{\mu \in [\mu', \mu]} V(\mu)}(\mu - \mu')^2,$$

stated in [12] and that follows from Lagrange equality, one can write

$$\begin{aligned}(B) &\leq K_T + \sum_{s=K_T+1}^{T} \mathbb{P}(\hat{\mu}_{a,s} > \mu^*(s)) + C_a \leq K_T + \sum_{s=K_T+1}^{T} e^{-sd(\mu^*(s),\mu_a)} + C_a \\ &\leq K_T + \sum_{s=K_T+1}^{T} e^{-s\frac{(\mu^*(s)-\mu_a)^2}{2V_a^2}} + C_a \leq K_T + \int_{K_T}^{\infty} e^{-s\frac{(\mu^*(s)-\mu_a)^2}{2V_a^2}} ds + C_a,\end{aligned}$$

where $V_a = \sup_{\mu\in]\mu_a,\mu_1[} V(\mu)$. Using the convexity of $x \mapsto d(x,\mu_1)$, a lower bound on $\mu^*(s) - \mu_a$ can be obtained, as in Appendix 2 of [12] :

$$\mu^*(s) - \mu_a \geq \frac{d(\mu_a,\mu_1) - \left[\frac{f(T)}{s} + r(K_T)\right]}{-d'(\mu_a,\mu_1)}$$

[12] provides a tight bound on the resulting integrals, and using a similar approach concludes the proof :

$$\begin{aligned}&\int_{K_T}^{\infty} e^{-s\frac{(\mu^*(s)-\mu_a)^2}{2V_a^2}} ds \leq \int_{K_T}^{\infty} \exp\left(-\frac{s}{2V_a^2 d'(\mu_a,\mu_1)^2}\left(d(\mu_a,\mu_1) - \left(\frac{f(T)}{s} + r(K_T)\right)\right)^2\right) ds \\ &\leq f(T)\int_{\frac{1+\varepsilon}{d(\mu_a,\mu_1)}}^{\infty} \exp\left(-\frac{uf(T)}{2V_a^2 d'(\mu_a,\mu_1)^2}\left(d(\mu_a,\mu_1) - \left(\frac{1}{u} + r(K_T)\right)\right)^2\right) du \\ &\leq f(T)\int_{\frac{1+\varepsilon}{d(\mu_a,\mu_1)}}^{\frac{2(1+\varepsilon)}{d(\mu_a,\mu_1)}} \exp\left(-\frac{(1+\varepsilon)\left(d(\mu_a,\mu_1) - \left(\frac{1}{u} + r(K_T)\right)\right)^2}{2V_a^2 d(\mu_a,\mu_1) d'(\mu_a,\mu_1)^2} f(T)\right) du \\ &\quad + f(T)\int_{\frac{2(1+\varepsilon)}{d(\mu_a,\mu_1)}}^{\infty} \exp\left(-\frac{uf(T)}{2V_a^2 d'(\mu_a,\mu_1)^2}\frac{d(\mu_a,\mu_1)^2}{4(1+\varepsilon)^2}\right) du \\ &\leq f(T)\frac{4(1+\varepsilon)^2}{d(\mu_a,\mu_1)^2}\int_0^{\infty} \exp\left(-\frac{(1+\varepsilon)v^2 f(T)}{2V_a^2 d(\mu_a,\mu_1) d'(\mu_a,\mu_1)^2}\right) dv + 8(1+\varepsilon)^2 V_a^2 \left(\frac{d'(\mu_a,\mu_1)}{d(\mu_a,\mu_1)}\right)^2 \\ &\leq \sqrt{f(T)}\sqrt{\frac{8V_a^2\pi(1+\varepsilon)^3 d'(\mu_a,\mu_1)^2}{d(\mu_a,\mu_1)^3}} + 8(1+\varepsilon)^2 V_a^2 \left(\frac{d'(\mu_a,\mu_1)}{d(\mu_a,\mu_1)}\right)^2.\end{aligned}$$