The Annals of Statistics
2018, Vol. 46, No. 2, 842–865
https://doi.org/10.1214/17-AOS1569

ON BAYESIAN INDEX POLICIES FOR SEQUENTIAL RESOURCE ALLOCATION

BY EMILIE KAUFMANN[1]

CNRS, Univ. Lille, Inria SequeL, UMR 9189—CRIStAL—Centre de Recherche en Informatique Signal et Automatique de Lille

This paper is about index policies for minimizing (frequentist) regret in a stochastic multi-armed bandit model, inspired by a Bayesian view on the problem. Our main contribution is to prove that the Bayes-UCB algorithm, which relies on quantiles of posterior distributions, is asymptotically optimal when the reward 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 provide a justification for the kl-UCB$^+$ and kl-UCB-H$^+$ algorithms, whose asymptotic optimality is also established.

1. Introduction. This paper presents 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, $\mathcal{V}_1, \dots, \mathcal{V}_K$, 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 $\mathcal{V}_{A_t}$, often interpreted as a reward. Indeed, the arms can be viewed as those of different slot machines, also called *one-armed bandits*, generating rewards according to some underlying probability distribution.

In several applications that range from the motivating example of clinical trials [38] to the more modern motivation of online advertisement (e.g., [16]), the goal of the agent is to adjust his strategy $\mathcal{A} = (A_t)_{t\in\mathbb{N}}$, also called a *bandit algorithm*, in order to maximize the rewards accumulated during his interaction with the bandit model. The adopted strategy has to be sequential, in the sense that the next arm to play is chosen based on past observations: letting $\mathcal{F}_t = \sigma(A_1, X_1, \dots, A_t, X_t)$

Received September 2016; revised March 2017.

[1] Supported by the French Agence Nationale de la Recherche (ANR) under grants ANR-13-BS01-0005 (SPADRO project) and ANR-11-JS02-005-01 (GAP project).

MSC2010 subject classifications. 62L05.

Key words and phrases. Multi-armed bandit problems, Bayesian methods, upper-confidence bounds, Gittins indices.

be the σ-field generated by the observations up to round t, A_t is $\sigma(\mathcal{F}_{t-1}, U_t)$-measurable, where U_t is a uniform random variable independent from $\mathcal{F}_{t-1}$ (as algorithms may be randomized).

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 minimizing the *regret*, defined as the expected difference between the reward accumulated by an oracle strategy always playing the best arm, and the reward accumulated by a strategy $\mathcal{A}$:

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

The expectation is taken with respect to the randomness in the sequence of successive rewards from each arm a, denoted by $(Y_{a,s})_{s\in\mathbb{N}}$, and the possible randomization of the algorithm, $(U_t)_t$. We denote by $N_a(t) = \sum_{s=1}^{t} \mathbb{1}_{(A_s=a)}$ the number of draws from arm a at the end of round t, so that $X_t = Y_{A_t, N_{A_t}(t)}$.

This paper focuses on good strategies in *parametric* bandit models, in which the distribution of arm a depends on some parameter θ_a: we write $\mathcal{V}_a = \nu_{\theta_a}$. 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. In the Bayesian view, $\boldsymbol{\theta}$ is a random variable, drawn from a prior distribution Π. More precisely, we define $\mathbb{P}_{\boldsymbol{\theta}}$ (resp., $\mathbb{E}_{\boldsymbol{\theta}}$) the probability (resp., expectation) under the probabilistic model in which for all a, $(Y_{a,s})_{s\in\mathbb{N}}$ is i.i.d. distributed under ν_{θ_a} and $\mathbb{P}^{\Pi}$ (resp., $\mathbb{E}^{\Pi}$) the probability (resp., expectation) under the probabilistic model in which for all a $(Y_{a,s})_{s\in\mathbb{N}}$ is i.i.d. conditionally to θ_a with conditional distribution ν_{θ_a}, and $\boldsymbol{\theta} \sim \Pi$. The expectation in (1) can thus be taken under either of these two probabilistic models. In the first case, this leads to the notion of frequentist regret, which depends on $\boldsymbol{\theta}$:

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

In the second case, this leads to the notion of Bayesian regret, sometimes called *Bayes risk* in the literature (see [27]), which depends on the prior distribution Π:

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

The first bandit strategy was introduced by Thompson in 1933 [38] in a Bayesian framework, and a large part of the early work on bandit models is adopting the same perspective [7, 8, 10, 19]. Indeed, as Bayes risk minimization has an *exact*—yet often intractable—solution, finding ways to efficiently compute this solution has been an important line of research. Since 1985 and the seminal work of Lai and Robbins [28], there is also a precise characterization of good bandit algorithms in a frequentist sense. They show that for any *uniformly efficient policy* $\mathcal{A}$

(i.e., such that for all $\boldsymbol{\theta}$, $\mathrm{R}_{\boldsymbol{\theta}}(T, \mathcal{A}) = o(T^{\alpha})$ for all $\alpha \in]0, 1]$), the number of draws of any suboptimal 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 yields 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, aimed at building *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 (UCB). The design of the confidence intervals has been successively improved [1, 4–6, 14, 21, 27] so as to obtain simple index policies for which nonasymptotic upper bound on the regret can be given. Among them, the kl-UCB algorithm [14] matches the lower bound (4) for exponential family bandit models. As they use confidence intervals on unknown parameters, all these index policies are based on *frequentist tools*. Nevertheless, it is interesting to note that the first index policy was introduced by Gittins in 1979 [19] to solve a Bayesian multi-armed bandit problem and is based on *Bayesian tools*, that is, on exploiting the posterior distribution on the parameter of each arm.

However, tools and objectives can 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, in particular the Bayes-UCB algorithm [24], which is based on quantiles of the posterior distributions on the means. Our main contribution is to prove that Bayes-UCB is asymptotically optimal, that is, it matches the lower bound (4), for any exponential bandit model and for a large class of prior distributions. Our analysis relies on two new ingredients: tight bounds on the tail of posterior distributions (Lemma 4), and a self-normalized deviation inequality featuring an exploration rate that decreases with the number of observations (Lemma 5). This last tool also allows 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.

The paper is structured as follows. Section 2 introduces the class of exponential family bandit models that we consider in the rest of the paper, and the associated frequentist and Bayesian tools. In Section 3, we present the Bayes-UCB algorithm, and give a proof of its asymptotic optimality. We introduce kl-UCB$^+$ and kl-UCB-H$^+$ in Section 4, in which we prove their asymptotic optimality and also exhibit connections with existing Bayesian policies. In Section 5, we illustrate numerically the good performance of our three asymptotically optimal, Bayesian-flavored index policies in terms of regret. We also investigate their ability to attain an optimal rate in terms of Bayes risk. Some proofs are provided in the Supplemtary Material [23].

1.1. *Notation.* Recall that $N_a(t) = \sum_{s=1}^{t} \mathbb{1}_{(A_s=a)}$ is the number of draws from arm a at the end of round 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 $\hat{\mu}_a(t) = 0$ if $N_a(t) = 0$, $\hat{\mu}_a(t) = \hat{\mu}_{a,N_a(t)}$ otherwise.

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 family bandit model.* A one-parameter canonical exponential family is a set $\mathcal{P}$ of probability distributions, indexed by a real parameter θ 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 $\Theta = (\theta^-, \theta^+) \subseteq \mathbb{R}$ 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 include Bernoulli distributions, Gaussian distributions with known variance, Poisson distributions or Gamma distributions with known shape parameter.

If $X \sim \nu_\theta$, it can be shown that $\mathbb{E}[X] = \dot{b}(\theta)$ and $\mathrm{Var}[X] = \ddot{b}(\theta) > 0$, where $\dot{b}$ (resp., $\ddot{b}$) is the derivative (resp., second derivative) of b with respect to the natural parameter θ. 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 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 $\mathrm{V}(\mu)$ of the distribution ν^μ is related to its mean in the following way:

$$\mathrm{V}(\mu) = \ddot{b}(\dot{b}^{-1}(\mu)). \tag{5}$$

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. When considering Bayesian bandit models, we restrict our attention to product prior distributions on $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$, such that 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. We let π_a^t be the posterior distribution on μ_a after the first t rounds of the bandit game. With a slight abuse of notation, we will identify π_a^t with its density, for which a more precise expression is provided in Section 2.3.

2.2. *Kullback–Leibler divergence and confidence intervals.* For distributions that belong to a one-parameter exponential family, the large deviation rate function has a simple and explicit form, featuring the Kullback–Leibler (KL) divergence, and one can build tight confidence intervals on their means. The KL-divergence between two distributions ν_θ and ν_λ in an exponential family has a closed-form expression as a function of the natural parameters θ and λ, given by

$$\mathrm{K}(\theta,\lambda) := \mathrm{KL}(\nu_\theta,\nu_\lambda) = \dot{b}(\theta)(\theta-\lambda) - b(\theta) + b(\lambda). \tag{6}$$

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

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

Applying the Cramér–Chernoff method (see, e.g., [9]) in an exponential family yields an explicit deviation inequality featuring this divergence function: if $\hat{\mu}_s$ is the empirical mean of s samples from ν^{μ} and $x > \mu$, one has $\mathbb{P}(\hat{\mu}_s > x) \leq \exp(-sd(x,\mu))$. This inequality can be used to build a confidence interval for μ based on a *fixed number of observations* s. Inside a bandit algorithm, computing a confidence interval on the mean of an arm a requires to take into account the *random number of observations* $N_a(t)$ available at round t. Using a self-normalized deviation inequality (see [14] and references therein), one can show that, at any round t of a bandit game, the kl-UCB index, defined as

$$u_a(t) := \sup\{q \in \mathrm{J} : N_a(t)d(\hat{\mu}_a(t),q) \leq \log(t\log^c(t))\}, \tag{7}$$

where $c \geq 3$ is a real parameter, satisfies $\mathbb{P}(u_a(t) > \mu_a) \gtrsim 1 - 1/(t\log^{c-2} t)$ and is thus an upper confidence bound on μ_a. The *exploration rate*, which is here $\log(t\log^c(t))$, controls the coverage probability of the interval.

Closed-form expressions for the divergence function d in the most common examples of exponential families are available (see [14]). Using the fact that $y \mapsto d(x,y)$ is increasing when $y > x$, an approximation of $u_a(t)$ can then be obtained using, for example, binary search.

2.3. *Posterior distributions in Bayesian exponential family bandits.* It is well known that the posterior distribution on the mean of a distribution that belongs to an exponential family depends on two sufficient statistics: the number of observations and the empirical means of these observations. With f_a the density of the prior distribution on μ_a, introducing

$$\pi_{a,n,x}(u) := \frac{\exp(n[\dot{b}^{-1}(u)x - b(\dot{b}^{-1}(u))])f_a(u)}{\int_{\mathrm{J}}\exp(n[\dot{b}^{-1}(u)x - b(\dot{b}^{-1}(u))])f_a(u)\,du} \qquad \text{for } u \in \mathrm{J},$$

the density of the posterior distribution on μ_a after t rounds of the bandit game can be written

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

While our analysis holds for any choice of prior distribution, in practice one may want to exploit the existence of families of conjugate priors (e.g., Beta distributions for Bernoulli rewards, Gaussian distributions for Gaussian rewards, Gamma distributions for Poisson rewards). With a prior distribution chosen in such a family, the associated posterior distribution is well known and its quantiles are easy to compute, which is of particular interest for the Bayes-UCB algorithm, described in the next section.

Finally, we give below a rewriting of the posterior distribution that will be very useful in the sequel to obtain tight bounds on its tails.

LEMMA 1.

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

PROOF. Let $u \in \mathrm{J}$. One has

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

using the closed-form expression (6) and the fact that $\theta = \dot{b}^{-1}(\mu)$. □

3. Bayes-UCB: A simple and optimal Bayesian index policy.

3.1. *Algorithm and main result.* The Bayes-UCB algorithm is an index policy that was introduced by [24] 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. For exponential family bandit models, given a product prior distribution on the means, the Bayes-UCB index is

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

where $Q(\alpha;\pi)$ is the quantile of order α of the distribution π [i.e., $\mathbb{P}_{X\sim\pi}(X \leq Q(\alpha;\pi)) = \alpha$] and c is a real parameter. In the particular case of bandit models with Gaussian arms, [33] have introduced a variant of Bayes-UCB with a slightly different tuning of the confidence level, under the name UCL (for Upper Credible Limit).

While the efficiency of Bayes-UCB has been demonstrated even beyond bandit models with independent arms, regret bounds are available only in very limited cases. For Bernoulli bandit models asymptotic optimality is established by [24] when a uniform prior distribution on the mean of each arm is used. For Gaussian bandit models, [33] give a logarithmic regret bound when an uninformative prior is used. In this section, we provide new finite-time regret bounds that hold in 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 let $\mu^- = \dot{b}(\theta^-)$ and $\mu^+ = \dot{b}(\theta^+)$ (μ^- may be equal to $-\infty$ and μ^+ to $+\infty$). We analyze Bayes-UCB for exponential bandit models satisfying the following assumption.

ASSUMPTION 2. There exist $\mu_0^- > \mu^-$ and $\mu_0^+ < \mu^+$ such that $\forall a \in \{1, \dots, K\}, \mu_0^- \leq \mu_a \leq \mu_0^+$.

For Poisson or exponential distributions, this assumption requires that the means of all arms are different from zero, while they should be included in]0, 1[for Bernoulli distributions. We now introduce a regularized version of the Bayes-UCB index that relies on the knowledge of μ_0^- and μ_0^+, as

$$\overline{q}_a(t) := Q\left(1 - \frac{1}{t(\log t)^c}; \pi_{a,N_a(t),\bar{\mu}_a(t)}\right), \tag{8}$$

where $\bar{\mu}_a(t) = \min(\max(\hat{\mu}_a(t), \mu_0^-), \mu_0^+)$. Note that μ_0^- and μ_0^+ can be chosen arbitrarily close to μ^- and μ^+, respectively, in which case $\overline{q}_a(t)$ often coincides with the original Bayes-UCB index $q_a(t)$.

THEOREM 3. *Let $\nu^{\boldsymbol{\mu}}$ be an exponential bandit model satisfying Assumption 2. 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 $c \geq 7$. 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 (8) satisfies, for all $\varepsilon > 0$,

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

From Theorem 3, taking the lim sup and letting ϵ go to zero show that (this slight variant of) Bayes-UCB satisfies

$$\forall a \neq a^*, \qquad \limsup_{T \to \infty} \frac{\mathbb{E}[N_a(T)]}{\log(T)} \leq \frac{1}{d(\mu_a, \mu^*)}.$$

Thus this index policy is asymptotically optimal, as it matches Lai and Robbins' lower bound (4). As we shall see in Section 5, from a practical point of view Bayes-UCB outperforms kl-UCB and performs similarly (sometimes slightly better, sometimes slightly worse) as Thompson Sampling, another popular Bayesian algorithm that we now discuss.

3.2. *Posterior quantiles versus posterior samples.* Over the past few years, another Bayesian algorithm, Thompson Sampling, has become increasingly popular for its good empirical performance, and we explain how Bayes-UCB is related to this alternative, randomized, Bayesian approach.

The Thompson Sampling algorithm, that draws each arm according to its posterior probability of being optimal, was introduced in 1933 as the very first bandit algorithm [38] and re-discovered recently for its good empirical performance [16, 36]. Thompson Sampling can be implemented in virtually any Bayesian bandit model in which one can sample the posterior distribution, by drawing *one* sample from the posterior on each arm and selecting the arm that yields the largest sample. In any such case, Bayes-UCB can be implemented as well and may appear as a more robust alternative as the quantiles can be estimated based on *several* samples in case there is no efficient algorithm to compute them.

Our experiments of Section 5 show that Bayes-UCB as well as the other Bayesian-flavored index policies presented in Section 4 are competitive with Thompson Sampling in general one-dimensional exponential families. Compared to Bayes-UCB, the theoretical understanding of Thompson Sampling is more limited: this algorithm is known to be asymptotically optimal in exponential family bandit models, yet only for specific choices of prior distributions [3, 25, 26].

In more complex bandit models, there are situations in which Bayes-UCB is indeed used over Thompson Sampling. When there is a potentially infinite number of arms and the mean reward function is assumed to be drawn from a Gaussian Process, the GP-UCB of [37], that coincides with Bayes-UCB, is very popular in the Bayesian optimization community [11].

3.3. *Tail bounds for posterior distributions.* Just like the analysis of [24], 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 [14].

In the particular case of Bernoulli bandit models, Bayes-UCB uses quantiles of Beta posterior distributions. In that case a specific argument, namely the fact that Beta(a, b) is the distribution of the ath order statistic among $a + b - 1$ uniform random variables, relates 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, 3, 25]. Note that this argument can only be used for Beta distributions with integer parameters, which rules out many possible prior distributions. The analysis of [33] in the Gaussian case also relies on specific tails bounds for the Gaussian posterior distributions. For exponential family bandit models, an upper bound on the tail of the posterior distribution was obtained by [26] using the Jeffreys prior.

Lemma 4 below presents 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 [26], which is not presented in this deterministic way, Lemma 4 is based on a different rewriting of the posterior distribution, given in Lemma 1.

LEMMA 4. *Let μ_0^-, μ_0^+ be defined in Assumption 2*:

1. *There exist 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 $a \in \{1, \dots, K\}$,*

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

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 $a \in \{1, \dots, K\}$,*

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

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

This result permits in particular to show that the quantile $\overline{q}_a(t)$ defined in (8) satisfies $\underline{\mathrm{U}}_a(t) \leq \overline{q}_a(t) \leq \overline{\mathrm{U}}_a(t)$, with

$$\underline{\mathrm{U}}_a(t) = \sup\bigl\{q < \mu_0^+ : N_a(t)d\bigl(\overline{\mu}_a(t), q\bigr) \leq \log\bigl(\bigl(At\log^c(t)\bigr)/N_a(t)\bigr)\bigr\},$$
$$\overline{\mathrm{U}}_a(t) = \sup\bigl\{q < \mu_0^+ : N_a(t)d\bigl(\overline{\mu}_a(t), q\bigr) \leq \log\bigl(Bt\log^c(t)\sqrt{N_a(t)}\bigr)\bigr\}.$$

Hence, despite their Bayesian nature, the indices used in Bayes-UCB are strongly related to frequentist kl-UCB type indices. However, compared to the index $u_a(t)$ defined in (7), the exploration rate that appears in $\underline{\mathrm{U}}_a(t)$ and $\overline{\mathrm{U}}_a(t)$ also features the current number of draws $N_a(t)$. Lai [27] gives an asymptotic analysis of any index strategy of the above form with an exploration function $g(T/N_a(t))$, where $g(t) \sim \log(t)$ when t goes to infinity. Yet neither $\underline{\mathrm{U}}_a(t)$ nor $\overline{\mathrm{U}}_a(t)$ are not exactly of that form, and we propose below a finite-time analysis that relies on new, nonasymptotic tools.

3.4. *Finite-time analysis*. We give here the proof of Theorem 3. To ease the notation, assume that arm 1 is an optimal arm, and let a be a suboptimal arm:

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

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 to 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

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

This yields

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

The posterior bounds established in Lemma 4 permit to further upper bound the two sums in the right-hand side of the above inequality. With C defined in Lemma 4, we introduce t_0, defined by

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

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

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

since by the lower bound in the second statement of Lemma 4,

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

Now using the lower bound in the first statement of Lemma 4,

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

On the other hand,

$$
\begin{aligned}
&\sum_{t=K}^{T-1}\mathbb{P}(\mu_1 - g_t \leq \bar{q}_a(t), A_{t+1}=a)\\
&\quad = \sum_{t=K}^{T-1}\mathbb{P}\biggl(\pi_{a,N_a(t),\bar{\mu}_a(t)}\bigl([\mu_1 - g_t, \mu^+[\bigr) \geq \frac{1}{t\log^c t}, A_{t+1}=a\biggr)\\
&\quad \leq \sum_{t=K}^{T-1}\mathbb{P}\biggl(\bar{\mu}_a(t) < \mu_1 - g_t,\\
&\qquad \pi_{a,N_a(t),\bar{\mu}_a(t)}\bigl([\mu_1 - g_t, \mu^+[\bigr) \geq \frac{1}{t\log^c t}, A_{t+1}=a\biggr)\\
&\qquad + \sum_{t=K}^{T-1}\mathbb{P}\bigl(\bar{\mu}_a(t) \geq \mu_1 - g_t, A_{t+1}=a\bigr).
\end{aligned}
\tag{9}
$$

Using Lemma 4, the first sum in (9) is upper bounded by

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

To third inequality follows from exchanging the sums over s and t and using that $\sum_{t=1}^{N}\mathbb{1}_{(N_a(t)=s)\cap(A_{t+1}=a)}$ is smaller than 1 for all s. The last inequality uses that if $\hat{\mu}_{a,s} \geq \mu_0$, $\overline{\mu}_{a,s} \leq \hat{\mu}_{a,s}$ and $d^+(\overline{\mu}_{a,s}, \mu_1 - g_s) \geq d^+(\hat{\mu}_{a,s}, \mu_1 - g_s)$. Then 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 (9) is upper bounded by

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

Putting things together, we showed that there exists some constant $N = \max(t_0, N_0 + (1 - e^{-d(\mu_0^-, \mu_a)})^{-1}) + 1$ such that

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

Term T_1 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 5 more intricate. In particular, Lemma 5 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 6 below.

LEMMA 5. *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 is larger than* $\exp(\max(\sqrt{3}, A^{-1/7}))$,

$$\begin{aligned} &\mathbb{P}\biggl(N_1(t) d^+(\hat{\mu}_1(t), \mu_1 - g_t) \ge \log \frac{At \log^c t}{N_1(t)}\biggr) \\ &\qquad \le e\biggl(\frac{1}{At \log t} + \frac{3 \log\log t + \log A}{At \log^2 t} + \frac{1}{At \log^3 t}\biggr) + \frac{1}{t^2}. \end{aligned}$$

From Lemma 5, 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 6. *Let f, g, h be three functions such that*

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

with g and $s\mapsto h(s)/s$ nonincreasing for s large enough.

For all $\varepsilon>0$, there exists a (problem-dependent) constant $N_a(\varepsilon)$ such that for all $T\ge N_a(\varepsilon)$,

$$\begin{aligned}&\sum_{s=1}^{T}\mathbb{P}\big(sd^+(\hat{\mu}_{a,s},\mu_1-g(s))\le f(T)+h(s)\big)\\ &\qquad\le\frac{1+\varepsilon}{d(\mu_a,\mu_1)}f(T)+\sqrt{f(T)}\sqrt{\frac{8\mathrm{V}_a^2\pi(1+\varepsilon)^3d'(\mu_a,\mu_1)^2}{d(\mu_a,\mu_1)^3}}\\ &\qquad\quad+8(1+\varepsilon)^2\mathrm{V}_a^2\left(\frac{d'(\mu_a,\mu_1)}{d(\mu_a,\mu_1)}\right)^2\frac{1}{1-e^{-d(\mu_0^-,\mu_a)}}+1,\end{aligned}$$

with $\mathrm{V}_a=\sup_{\mu\in[\mu_a,\mu_1]}\mathrm{V}(\mu)$, where the variance function is defined in (5).

Let $\varepsilon>0$. Using Lemma 6, with $f(s)=\log(s)+c\log\log(s)+\log(B)$, $g(s)=g_s$ defined in Lemma 5 and $h(s)=\frac{1}{2}\log(s)$, there exists problem dependent constants C_0 and $D_0(\varepsilon)$ such that

$$(T_2)\le\frac{1+\varepsilon}{d(\mu_a,\mu_1)}(\log T+c\log\log T)+C_0\sqrt{\log T+c\log\log T}+D_0(\varepsilon).$$

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

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

4. A Bayesian insight on alternative exploration rates. The kl-UCB index of an arm, $u_a(t)$, introduced in (7), uses the exploration rate $\log(t \log^c(t))$, that does not depend on arm a. Some alternatives to this universal exploration rate have been suggested in the literature, and we formally introduce two variants of kl-UCB, called kl-UCB$^+$ and kl-UCB-H$^+$ using an exploration rate that decreases with the number of draws of arm a. The tools developed for the analysis of Bayes-UCB allow us to prove the asymptotic optimality of both algorithms. We then show that the Bayesian literature on the multi-armed bandit problem provides a natural justification for these algorithms that are related to approximations of the Bayesian optimal optimal solution or the Gittins indices.

4.1. *The kl-UCB$^+$ and kl-UCB-H$^+$ algorithms.* We introduce in Definition 7 two new index policies, and prove their asymptotic optimality. The indices indices $u_a^{H,+}(t)$ and $u_a^+(t)$ both rely on an exploration rate that decreases with the number of plays of arm a. kl-UCB-H$^+$ additionally requires the knowledge of the horizon T. In practice, both algorithms outperform kl-UCB, as can be seen in Section 5.

DEFINITION 7. Let $c \geq 0$. We define kl-UCB-H$^+$ and kl-UCB$^+$ with parameter $c \geq 0$ as the index policies respectively based on the indices:

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

$$u_a^{+}(t) = \sup\left\{q : N_a(t)d(\hat{\mu}_a(t), q) \leq \log\left(\frac{t \log^c t}{N_a(t)}\right)\right\}. \tag{11}$$

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

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

in which an exploration rate of order $\log(t/N_1(t))$ appears. This control is obtained in Lemma 5, which can also be used to analyze the kl-UCB-H$^+$ and kl-UCB$^+$ algorithms that are based on such alternative exploration rates. The following theorem proves the asymptotic optimality of these two index policies. Its proof is provided in Appendix B of [23].

THEOREM 8. *Let $c \geq 7$. Each of the index policy associated to the indices defined by* (11) *and* (10) *satisfies, for all $\varepsilon > 0$,*

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

The use of alternative exploration rates in UCB-type algorithms has appeared before in the bandit literature, for example, the MOSS algorithm [4], based on the index

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

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 already noted by [17] 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. In this paper, additionally to proving the asymptotic optimality of these approaches, we now provide a new insight on the use of such alternative exploration rates by relating the kl-UCB-H$^+$ algorithm to other Bayesian policies.

4.2. *Bayesian optimal solution and Gittins indices.* The alternative exploration rate discussed in Section 4.1 happens to be related to two other Bayesian strategies for the multi-armed bandit problem: the Bayesian optimal solution and the Finite-Horizon Gittins index policy that we present here.

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, $X_t \sim \nu^{\mu}$ and $\mu \sim \Pi_t(A_t)$. 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

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

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

$$\mathbb{E}^{\Pi}\left[\sum_{t=1}^{T} X_t^g\right], \tag{12}$$

where $(X_t^g)_t$ is the sequence of rewards obtained under policy g. From the theory of MDPs (see, e.g., [32]), 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 [19] 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 maximizing (12) 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. This definition transposes to the nondiscounted case one of the equivalent definitions of the discounted Gittins index that can be found in [20].

DEFINITION 9. 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

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

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 (using, e.g., binary search). Each computation requires solving a MDP, but on a smaller state space: the possible posterior distributions on the mean of a single arm. Hence, the FH-Gittins algorithm, that is, the index policy based on the Finite-Horizon Gittins indices,

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

is a more practical algorithm than the Bayesian optimal solution. Although FH-Gittins 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, Ginebra and Clayton [18] propose a comparison of

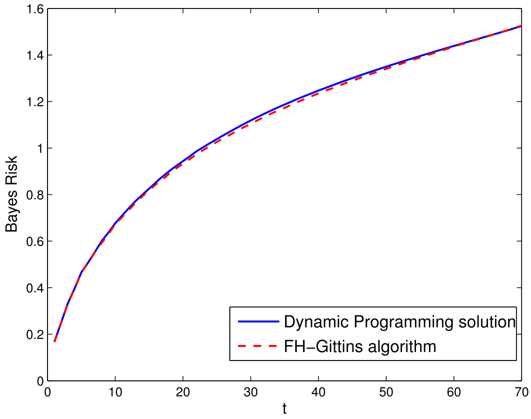

FIG. 1. *Bayes risk of the optimal strategy* (*blue*) *and FH-Gittins* (*dashed red*) *estimated using* $N = 10^6$ *replications of a bandit game, for which the means are drawn from* $\mathcal{U}([0, 1])$.

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.

Compared to a simple index policy like Bayes-UCB, the computational cost of the FH-Gittins algorithm (not to mention that of the Bayesian optimal strategy) is still very high. In particular, the complexity of these two approaches grows dramatically when the horizon T increases, which motivates some approximations that have been proposed for large horizons, described in the next sections.

However, when the FH-Gittins algorithm is efficiently implementable (i.e., for relatively small horizons), we would like to advocate its use for minimizing the frequentist regret. Indeed our experiments of Section 5 report good empirical performance in Bernoulli bandit models. In this particular case, using a uniform prior on the means, the set of (Beta) posterior is parametrized by two integers (the number of zeros and ones observed so far), and we could implement FH-Gittins up to horizon $T = 1000$. An efficient implementation of FH-Gittins for Gaussian bandits, up to horizon $T = 10{,}000$, has been recently given by [29]. More generally, finding efficient methods to compute Finite-Horizon Gittins indices is still an area of investigation [31]. Interestingly, [29] provides the first theoretical elements supporting the use of FH-Gittins for regret minimization, by giving the first logarithmic upper bound on its regret in the particular case of Gaussian bandit models. However, the asymptotic optimality of this algorithm for Gaussian bandits and more general models remains a conjecture.

4.3. *Approximation of the Bayesian optimal solution.* In the paper [27], 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, which implies in particular that the Bayes risk of the Bayesian optimal solution is of order $\log^2(T)$. Any strategy matching this lower bound can be viewed as an asymptotic approximation of the Bayesian optimal solution.

In the particular case of product prior distributions, we provide in Theorem 10 a Bayes risk lower bound 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 D of [23], follows however closely that of [27]. The lower bound is expressed in terms of the prior distribution on the natural parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ of the arms, with the following 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 10. *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} be 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 assumes 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 $[\mu_0^-, \mu_0^+] = \dot{b}(\Theta)$, he shows in particular that the index policy based on

$$I_a(t) = \sup\left\{q \in [\mu_0^-, \mu_0^+] : N_a(t)\overline{d}(\hat{\mu}_a(t), q) \leq \log\left(\frac{T}{N_a(t)}\right)\right\}, \tag{13}$$

where $\overline{d}(x, y) = d(\max(\mu_0^-, \min(\mu_0^+, x)), y)$, has a Bayes risk that asymptotically matches the lower bound of Theorem 10. This index policy is very similar to kl-UCB-H$^+$ and differs only from the use of a regularized version of the divergence function d.

While a recent line of research on Bayesian randomized algorithms (e.g., Thompson Sampling) has provided Bayes risk upper bounds in quite general settings ([34, 35]), 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. Bubeck and Liu [12] and Liu and Li [30] give 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. The joint prior distribution is thus supported on (μ_1, μ_2) and (μ_2, μ_1). In Section 5.2, we investigate numerically the optimality of the Bayesian index policies discussed in the paper with respect to the lower bound of Theorem 10.

4.4. *Approximation of the Finite-Horizon Gittins indices.* As discussed Section 4.2, the FH-Gittins algorithm, 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 9 that the Finite-Horizon Gittins index takes the form

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

where $V_\lambda^*(\pi, r)$ corresponds to the optimal value function associated to a calibration game $\mathcal{C}_\lambda$. In the paper [13], 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 $T - t$,

$$J_a(t) \simeq \sup\left\{q \in [\mu^-, \mu^+] : N_a(t)\tilde{d}(\hat{\mu}_a(t), q) \leq \log\left(\frac{T - t}{N_a(t)}\right)\right\}. \tag{14}$$

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 [15] 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 (14): 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 nonasymptotic lower bound on the Gaussian Gittins index obtained by [29].

These approximations of the Finite-Horizon Gittins indices provide another justification for exploration rates of the form $\log(h(t, T)/N_a(t))$, with some function h, which are also used by the kl-UCB-H$^+$ and kl-UCB$^+$ algorithms. These two algorithms can thus be viewed as Bayesian (inspired) index policies.

5. Numerical experiments.

5.1. *Regret minimization.* We first perform experiments with a moderate horizon $T = 1000$, which permits to include the Finite-Horizon Gittins algorithm discussed in Section 4.2. Figure 2 displays the regret of kl-UCB, Thompson Sampling

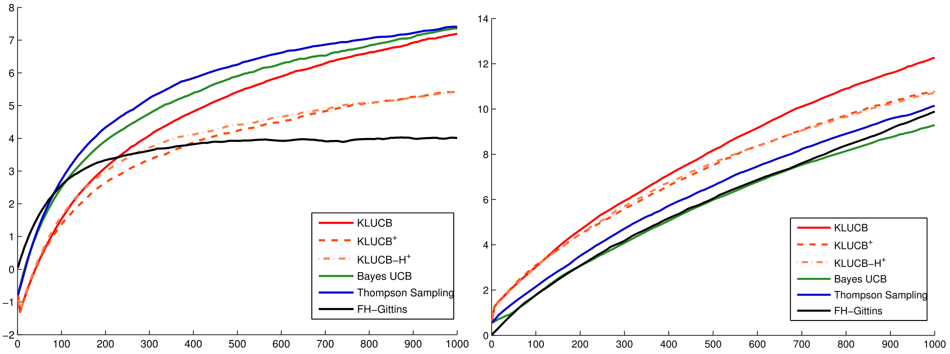

FIG. 2. *Regret on two-armed Bernoulli bandits* [$\boldsymbol{\mu} = [0.05\ 0.15]$ (*left*) $\boldsymbol{\mu} = [0.75\ 0.8]$ (*right*)] *up to horizon* $T = 1000$, *averaged over* $N = 10{,}000$ *simulations*.

and the four Bayesian (or Bayesian inspired) index policies discussed in this paper, in two instances of two-armed 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 $T = 20{,}000$, we then run experiments on a bandit model in which rewards follow an exponential distribution (which is a particular Gamma distribution). Bayes-UCB and Thompson Sampling are implemented using a conjugate InvGamma(1, 1) prior on the means. Results are displayed in Figure 3. 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., [14]), the Lai and Robbins lower bound, that is asymptotic, is quite pessimistic for finite (even large) horizons.

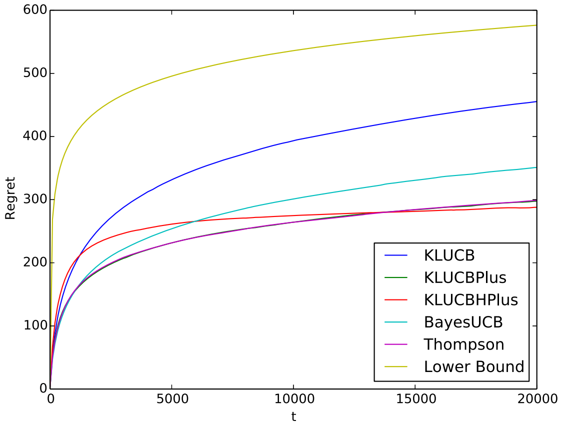

FIG. 3. *Regret on a five-armed bandit with Exponential distributions with means* $\boldsymbol{\mu} = [1\ 1.5\ 2\ 2.5\ 3]$ *up to horizon* $T = 20{,}000$, *averaged over* $N = 50{,}000$ *simulations*.

5.2. *Bayes risk minimization.* In this paper, Bayes risk minimization and its exact solution is mostly presented as a justification for improved algorithms for regret minimization. However, it is also interesting to understand whether the proposed algorithm are good approximations of the Bayesian solution, that is, whether they match the asymptotic lower bound of Theorem 10.

We report here results of experiments in Bernoulli bandit models with a uniform prior on the means. In this setting, some computations (see Appendix D.4 of the Supplemtary Material [23]) 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 $\log^2(T)$.

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$^+$

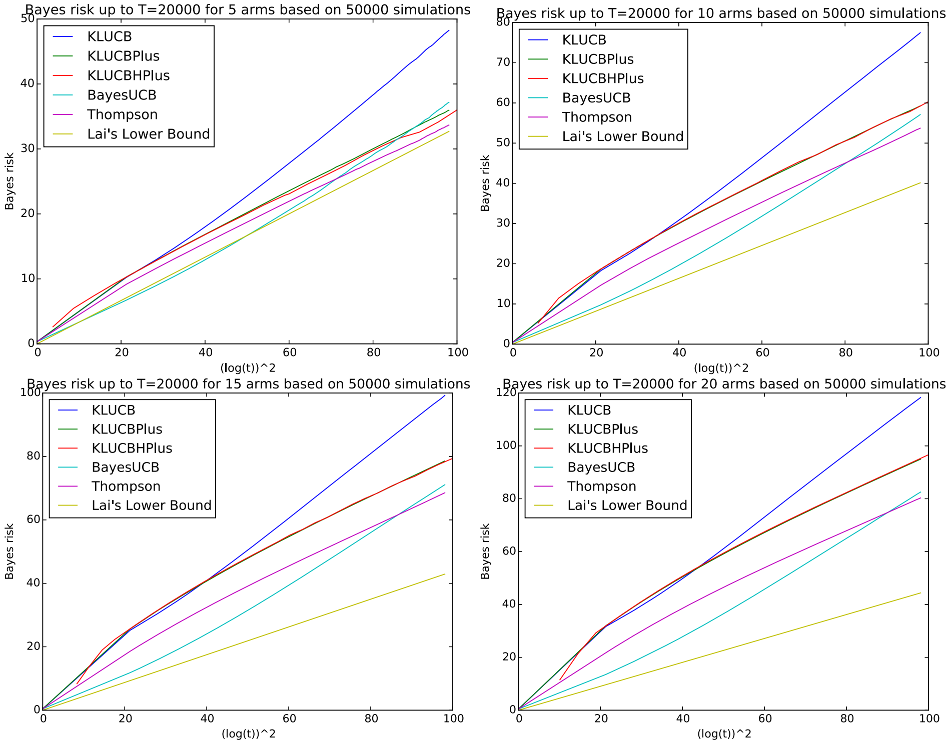

FIG. 4. *Bayes risk up to* $T = 20{,}000$ *on a Bernoulli bandit model with a uniform prior on the* K *arms, for* $K = 5, 10, 15, 20$, *averaged over* $N = 50{,}000$ *simulations.*

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.

6. Conclusion. This paper provides an analysis of the Bayes-UCB algorithm that does not rely on arguments specific to Bernoulli or Gaussian distributions, and is valid in any exponential family bandit model. It also brings theoretical justifications for the use of the kl-UCB-H$^+$ and kl-UCB$^+$ algorithms together with a new insight on the alternative exploration rate used by these algorithms. Finally, the proposed analysis holds for a wide class of prior distributions, namely all distributions that have positive density with respect to the Lebesgue measure. This shows that the choice of prior has no impact on the asymptotic optimality of Bayes-UCB, unlike what happens for Thompson Sampling in Gaussian bandit with unknown mean and variance [22]. Beyond asymptotic optimality, an interesting direction of future work would be to quantify the impact of the prior on second-order terms in the regret. Another important research direction is to better understand the Finite-Horizon Gittins strategy, which performs well in practice, but whose asymptotic optimality is still to be established.

SUPPLEMENTARY MATERIAL

Technical proofs (DOI: 10.1214/17-AOS1569SUPP; .pdf). The supplemental article contains the proofs of some results stated in the paper.

REFERENCES

[1] AGRAWAL, R. (1995). Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Adv. in Appl. Probab.* **27** 1054–1078. MR1358906

[2] AGRAWAL, S. and GOYAL, N. (2012). Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the* 25*th Conference on Learning Theory*.

[3] AGRAWAL, S. and GOYAL, N. (2013). Further optimal regret bounds for Thompson sampling. In *Proceedings of the* 16*th Conference on Artificial Intelligence and Statistics*.

[4] AUDIBERT, J.-Y. and BUBECK, S. (2010). Regret bounds and minimax policies under partial monitoring. *J. Mach. Learn. Res.* **11** 2785–2836. MR2738783

[5] AUDIBERT, J.-Y., MUNOS, R. and SZEPESVÁRI, C. (2009). Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoret. Comput. Sci.* **410** 1876–1902. MR2514714

[6] AUER, P., CESA-BIANCHI, N. and FISCHER, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* **47** 235–256.

[7] BELLMAN, R. (1956). A problem in the sequential design of experiments. *Sankhyā* **16** 221–229. MR0079386

[8] BERRY, D. A. and FRISTEDT, B. (1985). *Bandit Problems*: *Sequential Allocation of Experiments*. Chapman & Hall, London. MR0813698

[9] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193

[10] BRADT, R. N., JOHNSON, S. M. and KARLIN, S. (1956). On sequential designs for maximizing the sum of n observations. *Ann. Math. Stat.* **27** 1060–1074. MR0087288

[11] BROCHU, E., CORA, V. M. and DE FREITAS, N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. Technical report, Univ. British Columbia.

[12] BUBECK, S. and LIU, C.-Y. (2013). Prior-free and prior-dependent regret bounds for Thompson sampling. In *Advances in Neural Information Processing Systems*.

[13] BURNETAS, A. N. and KATEHAKIS, M. N. (2003). Asymptotic Bayes analysis for the finite-horizon one-armed-bandit problem. *Probab. Engrg. Inform. Sci.* **17** 53–82. MR1959385

[14] CAPPÉ, O., GARIVIER, A., MAILLARD, O.-A., MUNOS, R. and STOLTZ, G. (2013). Kullback–Leibler upper confidence bounds for optimal sequential allocation. *Ann. Statist.* **41** 1516–1541. MR3113820

[15] CHANG, F. and LAI, T. L. (1987). Optimal stopping and dynamic allocation. *Adv. in Appl. Probab.* **19** 829–853. MR0914595

[16] CHAPELLE, O. and LI, L. (2011). An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems*.

[17] GARIVIER, A. and CAPPÉ, O. (2011). The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the* 24*th Conference on Learning Theory*.

[18] GINEBRA, J. and CLAYTON, M. K. (1999). Small-sample performance of Bernoulli two-armed bandit Bayesian strategies. *J. Statist. Plann. Inference* **79** 107–122. MR1704187

[19] GITTINS, J. C. (1979). Bandit processes and dynamic allocation indices. *J. Roy. Statist. Soc. Ser. B* **41** 148–177. MR0547241

[20] GITTINS, J., GLAZEBROOK, K. and WEBER, R. (2011). *Multi-Armed Bandit Allocation Indices*, 2nd ed. Wiley, Chichester.

[21] HONDA, J. and TAKEMURA, A. (2010). An asymptotically optimal bandit algorithm for bounded support models. In *Proceedings of the* 23*rd Conference on Learning Theory*.

[22] HONDA, J. and TAKEMURA, A. (2014). Optimality of Thompson sampling for Gaussian bandits depends on priors. In *Proceedings of the* 17*th Conference on Artificial Intelligence and Statistics*.

[23] KAUFMANN, E. (2018). Supplement to "On Bayesian index policies for sequential resource allocation." DOI:10.1214/17-AOS1569SUPP.

[24] KAUFMANN, E., CAPPÉ, O. and GARIVIER, A. (2012). On Bayesian upper-confidence bounds for bandit problems. In *Proceedings of the* 15*th Conference on Artificial Intelligence and Statistics*.

[25] KAUFMANN, E., KORDA, N. and MUNOS, R. (2012). Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory*. *Lecture Notes in Computer Science* **7568** 199–213. Springer, Heidelberg. MR3042891

[26] KORDA, N., KAUFMANN, E. and MUNOS, R. (2013). Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems*.

[27] LAI, T. L. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *Ann. Statist.* **15** 1091–1114. MR0902248

[28] LAI, T. L. and ROBBINS, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.* **6** 4–22. MR0776826

[29] LATTIMORE, T. (2016). Regret analysis of the finite-horizon Gittins index strategy for multi-armed bandits. In *Proceedings of the* 29*th Conference on Learning Theory*, *COLT* 2016 1214–1245. Available at http://jmlr.org/proceedings/papers/v49/lattimore16.html.

[30] LIU, C.-Y. and LI, L. (2016). On the prior sensitivity of Thompson sampling. In *Algorithmic Learning Theory*. *Lecture Notes in Computer Science* **9925** 321–336. Springer, Cham. MR3591000

[31] NIÑO-MORA, J. (2011). Computing a classic index for finite-horizon bandits. *INFORMS J. Comput.* **23** 254–267. MR2816898

[32] PUTERMAN, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, New York. MR1270015

[33] REVERDY, P., SRIVASTAVA, V. and LEONARD, N. E. (2014). Modeling human decision making in generalized Gaussian multiarmed bandits. *Proc. IEEE* **102** 544–571.

[34] RUSSO, D. and VAN ROY, B. (2014). Learning to optimize via posterior sampling. *Math. Oper. Res.* **39** 1221–1243.

[35] RUSSO, D. and VAN ROY, B. (2014). Learning to optimize via information direct sampling. In *Advances in Neural Information Processing Systems*.

[36] SCOTT, S. L. (2010). A modern Bayesian look at the multi-armed bandit. *Appl. Stoch. Models Bus. Ind.* **26** 639–658. MR2752378

[37] SRINIVAS, N., KRAUSE, A., KAKADE, S. M. and SEEGER, M. W. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the International Conference on Machine Learning*.

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

INRIA LILLE NORD-EUROPE
40, AVENUE HALLEY
59650 VILLENEUVE D'ASCQ
FRANCE
E-MAIL: emilie.kaufmann@univ-lille1.fr