On Bayesian Upper Confidence Bounds for Bandit Problems

Emilie Kaufmann **Olivier Cappé** **Aurélien Garivier**
LTCI, CNRS & Telecom ParisTech

Abstract

Stochastic bandit problems have been analyzed from two different perspectives: a frequentist view, where the parameter is a deterministic unknown quantity, and a Bayesian approach, where the parameter is drawn from a prior distribution. We show in this paper that methods derived from this second perspective prove optimal when evaluated using the frequentist cumulated regret as a measure of performance. We give a general formulation for a class of Bayesian index policies that rely on quantiles of the posterior distribution. For binary bandits, we prove that the corresponding algorithm, termed Bayes-UCB, satisfies finite-time regret bounds that imply its asymptotic optimality. More generally, Bayes-UCB appears as an unifying framework for several variants of the UCB algorithm addressing different bandit problems (parametric multi-armed bandits, Gaussian bandits with unknown mean and variance, linear bandits). But the generality of the Bayesian approach makes it possible to address more challenging models. In particular, we show how to handle linear bandits with sparsity constraints by resorting to Gibbs sampling.

1 Introduction

The literature on stochastic multi-armed bandit problems is separated in two distinct approaches. In the frequentist view, the expected mean rewards corresponding to all arms are considered as unknown deterministic quantities and the goal of the algorithm is to achieve the best parameter-dependent performance. In contrast, in the Bayesian approach each arm is characterized by a parameter which is endowed with a prior distribution. The Bayesian performance is then defined as the average performance over all possible problem instances weighted by the prior on the parameters. In this work, we argue that algorithms derived from the Bayesian perspective also prove efficient when evaluated using frequentist measures of performance. Before exposing our contributions more precisely, we start by reviewing some aspects of these two alternative views.

In the classical parametric stochastic multi-armed bandit model, an agent faces K independent arms which depend on unknown parameters $\theta_1, \dots, \theta_K \in \Theta$. The draw of arm j at time t results in a reward X_t that is extracted from the i.i.d sequence $(Y_{j,t})_{t\geq 1}$ marginally distributed under ν_{θ_j}, whose expectation is denoted by μ_j. The agent sequentially draws the arms according to a strategy $(I_t)_{t\geq 1}$, where I_t denotes the arm chosen at round t, based on previous rewards $X_s = Y_{s,I_s}$ for $1 \leq s \leq t-1$. The agent's goal is to maximize the expected cumulated reward until time n, $\mathbb{E}_\theta\left[\sum_{t=1}^n X_t\right]$, or, equivalently, to minimize the cumulated regret

$$R_n(\theta) = \mathbb{E}_\theta\left[\sum_{t=1}^n \mu^* - \mu_{I_t}\right] = \sum_{j=1}^K (\mu^* - \mu_j)\mathbb{E}_\theta[N_n(j)]\,, \tag{1}$$

where $\mu^* = \max\{\mu_j : 1 \leq j \leq K\}$ and $N_n(j)$ denotes the number of draws of arm j up to time n.

Lai & Robbins [12], followed by Burnetas & Katehakis [3], have provided lower bounds on the number of suboptimal draws under any good strategy (having $o(n)$ regret for all bandit problems): for any arm j such that $\mu_j < \mu^*$,

$$\liminf_{n\to\infty} \frac{\mathbb{E}_\theta[N_n(j)]}{\log(n)} \geq \frac{1}{\inf_{\theta\in\Theta:\mu(\theta)>\mu^*} \mathrm{KL}(\nu_{\theta_j}, \nu_\theta)}\,, \tag{2}$$

where KL denotes the Kullback-Leibler divergence. For important classes of distributions, recent contributions have provided finite-time analysis of strategies that are asymptotically optimal in so far that they reach this lower bound. Following [10], [5] and [15] have analyzed algorithms based on the celebrated

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX. Copyright 2012 by the authors.

upper confidence bound (UCB) principle of [1] for, respectively, one-parameter exponential models and finitely-supported distributions.

When considering the multi-armed bandit model from a Bayesian point of view, one assumes that the parameter $\theta = (\theta_1, ..., \theta_K)$ is drawn from a prior distribution. More precisely, we will assume in the following that the parameters $(\theta_j)_{1\leq j\leq K}$ are drawn independently from prior distributions $(\pi_j)_{1\leq j\leq K}$ (usually chosen to be all equal), and that conditionally on $(\theta_j)_{1\leq j\leq K}$, the sequences $(Y_{1,t})_{t\geq 1}, \dots, (Y_{K,t})_{t\geq 1}$ are jointly independent and i.i.d. with marginal distributions $\nu_{\theta_1}, \dots, \nu_{\theta_K}$.

In this Bayesian setting, the goal is to maximize $\mathbb{E}\left[\sum_{t=1}^{n} X_t\right]$, where the expectation is relative to the entire probabilistic model, including the randomization over θ. Bayesian optimality can equivalently be measured considering the Bayesian regret $R_n^B = \mathbb{E}[R_n(\theta)]$ that averages the regret over the parameters. A major appeal of the Bayesian framework is the fact that a strategy with minimal Bayesian regret can be described, if not always computed.

To define the Bayesian strategy, let Π^t denote the posterior distribution of θ after t rounds of game, with Π^0 denoting the initial prior distribution. Due to our choice of independent priors on $(\theta_j)_{1\leq j\leq K}$, Π^t is a product distribution which is equivalently defined by the marginal posterior distributions $\pi_1^t, ..., \pi_K^t$. If at round t one chooses arm $I_t = j$ and consequently observes $X_t = Y_{j,t}$, the Bayesian update for arm j is

$$\pi_j^t(\theta_j) \propto \nu_{\theta_j}(X_t)\ \pi_j^{t-1}(\theta_j)\,, \tag{3}$$

whereas for $i \neq j$, $\pi_i^t = \pi_i^{t-1}$. A Bayesian algorithm is allowed to exploit the knowledge of the whole posterior Π^t to determine the next action I_{t+1}. In his seminal paper [7], Gittins showed that, for models admitting sufficient statistics, finding the Bayesian optimal strategy is equivalent to solving the planning problem in a related Markov decision model. Moreover, for several important cases, including that of Bernoulli rewards with Beta priors, the planning problem can be solved numerically thanks to a clever problem reduction using so-called 'Gittins indices'.

Gittins originally considered the infinite-horizon discounted problem in which one tries to maximize $\mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t X_t\right]$, where $0 < \gamma < 1$ is a real discount parameter. It is possible to show that the model reduction argument still holds when the horizon is known, making it possible to compute a finite-horizon variant of Gittins indices and, thus, to determine the finite-horizon Bayesian optimal strategy. The details of the corresponding algorithm are omitted here because of space limitations and we refer to the recent work of Niño-Mora [13] for discussion of the numerical complexity of this approach. However, we report in Section 4 some experiments on Bernoulli bandits that illustrate our finding that the corresponding policy constantly outperforms its frequentist UCB-like competitors on their own ground, that is, when evaluated using the parameter-dependent (frequentist) regret. Interestingly, Lai [11] established lower bounds for the Bayesian risk (depending on the prior) and showed (in particular cases) that algorithms reaching the lower-bound (2) were also Bayesian-optimal. Conversely, our finding that the Bayesian optimal strategy also achieves remarkable parameter-dependent performance for most (all?) value of the parameter θ is currently not supported by mathematical arguments.

Furthermore, computing the finite-horizon variant of the Gittins indices is only feasible for moderate horizons due to the need to repeatedly perform (and store the results of) dynamic programming recursions on reduced models. Even for small horizons, the associated computational load and memory footprint are orders of magnitude larger than those of the UCB-like algorithms considered in [1, 10, 5, 15].

Our objective is thus to propose a generic bandit algorithm, termed Bayes-UCB, that is inspired by the Bayesian interpretation of the problem but retains the simplicity of UCB-like algorithms. Our hope is that this algorithm is simple enough to be effectively implemented and yet sufficiently close to the Bayesian optimal policy to be able to reach the asymptotic lower bound of (2), including in cases that are currently not handled by UCB-like algorithms. In addition to promising simulation results reported in Section 4, we provide several significant elements that support our hopes. First, it is shown in Section 2 that instantiating the generic Bayes-UCB algorithm in different specific cases (one-parameter exponential families rewards, Gaussian-armed bandit with unknown means and variances, linear bandits, Gaussian process optimization) yields algorithms that share striking similarities with methods previously proposed in the literature. In the case of Bernoulli rewards, we provide in Section 3 (with corresponding proofs in appendix) a complete finite-time analysis of the Bayes-UCB algorithm that implies that it reaches the lower bound of (2). The proof of this result also reveals some enlightening facts about the construction of upper confidence bounds used in recently proposed variants of UCB such as those of [2, 5]. Finally, in Section 4, we consider the challenging setting of sparse linear bandits where we show how the Bayes-UCB strategy, using a sparsity inducing prior, can be numerically approximated using Markov Chain Monte Carlo simulations.

2 The Bayes-UCB algorithm

We start by presenting the rationale for the proposed algorithm before stating it more formally. First, being inspired by the Bayesian modeling of the bandit problem, the Bayes-UCB strategy is a function of the posteriors $(\pi_j^t)_{1\leq j\leq K}$. Due to the nature of our performance measure, the relevant aspect of θ_j is the expectation μ_j. Hence, denoting by λ_j^t, for $1 \leq j \leq K$, the posterior distribution of the mean μ_j induced by π_j^t, the proposed strategy is a function of $(\lambda_j^t)_{1\leq j\leq K}$ only. A similar principle is used in the so-called Thompson strategy which consists in drawing samples from the distributions $(\lambda_j^t)_{1\leq j\leq K}$ so as to select an arm j with probability equal to the posterior probability that its mean μ_j is the highest [18]. The Bayesian Learning Automaton advocated by [9] uses this idea of sampling the posterior, but no regret analysis is provided. The use of fixed-level quantiles of $(\lambda_j^t)_{1\leq j\leq K}$ as confidence indices appears in [14] as a special case of the Interval Estimation method. To be more specific, denote by $Q(t,\rho)$ the quantile function associated to the distribution ρ, such that $\mathbb{P}_\rho(X \leq Q(t,\rho)) = t$. [14] use indices of the form $Q(1-\alpha,\lambda_j^t)$ for $1 \leq j \leq K$, with α chosen to be equal to a few percents. In Bayes-UCB, we acknowledge the strong similarity between these posterior indices based on quantiles and the upper confidence bounds used in UCB and its variants: we consider indices of the form $Q(1-\alpha_t,\lambda_j^t)$, where α_t is of order $1/t$. As will be shown in Section 3 below for the case of binary rewards, this $1/t$ rate is deeply connected with the form of the upper confidence bounds used in variants of UCB that are known to reach the bound in (2). It is conjectured that no other rate can provide an algorithm that reaches the bound in (2) and that, furthermore, choices of the form $1/t^\beta$ with $\beta < 1$ do not even guarantee a finite-time logarithmic control of the regret. As a more pragmatic comment, we also observed in experiments not reported here that, in the case of binary rewards, the empirical performance of the method were superior when using $\alpha_t \equiv 1/t$. We are now ready to state the generic version of the Bayes-UCB algorithm.

In Algorithm 1, the horizon-dependent term $(\log n)^c$ is an artefact of the theoretical analysis that enables us, for $c \geq 5$, to both guarantee finite-time logarithmic regret bounds and achieve asymptotic optimality with respect to (2). But in simulations, the choice $c = 0$ actually proved to be the most satisfying. In cases where the prior Π^0 is chosen to correspond to an improper prior (see, e.g., the Gaussian models below), $q_j(t)$ is not defined when $t = 1$. In those cases it suffices, as is commonly done in most bandit algorithms, to make sure that initially one gathers a sufficient number of observations to guarantee that the posterior Π^t indeed becomes proper, for instance by drawing each arm a few times.

Algorithm 1 Bayes-UCB

```
Require: $n$ (horizon), $\Pi^0$ (initial prior on $\theta$)
    $c$ (parameters of the quantile)
for $t = 1$ to $n$ do
    for each arm $j = 1, \ldots, K$ do
        compute
            $q_j(t) = Q\left(1 - \frac{1}{t(\log n)^c}, \lambda_j^{t-1}\right)$
    end for
    draw arm $I_t = \arg\max_{j=1\ldots K} q_j(t)$
    get reward $X_t = Y_{I_t,t}$ and update $\Pi^t$ according
    to (3)
end for
```

As such, Algorithm 1 corresponds to a general principle that does not even require that the prior Π^0 be chosen as a product distribution: in fact, the GP-UCB algorithm for gaussian processes [17] can be seen as a variant of Bayes-UCB in which dependencies, in contrast, are of fundamental importance; but this is not a point that we emphasize in this article, and for the simplicity of notation we focus on the case where the coordinates of θ are independent. Implementing Algorithm 1 may require additional tools from the Bayesian computational toolbox to perform (or approximate) the Bayesian update of Π^t and/or to compute (or, again, approximate) the quantiles $q_j(t)$. We first discuss several important models for which Algorithm 1 corresponds to a a procedure that can be implemented exactly without the need to resort to numerical approximation (an example of the opposite situation will be considered in Section 4.3 below).

First consider the case where the reward distributions belong to a one-parameter exponential family, that is $\nu_{\theta_j}(x) = c(x)\exp(\phi(\theta_j)t(x) - a(\theta_j))$, with $\theta_j \in \mathbb{R}$. In this case, it is well known that the priors π_j^0 can be chosen to belong to the conjugate family so that the posteriors π_j^t are all members of the same conjugate family, indexed by their sufficient statistics. For Bernoulli rewards, for instance, using the prior Beta(a,b) for the probability of observing a non-zero reward, we have $\pi_j^t = \text{Beta}(a + S_t(j), b + N_t(j) - S_t(j))$, where $S_t(j) = \sum_{i=1}^t \mathbb{1}\{I_t = j\}X_t$. Likewise, for exponential rewards with a Gamma(c,d) prior on the parameter, $\pi_j^t = \text{Gamma}(c + N_t(j), d + S_t(j))$. In addition, in the single-parameter case, there is a one-to-one monotonic correspondence between the parameter θ_j and the expectation μ_j. Hence, $q_j(t)$ is obtained by computing the quantile of well-known parametric distributions (upper quantile in the case of Bernoulli rewards, as $\mu_j = \theta_j$, lower quantile for exponential rewards for which $\mu_j = 1/\theta_j$). In this case, as will be proved below for binary rewards, the resulting algorithm is surprisingly related to the KL-UCB algorithm of [5].

In general exponential family models, the Bayesian update is usually still computable explicitly (at least when using conjugate priors) but the relationship between the parameter θ_j and the expectation μ_j is less direct. A significant case where Bayes-UCB corresponds to a simple and efficient algorithm is when the rewards are assumed to be Gaussian, with both unknown mean μ_j and unknown variance σ_j^2. For simplicity, we consider improper non-informative priors on each arm, that is, $\pi_j^0(\mu_j, \sigma_j) = 1/\sigma_j^2$. It is well known that the marginal posterior distribution of μ_j at time t is then such that

$$\left.\frac{\mu_j - S_t(j)/N_t(j)}{\sqrt{S_t^{(2)}(j)/N_t(j)}}\right| X_1, ..., X_n \sim \mathcal{T}(N_t(j) - 1)\,,$$

where

$$S_t^{(2)}(j) = \frac{\left(\sum_{i=1}^t \mathbb{1}\{I_t = j\} X_t^2\right) - S_t^2(j)/N_t(j)}{N_t(j) - 1}\,,$$

and $\mathcal{T}(k)$ denote the Student-t distribution with k degrees of freedom. Therefore Bayes-UCB is the index policy associated to upper confidence bound

$$q_j(t) = \frac{S_j(t)}{N_j(t)} + \sqrt{\frac{S_t^{(2)}(j)}{N_j(t)}}\, Q\left(1 - \frac{1}{t}, \mathcal{T}(N_t(j) - 1)\right)\,,$$

omitting the $(\log n)^c$ constant for clarity. The Bayes-UCB index above is related to the index used in the UCB1-norm algorithm of Auer et al. in [1], where the quantile is replaced by $\sqrt{16\log(t-1)}$, which is obtained as an upper bound of $Q(1 - 1/t^4, \mathcal{T}(N_j(t) - 1))$. The practical performances of these two variants (Bayes-UCB and UCB1-norm) will be illustrated in Section 4 below.

We end this section with the more elaborate case of linear bandits in which the arms can be very numerous but share a strong common structure. Here again we will consider the case of Gaussian rewards with a multivariate Gaussian prior for the parameter θ that defines the model. The arms are fixed vectors $U_1, ..., U_K \in \mathbb{R}^d$. In this model, the choice of arm $I_t = j$ at time t results in the reward $y_t = U_j'\theta + \sigma^2\epsilon_t$. Following [16], our goal is to find strategies that minimize the frequentist regret

$$R_n = \mathbb{E}_\theta\left[\sum_{t=1}^n \left(\max_{1\le j\le K}(U_j'\theta) - U_{I_t}'\theta\right)\right]\,.$$

Denoting by $Y_t = [y_1, ..., y_t]'$ the vector of rewards and $X_t = [U_{I_1} ... U_{I_t}]'$ the design matrix, the problem rewrites:

$$Y_t = X_t\theta + \sigma^2 E_t\ ,\ \text{where}\ \ E_t \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_t)\,.$$

The Bayesian modeling here consists in a Gaussian $\mathcal{N}\left(0, \kappa^2 \mathrm{Id}_d\right)$ prior on θ, assuming the noise parameter σ^2 to be known. The posterior is

$$\begin{array}{rrcl} & \theta | X_t, Y_t & \sim & \mathcal{N}\left(M_t, \Sigma_t\right)\,, \\ \text{where} & M_t & = & (X_t'X_t + (\sigma/\kappa)^2 \mathrm{I}_d)^{-1} X_t' Y_t\,, \\ & \Sigma_t & = & \sigma^2 (X_t'X_t + (\sigma/\kappa)^2 \mathrm{I}_d)^{-1}\,. \end{array}$$

The posterior distribution λ_j^t on $\mu_j = U_j'\theta$ is therefore $\mathcal{N}(U_j'M_t, U_j'\Sigma_t U_j)$. Hence, Bayes-UCB selects the arm by maximising the index:

$$q_j(t) = U_j'M_t + ||U_j||_{\Sigma_t} Q\left(1 - \frac{1}{t}, \mathcal{N}(0,1)\right)\,.$$

[4] and [16] propose an optimistic approach for this problem based on a confidence ellipsoid located around the least-square estimate $\hat{\theta}_t$. This method is equivalent to choosing arm j such that $U_j\hat{\theta}_t + \rho(t)||U_j||_{(X_t'X_t)^{-1}}$ is maximal. For an improper prior ($\kappa = \infty$), we have $M_t = \hat{\theta}_t$ and $\Sigma_t = \sigma^2(X_t'X_t)^{-1}$. Thus, this approach can again be interpreted as a particular case of Bayes-UCB. In Section 4.3, we consider the case where θ is a sparse vector. It is not obvious how to design an UCB algorithm for this case. Yet, we show that one can implement the Bayes-UCB algorithm with the help of Gibbs sampling.

3 Analysis of the Bayes-UCB algorithm for binary rewards

In this section, we focus on the case where the rewards have a Bernoulli distribution, and when the prior is the Beta$(1,1)$, or uniform, law . We show that the Bayes-UCB algorithm is optimal, in the sense that it reaches the lower-bound (2) of Lai and Robbins.

Theorem 1 *For any $\epsilon > 0$, choosing the parameter $c \geq 5$ in the Bayes-UCB algorithm, the number of draws of any sub-optimal arm j is upper-bounded by*

$$\mathbb{E}[N_n(j)] \leq \frac{1+\epsilon}{d(\mu_j, \mu^*)} \log(n) + o_{\epsilon,c}\left(\log(n)\right)\,.$$

A non-asymptotic form of Theorem 1 is proved in the appendix. The analysis relies on tight bounds of the quantiles of the Beta distributions, which are summarized in the following lemma.

Lemma 1 *Denoting by $d(x, y)$ the KL divergence between Bernoulli distributions with parameters x and y, the posterior quantile $q_j(t)$ used by the Bayes-UCB algorithm satisfies*

$$\tilde{u}_j(t) \leq q_j(t) \leq u_j(t) \ ,$$

where

$$u_j(t) = \underset{x > \frac{S_t(j)}{N_j(t)}}{\textit{argmax}} \left\{ d\left(\frac{S_t(j)}{N_t(j)}, x\right) \leq \frac{\log(t) + c\log(\log(n))}{N_t(j)} \right\} ,$$

$$\tilde{u}_j(t) = \underset{x > \frac{S_t(j)}{N_t(j)+1}}{\textit{argmax}} \left\{ d\left(\frac{S_t(j)}{N_t(j)+1}, x\right) \leq \frac{\log\left(\frac{t}{N_t(j)+2}\right) + c\log(\log(n))}{(N_t(j)+1)} \right\} .$$

Surprisingly, the Bayesian quantiles match the upper confidence bound used by the two variants KL-UCB and KL-UCB+ of the (Bernoulli-optimal) algorithm for bounded bandit problems analyzed in [5], showing thus a very similar behavior. The fact that, in $\tilde{u}_j(t)$, the current time t is divided by $N_t(j)$ in the logarithmic bonus is unexpectedly reminiscent of the MOSS algorithm of [2]. The proof of Lemma 1, given in the appendix, relies on the following remark: for any integers a, b, the distribution $\text{Beta}(a, b)$ is the law of the a-th order statistic among $a + b - 1$ uniform random variables, so that

$$\mathbb{P}(X \geq x) = \mathbb{P}(S_{a+b-1,x} \leq a-1) = \mathbb{P}(S_{a+b-1,1-x} \geq b) \ ,$$

where $S_{n,x}$ denotes a binomial distribution with parameters n and x. Bounding the beta quantiles boils down to controlling the binomial tails, which is achieved using Sanov's inequality:

$$\frac{e^{-nd\left(\frac{k}{n},x\right)}}{n+1} \leq \mathbb{P}(S_{n,x} \geq k) \leq e^{-nd\left(\frac{k}{n},x\right)} \ , \qquad (4)$$

where the rightmost inequality holds for $k \geq nx$.

4 Numerical experiments

4.1 Binary bandits

Numerical experiments have been carried out in a frequentist setting for bandits with Bernoulli rewards: for a fixed parameter θ and an horizon n, N bandit games with Bernoulli rewards are repeated for a given strategy. The main purpose of these numerical experiments is to compare the performance in terms of cumulated regret of Bayes-UCB with those of UCB and KL-UCB. These are presented on Figure 1, where the regret is averaged over $N = 5000$ simulations for two different two-armed bandit problems with horizon $n = 500$. We also included in the comparison the Bayesian algorithm based on Finite-Horizon Gittins indices (FH-Gittins). Whereas the performance of FH-Gittins are more striking in the top situation (0.1/0.2) than in the bottom one (0.45/0.55), Bayes-UCB also improves equally over KL-UCB in all scenarios.

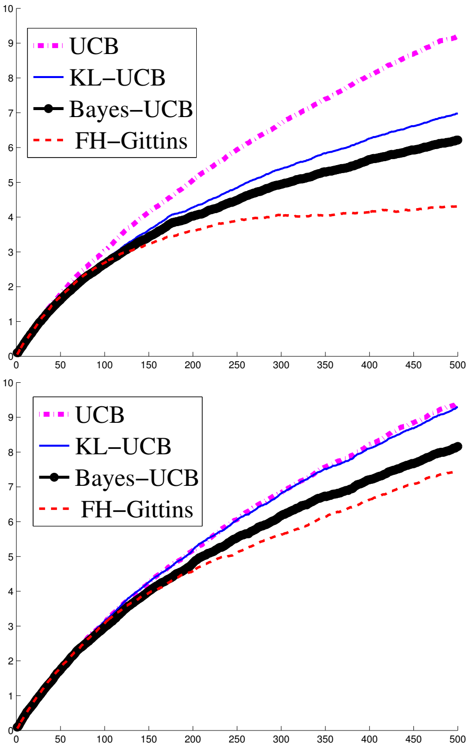

Figure 1: Cumulated regret for the two armed-bandit problem with $\mu_1 = 0.1, \mu_2 = 0.2$ (top) and $\mu_1 = 0.45, \mu_2 = 0.55$ (bottom).

4.2 Gaussian rewards with unknown means and variances

For the bandit problem with Gaussian rewards with unknown mean and variance, few algorithms have been proposed. We compare Bayes-UCB with UCB1-norm and UCB-Tuned (see [1]). Figure 2 presents the regret in a 4-arms problem, on a horizon $n = 10000$, averaged over $N = 1000$ simulations. UCB-Tuned seems unadapted to the problem, whereas UCB1-norm and Bayes-UCB achieve a regret proving that the asymptotic lower bound of Burnetas & Katehakis is pessimistic for such short horizons (see also [5]). Bayes-UCB outperforms UCB1-norm, mostly because of the more appropriate choice of a quantile of order $1 - 1/t$.

4.3 Sparse linear bandits

The linear bandit model presented in Section 2 relies on linear regression. Many recent works have high-

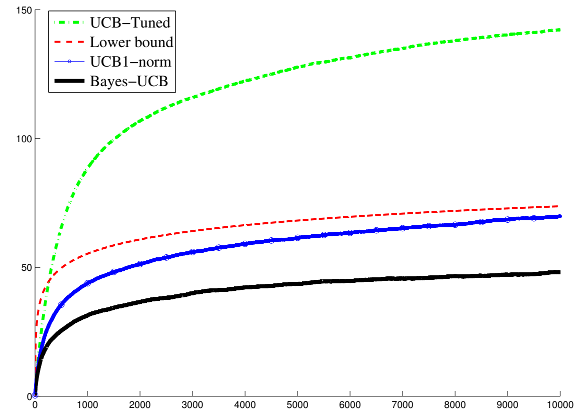

Figure 2: Regret in a 4-arms problem with parameters $\mu = [1.8\ 2\ 1.5\ 2.2]$, $\sigma = [0.5\ 0.7\ 0.5\ 0.3]$.

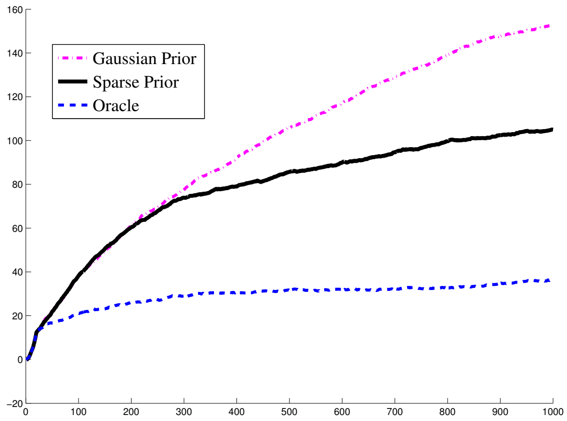

Figure 3: Cumulated regret in a 20 arms problem for Bayes-UCB with different prior distributions.

lighted the importance of sparsity issues in this context. We show that Bayes-UCB can address sparse linear bandit problems by using a prior that encourages sparsity of the parameter θ. This 'spike-and-slab' prior is defined as follows: the coordinates of θ are independent, with distribution

$$\theta_j \sim \epsilon\delta_0 + (1-\epsilon)\mathcal{N}(0,\kappa^2)\ .$$

Let C be the random vector in $\mathbb{R}^d$ indicating the non-zero coordinates of θ: $C_j = \mathbb{1}_{(\theta_j \neq 0)}$. If J denotes a set of indices, let $X_{t,J} \in \mathcal{M}_{t,|J|}(\mathbb{R})$ be the submatrix of X_t with columns in J only and $\theta_J \in \mathbb{R}^{|J|}$ the subvector with coordinates in J.

Given C and Y_t, denote by J_1 the set of non-zeros coordinates in C. The subvector θ_{J_1} is the solution of a Bayesian regression problem with prior $\mathcal{N}(0, \kappa^2 I_{|J_1|})$, hence

$$\begin{aligned}\theta_{J_1}|C,Y_t \ \sim\ & \mathcal{N}\left((X'_{t,J_1}X_{t,J_1} + (\sigma/\kappa)^2\mathrm{I}_{|J_1|})^{-1}X'_{t,J_1}Y_t\right.\\ & \left.;\ \sigma^2(X'_{t,J_1}X_{t,J_1} + (\sigma/\kappa)^2\mathrm{I}_{|J_1|})^{-1}\right)\ .\end{aligned}$$

The marginal distribution of C given Y is

$$P(C|Y) \propto \epsilon^{[J_0|}(1-\epsilon)^{|J_1|}\mathcal{N}\left(Y_t|0, \kappa^2 X_{t,J_1}X'_{t,J_1} + \sigma^2\mathrm{I}_t\right)\ .$$

The normalization term involves a sum over 2^d possible configurations of C. When d is small, the exact Bayes-UCB indices can be computed, as the dot-product $U'_j\theta$ follows a mixture of Gaussian distributions. For higher dimensions, one can use Gibbs sampling to sample from $C|Y$, and produce samples from $\theta|Y$ that lead to approximated values of $q_j(t)$.

Numerical simulation have been carried out for a sparse problem in dimension $d = 10$ where θ only has two non-zero coordinates. On Figure 3 we compare the regret of Bayes-UCB for three different priors: the general multivariate Gaussian prior discussed in Section 2, an oracle Gaussian prior on the first two coordinates only (meaning that the sparsity pattern is known) and Bayes-UCB with a sparse prior. The 20 arms of the problem are chosen randomly on the unit sphere and the regret is averaged over $N = 100$ simulations for an horizon $n = 1000$. As expected, the use of a sparsity-inducing prior in this case results in an algorithm with greatly enhanced performance.

5 Conclusion

Although frequentist and Bayesian bandits correspond to two different probabilistic frameworks, we have observed that using Bayesian ideas often provides efficient algorithms for the frequentist bandit setting. The proposed Bayes-UCB approach appears to provide a generic and efficient solution for various bandit problems, including challenging ones such as sparse linear bandits. At this point, finite-time regret bounds and asymptotic optimality of the Bayes-UCB strategy have only been proved for binary multi-armed bandits. However, a similar proof can be given for the case of Gaussian multi-armed bandits, when the arm variances are known. We believe that those results can also be extended to more general cases and, in particular, to exponential family distributions and Gaussian linear regression models.

A Proof of Lemma 1

If $X \sim \mathrm{Beta}(a,b)$, equation (4) gives, for $x > \frac{a-1}{a+b-1}$,

$$\frac{e^{-(a+b-1)d(\frac{a-1}{a+b-1},x)}}{a+b} \leq \mathbb{P}(X \geq x) \leq e^{-(a+b-1)d(\frac{a-1}{a+b-1},x)}$$

Let $q_{1-\gamma} = Q(1-\gamma, \mathrm{Beta}(a,b))$. Since :

$$(a+b-1)d\left(\frac{a-1}{a+b-1},x\right) \geq \log(1/\gamma) \quad \Rightarrow \quad x \geq q_{1-\gamma}$$

we have that :

$$\begin{aligned} x^*_+ &= \underset{x > \frac{a-1}{a+b-1}}{\operatorname{argmin}} \{(a+b-1)d\left(\frac{a-1}{a+b-1}, x\right) \geq \log(1/\gamma)\} \\ &= \underset{x > \frac{a-1}{a+b-1}}{\operatorname{argmax}} \{(a+b-1)d\left(\frac{a-1}{a+b-1}, x\right) \leq \log(1/\gamma)\} \end{aligned}$$

is still an upper bound for the quantile $q_{1-\gamma}$. The same reasoning shows $q_{1-\gamma}$ is lower-bounded by

$$x^*_- = \underset{x > \frac{a-1}{a+b-1}}{\operatorname{argmax}} \left\{(a+b-1)d\left(\frac{a-1}{a+b-1}, x\right) \leq \log\left(\frac{1}{\gamma(a+b)}\right)\right\}$$

Moreover we can easily show that

$$x^*_+ \leq \underset{x > \frac{a-1}{a+b-2}}{\operatorname{argmax}} \{(a+b-2)d\left(\frac{a-1}{a+b-2}, x\right) \leq \log(1/\gamma)\}$$

using mainly the fact that $y \mapsto d(y,x)$ is decreasing for $y < x$.We get the final result using $a = S_j(t)+1$,$b = N_j(t) - S_j(t) + 1$ and $\gamma = 1/(t\log(n)^c)$.

B Proof of Theorem 1

Without loss of generality, one supposes arm 1 is optimal and arm 2 is suboptimal. To prove Theorem 1, we show more precisely that there exists $N(\epsilon)$ and $K_c > 0$ such that for $n \geq N(\epsilon)$:

$$\begin{aligned} \mathbb{E}[N_n(2)] \leq & \frac{(1+\epsilon)(\log(n) + c\log(\log(n)))}{d(\mu_2,\mu_1)} \\ & + 1 + K_c(\log(\log(n)))^2 + \frac{1}{n-1} \\ & \frac{(1+\epsilon/2)^2}{\epsilon^2\left(\min\left(\mu_2(1-\mu_2);\mu_1(1-\mu_1)\right)\right)^2} . \end{aligned}$$

Let

$$\beta_n = \sqrt{\frac{1}{\log(n)}}$$

One starts with the following decomposition

$$N_n(2) \leq \sum_{t=1}^{n} \mathbb{1}_{(\mu_1 - \beta_n > q_1(t))} + \sum_{t=1}^{n} \mathbb{1}_{(\mu_1 - \beta_n \leq q_1(t)) \cap (I_t = 2)} . \tag{5}$$

This decomposition is motivated by the one used for KL-UCB in [5], but to evaluate the over-estimation of the optimal arm, we no longer compare $q_1(t)$ to μ_1 but to $\mu_1 - \beta_n$. The influence of β_n makes the left-term (under-estimation term) smaller and the right-term bigger. Now recall that the indices $q_1(t), q_2(t)$ used in Bayes-UCB are close to KL-UCB-like indices. We indeed use the fact that : (i) $\tilde{u}_1(t) \leq q_1(t)$, and, (ii) $q_2(t) \leq u_2(t)$.

Lemma 2

$$\begin{aligned} \mathbb{E}[N_j(2)] \leq & \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(\mu_1 - \beta_n > \tilde{u}_1(t)\right)}_{A} \\ & + \underbrace{\sum_{s=1}^{n} \mathbb{P}\left(sd^+\left(\hat{\mu_2}(s), \mu_1 - \beta_n\right) \leq \log(n) + c\log(\log(n))\right)}_{B} . \end{aligned}$$

where $d^+(x,y) = d(x,y)\mathbb{1}_{(x \leq y)}$

Proof of lemma 2 Term A follows from (i). By (ii) if $I_t = 2$, $q_1(t) \leq q_2(t) \leq u_2(t)$ so the most right term in (5) is upper-bounded as

$$\sum_{t=1}^{n} \mathbb{1}_{(\mu_1 - \beta_n \leq q_1(t)) \cap (I_t = 2)} \leq \sum_{t=1}^{n} \mathbb{1}_{(I_t = 2) \cap (\mu_1 - \beta_n \leq u_2(t))} .$$

Summing over the values of $N_t(2)$, and using the same trick as in lemma 7 in [5], the last term is bounded by

$$\sum_{s=1}^{n} \mathbb{1}_{(sd^+(\hat{\mu_2}(s), \mu_1 - \beta_n) \leq \log(n) + c\log(\log(n)))}$$

and the result follows by taking the expectation.

□

Now we have to upper bound separately A and B.

Study of term A To deal with term A, we write a new decomposition, depending on the number of draws of the optimal arm:

$$\begin{aligned} (A) \leq & \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(\mu_1 - \beta_n > \tilde{u_1}(t)\ ,\ N_t(1) + 2 \leq \log^2(n)\right)}_{A_1} \\ + & \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(\mu_1 - \beta_n > \tilde{u}_1(t)\ ,\ N_t(1) + 2 \geq \log^2(n)\right)}_{A_2} . \end{aligned}$$

Study of term A_1 Using that the term $\log\left(\frac{t}{N_t(j)+2}\right)$ in $\tilde{u_1}(t)$ is lower-bounded by $\log\left(\frac{t}{\log(n)^2}\right)$, we show

$$(\mu_1 - \beta_n > \tilde{u}_1(t)\ ,\ N_t(1) + 2 \leq \log^2(n)) \subseteq \left(\mu_1 > \bar{u}^t_{1,\delta}\right)$$

where

$$\begin{aligned} \bar{u}^t_{1,\delta} &= \underset{x > \frac{S_t(j)}{N_j(t)+1}}{\operatorname{argmax}} \left\{(N_t(1)+1)d\left(\frac{S_t(1)}{N_t(1)+1}, x\right) \leq \delta\right\} , \\ \delta &= \log(t) + (c-2)\log(\log(n)) . \end{aligned}$$

This appears to be the under-estimation term in a biased version of KL-UCB with the parameter $c' = c-2$ instead of c. With a straightforward adaptation (omitted here) of the proof of theorem 10 in [5] we obtain the following self-normalized inequality.

Lemma 3

$$\mathbb{P}\left(\mu_1 > \bar{u}_{1,\delta}(t)\right) \leq (\delta \log(t) + 1)\exp(-\delta + 1)$$

And lemma 3 leads to the upper-bound

$$\begin{aligned}(A_1) &\leq 1 + \sum_{t=2}^{n} \frac{e\left(\log^2(t) + (c-2)\log(t)\log(\log(n)) + 1\right)}{t(\log(n))^{c-2}} \\ &\leq 1 + (2e + e(c-2)\log(\log(n)))\sum_{t=1}^{n}\frac{1}{t\log(t)^{c-4}} \\ &\leq 1 + K_c(\log(\log(n))^2 \text{ for } c \geq 5 \ .\end{aligned}$$

Study of term A_2 In this term, the optimal arm has been sufficiently drawn to be well estimated, so we can use that

$$(\mu_1 - \beta_n > \tilde{u}_1(t)) \subset \left(\mu_1 - \beta_n > \frac{S_t(1)}{N_t(1)+1}\right)$$

and bound the deviation of the Bayesian empirical mean. Note that as

$$\frac{S_t(1)}{N_t(1)+1} \geq \frac{S_t(1)}{N_t(1)} - \frac{1}{N_t(1)+1} \ ,$$

dealing with the bias leads to (denoting by $t' = \lceil \log(n)^2 - 2 \rceil$):

$$(A_2) \leq \sum_{t=t'}^{n} \mathbb{P}\left(\exists s \leq t : \sum_{r=1}^{s} \tilde{Y}_{1,r} \geq \beta_n(\log(n)^2 - 2) - 1\right) \ ,$$

where $\tilde{Y}_{1,r} = \mu_1 - Y_{1,r}$ is the deviation from the mean of $Y_{1,r} \sim \mathcal{B}(\theta_1)$. Then we use a maximal inequality and get, replacing β_n by its value :

$$\begin{aligned}(A_2) &\leq \sum_{t=t'}^{n} e^{-2t\left(\frac{\log(n)^2-2}{\sqrt{\log(n)}}-1\right)^2} \leq \sum_{t=1}^{\infty} e^{-2t\left(\frac{\log(n)^2-2}{\sqrt{\log(n)}}-1\right)^2} \\ &= \frac{1}{e^{2\left(\frac{\log(n)^2-2}{\sqrt{\log(n)}}-1\right)^2} - 1} \text{ for } n \text{ s.t. } \log(n)^2 \geq 3 \ .\end{aligned}$$

Note that $2\left(\frac{\log(n)^2-2}{\sqrt{\log(n)}} - 1\right)^2 \geq \frac{(\log(n)^2-2)^2}{\log(n)}$ for n such that $\frac{(\log(n)^2-2)}{\sqrt{\log(n)}} \geq \frac{\sqrt{2}}{\sqrt{2}-1}(*)$. For such n we obtain

$$(A_2) \leq \frac{1}{e^{\log(n)\frac{(\log(n)^2-2)^2}{\log(n)^2}} - 1} \underset{(**)}{\leq} \frac{1}{n-1},$$

where $(*)$ and $(**)$ hold for $\log(n) \geq 4$. Finally, for $n \geq \exp(4)$,

$$(A_2) \leq \frac{1}{n-1} \ .$$

Study of term B Introducing, for $\epsilon > 0$

$$K_n = \frac{(1+\epsilon)(\log(n) + c\log(\log(n)))}{d(\mu_2, \mu_1)},$$

term B can be rewritten as

$$\begin{aligned}(B) \leq K_n + &\\ &\sum_{\lfloor K_n \rfloor + 1}^{n} \mathbb{P}\left(d^+(\hat{\mu}_2(s), \mu_1 - \beta_n) \leq \frac{\log(n) + c\log(\log(n))}{K_n}\right) \\ &\leq K_n + \sum_{\lfloor K_n \rfloor + 1}^{n} \mathbb{P}\left(d^+(\hat{\mu}_2(s), \mu_1 - \beta_n) \leq \frac{d(\mu_2,\mu_1)}{1+\epsilon}\right) \ .\end{aligned}$$

The function $g(q) = d^+(\hat{\mu}_2(s), q)$ is convex and differentiable and $g'(q) = \frac{q - \hat{\mu}_2(s)}{q(1-q)}\mathbb{1}_{(q > \hat{\mu}_2(s))}$, thus

$$d^+(\hat{\mu}_2(s), \mu_1) \leq d^+(\hat{\mu}_2(s), \mu_1 - \beta_n) + \beta_n \frac{\mu_1 - \hat{\mu}_2(s)}{\mu_1(1-\mu_1)} \ .$$

And therefore (bounding $\mu_1 - \hat{\mu}_2(s)$ by 2) :

$$\begin{aligned}&\left(d^+(\hat{\mu}_2(s), \mu_1 - \beta_n) \leq \frac{d(\mu_2,\mu_1)}{1+\epsilon}\right) \\ \subset &\left(d^+(\hat{\mu}_2(s), \mu_1) \leq \frac{d(\mu_2,\mu_1)}{1+\epsilon} + \beta_n \frac{2}{\mu_1(1-\mu_1)}\right) \ .\end{aligned}$$

Thus for $n \geq \exp\left(\left(\frac{2(1+\epsilon)(1+\epsilon/2)}{\epsilon\mu_1(1-\mu_1)d(\mu_2,\mu_1)}\right)^2\right)$ we obtain $\frac{d(\mu_2,\mu_1)}{1+\epsilon} + \beta_n\frac{2}{\mu_1(1-\mu_1)} \leq \frac{d(\mu_2,\mu_1)}{1+\epsilon/2}$ and

$$(B) \leq K_n + \sum_{s=\lfloor K_n \rfloor + 1}^{n} \mathbb{P}\left(d^+(\hat{\mu}_2(s), \mu_1) \leq \frac{d(\mu_2,\mu_1)}{(1+\epsilon/2)}\right) \ .$$

This term is upper-bounded precisely in [15], by

$$(B) \leq K_n + \frac{(1+\epsilon/2)^2}{\epsilon^2\left(\min\left(\mu_2(1-\mu_2); \mu_1(1-\mu_1)\right)\right)^2} \ .$$

Conclusion For $\epsilon > 0$ let

$$N(\epsilon) = \max\left\{e^4; \exp\left(\left(\frac{2(1+\epsilon)(1+\epsilon/2)}{\epsilon\mu_1(1-\mu_1)d(\mu_2,\mu_1)}\right)^2\right)\right\} \ .$$

Then, for $n \geq N(\epsilon)$ the following bound holds for $c \geq 5$:

$$\begin{aligned}\mathbb{E}[N_n(2)] \leq &\frac{(1+\epsilon)(\log(n) + c\log(\log(n)))}{d(\mu_2,\mu_1)} \\ &+1 + K_c(\log(\log(n)))^2 + \frac{1}{n-1} \\ &+\frac{(1+\epsilon/2)^2}{\epsilon^2\left(\min\left(\mu_2(1-\mu_2); \mu_1(1-\mu_1)\right)\right)^2} \ ,\end{aligned}$$

that is,

$$\mathbb{E}[N_n(2)] \leq \frac{(1+\epsilon)\log(n)}{d(\mu_2,\mu_1)} + R_n(\epsilon, c) \ ,$$

with $R_n(\epsilon, c) = o(\log(n))$, for every $\epsilon > 0$.

□

Aknowledgement Work supported by ANR Grant *Banhdits* ANR-2010-BLAN-0113-03.

References

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

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

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

[4] V. Dani, T.P. Hayes, S.M. Kakade. Stochastic linear optimization under bandit feedback. In *Conference On Learning Theory COLT* 2008.

[5] A. Garivier, O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond In *Conference On Learning Theory COLT* , 2011.

[6] E. Frostig, G. Weiss, Four proofs of Gittins' multiarmed bandit theorem Preprint, 1999.

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

[8] J. Gittins, K. Glazebrook and R. Weber. Multi-armed bandit allocation indices (2nd Edition) Wiley, 2011.

[9] O.C. Granmo. Solving Two-Armed Bernoulli Bandit Problems Using a Bayesian Learning Automaton in *International Journal of Intelligent Computing and Cybernetics (IJICC)* 3(2):207-234, 2010.

[10] J. Honda and A. Takemura. An asymptotically optimal bandit algorithm for bounded support models. In T. Kalai and M. Mohri, editors, *Conference On Learning Theory COLT*, 2010.

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

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

[13] J. Niño-Mora. Computing a classic index for Finite-Horizon bandits Journal on Computing 23(2):254-267, 2011

[14] N.G Pavlidis, D.K. Tasoulis and D.J. Hand. Simulation studies of multi-armed bandits with covariates In *Proc. 10th International Conference on Computer Modelling*, Cambridge, UK, 2008.

[15] O. Maillard, R. Munos, G. Stoltz. A finite-time analysis of Multi-armed bandits problems with Kullback-Leibler Divergence In *Conference On Learning Theory COLT* , 2011.

[16] P.Rusmevichientong, J.N. Tsitsiklis. linearly Parameterized Bandits In *Mathematics of Operations Research* 32(2):395-411, 2010.

[17] N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning ICML 10*, 2010.

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