On the efficiency of Bayesian bandit algorithms from a frequentist point of view

Emilie Kaufmann
Telecom ParisTech
kaufmann@enst.fr

Olivier Cappé
CNRS & Telecom ParisTech
cappe@enst.fr

Aurélien Garivier
CNRS & Telecom ParisTech
garivier@enst.fr

Abstract

In this contribution, we argue that algorithms derived from the Bayesian modelling of the multiarmed bandit problem are also optimal when evaluated using the frequentist cumulated regret as a measure of performance. We first show that the classical Gittins argument can be applied to convert the finite-horizon Bayesian multiarmed bandit problem into an MDP planning task that is numerically solvable for moderate horizons. The corresponding strategy is shown in simulations to outperform its competitors, including approaches, such as the recently proposed KL-UCB algorithm, that are known to be asymptotically optimal in the sense of reaching the lower bound of Lai and Robbins for the regret. Motivated by this observation, we propose a crude approximation of the optimal Bayesian decision rule in the form of a simple index policy using, for each arm, a suitably chosen quantile of the posterior distribution of the mean. For Bernoulli rewards, this algorithm is shown to be asymptotically optimal and turns out to show interesting connections with several recently proposed refinements of the UCB algorithm.

1 Two points of view on the multiarmed bandit problem

In the parametric stochastic multiarmed bandit model, an agent faces K independent arms which depend on unknown parameters $\theta_1, \dots, \theta_K$. The draw of arm j results in a reward that is read from the i.i.d sequence $(Y_{j,t})_{t\in\mathbb{N}}$ with mean μ_j. The agent sequentially draws the arms and his aim is to find a strategy I_t, where I_t is the arm chosen after t rounds, based on previous rewards $(X_s = Y_{s,I_{s-1}})_{s\leq t}$, that maximizes the expected rewards until time n, $\mathbb{E}_\theta\left[\sum_{t=1}^n X_t\right]$, or equivalently minimizes the cumulated regret :

$$R_n(\theta) = \mathbb{E}_\theta\left[\sum_{t=1}^{n} \mu^* - \mu_{I_{t-1}}\right] \tag{1}$$

where $\mu^* = \max_{j=1\dots K} \mu_j$ is the mean of the optimal arm. For the sake of simplicity, we focus in the sequel on bandits with Bernoulli rewards for which $\mu_j(\theta_j) = \theta_j$ and denote by θ^* the parameter associated to μ^*. Many 'frequentist' algorithms have been developed for this setting. UCB (see [1]) and more recently KL-UCB (see [3] or [11]) are examples of index policies : at each round an index is computed for each arm and the arm with highest index is chosen. The common idea of these algorithms is to use an upper-confidence bound for the empirical mean of rewards received from each arm, in order to balance exploration and exploitation.

In [8], Lai & Robbins show that every strategy draws infinitely often any suboptimal arm j such that

$$\liminf_{n\to\infty} \frac{\mathbb{E}_\theta[N_n(j)]}{\log(n)} \geq \frac{1}{KL(\mathcal{B}(\theta_j), \mathcal{B}(\theta^*))}$$

where KL stands for the Kullback-Leibler divergence and $N_n(j)$ the number of pulls of arm j before time n. An algorithm such that $\limsup_{n\to\infty} \mathbb{E}_\theta[N_n(j)]/\log(n) \leq (KL(\mathcal{B}(\theta_j), \mathcal{B}(\theta^*)))^{-1}$

is therefore asymptotically optimal. So a good regret bound, or equivalently a bound on the average draws of a suboptimal arm, gives us the optimality of an algorithm in this frequentist setting.

The Bayesian formulation Now we adopt a Bayesian point of view and put a prior distribution on the parameter $\theta = (\theta_1, ..., \theta_K)$. The new probabilistic model is :

- $\theta_j \sim \pi_j$ and the π_j are independent
- $\forall j$,the $(Y_{j,t})_t$ are i.i.d conditionally to θ_j with Bernoulli distribution $\mathcal{B}(\theta_j)$
- if $i \neq j$, $\forall t, t'$, $Y_{j,t}$ and $Y_{i,t'}$ are independent

In the frequentist definition (1) of the regret, the expectation is that of the frequentist model, given the fixed parameter θ. In the Bayesian setting, we want to maximize $\mathbb{E}\left[\sum_{t=1}^{n} X_t\right]$, where the expectation is relative to the above probabilistic model. As shown by Gittins, Bayesian optimal policies are index policies, that is a special class of policies where at each round an index is computed for each arm, and the arm with highest index is chosen.

Bayesian algorithms : Bayesian or frequentist optimality ? Let $\Pi_t = (\pi_1^t, ..., \pi_K^t)$ be the current posterior on the arms after t rounds of game. $\Pi_0 = (\pi_1, ..., \pi_K)$, the initial prior distribution. If at round t one chooses $(I_t = j)$ and then observe $X_{t+1} = Y_{j,t+1}$ the Bayesian update for arm j is :

$$\pi_j^{t+1} \propto f(X_{t+1}; \theta_j)\pi_j^t$$

whereas for $i \neq j$, $\pi_i^{t+1} = \pi_i^t$. A Bayesian algorithm is an algorithm that uses states Π_t to determine action I_t. Here we present two Bayesian algorithms. The first, described in Section 2, returns the solution to the Bayesian finite-horizon bandit problem, this means a strategy that maximizes $\mathbb{E}\left[\sum_{t=1}^{n} X_t\right]$ among all strategies for the Bayesian model above. But we also investigate its performance in terms of regret in the frequentist setting. The second, Bayes-UCB, described in Section 3 is a new index policy, with indices inspired by Bayesian ideas since they are posterior distribution quantiles. We show that this Bayesian-motivated algorithm is optimal in the frequentist setting and more strikingly that it is very similar to KL-UCB.

2 Solving the Bayesian problem in the Bernoulli case : the FHG algorithm

In his landmark paper [5], Gittins gives a Bayesian resolution of the multiarmed bandit problem in the special case of Bernoulli rewards. But his solution, an index policy based on dynamic allocation indices, holds for an infinite game with discounted reward. Unlike Gittins, we solve the MDP modelling of our problem for a finite horizon n, which leads us to a Bayesian-optimal solution of the finite-horizon bandit problem. This involves the definition of a new finite-horizon Gittins index (FHG index), and the associated index policy is the finite-horizon Gittins algorithm (FH-Gittins).

The MDP formulation for Bernoulli bandits In the Bayesian model above, we choose $\theta_j \sim$ Beta(a, b), a conjugate prior to the Bernoulli likelihood. The game until the end of round t is summarized by the matrix $S_t \in \mathcal{M}_{K,2}$ where $S_t(j, 1)$ (resp.$S_t(j, 2)$) denotes the number of ones (resp. of zeros) observed from the draws of arm j until time t. Because of the simple Bayesian update with this prior, π_j^t is a Beta distribution with parameter $S_t(j, 1) + a$ and $S_t(j, 2) + b$, with mean $(S_t(j, 1) + a)/(S(j, 1) + S(j, 2) + a + b)$ and therefore ($E_{j,k} \in \mathcal{M}_{K,2}$ being an elementary matrix with a single one at (j, k)) :

$$\mathbb{P}(S_{t+1} = S + E_{j,1} | S_t = S, I_t = j) = \mathbb{E}[X_{t+1} | S_t = S, I_t = j] = \frac{S(j, 1) + a}{S(j, 1) + S(j, 2) + a + b} \quad (2)$$

Equation (2) shows us that each draw of an arm gives us a transition $(S_t, I_t) \rightsquigarrow (S_{t+1}, X_{t+1})$ in a Markov Decision Process (MDP) with states $S \in \mathcal{M}_{K,2}(\mathbb{N})$ and actions $1, ..., K$.

Solving the planning problem for this MDP for a finite horizon n is equivalent to finding a Bayesian-optimal strategy for horizon n. Even if the action and state spaces (included in $\mathcal{M}_{K,2}(\{0, ..., n\})$) are finite, they are too big for a direct resolution using dynamic programming. We use for our finite problem the main idea of Gittins : a reduction of the dimension, by focusing on each arm separately.

Finite-horizon Gittins index : solving a calibration problem Consider the one arm situation where you can alternatively play the arm and get the associated reward or not play and get a fix reward λ. In this auxiliary problem, called $\mathcal{B}_\lambda$, λ represents to cost of playing. Intuitively, at a given time of this finite-horizon game, and given past observation ((s_1, s_2) : ones and zeros obtained from the arm) the higher λ his, the less one should be willing to take the risk of playing. The critical value of λ for which one would still play, even at the cost λ, is the finite-horizon Gittins index, denoted $\nu(t, (s_1, s_2))$. This index can be interpreted as the highest price worth paying for playing the arm. The associated index policy can be proven to be optimal by adapting a proof by Weber for the infinite-discounted case (see for example [6] or [4]).

Implementation of the FHG algorithm Practical computation of index $\nu(t, (s_1, s_2))$ involves repeated resolutions of the $\mathcal{B}_\lambda$ problem using dynamic programming (each resolution is achieved in $O((n-t)^2)$ operations). As the critical value of λ appears as the first zero of a convex function, the number of problems to solve can be reduced. Alternative computational methods are discussed in the recent paper [9]. For moderate values of K and n where these computations can be performed, numerical experiments (see Section 4 below) show that the FHG algorithm regularly outperforms its competitors, when using the frequentist cumulated regret as the measure of performance.

FHG index versus UCB index The Gittins index can also be interpreted as an upper confidence bound, as it can be shown that $\nu(t, s_1, s_2) \geq \frac{s_1+a}{s_1+s_2+a+b}$, the right term being the mean of the posterior. The FHG index however incorporate the knowledge of the horizon n to progressively reduce exploration. In UCB for instance, the index is defined as $\hat{\mu}_{j,N_t(j)} + (\log(t)/(2N_t(j)))^{1/2}$, where $N_t(j) = S_t(j,1) + S_t(j,2)$, and thus increases whenever arm j is not played (i.e. $N_t(j) = N_{t+1}(j)$). In contrast, the FHG index $\nu(t, S_t(j,1), S_t(j,2))$ decreases with time, becoming more greedy as the remaining time decreases. Perhaps more importantly, the FHG index adapts itself to the estimated value of the mean reward, which makes it efficient also in situations where UCB is clearly sub-optimal (with mean rewards close to 0 or 1).

3 A simplified Bayesian algorithm: Bayes-UCB

Practically FHG seems very good for a given frequentist problem, although we have only proven its Bayesian optimality yet and have not analysed its frequentist regret. We focus here on a much simpler algorithm, Bayes-UCB, inspired by the Bayesian model, but for which we exhibit a frequentist regret bound for the Bernoulli case. The idea of exploiting a posterior distribution on the arm is not new, for example ideas presented by Thompson, forerunner of clinical trials, in [12] suggest to draw an arm according to samples from the posterior distribution. The Bayesian Learning Automaton advocated by Granmo in [7] uses this idea, but in a two-armed setting and no regret analysis is proposed. Here we won't use samples but quantiles, which will lead to a Bayesian-inspired index policy. The use of quantile is not new either, since it appears in Interval Estimation methods mentionned in [10] for example, but only fixed quantiles are used, and again there is no regret analysis.

Recall that Π_0 is the initial prior on θ and Π_t the current posterior. Bayes-UCB is the index policy associated to the index $q_j(t) = (1 - 1/(t\log(n)^c))$ − quantile of distribution π_j^t. For the Bernoulli case with Beta$(1,1)$ (uniform) prior :

$$q_j(t) = \left(1 - \frac{1}{t\log(n)^c}\right) - \text{quantile of the distribution Beta}(S_t(j)+1, N_t(j) - S_t(j) + 1)$$

Bounding the quantile Since the Bayes-UCB algorithm for Bernoulli uses beta-quantile, we want to bound them as tightly as possible to have a good estimation of the Bayesian-based index $q_j(t)$. The following connection between Beta and Binomial distributions is useful : a Beta(a,b) (with a,b integers) can be seen as the a-th order statistic among $a+b-1$ uniform random variables, so

$$\mathbb{P}(X \geq x) = \mathbb{P}(\text{less than } a-1 \text{ r.v. are} \leq x) = \mathbb{P}(S_{a+b-1,x} \leq a-1)$$

where $S_{n,x}$ denotes a binomial distribution with parameters n and x. Bounding the beta quantiles boils down to controlling the binomial tail, which can be done with a Sanov inequality:

$$\frac{1}{n+1} e^{-nKL(\mathcal{B}(\frac{k}{n}),\mathcal{B}(x))} \leq \mathbb{P}(S_{n,x} \geq k) \leq e^{-nKL(\mathcal{B}(\frac{k}{n}),\mathcal{B}(x))}$$

This leads to a tight bound for beta quantiles and, as a result, for $q_j(t)$. Denoting $d(x,y) = KL\left(\mathcal{B}(x), \mathcal{B}(y)\right)$, we have $\tilde{u}_j(t) \leq q_j(t) \leq u_j(t)$ with :

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

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

Interestingly, the upper bound $u_j(t)$ is exactly the index used in the KL-UCB algorithm, whereas $\tilde{u}_j(t)$ corresponds to a biased version of KL-UCB, where, additionally, t is replaced by $t/(N_j(t)+2)$ in the logarithmic term. This latter alternative form of the exploration bonus, which appears naturally in this Bayesian-inspired setting, has been suggested before in the literature. For example the MOSS algorithm [2] is inspired by UCB and uses $n/(KN_t(j))$ instead of t in the exploration term. Similarly in [3], the KL-UCB+ version, using $t/N_t(j)$ instead of t is found to be practically more efficient. For Bayes-UCB, we were able to bound the regret as follows, thus showing that the algorithm is asymptotically optimal.

Theorem 1 *Let $\epsilon > 0$. For the Bayes-UCB algorithm with parameter $c \geq 5$, the number of draws of a sub-optimal arm j is such that :*

$$\mathbb{E}[N_n(j)] \leq \frac{(1+\epsilon)}{KL\left(\mathcal{B}(\theta_j), \mathcal{B}(\theta^*)\right)} \log(n) + o_{\epsilon,c}\left(\log(n)\right)$$

4 Experiments and conclusions

Numerical experiments have been carried out in a frequentist setting : for a fixed parameter θ and an horizon n, N bandit games with Bernoulli rewards are repeated for a given strategy. Some of these simulations (not shown here) illustrate the behaviour of our two algorithms : for instance empirical distribution of the number of draws of the optimal arm or of the regret confirms that FH-Gittins is more risky than its frequentist counterparts and that KL-UCB and Bayes-UCB play in a very similar way whatever θ is. But the main purpose of our numerical experiments is to compare the performance in terms of cumulated regret of our two algorithms 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 $n = 500$. In the 0.45/0.55 (right) situation, where UCB and KL-UCB behave quite similarly, FH-Gittins already outperforms KL-UCB, but the difference is even more significant in the 0.1/0.2 (left) situation, where UCB is (provably) worse than KL-UCB. In these two settings, Bayes-UCB also improves over KL-UCB but its performances are less striking than those of the FH-Gittins algorithm.

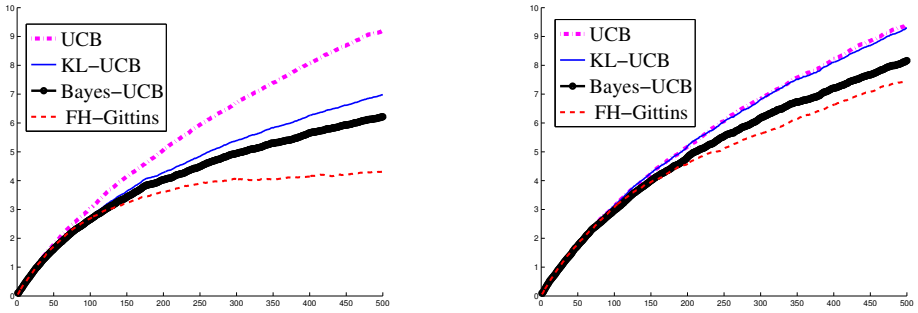

Figure 1: Cumulated regret for the two armed-bandit problem 0.1/0.2 (left) and 0.45/0.55 (right)

Although frequentist and Bayesian bandits are based on two different probabilistic frameworks, using Bayesian ideas leads to efficient algorithms for solving the frequentist multiarmed bandit problem. Gittins algorithm displays impressive performance but is only implementable for moderate values of the horizon. In contrast, Bayes-UCB is easy to implement and can be proven to be asymptotically optimal. Although we have focussed here on the case of Bernoulli rewards, these ideas can be extended to other cases such as Gaussian or, more generally, exponential family rewards.

References

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

[2] Jean-Yves Audibert, Sbastien Bubeck, Regret Bounds and Minimax Policies under Partial Monitoring *Journal of Machine Learning Research*, 2010

[3] Aurélien Garivier, Olivier Cappé, The KL-UCB algorithm for bounded stochastic bandits and beyond COLT, 2011

[4] Esther Frostig, Gideon Weiss, Four proofs of Gittins' multiarmed bandit theorem In *Applied Probability Trust*, 1999

[5] John Gittins, Bandit Processes and Dynamic Allocation Indices In *Journal of the Royal Statistical Society*, 1979

[6] John Gittins, Kevin Glazebrook and Richard Weber, Multi-armed bandit allocation indices (2nd Edition) Wiley, 2011

[7] Ole Christopher Granmo, Solving Two-Armed Bernoulli Bandit Problems Using a Bayesian Learning Automaton in *International Journal of Intelligent Computing and Cybernetics (IJICC)* Volume 3, Issue 2, 2010, pp. 207 - 234, 2010

[8] T.L. Lai, Herbert Robbins, Asymptotically efficient adaptive allocation rules in *Advances in applied mathematics*, 1985

[9] J. Niño-Mora, Computing a Classic Index for Finite-Horizon Bandits in INFORMS Journal on Computing 23(2) 254-267, 2011

[10] 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

[11] Odalric-Ambrym Maillard, Rémi Munos, Gilles Stoltz, A finite-time analysis of Multi-armed bandits problems with Kullback-Leibler Divergence COLT, 2011

[12] W.R. Thompson, On the Likelihood That one unknown Probability Exceeds Another in View of the Evidence of Two Samples *Biometrika* 25 285-294, 1933