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

Emilie Kaufmann, Olivier Cappé and Aurélien Garivier (name@telecom-paristech.fr)

TELECOM ParisTech

Introduction

Are bayesian algorithms for the multiarmed bandit (MAB) problem optimal towards frequentist measure of performance? Our answer relies on both numerical experiments for a bayesian optimal policy, adapted from Gittins ideas, and an optimal regret bound for the Bayesian-inspired Bayes-UCB

Bayesian vs. Frequentist Model for MAB

K independent arms depending on a parameter θ (Bernoulli distribution for the sake of simplicity); optimal arm is $j^* = \text{argmax}\ \theta_j$ and $\theta^* = \theta_{j^*}$ is the highest expectation of reward associated

Two probabilistic modellings

Frequentist :

- $\theta_1, \dots, \theta_K$ unknown parameters
- $(Y_{j,t})_t$ i.i.d. with Bernoulli distribution $\mathcal{B}(\theta_j)$

Bayesian :

- $\theta_j \overset{i.i.d.}{\sim} \pi_j$
- $(Y_{j,t})_t$ are i.i.d. conditionally to θ_j with distribution $\mathcal{B}(\theta_j)$

At time $t+1$, arm I_t is chosen and reward $X_{t+1} = Y_{I_t,t+1}$ is observed

Two measures of performance

- Minimize (classic) regret

$$R_n(\theta) = \mathbb{E}_\theta[\sum_{t=1}^{n} \theta^* - \theta_{I_{t-1}}]$$

- Minimize "bayesian" regret

$$R_n = \int R_n(\theta) d\pi(\theta)$$

Bayes-UCB : a simple Bayesian strategy

Some ideas for using the posterior :

- sampling from the posterior (Thompson Sampling)
- using quantiles : fixed or adaptive (Bayes-UCB)
- adapt the bayesian exact solution from Gittins (FHG-algorithm)

Bayes-UCB algorithm is the index policy associated to

$$q_j(t) = \left(1 - \frac{1}{t(\log n)^c}\right) - \text{quantile of the posterior distribution}$$
$$\text{Beta}(S_t(j,1)+1, S_t(j,2)+1)$$

(in practice, we take $c = 0$)

- **Theoretical guarantee:** frequentist 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 :*

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

This leads to an upper-bound for the regret matching the Lai&Robbins lower bound on the number of draws of suboptimal arms

Bayes-UCB versus frequentist algorithms

The Bayes-UCB index appears to be very close to the recently-proposed KL-UCB algorithm [1]: $\tilde{u}_j(t) \leq q_j(t) \leq u_j(t)$ with :

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

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

where $d(x,y) = KL\left(\mathcal{B}(x), \mathcal{B}(y)\right) = x \log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$

Bayes-UCB appears to build automatically confidence intervals based on Kullback-Leibler divergence, that are adapted to the geometry of the problem in this specific case

References

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

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

Generic Bayesian algorithm

Let $\Pi_t = (\pi_1^t, \dots, \pi_K^t)$ be the current posterior on the arms after t rounds of game. 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 :

$$\pi_j^{t+1} \propto f(X_{t+1}; \theta_j)\pi_j^t \quad \text{and for} \quad i \neq j,\ \pi_i^{t+1} = \pi_i^t$$

A Bayesian algorithm uses Π_t to determine action I_t. We focus on Bayesian Index Policies : an index is computed using Π_t and arm with highest index is chosen

MDP Formulation for Bernoulli Bandits

$$\begin{array}{cccc} & \text{ones observed} & \text{zero observed} & \\ S_{11} = & \multicolumn{2}{c}{\begin{pmatrix} 1 & 2 \\ 5 & 1 \\ 0 & 2 \end{pmatrix}} & \longleftarrow \text{index of the arm} \end{array}$$

Beta(a,b) the prior over each arm

Matrix $S_t \in \mathcal{M}_{K,2}$ summarizes the game :

- $S_t(j,1)$ (resp. $S_t(j,2)$) is the number of ones (resp. zeros) observed from arm j until time t
- Line j gives the parameters of the Beta posterior over arm j, π_j^t

(S_t, X_t) is a trajectory in a MDP with transition and reward function :

$$\mathbb{P}(S_{t+1} = S + E_{j,1} | S_t = S, I_t = j) =$$

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

Gittins solves this MDP by resorting to a calibration problem for each arm (reduction of the dimension), but as an infite-horizon, discounted MDP

Finite-Horizon-Gittins algorithm

The solution of the above MDP, for a given horizon n, and without discount, also reduces to an index policy

Finite-Horizon Gittins index For a given arm, at time t of game and given the observation of (s_1, s_2)

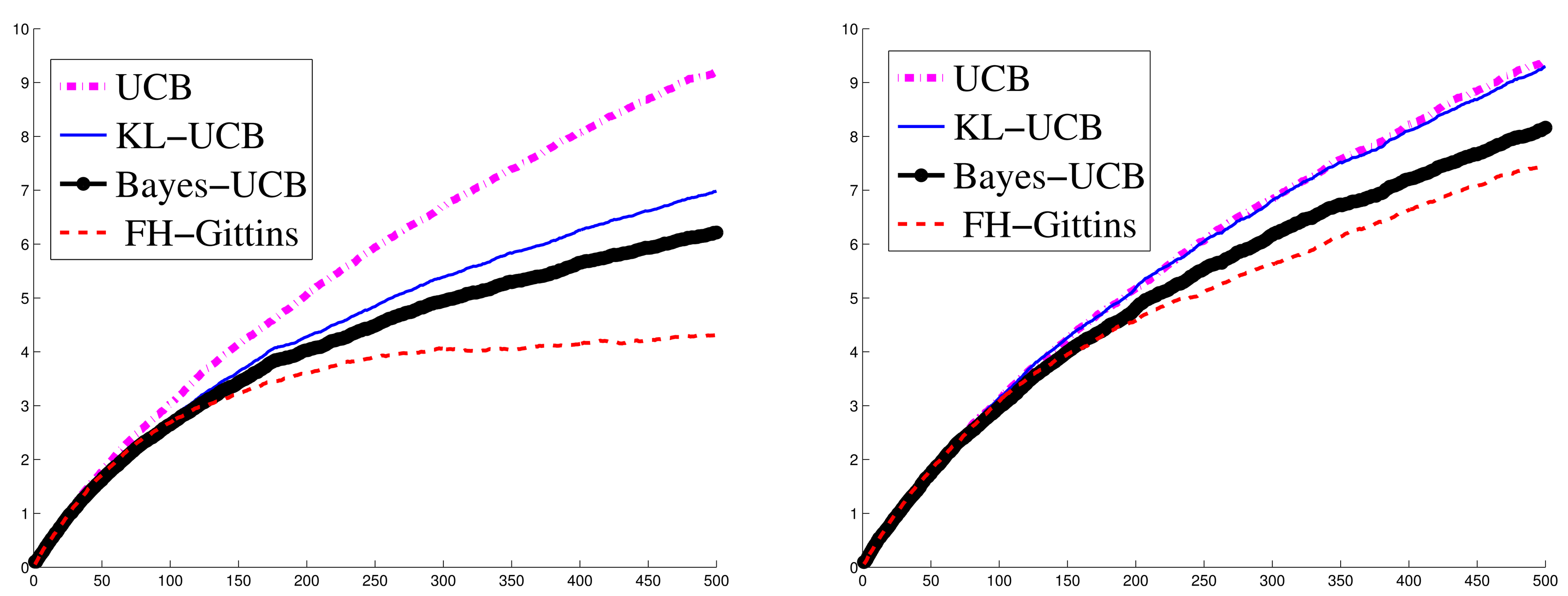
$$\nu(t, (s_1, s_2)) = \sup_{\substack{0 < \tau \leq n-t \\ \text{stopping time}}} \frac{\mathbb{E}_{(s_1,s_2)}\left[\sum_{k=0}^{\tau-1} X_{t+1}\right]}{\mathbb{E}_{(s_1,s_2)}[\tau]}$$

where $(X_t)_t$ denote the successive rewards obtained from this arm, and the expectation involves a prior Beta(s_1+a, s_2+b) on its parameter

FH-Gittins algorithm is the index policy associated to $\nu(t, S_t(j,:))$
The computation of the index is more involved

- **Theoretical guarantee:** Bayesian optimal

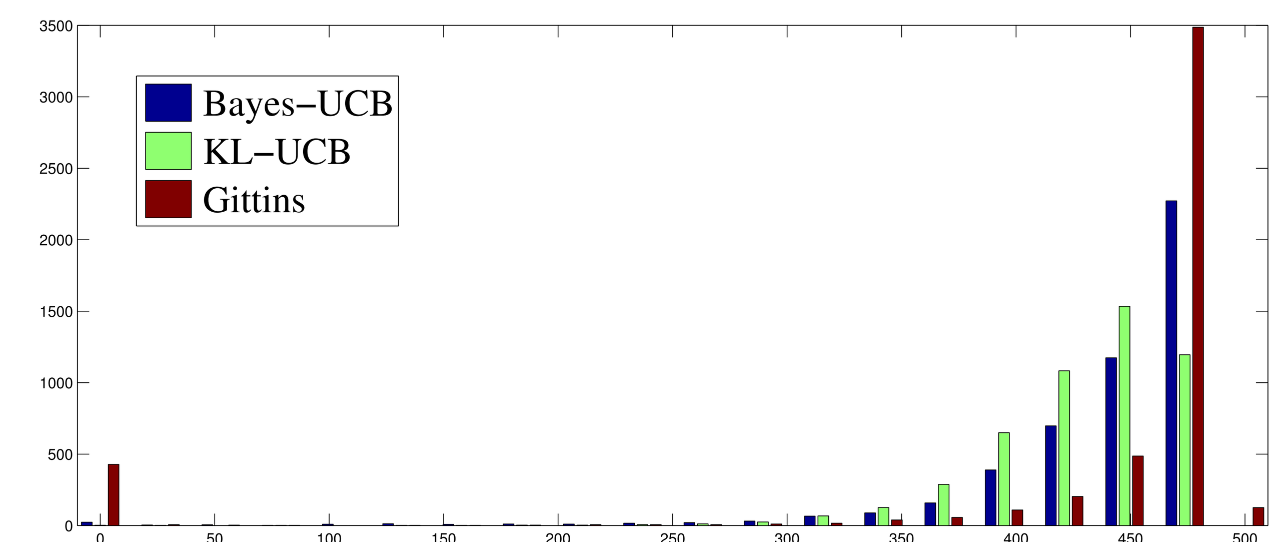
Comparison and numerical experiments



$\theta_1 = 0.1, \theta_2 = 0.2$ $\qquad$ $\theta_1 = 0.45, \theta_2 = 0.55$

Cumulated regret curves for several strategies (estimated with $N = 5000$ repetition of the bandit game with horizon $n = 500$) in a low-reward (left) or an average reward (right) problems

- Gittins improves significantly over all algorithms
- Bayes-UCB and KL-UCB have a very similar behaviour regardless of the problem, Bayes-UCB beeing slightly better
- Gittins is more risky (less explorative) than other algorithms



Histogram (for $N = 5000$ bandit-games) of the number of draws of optimal arm in a two armed problem with $\theta_1 = 0.8, \theta_2 = 0.9$ and horizon $n = 500$.