On Bayesian Upper Confidence Bounds for bandit problems

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

TELECOM ParisTech

In a nutshell

What is the performance of Bayesian bandit algorithms from a frequentist point of view? Not only does Bayes-UCB show striking similarities with its frequentist counterparts, but it appears to outperform them on their own ground, which is supported by an optimal regret bound for the Bernoulli case.

Bayesian vs. Frequentist Model for MAB

K independent arms. Arm j depends on parameter θ_j and has expectation μ_j; optimal arm is $j^* = \text{argmax}\ \mu_j$ and $\mu^* = \mu_{j^*}$ is the highest expectation of reward associated.

Two probabilistic modelings

Frequentist :

- $\theta_1, \dots, \theta_K$ unknown parameters
- $(Y_{j,t})_t$ is i.i.d. with distribution ν_{θ_j}

Bayesian :

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

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

- Minimize "bayesian" regret

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

Background

- $\Pi_t = (\pi_1^t, \dots, \pi_K^t)$ the current posterior over $(\theta_1, ..., \theta_K)$
- $\Lambda_t = (\lambda_1^t, \dots, \lambda_K^t)$ the current posterior over the means $(\mu_1, ..., \mu_K)$

A Bayesian algorithm uses Π_{t-1} to determine action I_t.

Our inspiration: frequentist index policies using:

- Upper Confidence Bound for the empirical mean... (UCB)
- ... built using KL-divergence (KL-UCB, frequentist optimal)

Some ideas to design Bayesian bandit algorithms:

- adapt the Bayesian exact solution from Gittins (Finite-Horizon Gittins algorithm, Bayesian optimal)
- sample from the posterior (Thompson Sampling: dates back to 1933, recent upper bound on its frequentist regret by Agrawal and Goyal)
- **use quantiles: fixed or adaptive** (Bayes-UCB)

Our algorithm: Bayes-UCB

Bayes-UCB algorithm is the index policy associated to:

$$q_j(t) \quad = \quad Q\left(1 - \frac{1}{t(\log t)^c}, \lambda_j^{t-1}\right)$$

This means at time t choose $I_t = \underset{j=1\dots K}{\text{argmax}}\ q_j(t)$

Parameters : c (in practice, take $c=0$), initial prior Π_0

Case 1: Binary bandits

ν_{θ_j} is the Bernoulli distribution $\mathcal{B}(\theta_j)$, π_j^0 the (conjugate) prior Beta$(1,1)$

- **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.

- **Link to a frequentist algorithm:**

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

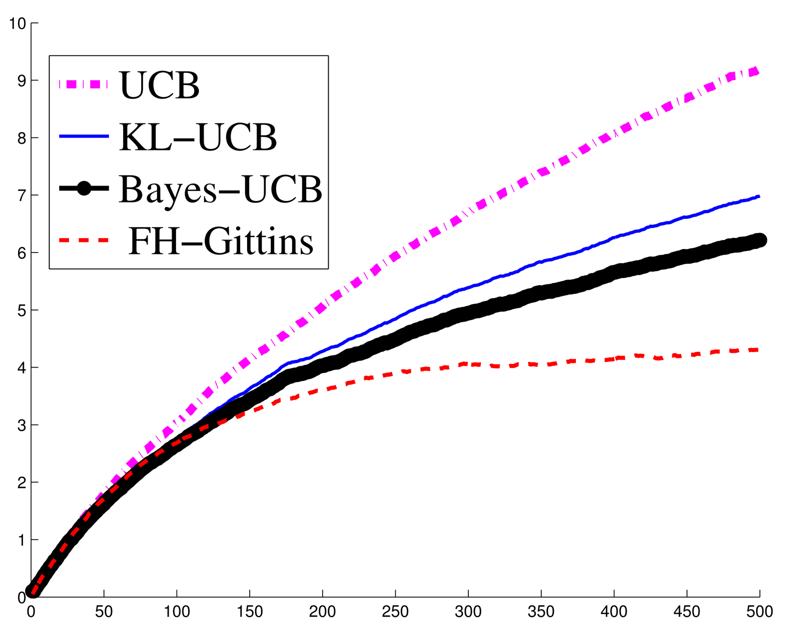
$$u_j(t) \quad = \quad \underset{x > \frac{S_t(j)}{N_j(t)}}{\text{argmax}} \left\{ d\left(\frac{S_t(j)}{N_t(j)}, x\right) \leq \frac{\log(t) + c\log(\log(n))}{N_t(j)} \right\}$$

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

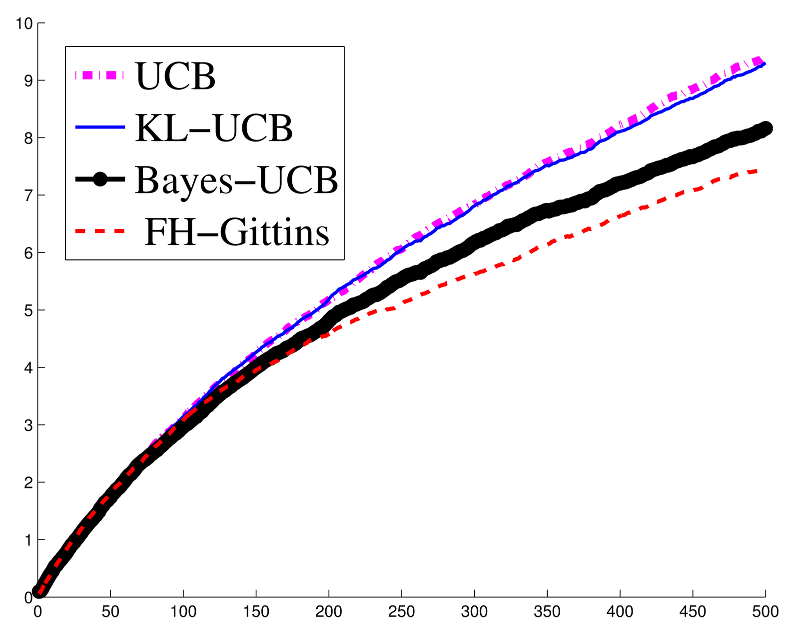
where $d(x,y) = KL\left(\mathcal{B}(x), \mathcal{B}(y)\right) = x \log \frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$

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.

- **Numerical experiments:**



$\theta_1 = 0.1, \theta_2 = 0.2$



$\theta_1 = 0.45, \theta_2 = 0.55$

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

Case 2: the exponential family

- **Canonical exponential family:** we observe empirically that the link between the Bayes-UCB and the KL-UCB index generalizes, and we obtain theoretical guarantees for Gaussian bandits $\nu_\theta = \mathcal{N}(\theta, 1)$
- **A two-dimensional example:** Gaussian distribution $\nu_{\theta_j} = \mathcal{N}(\mu_j, \sigma_j^2)$, with both mean μ_j and variance σ_j^2 unknown

$$q_j(t) = \frac{S_j(t)}{N_j(t)} + \sqrt{\frac{S_t^{(2)}(j)}{N_j(t)}} Q\left(1 - \frac{1}{t}, \mathcal{T}(N_t(j) - 1)\right) \text{ with } \pi_j^0(\mu_j, \sigma_j) = \frac{1}{\sigma_j^2}$$

$\rightarrow$ empirically better than Auer UCB1-norm, very similar index

Case 3: linear bandit problem

- arms : fixed vectors $U_1, ..., U_K \in \mathbb{R}^d$
- parameter of the model : $\theta \in \mathbb{R}^d$
- reward : $y_t = U_{I_t}'\theta + \sigma\epsilon_t$ with $\epsilon_t \sim \mathcal{N}(0,1)$
- goal : minimize regret $\mathbb{E}_\theta\left[\sum_{t=1}^n \left(\max_{1\leq j\leq K}(U_j'\theta) - U_{I_t}'\theta\right)\right]$

With a Gaussian prior: $\boldsymbol{\theta \sim \mathcal{N}\left(0, \kappa^2 \mathrm{I}_d\right)}$ The posterior is

$$\theta | X_t, Y_t \ \sim \mathcal{N}(\underbrace{(X_t'X_t + (\sigma/\kappa)^2 \mathrm{I}_d)^{-1} X_t' Y_t}_{\hat{\theta}_t}, \underbrace{\sigma^2 (X_t'X_t + (\sigma/\kappa)^2 \mathrm{I}_d)^{-1}}_{\Sigma_t})$$

Therefore

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

While a frequentist approach based on uncertainty ellipsoids leads to:

$$q_j(t) = U_j'\hat{\theta}_t + ||U_j||_{\Sigma_t}\beta_t(\delta) \ \text{ with } \ \mathbb{P}\left((\theta - \hat{\theta}_t)\Sigma_t^{-1}(\theta - \hat{\theta}_t) \leq \beta_t(\delta)\right) \geq 1 - \delta$$

With a sparsity-inducing prior: $\boldsymbol{\theta_j \sim \epsilon\delta_0 + (1-\epsilon)\mathcal{N}(0, \kappa^2)}$

In this case we can sample from the posterior using a Gibbs sampler, and estimate the quantiles used in Bayes-UCB. Here is the cumulated regret in a sparse problem with 20 arms and $d = 10$ for Bayes-UCB with different prior distributions. The oracle uses a Gaussian prior on the known non-zero components of θ.