On Bayesian Upper Confidence Bounds for bandit problems

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

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 $\theta_j$ and has expectation $\mu_j$; optimal arm is $j^\star = \arg \max_j \mu_j$ and $\mu^\star = \mu_{j^\star}$ is the highest expectation of reward associated.

Two probabilistic modelings

<table>
<thead>
<tr>
<th>Frequentist</th>
<th>Bayesian</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1, \ldots, \theta_K$ unknown parameters</td>
<td>$\theta_j \sim \pi_j$</td>
</tr>
<tr>
<td>$(Y_{t,j})<em>j$ is i.i.d. with distribution $\nu</em>{\theta_j}$</td>
<td>$(Y_{t,j})<em>j$ is i.i.d. conditionally to $\theta_j$ with distribution $\nu</em>{\theta_j}$</td>
</tr>
</tbody>
</table>

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 classical regret
- Minimize “bayesian” regret

CASE 1: BINARY BANDITS

$\nu_0$ is the Bernoulli distribution $\mathcal{B}(\theta_j)$, $\pi_j$ the (conjugate) prior $\text{Beta}(1,1)$.

Theoretical guarantee: frequentist optimal

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

$$\mathbb{E}_\theta[\mathbf{N}_t(j)] \leq \frac{1 + \epsilon}{KL(B(\alpha_j), B(\theta_j))} \log(n) + o(t, \log(n))$$

This leads to an upper-bound for the regret matching the LaïkRobbins 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:

$$u_j(t) = \arg \max_{\theta \sim \pi_j} d \left( \frac{S_j(t) / N_j(t)}{N_j(t)} \right) \leq \frac{\log(t) + c \log(\log(n))}{N_j(t)}$$

$$\tilde{u}_j(t) = \arg \max_{\theta \sim \pi_j} d \left( \frac{S_j(t) / (N_j(t) + 1)}{N_j(t) + 1} \right) \leq \frac{\log(\log(n)) + c \log(\log(n))}{(N_j(t) + 1)}$$

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

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:

**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 $\nu_0 = \mathcal{N}(\theta, 1)$.

- A two-dimensional example: Gaussian distribution $\nu_{\theta_j} = \mathcal{N}(\mu_j, \sigma_j^2)$, with both mean $\mu_j$ and variance $\sigma_j^2$ unknown

$$q_j(t) = \frac{S_j(t)}{N_j(t)} + \sqrt{\frac{S_j(t)}{N_j(t)^2} \frac{\log(\log(n)) + c \log(\log(n))}{N_j(t) + 1}}$$

This means at time $t$ choose $I_t = \arg \max_j q_j(t)$ with $\pi_0(\mu_j, \sigma_j) = \frac{1}{\sigma_j^2}$.

PARAMETERS: $c$ (in practice, take $c = 0$), initial prior $\Pi_0$.

**CASE 3: LINEAR BANDIT PROBLEM**

-arms : fixed vectors $U_1, \ldots, U_K \in \mathbb{R}^d$

- parameter of the model : $\theta \in \mathbb{R}^d$

-reward : $Y_i = U_i^\top \theta + \sigma \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, 1)$

-goal : minimize regret $\mathbb{E}_\theta[\max_{s \leq t} \{U_\star^\top \theta - U_s^\top \theta\}]$

With a Gaussian prior:

$$\theta | X_i, Y_i \sim \mathcal{N}(X_i^\top X_{\pi}, \sigma^2 X_{\pi}^2 + \sigma^2 \epsilon_i^2)^{-1} X_i^\top Y_i, \sigma^2 (X_i^\top X_i + \sigma^2 \epsilon_i^2)^{-1}$$

Therefore

$$q_j(t) = U_j^\top \hat{\theta}_t + ||U_j||_{\Sigma_{\pi}} \sqrt{\frac{t}{2}} \mathcal{N}(0, 1)$$

While a frequentist approach based on uncertainty ellipsoids leads to:

$$q_j(t) = U_j^\top \hat{\theta}_t + \frac{||U_j||_{\Sigma_{\pi}}}{\sqrt{t}} \beta_t(\delta)$$

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

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 $\theta$. 

**OUR ALGORITHM: BAYES-UCB**

Bayes-UCB algorithm is the index policy associated to:

$$q_j(t) = Q \left( 1 - \frac{1}{t \log(\log(n))} \right)$$

This means at time $t$ choose $I_t = \arg \max_j q_j(t)$ with $\pi_0(\mu_j, \sigma_j) = \frac{1}{\sigma_j^2}$

Parameters: $c$ (in practice, take $c = 0$), initial prior $\Pi_0$.

**BACKGROUND**

- $I_t = \{\theta_1, \ldots, \theta_K\}$ the current posterior over $(\theta_1, \ldots, \theta_K)$
- $A^*_t = \{\lambda_1, \ldots, \lambda_K\}$ the current posterior over the means $(\mu_1, \ldots, \mu_K)$

A Bayesian algorithm uses $I_{t-1}$ 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)