On Bayesian bandit algorithms

Emilie Kaufmann

joint work with

Olivier Cappé, Aurélien Garivier, Nathaniel Korda and Rémi Munos

TELECOM ParisTech

July 1st, 2012

1. Bayesian bandits versus Frequentist Bandit

1. Bayesian bandits versus Frequentist Bandit
2. Gittins' Bayesian solution
3. The Bayes-UCB algorithm

Two probabilistic modelling

K independent arms. $\mu^* = \mu_{j^*}$ highest expectation of reward.

Frequentist :

- $\theta_1, \ldots, \theta_K$ unknown parameters
- $(Y_{j,t})_t$ is i.i.d. with distribution ν_{θ_j} with mean μ_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, arm I_t is chosen and reward $X_t = Y_{I_t,t}$ is observed

Two measures of performance

- Minimize (classic) regret

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

- Minimize bayesian regret

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

Our goal

Design Bayesian bandit algorithms
which are optimal in terms of frequentist regret

- Lai and Robbins asymptotic rate for the regret:

$$\liminf \frac{\mathbb{E}_\theta[N_n(j)]}{\log(n)} \geq \frac{1}{\mathrm{KL}(\nu_{\theta_j}, \nu_{\theta^*})} \quad \text{if } j \text{ is non optimal}$$

$$\liminf \frac{\mathbb{E}_\theta[R_n]}{\log(n)} \geq \sum_{j \text{ non optimal}} \frac{\mu^* - \mu_j}{\mathrm{KL}(\nu_{\theta_j}, \nu_{\theta^*})}$$

Some Bayesian and frequentist algorithms

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

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

Frequentist algorithms:

- upper confidence bound on the empirical mean (UCB) [Auer at al. 2002]
- UCB based on KL-divergence (KL-UCB) [Garivier, Cappé 2011]

Bayesian algorithms:

- Gittins indices [Gittins, 1979]
- quantiles of the posterior
- samples from the posterior [Thompson, 1933]

The Finite-Horizon Gittins algorithm

Often heard : Gittins solved the Bayesian MAB
ONLY PARTIALLY TRUE

- gives an optimal policy for bayesian **discounted** regret
- only for simple parametric cases

Finite-Horizon Gittins algorithm :

- is **Bayesian optimal** for the **finite horizon problem**
- involves indices hard to compute
- is heavily horizon-dependent
- no theoretical proof of its frequentist optimality

The general algorithm

Recall :

- $\Lambda_t = (\lambda_1^t, \dots, \lambda_K^t)$ is the current posterior over the means $(\mu_1, ..., \mu_K)$

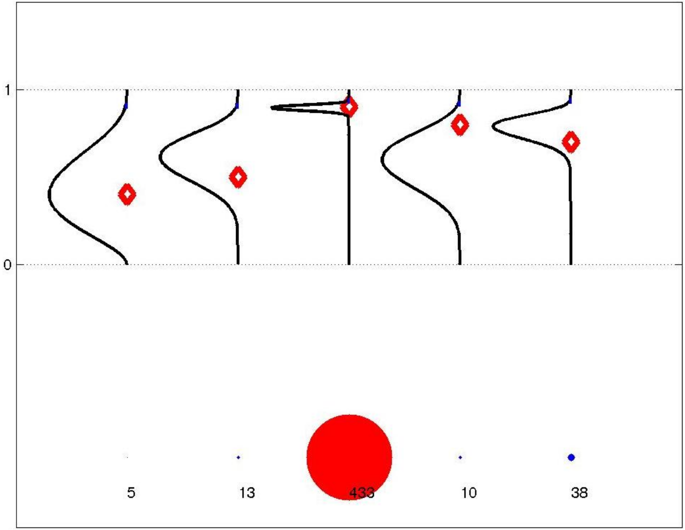
The **Bayes-UCB algorithm** is the index policy associated with:

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

ie, at time t choose

$$I_t = \underset{j=1...K}{\text{argmax}}\ q_j(t)$$

An illustration for Bernoulli bandits



Theoretical results for the Bernoulli case

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

- **Bayes-UCB is frequentist optimal in this case**

Theorem (Kaufmann,Cappé,Garivier 2012)

Let $\epsilon > 0$; for the Bayes-UCB algorithm with parameter $c \geq 5$, the number of draws of a suboptimal arm j is such that :

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

- **Link to a frequentist algorithm:**

Bayes-UCB index is close to KL-UCB index: $\tilde{u}_j(t) \leq q_j(t) \leq u_j(t)$ with:

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

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

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.

Where does it come from?

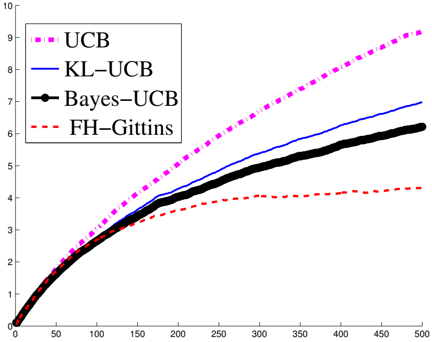
- First element: link between Beta and Binomial distribution:

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

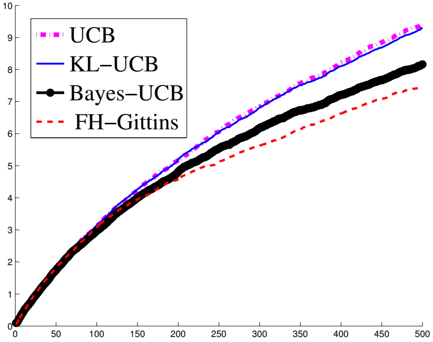
- Second element: Sanov inequality leads to the following inequality:

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

Experimental results



$\theta_1 = 0.1, \theta_2 = 0.2$



$\theta_1 = 0.45, \theta_2 = 0.55$

Cumulated regret curves for several strategies (estimated with $N = 5000$ repetitions of the bandit game with horizon $n = 500$) for two different problems

Beyond the Bernoulli case

In more general cases, **the Bayes-UCB algorithm is very close to existing frequentist algorithms**:

- bandits with rewards in a one parameter exponential familly
- Gaussian bandits with unknown mean and variance
- Linear Bandit setting with prior over the parameter

The algorithm

- A very simple algorithm:

$$\forall j \in \{1..K\}, \quad s_{j,t} \sim \lambda_j^t$$
$$I_t = \text{argmax}_j \; s_{j,t}$$

- (Recent) interest for this algorithm:
  - a very old algorithm : dates back to 1933
  - partial analysis proposed [Granmo 2010][May, Korda, Lee, Leslie 2011]
  - extensive numerical study beyond the Bernoulli case [Chapelle, Li 2011]
  - first logarithmic upper bound on the regret [Agrawal,Goyal COLT 2012]

An optimal regret bound for the Bernoulli case

Assume the first arm is the unique optimal and $\Delta_a = \mu_1 - \mu_a$.

- First upper bound :

Theorem (Agrawal,Goyal, 2012)

$$\mathbb{E}[R_n] \leq C \left(\sum_{j=2}^{K} \frac{1}{\Delta_j} \right) \log(n) + o_\theta(\log(n))$$

- First optimal upper bound :

Theorem (Kaufmann,Korda,Munos 2012)

$\forall \epsilon > 0$

$$\mathbb{E}[R_n] \leq (1+\epsilon) \left(\sum_{j=2}^{K} \frac{\Delta_j}{KL\left(\mathcal{B}(\mu_j), \mathcal{B}(\mu_1)\right)} \right) \log(n) + o_{\theta,\epsilon}(\log(n))$$

Sketch of the analysis

- Bound the expected number of draws of a suboptimal arm j (1 is optimal)
- A usual decomposition in an index policy analysis is

$$\mathbb{E}[N_t(j)] \leq \sum_{t=1}^{T} \mathbb{P}\left(ind_{1,t} < \mu_1\right) + \sum_{t=1}^{T} \mathbb{P}\left(ind_{j,t} \geq ind_{1,t} > \mu_1, I_t = j\right)$$

- Decomposisition used for Thompson Sampling is

$$\mathbb{E}[N_t(j)] \leq \sum_{t=1}^{T} \mathbb{P}\left(s_{1,t} \leq \mu_1 - \sqrt{\frac{6\ln(t)}{N_t(1)}}\right) + \sum_{t=1}^{T} \mathbb{P}\left(s_{j,t} > \mu_1 - \sqrt{\frac{6\ln(t)}{N_t(1)}}, I_t = j\right)$$

Sketch of the analysis

- An extra deviation inequality is needed

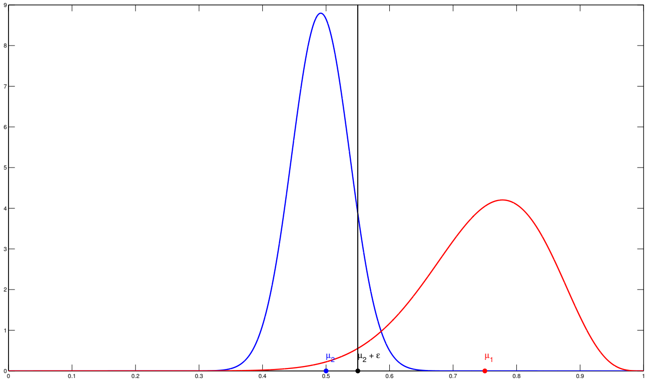
Proposition

There exists constants $b = b(\mu_1, \mu_j) \in (0, 1)$ *and* $C_b = C_b(\mu_1, \mu_j) < \infty$ *such that*

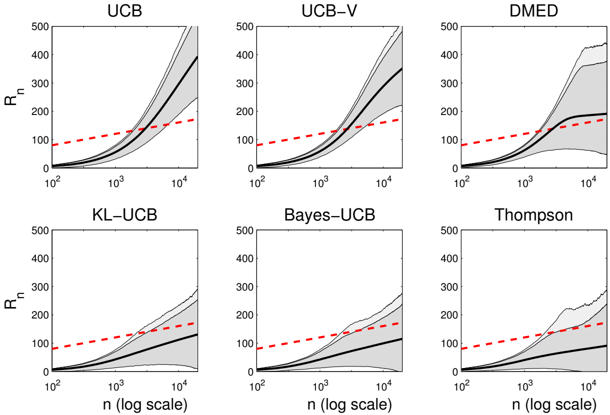
$$\sum_{t=1}^{\infty} \mathbb{P}\left(N_t(1) \leq t^b\right) \leq C_b.$$

Where does it come from?

$$\left(N_t(1) \leq t^b\right) = \Big(\text{ exists a time range of length at least } t^{1-b} - 1 \text{ with no draw of arm 1}\Big)$$



Numerical summary



Regret as a function of time (on a log scale) in a ten arms problem with low rewards, horizon $n = 20000$, average over $N = 50000$ trials.

Conclusion and perspectives

You are now aware that:

- Bayesian algorithms are efficient for the frequentist MAB problem
- Bayes-UCB show striking similarity with frequentist algorithms
- Bayes-UCB and Thompson Sampling are optimal for Bernoulli bandits

Some perspectives:

- A better understanding of the Finite-Horizon Gittins indices
- Using Thompson with more involved priors
- A more general analysis of Bayes-UCB and Thompson Sampling

Summary of the contributions

- Gittins and Bayes-UCB algorithm:

  Emilie Kaufmann, Olivier Cappé and Aurélien Garivier
  On Bayesian upper confidence bounds for bandits problem
  AISTATS 2012.

- Analysis of Thompson Sampling:

  Emilie Kaufmann, Nathaniel Korda and Rémi Munos
  Thompson Sampling : an optimal finite time analysis
  Submitted.