Un point de vue bayésien pour des algorithmes de bandits plus performants

Emilie Kaufmann

SMILE, 19 novembre 2012

1. Bayesian Bandits, Frequentist Bandits

Two probabilistic modellings

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

Frequentist :

- $\theta_1, \dots, \theta_K$ unknown parameters
- $(Y_{a,t})_t$ is i.i.d. with distribution ν_{θ_a} with mean μ_a

Bayesian :

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

At time t, arm A_t is chosen and reward $X_t = Y_{A_t,t}$ is observed

Two measures of performance

- Minimize (classic) regret

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

- Minimize bayesian regret

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

Asymptotically optimal algorithms in the frequentist setting

$N_a(t)$ the number of draws of arm a up to time t

$$R_n(\theta) = \sum_{a=1}^{K} (\mu^* - \mu_a)\mathbb{E}_\theta[N_a(n)]$$

- Lai and Robbins,1985 : every consistent policy satisfies

$$\mu_a < \mu^* \Rightarrow \liminf_{n\to\infty} \frac{\mathbb{E}_\theta[N_a(n)]}{\ln n} \geq \frac{1}{\mathsf{KL}(\nu_{\theta_a}, \nu_{\theta^*})}$$

- A bandit algorithm is **asymptotically optimal** if

$$\mu_a < \mu^* \Rightarrow \limsup_{n\to\infty} \frac{\mathbb{E}_\theta[N_a(n)]}{\ln n} \leq \frac{1}{\mathsf{KL}(\nu_{\theta_a}, \nu_{\theta^*})}$$

Our goal

Design Bayesian bandit algorithms
that are asymptotically optimal in terms
of frequentist regret

Some sucessfull frequentist algorithms

The following heuristic defines a family of optimistic index policies:

- For each arm a, compute a confidence interval on the unknown parameter μ_a:

$$\mu_a \leq UCB_a(t) \quad w.h.p$$

- Use the *optimism-in-face-of-uncertainty principle*:

'act as if the best possible model was the true model'

The algorithm chooses at time t arm with highest Upper Confidence Bound

$$A_t = \arg\max_a \ UCB_a(t)$$

Some sucessfull frequentist algorithms

Example for Bernoulli rewards:

- UCB [Auer et al. 02] uses Hoeffding bounds:

$$UCB_a(t) = \frac{S_a(t)}{N_a(t)} + \sqrt{\frac{\alpha \log(t)}{2N_a(t)}}$$

and one has:

$$\mathbb{E}[N_a(n)] \leq \frac{K_1}{2(\mu_a - \mu^*)^2} \ln n + K_2, \quad \text{with } K_1 > 1.$$

Some sucessfull frequentist algorithms

Example for Bernoulli rewards:

- KL-UCB [Cappé, Garivier, Maillard, Stoltz, Munos 11-12] uses the index:

$$u_a(t) = \underset{x > \frac{S_a(t)}{N_a(t)}}{\text{argmax}} \left\{ K\left(\frac{S_a(t)}{N_a(t)}, x\right) \leq \frac{\ln(t) + c \ln\ln(t)}{N_a(t)} \right\}$$

  with

$$K(p,q) = \mathrm{KL}\left(\mathcal{B}(p), \mathcal{B}(q)\right) = p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right)$$

  and one has

$$\mathbb{E}[N_a(n)] \leq \frac{1}{K(\mu_a, \mu^*)} \ln n + K$$

Bayesian algorithms

At the end of round t,

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

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

In the Bernoulli case, $\theta = \mu$ and $\Pi_t = \Lambda_t$

- $\mu_a \sim \mathcal{U}([0,1]) = \text{Beta}(1,1)$
- $\pi_a^t = \text{Beta}(S_a(t)+1, N_a(t) - S_a(t) + 1)$

Some ideas for Bayesian algorithms:

- Gittins indices [Gittins, 1979]
- quantiles of the posterior
- samples from the posterior [Thompson, 33]

MDP formulation of the Bernoulli bandit game

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

- $\mathcal{S}_t(a,1)$ is the number of ones observed from arm a until time t
- $\mathcal{S}_t(a,2)$ is the number of ones observed from arm a until time t
- Line a gives the parameters of the Beta posterior over arm a, π_a^t

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

Gittins's ideas

- $\mathcal{S}_t$ can be seen as a state in a Markov Decision Process
- the optimal policy in this MDP is an index policy

$$\underset{(A_t)}{\arg\max}\ \mathbb{E}\left[\sum_{t=1}^{n} X_t\right]$$

The Finite-Horizon Gittins algorithm

The Finite-Horizon Gittins algorithm

- is Bayesian optimal for the finite horizon problem
- consists in a index policy
- display very good performance on frequentist problems !

But...

- FH-Gittins indices are hard to compute
- the algorithm is heavily horizon-dependent
- there is 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 $(\mu_1, ..., \mu_K)$

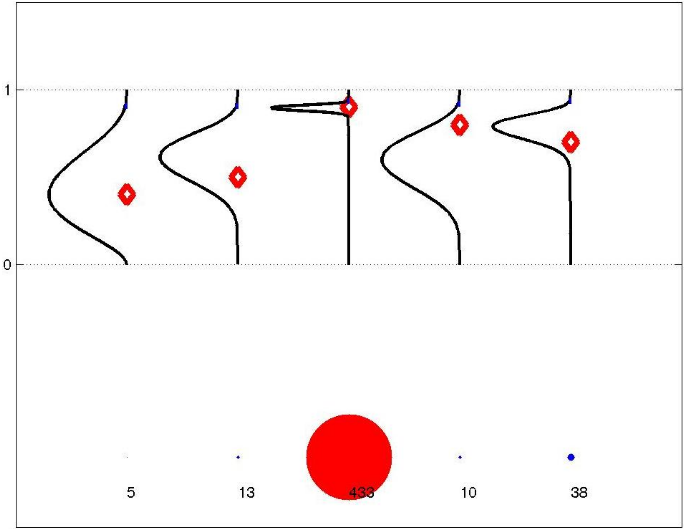
The **Bayes-UCB algorithm** is the index policy associated with:

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

ie, at time t choose

$$A_t = \underset{a=1...K}{\text{argmax}}\ q_a(t)$$

An illustration for Bernoulli bandits



Theoretical results for the Bernoulli case

- **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 a is such that :

$$\mathbb{E}_\theta[N_a(n)] \leq \frac{1+\epsilon}{K(\mu_a, \mu^*)} \log(n) + o_{\epsilon,c}\left(\log(n)\right)$$

- **Link to a frequentist algorithm:**

Bayes-UCB index is close to KL-UCB index: $\tilde{u}_a(t) \leq q_a(t) \leq u_a(t)$ with:

$$u_a(t) = \underset{x > \frac{S_a(t)}{N_a(t)}}{\text{argmax}} \left\{ K\left(\frac{S_a(t)}{N_a(t)}, x\right) \leq \frac{\log(t) + c\log(\log(t))}{N_a(t)} \right\}$$

$$\tilde{u}_a(t) = \underset{x > \frac{S_a(t)}{N_a(t)+1}}{\text{argmax}} \left\{ K\left(\frac{S_a(t)}{N_a(t)+1}, x\right) \leq \frac{\log\left(\frac{t}{N_a(t)+2}\right) + c\log(\log(t))}{(N_a(t)+1)} \right\}$$

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?

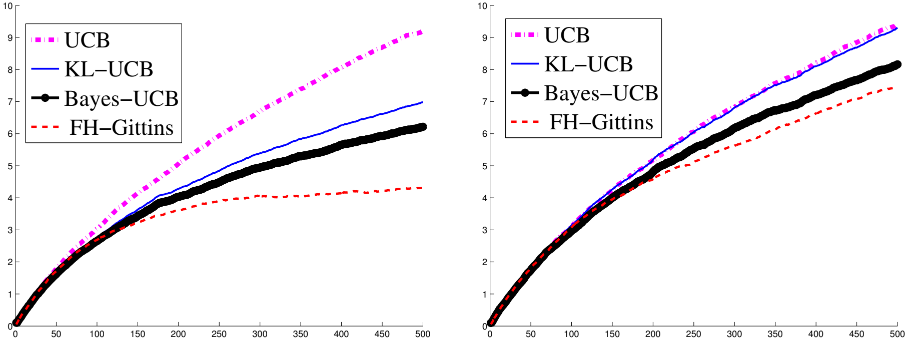
- First element: link between Beta and Binomial distribution:

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

- 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

Thompson Sampling

- A very simple algorithm:

$$\forall a \in \{1..K\}, \quad \theta_a(t) \sim \lambda_a^t$$
$$A_t = \text{argmax}_a \; \theta_a(t)$$

This algorithm is not optimistic any more.

- (Recent) interest for this algorithm:
  - a very old algorithm [Thompson 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 2012]

An optimal regret bound for Bernoulli bandits

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

- Known result : [Agrawal,Goyal 2012]

$$\mathbb{E}[R_n] \leq C \left(\sum_{a=2}^{K} \frac{1}{\Delta_a} \right) \ln(n) + o_\mu(\ln(n))$$

An optimal regret bound for Bernoulli bandits

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

- Known result : [Agrawal,Goyal 2012]

$$\mathbb{E}[R_n] \leq C \left(\sum_{a=2}^{K} \frac{1}{\Delta_a} \right) \ln(n) + o_\mu(\ln(n))$$

- Our improvement : [Kaufmann,Korda,Munos 2012]

Theorem $\forall \epsilon > 0$,

$$\mathbb{E}[R_n] \leq (1 + \epsilon) \left(\sum_{a=2}^{K} \frac{\Delta_a}{K(\mu_a, \mu^*)} \right) \ln(n) + o_{\mu,\epsilon}(\ln(n))$$

Step 1: Decomposition

- We adapt an analysis working for optimistic index policies:

$$A_t = \text{argmax}_a l_a(t)$$

$$\mathbb{E}[N_a(n)] \leq \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(l_1(t) < \mu_1\right)}_{o(\ln(n))} + \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(l_a(t) \geq l_1(t) > \mu_1, A_t = a\right)}_{\ln(n)/K(\mu_a,\mu_1)+o(\ln(n))}$$

Step 1: Decomposition

- We adapt an analysis working for optimistic index policies:

$$A_t = \text{argmax}_a l_a(t)$$

$$\mathbb{E}[N_a(n)] \leq \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(l_1(t) < \mu_1\right)}_{o(\ln(n))} + \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(l_a(t) \geq l_1(t) > \mu_1, A_t = a\right)}_{\ln(n)/K(\mu_a,\mu_1)+o(\ln(n))}$$

$\Rightarrow$ Does NOT work for Thompson Sampling

Step 1: Decomposition

- We adapt an analysis working for optimistic index policies:

$$A_t = \text{argmax}_a l_a(t)$$

$$\mathbb{E}[N_a(n)] \leq \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(l_1(t) < \mu_1\right)}_{o(\ln(n))} + \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(l_a(t) \geq l_1(t) > \mu_1, A_t = a\right)}_{\ln(n)/K(\mu_a,\mu_1)+o(\ln(n))}$$

$\Rightarrow$ Does NOT work for Thompson Sampling

- Our decomposition for Thompson Sampling is

$$\mathbb{E}[N_a(n)] \leq \sum_{t=1}^{n} \mathbb{P}\left(\theta_1(t) \leq \mu_1 - \sqrt{\frac{6\ln t}{N_1(t)}}\right) + \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(\theta_a(t) > \mu_1 - \sqrt{\frac{6\ln t}{N_1(t)}}\right)}_{(*)}$$

Step 2: Linking samples to other known indices

- We introduce the following quantile:

$$q_a(t) := Q\left(1 - \frac{1}{t\ln(n)}, \pi_a^t\right)$$

Step 2: Linking samples to other known indices

- We introduce the following quantile:

$$q_a(t) := Q\left(1 - \frac{1}{t\ln(n)}, \pi_a^t\right)$$

- And the corresponding KL-UCB index:

$$u_a(t) := \underset{x > \frac{S_a(t)}{N_a(t)}}{\text{argmax}} \left\{ K\left(\frac{S_a(t)}{N_a(t)}, x\right) \leq \frac{\ln(t) + \ln(\ln(n))}{N_a(t)} \right\}$$

Step 2: Linking samples to other known indices

- We introduce the following quantile:

$$q_a(t) := Q\left(1 - \frac{1}{t\ln(n)}, \pi_a^t\right)$$

- And the corresponding KL-UCB index:

$$u_a(t) := \underset{x > \frac{S_a(t)}{N_a(t)}}{\text{argmax}} \left\{ K\left(\frac{S_a(t)}{N_a(t)}, x\right) \leq \frac{\ln(t) + \ln(\ln(n))}{N_a(t)} \right\}$$

- We have already seen that:

$$q_a(t) < u_a(t)$$

Step 2: Linking samples to other known indices

- Introducing the quantile $q_{a(t)}$:

$$\sum_{t=1}^{n} \mathbb{P}\left(\theta_a(t) > \mu_1 - \sqrt{\frac{6\ln t}{N_1(t)}}, A_t = a\right)$$

$$\leq \sum_{t=1}^{n} \mathbb{P}\left(q_a(t) > \mu_1 - \sqrt{\frac{6\ln t}{N_1(t)}}, A_t = a\right) + \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(\theta_a(t) > q_a(t)\right)}_{\leq 2}$$

Step 2: Linking samples to other known indices

- Introducing the quantile $q_{a(t)}$:

$$\sum_{t=1}^{n} \mathbb{P}\left(\theta_a(t) > \mu_1 - \sqrt{\frac{6\ln t}{N_1(t)}}, A_t = a\right)$$

$$\leq \sum_{t=1}^{n} \mathbb{P}\left(q_a(t) > \mu_1 - \sqrt{\frac{6\ln t}{N_1(t)}}, A_t = a\right) + \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(\theta_a(t) > q_a(t)\right)}_{\leq 2}$$

- Then the KL-UCB index $u_a(t)$:

$$\sum_{t=1}^{n} \mathbb{P}\left(\theta_a(t) > \mu_1 - \sqrt{\frac{6\ln t}{N_1(t)}}, A_t = a\right)$$

$$\leq \sum_{t=1}^{n} \mathbb{P}\left(u_a(t) > \mu_1 - \sqrt{\frac{6\ln t}{N_1(t)}}, A_t = a\right) + 2$$

Step 3: An extra deviation result

- The current decomposition is:

$$\mathbb{E}[N_a(n)] \leq \sum_{t=1}^{n} \mathbb{P}\left(\theta_1(t) \leq \mu_1 - \sqrt{\frac{6\ln t}{N_1(t)}}\right) + \sum_{t=1}^{n} \mathbb{P}\left(u_a(t) > \mu_1 - \sqrt{\frac{6\ln t}{N_1(t)}}, A_t = a\right) + 2$$

Step 3: An extra deviation result

- The current decomposition is:

$$\mathbb{E}[N_a(n)] \leq \sum_{t=1}^{n} \mathbb{P}\left(\theta_1(t) \leq \mu_1 - \sqrt{\frac{6\ln t}{N_1(t)}}\right) + \sum_{t=1}^{n} \mathbb{P}\left(u_a(t) > \mu_1 - \sqrt{\frac{6\ln t}{N_1(t)}}, A_t = a\right) + 2$$

- We prove a deviation result:

Proposition

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

$$\sum_{t=1}^{\infty} \mathbb{P}\left(N_1(t) \leq t^b\right) \leq C_b.$$

Step 3: An extra deviation result

- The current decomposition is:

$$\mathbb{E}[N_a(n)] \leq \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(\theta_1(t) \leq \mu_1 - \sqrt{\frac{6 \ln t}{N_1(t)}}, N_1(t) > t^b\right)}_{A}$$

$$+ \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(u_a(t) > \mu_1 - \sqrt{\frac{6 \ln t}{N_1(t)}}, N_1(t) > t^b, A_t = a\right)}_{B} + 2 + 2C_b$$

- We prove a deviation result:

Proposition

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

$$\sum_{t=1}^{\infty} \mathbb{P}\left(N_1(t) \leq t^b\right) \leq C_b.$$

Step 4: Final decomposition

- The final decomposition is:

$$\mathbb{E}[N_a(n)] \leq \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(\theta_1(t) \leq \mu_1 - \sqrt{\frac{6\ln t}{N_1(t)}}, N_1(t) > t^b\right)}_{A}$$

$$+ \underbrace{\sum_{t=1}^{n} \mathbb{P}\left(u_a(t) > \mu_1 - \sqrt{\frac{6\ln t}{t^b}}, A_t = a\right)}_{B} + 2 + 2C_b$$

One can show :

- $A = o(\ln(n))$
- $B = \frac{\ln(n)}{K(\mu_a,\mu^*)} + o(\ln(n))$

Understanding the deviation result

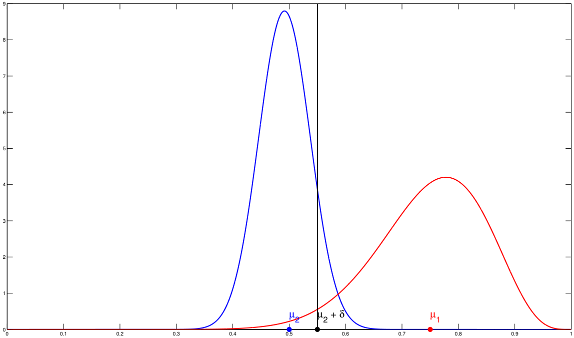
- Recall the result

There exists constants $b = b(\mu_1, \mu_2) \in (0, 1)$ and $C_b < \infty$ such that

$$\sum_{t=1}^{\infty} \mathbb{P}\left(N_1(t) \leq t^b\right) \leq C_b.$$

- Where does it come from?

$$\left\{N_1(t) \leq t^b\right\} = \{\text{there exists a time range of length at least } t^{1-b} - 1 \text{ with no draw of arm } 1\}$$



Assume that :

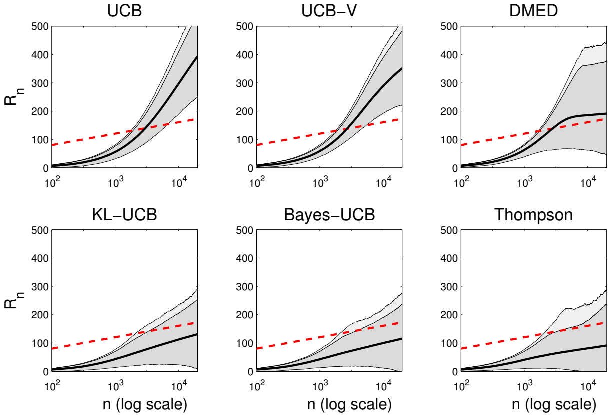
- on $\mathcal{I}_j = [\tau_j, \tau_j + \lceil t^{1-b} - 1 \rceil]$ there is no draw of arm 1
- there exists $\mathcal{J}_j \subset \mathcal{I}_j$ such that $\forall s \in \mathcal{J}_j, \forall a \neq 1$, $\theta_a(s) \leq \mu_2 + \delta$

Then :

- $\forall s \in \mathcal{J}_j$, $\theta_1(s) \leq \mu_2 + \delta$

$\Rightarrow$ This only happens with small probability

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.

In practise

In the Bernoulli case, for each arm,

- KL-UCB requires to solve an optimization problem:

$$u_a(t) = \underset{x > \frac{S_a(t)}{N_a(t)}}{\text{argmax}} \left\{ K\left(\frac{S_a(t)}{N_a(t)}, x\right) \leq \frac{\ln(t) + c \ln\ln(t)}{N_a(t)} \right\}$$

- Bayes-UCB requires to compute one quantile of a Beta distribution
- Thompson requires to compute one sample of Beta distribution

In practise

In the Bernoulli case, for each arm,

- KL-UCB requires to solve an optimization problem:

$$u_a(t) = \underset{x > \frac{S_a(t)}{N_a(t)}}{\text{argmax}} \left\{ K\left(\frac{S_a(t)}{N_a(t)}, x\right) \leq \frac{\ln(t) + c \ln\ln(t)}{N_a(t)} \right\}$$

- Bayes-UCB requires to compute one quantile of a Beta distribution
- Thompson requires to compute one sample of Beta distribution

Some extensions:

- rewards in exponential families
- algorithms for bounded reward (using the Bernoulli case)
- (Bayesian algorithms only!) More general cases where posterior is not directly computable (MCMC simulation)

The Linear Bandit model

The model

- arms : fixed vectors $U_1, ..., U_K \in \mathbb{R}^d$
- parameter of the model : $\theta^* \in \mathbb{R}^d$
- when arm A_t is drawn, one observes reward

$$y_t = U'_{A_t}\theta^* + \sigma\epsilon_t \text{ with } \epsilon_t \text{ some noise}$$

- goal : design a strategy minimizing regret

$$\mathbb{E}^*_\theta\left[\sum_{t=1}^{n}\left(\max_{1\leq a\leq K}(U'_a\theta) - U'_{A_t}\theta\right)\right]$$

Applications:

- Stochastic shortest path problem
- Contextual advertisement

The Linear Bandit model

$$y_t = U'_{A_t}\theta^* + \sigma\epsilon_t$$

Optimistic algorithms for this setting

- use $\hat{\theta}_t$, some least square estimate of θ^*
- build a confidence ellipsoid around $\hat{\theta}_t$:

$$\mathcal{E}_t = \{\theta : ||\theta - \hat{\theta}_t||_{\Sigma_t^{-1}} \leq \beta(t)\}$$

- choose

$$A_t = \arg\max_a \max_{\theta \in \mathcal{E}_t} U_a^T \theta$$

- which rewrites:

$$A_t = \arg\max_a U_a^T \hat{\theta}_t + ||U_a||_{\Sigma_t} \beta(t)$$

Bayesian algorithms with Gaussian Prior

$$\begin{aligned} y_t &= U'_{A_t}\theta^* + \sigma\epsilon_t \quad \epsilon_t \sim \mathcal{N}(0,1) \\ Y_t &= X_t\theta^* + \sigma\mathcal{E}_t \quad \mathcal{E}_t \sim \mathcal{N}(0,\mathsf{I}_t) \end{aligned}$$

Gaussian prior: $\theta^* \sim \mathcal{N}\left(0, \kappa^2 \mathsf{I}_d\right)$

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

Bayes-UCB for Linear Bandits

The posterior on the means of each arms are:

$$U_a'\theta^*|X_t, Y_t \sim \mathcal{N}\left(U_a'\hat{\theta}_t, ||U_a||^2_{\Sigma_t}\right)$$

Bayes-UCB is therefore the index policy associated with:

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

- very similar to frequentist optimistic approaches
- here the arms are not independent and we used marginal distributions

Thompson Sampling for Linear Bandits

$$\theta^*|X_t, Y_t \sim \mathcal{N}(\hat{\theta}_t, \Sigma_t) \qquad U_a'\theta^*|X_t, Y_t \sim \mathcal{N}\left(U_a'\hat{\theta}_t, ||U_a||^2_{\Sigma_t}\right)$$

'Marginal' Thompson Sampling At time t

- $\forall a = 1...K$, draw independent samples $\theta_a \sim \mathcal{N}\left(U_a'\hat{\theta}_t, ||U_a||^2_{\Sigma_t}\right)$
- choose $A_t = \underset{a}{\text{argmax}}\ \theta_a$

First elements of theoretical analysis in [Agrawal, Goyal, sept 2012]

'Joint' Thompson Sampling At time t

- draw $\theta \sim \mathcal{N}(\hat{\theta}_t, \Sigma_t)$
- choose $A_t = \underset{a}{\text{argmax}}\ U_a'\theta$

Open question : Which approach is more suitable?

Conclusion and perspectives

You are now aware that:

- Bayesian algorithm are efficient for the frequentist MAB problem
- Bayes-UCB show striking similarity with frequentist algorithms
- Thompson Sampling is an easy-to-implement alternative to optimistic algorithms
- Bayes-UCB and Thompson Sampling are optimal for Bernoulli bandits

Future work:

- A better understanding of the Finite-Horizon Gittins indices
- Regret analysis of Bayes-UCB and Thompson Sampling
  - for rewards in exponential families
  - in the Linear Bandit model
- Thompson Sampling in model-based reinforcement learning

References:

- Bayes-UCB algorithm:

  Emilie Kaufmann, Olivier Cappé and Aurélien Garivier
  On Bayesian upper confidence bounds for bandit problems
  AISTATS 2012

- Analysis of Thompson Sampling:

  Emilie Kaufmann, Nathaniel Korda and Rémi Munos
  Thompson Sampling : an asymptotically optimal finite-time analysis
  ALT 2012