Bayesian and Frequentist Methods in Bandit Models

Emilie Kaufmann, Telecom ParisTech

Bayes In Paris, ENSAE, October 24th, 2013

Bandit model

A **multi-armed bandit model** is a set of K arms where

- Each arm a is a probability distribution ν_a of mean μ_a
- Drawing arm a is observing a realization of ν_a
- Arms are assumed to be independent

In a **bandit game**, at round t, a forecaster

- chooses arm A_t to draw based on past observations, according to its **sampling strategy**
- observes 'reward' $X_t \sim \nu_{A_t}$

Our bandit problem: regret minimization

$$\mu^* = \max_a \mu_a \quad \text{and} \quad a^* = \underset{a}{\text{argmax}}\ \mu_a$$

The forecaster wants to **maximize the reward accumulated during learning** or equivalentely minimize its **regret**:

$$R_n = n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right]$$

He has to find a sampling strategy (or bandit algorithm) that

- realizes a **tradeoff between exploration and exploitation**

Applications (with arms beeing Bernoulli random variables)

- Finding the best slot machine in a 'casino' (just for the name!)
- Initial motivation: Sequential allocation of medical treatments

A recent motivation for bandits: Online advertisement

Yahoo!(c) has to choose between K different advertisements the one to display on its webpage for each user (indexed by $t \in \mathbb{N}$).

- Ad number $a \rightarrow$ **unknown** probability of click p_a
- **Unknown** best advertisement $a^* = \text{argmax}_a \, p_a$
- $X_{t,a} \sim \mathcal{B}(p_a)$: $(X_{t,a} = 1)$=(user t has clicked on ad a)

Yahoo!(c):

- chooses ad A_t to display for user number t
- observes whether the user has clicked or not: X_{t,A_t}
- wants to maximize the click-through-rate

$\Rightarrow$ How should Yahoo!(c) choose ad A_t to display depending on the previous clicks of the $(t-1)$ first users?

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 :

- $\theta_a \overset{i.i.d.}{\sim} \pi_a$
- $(Y_{a,t})_t$ 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 regret

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

- Minimize Bayes risk

$$\text{Risk}_n = \mathbb{E} \left[\sum_{t=1}^n \mu^* - \mu_{A_t} \right] = \int R_n(\theta) d\pi(\theta)$$

Frequentist tools, Bayesian tools

Bandit algorithms based on frequentist tools use:

- MLE for the mean parameter of each arm
- confidence intervals for the parameter of each arm

Bandit algorithms based on Bayesian tools use:

- $\Pi_t = (\pi_1^t, \ldots, \pi_K^t)$ the current posterior over $(\theta_1, ..., \theta_K)$

Frequentist tools, Bayesian tools

Bandit algorithms based on frequentist tools use:

- MLE for the mean parameter of each arm
- confidence intervals for the parameter of each arm

Bandit algorithms based on Bayesian tools use:

- $\Pi_t = (\pi_1^t, \ldots, \pi_K^t)$ the current posterior over $(\theta_1, ..., \theta_K)$

One can **separate tools and objectives**:

Objective	Frequentist algorithms	Bayesian algorithms
Regret	?	?
Bayes risk	?	?

Bayesian algorithms minimizing Bayes risk

Objective	Frequentist algorithms	Bayesian algorithms
Regret	?	?
Bayes risk	?	?

MDP formulation of the Bernoulli bandit game

Benoulli bandits with uniform prior on the means: $\theta_a = \mu_a$

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

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

- Line a gives the parameters of the Beta posterior over arm a, π_a^t

$$S_{11} = \begin{pmatrix} 1 & 2 \\ 5 & 1 \\ 0 & 2 \end{pmatrix}$$

(ones observed; zero observed; index of the arm)

$\mathcal{S}_t$ can be seen as a state in a Markov Decision Process, and the optimal policy is (depending on the criterion)

$$\underset{(A_t)}{\arg\max}\ \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} X_t\right] \quad \text{or} \quad \underset{(A_t)}{\arg\max}\, \mathbb{E}\left[\sum_{t=1}^{n} X_t\right]$$

The Finite-Horizon Gittins algorithm

Gittins ([1979]) shows the optimal policy in the discounted case is an index policy:

$$A_t = \underset{a}{\text{argmax}}\ \nu_{Disc}(\pi_t(a)).$$

Similarly, in the finite-horizon case (our setting), the optimal policy has an explicit formulation

$$A_t = \underset{a}{\text{argmax}}\ \nu_{FH}(\pi_t(a), n-t)$$

The Finite-Horizon Gittins algorithm

- minimizes minimizes the Bayes risk Risk_n
- and display very good performance on frequentist problems !

But...

- FH-Gittins indices are hard to compute
- the algorithm is heavily horizon-dependent

Frequentist algorithms minimizing regret

Objective	Frequentist algorithms	Bayesian algorithms
Regret	?	?
Bayes risk	?	Finite-Horizon Gittins algorithm

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

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

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

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

A family of 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

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

Towards optimal 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:

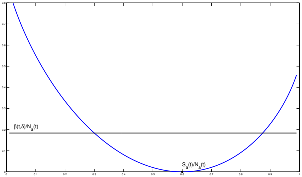
$$\mathbb{E}[N_a(n)] \leq \frac{K_1}{2(\mu_a - \mu^*)^2} \ln n + K_2, \quad \text{with } K_1 > 1.$$

KL-UCB: and asymptotically optimal frequentist algorithm

Example for Bernoulli rewards:

- KL-UCB [Cappé et al. 2013] 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 $\quad K(p,q) = \text{KL}\left(\mathcal{B}(p), \mathcal{B}(q)\right) = p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right)$

and one has

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

Frequentist algorithms optimal minimizing Bayes risk

Objective	Frequentist algorithms	Bayesian algorithms
Regret	KL-UCB	?
Bayes risk	KL-UCB	Finite-Horizon Gittins algorithm

(at least in an asymptotic sense, see [Lai 1987])

Bayesian algorithms minimizing regret

Objective	Frequentist algorithms	Bayesian algorithms
Regret	KL-UCB	?
Bayes risk	KL-UCB	Finite-Horizon Gittins algorithm

We want to design Bayesian algorithm that are optimal with respect to the frequentist regret

UCBs versus Bayesian algorithms

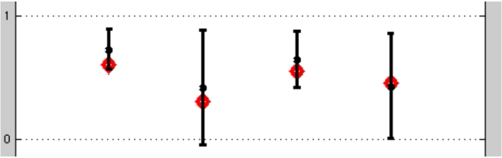


Figure: Confidence intervals on the arms means after t rounds of a bandit game

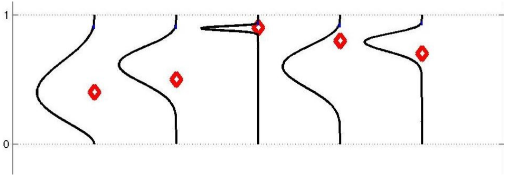


Figure: Posterior over the means of the arms after t rounds of a bandit game

UCBs versus Bayesian algorithms

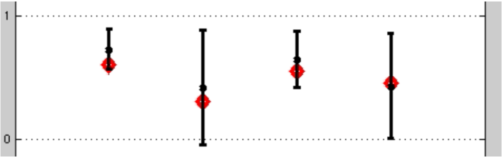


Figure: Confidence intervals on the arms means after t rounds of a bandit game

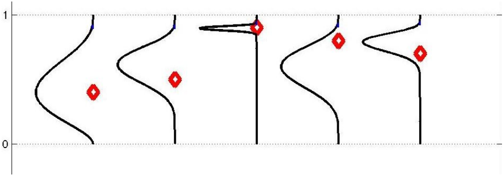


Figure: Posterior over the means of the arms after t rounds of a bandit game

$\Rightarrow$ How do we exploit the posterior in a Bayesian bandit algorithm?

The Bayes-UCB algorithm

Let :

- $\Pi_0 = (\pi_1^0, \ldots, \pi_K^0)$ be a prior distribution over $(\theta_1, ..., \theta_K)$
- $\Lambda_t = (\lambda_1^t, \ldots, \lambda_K^t)$ be the posterior over the means $(\mu_1, ..., \mu_K)$ a the end of round t

The **Bayes-UCB algorithm** chooses at time t

$$A_t = \underset{a}{\text{argmax}}\ Q\left(1 - \frac{1}{t(\log t)^c}, \lambda_a^{t-1}\right)$$

where $Q(\alpha, \pi)$ is the quantile of order α of the distribution π.

The Bayes-UCB algorithm

Let :

- $\Pi_0 = (\pi_1^0, \ldots, \pi_K^0)$ be a prior distribution over $(\theta_1, ..., \theta_K)$
- $\Lambda_t = (\lambda_1^t, \ldots, \lambda_K^t)$ be the posterior over the means $(\mu_1, ..., \mu_K)$ a the end of round t

The **Bayes-UCB algorithm** chooses at time t

$$A_t = \underset{a}{\text{argmax}}\ Q\left(1 - \frac{1}{t(\log t)^c}, \lambda_a^{t-1}\right)$$

where $Q(\alpha, \pi)$ is the quantile of order α of the distribution π.

Bernoulli reward with uniform prior: $\theta = \mu$ and $\Pi_t = \Lambda_t$

$$A_t = \underset{a}{\text{argmax}}\ Q\left(1 - \frac{1}{t(\log t)^c}, \text{Beta}(S_a(t) + 1, N_a(t) - S_a(t) + 1)\right)$$

Theoretical results for the Bernoulli case

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

Theorem [K.,Cappé,Garivier 2012]
Let $\epsilon > 0$. The Bayes-UCB algorithm using a uniform prior over the arms and with parameter $c \geq 5$ satisfies

$$\mathbb{E}_\theta[N_a(n)] \leq \frac{1+\epsilon}{\mathrm{KL}(\mathcal{B}(\mu_a), \mathcal{B}(\mu^*))} \log(n) + o_{\epsilon,c}(\log(n)).$$

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?

We have a tight bound on the tail of posterior distributions (Beta distributions)

- First element: link between Beta and Binomial distribution:

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

- Second element: Sanov inequality: for $k > nx$,

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

Thompson Sampling

- A randomized Bayesian algorithm:

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

- (Recent) interest for this algorithm:
  - a very old algorithm
    [Thompson 1933]
  - partial analysis proposed
    [Granmo 2010][May, Korda, Lee, Leslie 2012]
  - 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 : [K.,Korda,Munos 2012]

Theorem $\forall \epsilon > 0,$

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

Two key elements in the proof

- Introduce a quantile to replace the sample:

$$q_a(t) := Q\left(1 - \frac{1}{t\ln(n)}, \pi_a^t\right) \text{ such that } \sum_{t=1}^{n} \mathbb{P}\left(\theta_a(t) > q_a(t)\right) \leq 2$$

and use what we know about quantiles (cf. Bayes-UCB)

Two key elements in the proof

- Introduce a quantile to replace the sample:

$$q_a(t) := Q\left(1 - \frac{1}{t\ln(n)}, \pi_a^t\right) \text{ such that } \sum_{t=1}^{n} \mathbb{P}\left(\theta_a(t) > q_a(t)\right) \leq 2$$

  and use what we know about quantiles (cf. Bayes-UCB)

- Proove separately that the optimal arm has to be drawn a lot

Proposition

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

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

Thompson Sampling in Exponential Family bandits

- Arm a has distribution ν_{θ_a} with density

$$f(x|\theta_a) = A(x)\exp(T(x)\theta_a - F(\theta_a)).$$

- The Jeffreys' prior on arm a is

$$\pi_J(\theta) \propto \sqrt{|F''(\theta)|}.$$

- Practical implementation of TS with Jeffreys' prior

Name	Distribution	Prior on λ	Posterior on λ
$\mathcal{B}(\lambda)$	$\lambda^x(1-\lambda)^{1-x}\delta_{0,1}$	$\text{B}\left(\frac{1}{2},\frac{1}{2}\right)$	$\text{B}\left(\frac{1}{2}+s,\frac{1}{2}+n-s\right)$
$\mathcal{N}(\lambda,\sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\lambda)^2}{2\sigma^2}}$	$\propto 1$	$\mathcal{N}\left(\frac{s}{n},\frac{\sigma^2}{n}\right)$
$\Gamma(k,\lambda)$	$\frac{\lambda^k}{\Gamma(k)}x^{k-1}e^{-\lambda x}\mathbb{1}_{[0,+\infty[}$	$\propto \frac{1}{\lambda}$	$\Gamma(kn,s)$
Pareto(x_m,λ)	$\frac{\lambda x_m^\lambda}{x^{\lambda+1}}\mathbb{1}_{[x_m,+\infty[}$	$\propto \frac{1}{\lambda}$	$\Gamma\left(n+1, s-n\log x_m\right)$

Posterior after n observations $y_1,\dots,y_n$, with $s=\sum_{s=1}^n T(y_s)$.

Thompson Sampling in Exponential Family bandits

Theorem [Korda,K.,Munos 13]
If the rewards distributions belong to a 1-dimensional canonical exponential family, Thompson sampling with Jeffreys' prior π_J satisfies

$$\lim_{n\to\infty} \frac{\mathbb{E}[N_a(n)]}{\ln n} = \frac{1}{\mathrm{KL}(\nu_{\theta_a}, \nu_{\theta_{a^*}})}.$$

Thompson Sampling in Exponential Family bandits

An idea of the proof

$\theta_a(t)$ be a sample of the posterior $\pi_a(t)$ on θ_a. TS samples at round t
$A_t = \arg\max_a \theta_a(t)$.

$$E_{a,t} = \left(\left| \frac{1}{N_a(t)} \sum_{s=1}^{N_a(t)} T(Y_{a,s}) - F'(\theta_a) \right| \leq \delta_a \right), \; E^{\theta}_{a,t} = (\mu(\theta_a(t)) \leq \mu_a + \Delta_a)$$

$$\begin{aligned} \mathbb{E}[N_a(n)] &= \sum_{t=1}^{n} \mathbb{P}(A_t = a, E_{a,t}, E^{\theta}_{a,t}) + \sum_{t=1}^{n} \mathbb{P}(A_t = a, E_{a,t}, (E^{\theta}_{a,t})^c) \\ &\quad + \sum_{t=1}^{n} \mathbb{P}(A_t = a, E^{c}_{a,t}) \end{aligned}$$

Thompson Sampling in Exponential Family bandits

Two key ingredients

Theorem (posterior concentration)
There exists two constants $C_{1,a}, C_{2,a}$ such that

$$\mathbb{P}((E_{a,t}^{\theta})^c|\mathcal{F}_t)\mathbb{1}_{E_{a,t}} \leq C_{1,a}N_{a,t}e^{-N_{a,t}(1-\delta_a C_{2,a})\mathrm{KL}(\nu_{\theta_a},\nu_{\mu^{-1}(\mu_a+\Delta_a)})}$$

Proposition (number of draws of the optimal arm)
For every $b \in]0,1[$, there exists $C_b < \infty$ such that

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

Understanding the deviation result

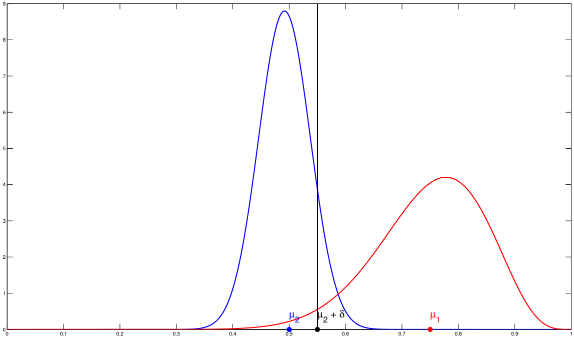
- Recall the result

For every $b \in]0,1[$ there exists a constant $C_b < \infty$ 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 :

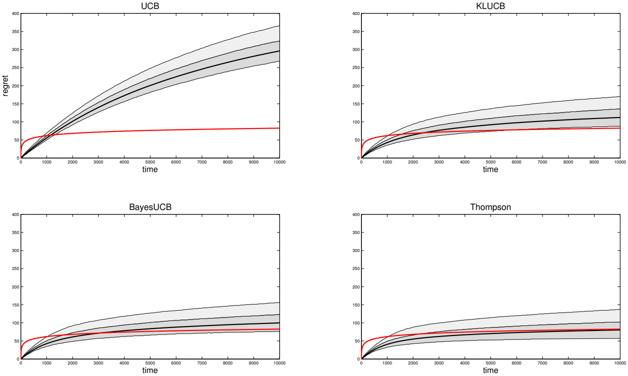
- on $\mathcal{I}_j = [\tau_j, \tau_j + \lceil t^{1-b} - 1 \rceil]$ there is no draw of arm 1
- there exists $\mathcal{J}_j \subset \mathcal{I}_j$ such that $\forall s \in \mathcal{J}_j, \forall a \neq 1$, $\mu(\theta_a(s)) \leq \mu_2 + \delta$

Then :

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

$\Rightarrow$ This only happens with small probability

Why using Bayesian algorithm in the frequentist setting?



Regret as a function of time in a ten arms Bernoulli bandit problem with low rewards, horizon $n = 20000$, average over $N = 50000$ trials.

Why using Bayesian algorithm in the frequentist setting?

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 Sampling requires to compute one sample of a Beta distribution

Why using Bayesian algorithm in the frequentist setting?

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 Sampling requires to compute one sample of a Beta distribution

Other advantages of Bayesian algorithms:

- they easily generalize to more complex models...
- ...even when the posterior is not directly computable (using MCMC)
- the prior can incorporate correlation between arms

One bandit model, two bandit problems

Recall a bandit model is simply a set of K unknown distributions. Assume

$$\underbrace{\mu_1 \geq \cdots \geq \mu_m}_{\mathcal{S}_m^*} > \mu_{m+1} \geq \cdots \geq \mu_K.$$

- We have seen sofar one bandit problem: regret minimization
- We introduce here another bandit problem: pure-exploration:

The forecaster has to **find the set of m best arms**, using as few observations of the arms as possible, but *without suffering a loss when drawing a bad arm*.

The forecaster:

- draws the arms according to an *exploration strategy*
- stops at time τ (*stopping strategy*) and recommends a set S of m arms

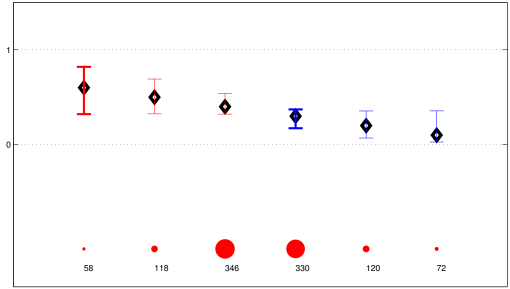
His goal:

$$\mathbb{P}(S = \mathcal{S}_m^*) \geq 1 - \delta \text{ and } \mathbb{E}[\tau] \text{ as small as possible}$$

KL-UCB: an algorithm for finding the m best arms

At round t, the KL-LUCB algorithm ([K., Kalyanakrishnan, 13])

- draws two well-chosen arms: u_t and l_t
- stops when CI for arms in $\hat{\mathcal{S}}_m(t)$ and $(\hat{\mathcal{S}}_m)^c(t)$ are separated
- recommends the set $\hat{\mathcal{S}}_m(\tau)$ of m empirical best arms



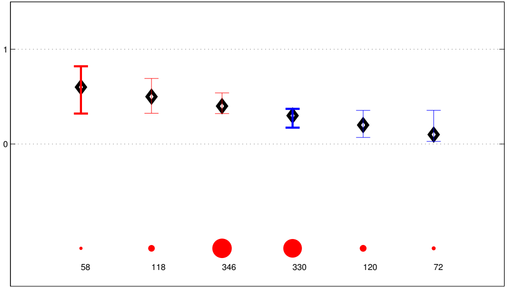
K=6,m=3. Set $\hat{\mathcal{S}}_m(t)$, arm l_t in bold Set $(\hat{\mathcal{S}}_m(t))^c$, arm u_t in bold

Bayesian algorithms for finding the m best?

KL-LUCB uses KL-confidence intervals:

$$
\begin{aligned}
L_a(t) &= \min\left\{q \leq \hat{p}_a(t) : N_a(t)K(\hat{p}_a(t), q) \leq \beta(t,\delta)\right\}, \\
U_a(t) &= \max\left\{q \geq \hat{p}_a(t) : N_a(t)K(\hat{p}_a(t), q) \leq \beta(t,\delta)\right\}.
\end{aligned}
$$

We use $\beta(t,\delta) = \log\left(\frac{k_1 K t^\alpha}{\delta}\right)$ to make sure $\mathbb{P}(S = \mathcal{S}_m^*) \geq 1 - \delta$.

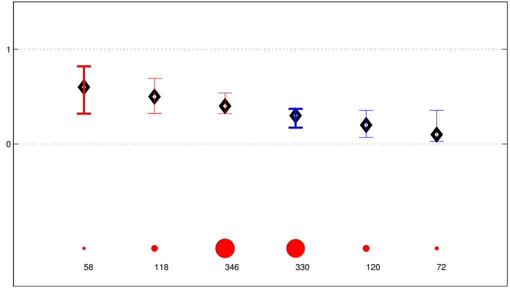


Bayesian algorithms for finding the m best?

KL-LUCB uses KL-confidence intervals:

$$L_a(t) = \min\{q \leq \hat{p}_a(t) : N_a(t)K(\hat{p}_a(t), q) \leq \beta(t,\delta)\},$$
$$U_a(t) = \max\{q \geq \hat{p}_a(t) : N_a(t)K(\hat{p}_a(t), q) \leq \beta(t,\delta)\}.$$

We use $\beta(t,\delta) = \log\left(\frac{k_1 K t^\alpha}{\delta}\right)$ to make sure $\mathbb{P}(S = \mathcal{S}_m^*) \geq 1 - \delta$.



$\Rightarrow$ How to propose a Bayesian algorithm that adapts to δ?

1. The multiarmed bandit problem
2. Bayesian bandits, frequentist bandits
3. Two Bayesian bandit algorithms
   - Bayes-UCB
   - Thompson Sampling
4. Bayesian algorithms for pure exploration?
5. Conclusion

Conclusion for regret minimization

Objective	Frequentist algorithms	Bayesian algorithms
Regret	KL-UCB	Bayes-UCB Thompson Sampling
Bayes risk	KL-UCB	Gittins algorithm for finite horizon

Work in progress

Objective	Frequentist algorithms	Bayesian algorithms
Regret	KL-UCB	Bayes-UCB ↑Thompson Sampling ↓
Bayes risk	KL-UCB	Gittins algorithm for finite horizon

?

Conclusion

Regret minimization: Go Bayesian!

- Bayes-UCB show striking similarities with KL-UCB
- Thompson Sampling is an easy-to-implement alternative to the optimistic approach
- both algorithms are asymptotically optimal towards frequentist regret (and more efficient in practise) in the Bernoulli case
- Thompson Sampling with Jeffreys' prior is asymptotically optimal when rewards belong to a one-dimensional exponential family, which matches the guarantees of the KL-UCB algorithm

Conclusion

Regret minimization: Go Bayesian!

- Bayes-UCB show striking similarities with KL-UCB
- Thompson Sampling is an easy-to-implement alternative to the optimistic approach
- both algorithms are asymptotically optimal towards frequentist regret (and more efficient in practise) in the Bernoulli case
- Thompson Sampling with Jeffreys' prior is asymptotically optimal when rewards belong to a one-dimensional exponential family, which matches the guarantees of the KL-UCB algorithm

Natural open question:

- Can Bayesian tools be used to build efficient algorithms for the pure-exploration objective?

References (1/2)

- S.Agrawal and N.Goyal, *Analysis of Thompson Sampling for the multi-armed bandit problem*, COLT 2012
- O.Cappé, A.Garivier, O. Maillard, R. Munos, G. Stoltz. *Kullback-Leibler upper confidence bound for optimal sequential allocation*, Annals of Statistics, 2013
- J.C. Gittins. *Bandit processes and dynamic allocation indices*, Journal of the Royal Statistical Society, Serie B, 1979
- E. Kaufmann, O. Cappé, and A. Garivier. *On Bayesian upper confidence bounds for bandit problems*. AISTATS 2012
- E.Kaufmann, N.Korda, and R.Munos. *Thompson Sampling: an asymptotically optimal finite-time analysis*. ALT 2012

References (2/2)

- E.Kaufmann and S.Kalyanakrishnan. *Information Complexity in Bandit Subset Selection*, COLT 2013
- N.Korda, E.Kaufmann, R.Munos. *Thompson Sampling for one-dimensional exponential families*, to appear in NIPS 2013
- T.Lai, *Adaptive treatment allocation and the multi-armed bandit problem*, Annals of Statistics, 1987
- T.Lai and H.Robbins. *Asymptotically efficient adaptive allocation rules*, Advances in applied mathematics, 1985
- W. Thompson, *On the likelihood that one unknown probability exceeds another in view of the evidence of two samples*, Biometrika, 1933