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

Emilie Kaufmann, Telecom ParisTech

Rencontre des Jeunes Statisticiens, Aussois, 28 août 2013

Bandit model

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

- Arm a is an unknown probability distribution ν_a with 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** (or **bandit algorithm**)
- observes 'reward' $X_t \sim \nu_{A_t}$

The forecaster is to **learn which arm(s) is (are) the best**

$$a^* = \text{argmax}_a \; \mu_a$$

Bernoulli bandit model

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

- Arm a is a Bernoulli distribution $\mathcal{B}(p_a)$ with unknown mean $\mu_a = p_a$
- Drawing arm a is observing a realization of $\mathcal{B}(p_a)$ (0 or 1)
- Arms are assumed to be <u>independent</u>

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

- chooses arm A_t to draw based on past observations, according to its **sampling strategy** (or **bandit algorithm**)
- observes 'reward' $X_t \sim \mathcal{B}(p_{A_t})$

The forecaster is to **learn which arm(s) is (are) the best**

$$a^* = \text{argmax}_a \, p_a$$

The classical bandit problem: regret minimization

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

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

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

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

An alternative: 'pure-exploration'

The forecaster has to **find the best arm(s)**, and does not suffer a loss when drawing 'bad arms'.

He has to find a sampling strategy that

- **optimaly explores** the environnement,

together with a stopping criterion and to recommand a set $\mathcal{S}$ of m arms such that

$$\mathbb{P}(\mathcal{S} \text{ is the set of } m \text{ best arms}) \geq 1 - \delta.$$

Zoom on an application: Online advertisement

Yahoo!(c) has to choose between K different advertisement 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 $a^* = \text{argmax}_a \ p_a$
- If ad a is displayed for user t, he clicks on it with probability p_a

Yahoo!(c):

- chooses ad A_t to display for user number t
- observes whether the user has clicked or not: $X_t \sim \mathcal{B}(p_{A_t})$

Zoom on an application: Online advertisement

Yahoo!(c) has to choose between K different advertisement 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 $a^* = \text{argmax}_a\, p_a$
- If ad a is displayed for user t, he clicks on it with probability p_a

Yahoo!(c):

- chooses ad A_t to display for user number t
- observes whether the user has clicked or not: $X_t \sim \mathcal{B}(p_{A_t})$

Yahoo!(c) can ajust its strategy (A_t) so as to

Regret minimization	Pure-exploration
Maximize the number of clicks during n interactions	Identify the best advertisement with probability at least $1 - \delta$

Zoom on an application: Online advertisement

Yahoo!(c) has to choose between K different advertisement 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 $a^* = \text{argmax}_a \, p_a$
- If ad a is displayed for user t, he clicks on it with probability p_a

Yahoo!(c):

- chooses ad A_t to display for user number t
- observes whether the user has clicked or not: $X_t \sim \mathcal{B}(p_{A_t})$

Yahoo!(c) can ajust its strategy (A_t) so as to

Regret minimization	Pure-exploration
Maximize the number of clicks during n interactions	Identify the best advertisement with probability at least $1 - \delta$

Two probabilistic modellings

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

Frequentist :

- $\theta = (\theta_1, \dots, \theta_K)$ unknown parameter
- $(Y_{a,t})_t$ is i.i.d. with distribution ν_{θ_a} with mean $\mu_a = \mu(\theta_a)$

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

$$\text{Risk}_n = \int R_n(\theta) d\pi(\theta)$$

Frequentist tools, Bayesian tools

Bandit algorithms based on frequentist tools use:

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

Bandit algorithms based on Bayesian tools use:

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

$$\pi_a^t = p(\theta_a | \text{past observations from arm } a)$$

Frequentist tools, Bayesian tools

Bandit algorithms based on frequentist tools use:

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

Bandit algorithms based on Bayesian tools use:

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

$$\pi_a^t = p(\theta_a | \text{past observations from arm } a)$$

One can **separate tools and objectives**:

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

Frequentist tools, Bayesian tools

Bandit algorithms based on frequentist tools use:

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

Bandit algorithms based on Bayesian tools use:

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

$$\pi_a^t = p(\theta_a | \text{past observations from arm } a)$$

One can **separate tools and objectives**:

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

Our goal

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

Asymptotically optimal algorithms towards the regret

$N_a(t)$ 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 algorithm 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 mean:

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

- 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 for Bernoulli bandits

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

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

where $\hat{p}_a(t) = \frac{S_a(t)}{N_a(t)}$ is the empirical mean of arm a.

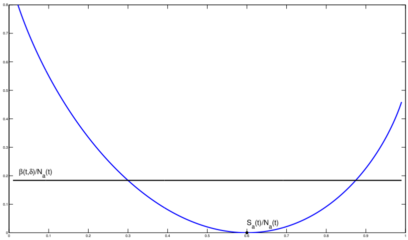
- Finite-time bound:

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

Towards optimal algorithms for Bernoulli bandits

- KL-UCB[Cappé et al. 2013] uses the index:

$$u_a(t) = \max\left\{q \geq \hat{p}_a(t) : N_a(t)\mathsf{K}(\hat{p}_a(t), q) \leq \log t + c \log\log t\right\}$$



with

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

- Finite-time bound:

$$\mathbb{E}[N_a(n)] \leq \frac{1}{\mathsf{K}(p_a, p^*)} \ln n + C$$

UCBs versus Bayesian algorithms

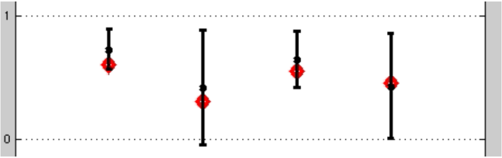


Figure : Confidence intervals for the arms means after t rounds of a bandit game

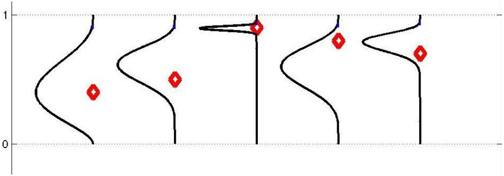


Figure : Posterior over the arms means after t rounds of a bandit game

UCBs versus Bayesian algorithms

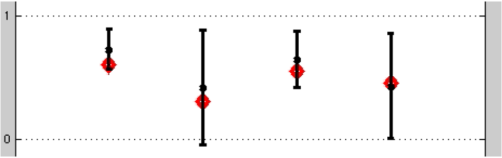


Figure : Confidence intervals for the arms means after t rounds of a bandit game

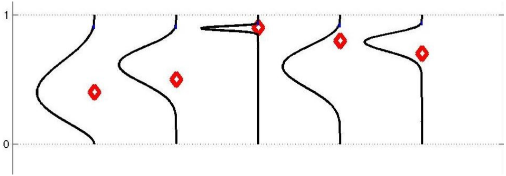


Figure : Posterior over the arms means after t rounds of a bandit game

⇒ 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.

For Benoulli bandits with uniform prior on the means:

$\theta_a = \mu_a = p_a \quad \Lambda_t = \Pi_t$

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

Theoretical results for Bernoulli bandits

- **Bayes-UCB is asymptotically optimal**

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

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

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:

$$
\begin{aligned}
u_a(t) &= \max\left\{q \geq \frac{S_a(t)}{N_a(t)} : N_a(t)\mathsf{K}\left(\frac{S_a(t)}{N_a(t)}, q\right) \leq \log t + c\log\log t\right\} \\
\tilde{u}_a(t) &= \max\left\{q \geq \frac{S_a(t)}{N_a(t)+1} : (N_a(t)+1)\mathsf{K}\left(\frac{S_a(t)}{N_a(t)+1}, q\right)\right. \\
&\qquad\qquad \left.\leq \log\left(\frac{t}{N_a(t)+2}\right) + c\log\log t\right\}
\end{aligned}
$$

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.

1. Two bandit problems
2. Regret minimization: Bayesian bandits, frequentist bandits
3. Two Bayesian bandit algorithms
   - The Bayes-UCB algorithm
   - Thompson Sampling
4. Conclusion and perspectives

Thompson Sampling

- A randomized Bayesian algorithm:

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

- (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]

Thompson Sampling (Bernoulli bandits)

- A randomized Bayesian algorithm:

$$\forall a \in \{1..K\}, \quad \theta_a(t) \sim \text{Beta}(S_a(t) + 1, N_a(t) - S_a(t) + 1)$$
$$A_t = \text{argmax}_a \; \theta_a(t)$$

- (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 arm.

- Known result : [Agrawal,Goyal 2012]

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

An optimal regret bound for Bernoulli bandits

Assume the first arm is the unique optimal arm.

- Known result : [Agrawal,Goyal 2012]

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

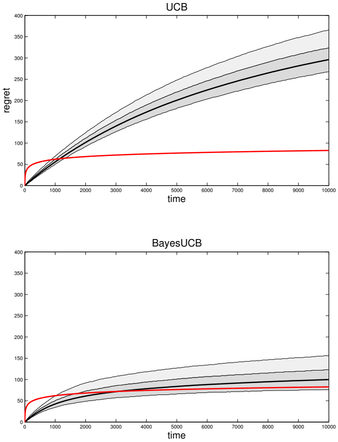
- Our improvement : [K.,Korda,Munos 2012]

Theorem $\forall \epsilon > 0$,

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

In practise

$$\theta = [0.1\ 0.05\ 0.05\ 0.05\ 0.02\ 0.02\ 0.02\ 0.01\ 0.01\ 0.01]$$



In practise

In the Bernoulli case, for each arm,

- KL-UCB requires to solve an optimization problem:

$$u_a(t) = \max \{q \geq \hat{p}_a(t) : N_a(t)\mathsf{K}(\hat{p}_a(t), q) \leq \log t + c \log \log t\}$$

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

In practise

In the Bernoulli case, for each arm,

- KL-UCB requires to solve an optimization problem:

$$u_a(t) = \max\left\{q \geq \hat{p}_a(t) : N_a(t)\mathsf{K}(\hat{p}_a(t), q) \leq \log t + c \log\log t\right\}$$

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

Summary for regret minimization

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

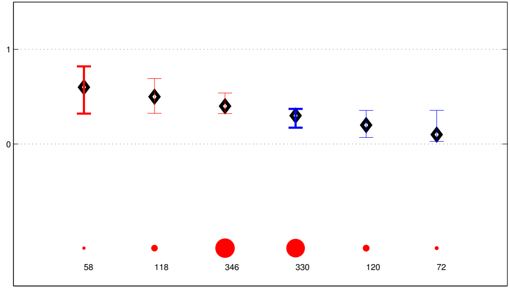
Future work:

- Is Gittins algorithm optimal with respect to the regret?
- Are our Bayesian algorithms efficient with respect to the Bayes risk?

Bayesian algorithm for pure-exploration?

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 $J(t)$ and $J(t)^c$ are separated
- recommends the set of m empirical best arms



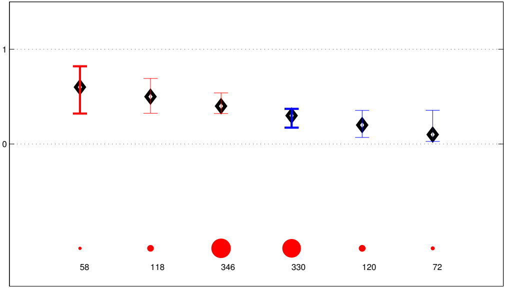
m=3. Set $J(t)$, arm l_t in bold Set $J(t)^c$, arm u_t in bold

Bayesian algorithm for pure-exploration?

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 $\mathbb{P}(\mathcal{S} = \mathcal{S}_m^*) \geq 1 - \delta$.

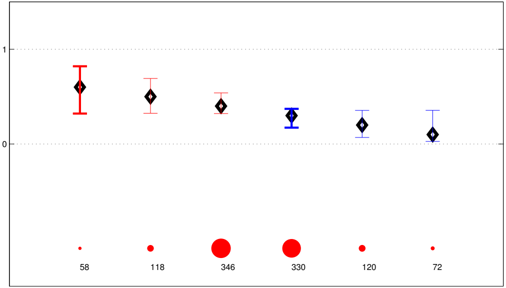


Bayesian algorithm for pure-exploration?

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 $\mathbb{P}(\mathcal{S} = \mathcal{S}_m^*) \geq 1 - \delta$.



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

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)

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)

TODO list:

- Go deeper into the link between Bayes risk and (frequentist) regret (Gittins' frequentist optimality?)
- Obtain theoretical guarantees for Bayes-UCB and Thompson Sampling beyond Bernoulli bandit models (e.g. when rewards belong to the exponential family)
- Develop Bayesian algorithm for the pure-exploration objective?

References

- 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
- E.Kaufmann, S.Kalyanakrishnan. *Information Complexity in Bandit Subset Selection*, COLT 2013