Méthodes bayésiennes et fréquentistes dans les modèles de bandit

Emilie Kaufmann, Telecom ParisTech
joint work with
Olivier Cappé, Aurélien Garivier, Rémi Munos, Nathaniel Korda
and Shivaram Kalyanakrishnan

Séminaire de statistiques, Orsay, 22 mai 2014

Outline

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** (or bandit algorithm)
- observes a sample $X_t \sim \nu_{A_t}$

The agent wants to **learn which arm(s) have highest means**

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

The (classical) bandit problem: regret minimization

Samples are seen as *rewards* (as in reinforcement learning)

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

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

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

Best arm identification (or 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 environment,

together with a stopping criterion, and then **recommends a set $\hat{\mathcal{S}}$ of m arms** such that

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

Zoom on an application: Medical trials

A doctor can choose between K different treatments for a given symptom.

- treatment number a has unknown probability of success μ_a
- **Unknown** best treatment $a^* = \text{argmax}_a \, \mu_a$
- If treatment a is given to patient t, he is cured with probability μ_a

The doctor:

- chooses treatment A_t to give to patient t
- observes whether the patient is healed : $X_t \sim \mathcal{B}(\mu_{A_t})$

Zoom on an application: Medical trials

A doctor can choose between K different treatments for a given symptom.

- treatment number a has unknown probability of success μ_a
- **Unknown** best treatment $a^* = \text{argmax}_a \ \mu_a$
- If treatment a is given to patient t, he is cured with probability μ_a

The doctor:

- chooses treatment A_t to give to patient t
- observes whether the patient is healed : $X_t \sim \mathcal{B}(\mu_{A_t})$

The doctor can ajust his strategy (A_t) so as to

Regret minimization	Pure-exploration
Maximize the number of patient healed during a study involving T patients	Identify the best treatment with probability at least $1-\delta$ (and always give this one later)

Zoom on an application: Medical trials

A doctor can choose between K different treatments for a given symptom.

- treatment number a has unknown probability of success μ_a
- **Unknown** best treatment $a^* = \text{argmax}_a\ \mu_a$
- If treatment a is given to patient t, he is cured with probability μ_a

The doctor:

- chooses treatment A_t to give to patient t
- observes whether the patient is healed : $X_t \sim \mathcal{B}(\mu_{A_t})$

The doctor can ajust his strategy (A_t) so as to

Regret minimization	Pure-exploration
Maximize the number of patient healed during a study involving T patients	Identify the best treatment with probability at least $1-\delta$ (and always give this one later)

Two probabilistic modelings

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

Frequentist :

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

Bayesian :

- $\theta_a \overset{i.i.d.}{\sim} \pi_a$
- $(X_{a,t})_t$ is i.i.d. conditionally to θ_a with distribution ν_{θ_a}

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

Two measures of performance

- Minimize regret

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

- Minimize Bayes risk

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

Frequentist tools, Bayesian tools

Bandit algorithms based on frequentist tools use:

- Maximum Likelihood Estimator of the mean of each arms
- Confidence Intervals on the mean of each arms

Bandit algorithms based on Bayesian tools use:

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

Frequentist tools, Bayesian tools

Bandit algorithms based on frequentist tools use:

- Maximum Likelihood Estimator of the mean of each arms
- Confidence Intervals on the mean of each arms

Bandit algorithms based on Bayesian tools use:

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

One can **separate tools and objectives**:

Performance criterion	Frequentist algorithms	Bayesian algorithms
Regret	?	?
Bayes risk	?	?

Frequentist tools, Bayesian tools

Bandit algorithms based on frequentist tools use:

- Maximum Likelihood Estimator of the mean of each arms
- Confidence Intervals on the mean of each arms

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**:

Performance criterion	Frequentist algorithms	Bayesian algorithms
Regret	?	?
Bayes risk	?	?

Bayesian algorithm optimal with respect to the Bayes risk

There exists a Bayesian optimal solution to Bayes risk minimization, obtained by dynamic programming.

Bernoulli bandit model $\nu = (\mathcal{B}(\theta_1), \dots, \mathcal{B}(\theta_K))$

- $\theta_a \sim \mathcal{U}([0,1])$
- $\pi_a^t = \text{Beta}(\#|\text{ones observed}| + 1, \#|\text{zeros observed}| + 1)$

The game is summarized by a 'posterior matrix' $\mathcal{S}_t \in \mathcal{M}_{K,2}$

$$\mathcal{S}_t = \begin{pmatrix} 1 & 2 \\ 5 & 1 \\ 0 & 2 \end{pmatrix} \xrightarrow{A_t=2} \begin{cases} \begin{pmatrix} 1 & 2 \\ 6 & 1 \\ 0 & 2 \end{pmatrix} & \textit{if } X_{A_t,t} = 1 \\ \\ \begin{pmatrix} 1 & 2 \\ 5 & 2 \\ 0 & 2 \end{pmatrix} & \textit{if } X_{A_t,t} = 0 \end{cases}$$

$\mathcal{S}_t$ can be seen as a state in a **Markov Decision Process**.

Bayesian algorithm optimal with respect to the Bayes risk

There exists a Bayesian optimal solution to Bayes risk minimization, obtained by dynamic programming.

There exists an optimal policy (A_t) in this MDP satisfying

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

- NOT tracable for large horizon
- with the discounted criterion,
  [Gittins'79] shows the optimal policy reduces to an index policy
- with a finite horizon,
  it does *not* reduce to an index policy

Asymptotically optimal algorithms in the frequentist setting

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

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

- Lai and Robbins,1985 : every consistent policy satisfies

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

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

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

Algorithms: a family of optimistic index policies

- For each arm a, compute an Upper Confidence Bound on μ_a:

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

- Act as if the best possible model was the true model (*optimism-in-face-of-uncertainty*):

$$A_{t+1} = \underset{a}{\arg\max} \; UCB_a(t)$$

Algorithms: a family of optimistic index policies

- For each arm a, compute an Upper Confidence Bound on μ_a:

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

- Act as if the best possible model was the true model (*optimism-in-face-of-uncertainty*):

$$A_{t+1} = \arg\max_a \; UCB_a(t)$$

Example UCB1 [Auer et al. 02] uses Hoeffding bounds:

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

$S_a(t)$: sum of the rewards collected from arm a up to time t.

UCB1 satisfies, for bounded rewards,

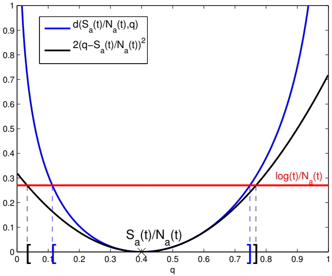
$$\mathbb{E}[N_a(T)] \leq \frac{K_1}{2(\mu_a - \mu^*)^2} \log T + K_2, \quad \text{with } K_1 > 1.$$

KL-UCB: an asymptotically optimal frequentist algorithm

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

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

with $d(p,q) = \text{KL}\left(\mathcal{B}(p), \mathcal{B}(q)\right) = p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right)$.



$$\mathbb{E}[N_a(T)] \leq \frac{1}{d(\mu_a, \mu^*)} \log T + C$$

Summary so far

Objective	Frequentist algorithms	Bayesian algorithms
Regret	KL-UCB	?
Bayes risk	?	Dynamic Programming (not tractable)

Summary so far

Objective	Frequentist algorithms	Bayesian algorithms
Regret	KL-UCB	?
Bayes risk	$\simeq$ KL-UCB [Lai 87]	Dynamic Programming (not tractable)

Summary so far

Objective	Frequentist algorithms	Bayesian algorithms
Regret	KL-UCB	?
Bayes risk	$\simeq$ KL-UCB [Lai 87]	Dynamic Programming (not tractable)

Our objective

We aim at designing algorithms using Bayesian tools that are optimal with respect to (frequentist) regret

UCBs versus Bayesian algorithms

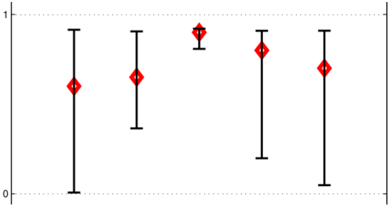


Figure: Confidence intervals on the arms means after t rounds of a bandit game

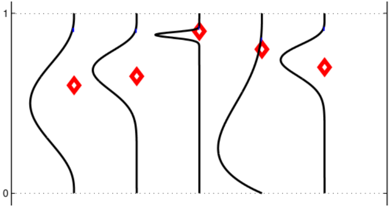


Figure: Posterior over the means of the arms after t rounds of a bandit game

$\Rightarrow$ 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

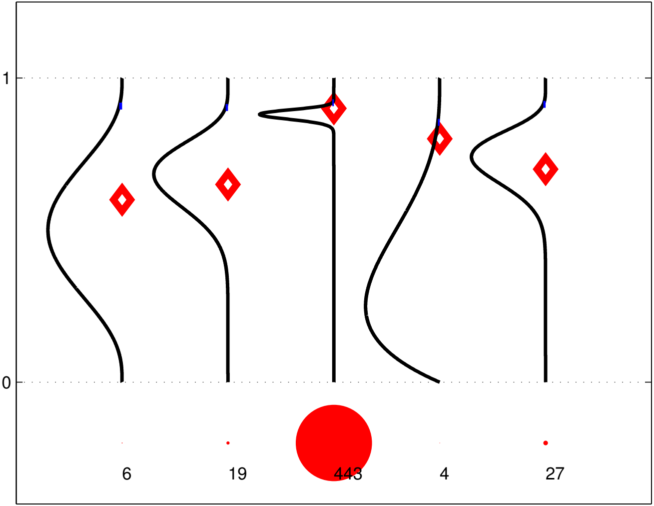
$$A_t = \underset{a}{\text{argmax}}\, Q\left(1 - \frac{1}{t(\log t)^c}, \lambda_a^{t-1}\right)$$

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

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

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

Bayes UCB in action !



Theoretical results for the Bernoulli case

- **Bayes-UCB is asymptotically optimal for Bernoulli rewards**

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

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

Links with a frequentist algorithm

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

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

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.

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

$\Pi^t = (\pi_1^t, \dots, \pi_K^t)$ the posterior distribution on $(\theta_1, \dots, \theta_K)$ at the end of round t.

- A randomized bayesian algorithm:

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

$\Rightarrow$ Each arm is drawn according to its posterior probability of being optimal

- (Recent) interest for this algorithm:
  - TS is the first bandit algorithm proposed [Thompson 1933]
  - Partial analysis were proposed by [Granmo 2010][May, Korda, Lee, Leslie 2012]
  - Numerical studies assess its performance beyond the Bernoulli case [Scott, 2010],[Chapelle, Li 2011]
  - The first logarithmic upper bound on the regret was given by [Agrawal,Goyal 2012]

An optimal regret bound for Bernoulli bandits

Assume arm 1 is the unique optimal arm and let $\Delta_a = \mu_1 - \mu_a$.

- Known result : [Agrawal,Goyal 2012]

$$R_T(\theta) \leq C \left(\sum_{a=2}^{K} \frac{1}{\Delta_a^2} \right)^2 \log(T) + o_\mu(\log(T))$$

An optimal regret bound for Bernoulli bandits

Assume arm 1 is the unique optimal arm and let $\Delta_a = \mu_1 - \mu_a$.

- Known result : [Agrawal,Goyal 2012]

$$R_T(\theta) \leq C \left(\sum_{a=2}^{K} \frac{1}{\Delta_a^2} \right)^2 \log(T) + o_\mu(\log(T))$$

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

Theorem $\forall \epsilon > 0$,

$$R_T(\theta) \leq (1+\epsilon) \left(\sum_{a=2}^{K} \frac{\Delta_a}{d(\mu_a, \mu^*)} \right) \log(T) + o_{\mu,\epsilon}(\log(T))$$

Two key elements in the proof

- Introduce a quantile to replace the sample:

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

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

Two key elements in the proof

- Introduce a quantile to replace the sample:

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

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

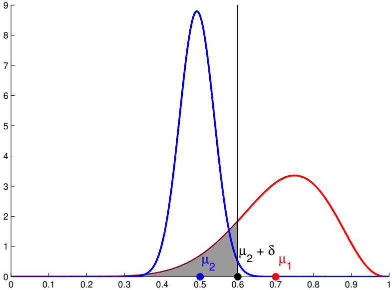
- Proove separately that the optimal arm has to be drawn a lot

Proposition

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

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

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



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 :

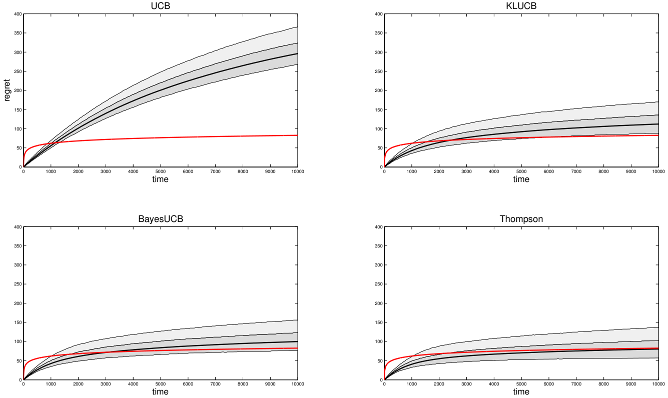
- $\forall s \in \mathcal{J}_j$, $\theta_1(s) \leq \mu_2 + \delta$

$\Rightarrow$ This only happens with small probability

Summary

Objective	Frequentist algorithms	Bayesian algorithms
Regret	KL-UCB	Bayes-UCB Thompson Sampling
Bayes risk	$\simeq$ KL-UCB [Lai 87]	Dynamic Programming (not tractable)

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 $T = 20000$, 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:

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

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

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

- 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

m best arms identification

Assume $\mu_1 \geq \dots \geq \mu_m > \mu_{m+1} \geq \dots \mu_K$ (Bernoulli bandit model)

Parameters and notations

- m the number of arms to find
- $\delta \in]0,1[$ a risk parameter
- $\mathcal{S}_m^* = \{1, \dots, m\}$ the set of m optimal arms

m best arms identification

Assume $\mu_1 \geq \cdots \geq \mu_m > \mu_{m+1} \geq \ldots \mu_K$ (Bernoulli bandit model)

Parameters and notations

- m the number of arms to find
- $\delta \in]0,1[$ a risk parameter
- $\mathcal{S}_m^* = \{1, \ldots, m\}$ the set of m optimal arms

The forecaster

- chooses at time t one (or several) arms to draw
- decides to stop after a (possibly random) total number of samples from the arms τ
- recommends a set $\hat{\mathcal{S}}$ of m arms

m best arms identification

Assume $\mu_1 \geq \cdots \geq \mu_m > \mu_{m+1} \geq \ldots \mu_K$ (Bernoulli bandit model)

Parameters and notations

- m the number of arms to find
- $\delta \in]0,1[$ a risk parameter
- $\mathcal{S}_m^* = \{1, \ldots, m\}$ the set of m optimal arms

The forecaster

- chooses at time t one (or several) arms to draw
- decides to stop after a (possibly random) total number of samples from the arms τ
- recommends a set $\hat{\mathcal{S}}$ of m arms

His goal (in the *fixed-confidence setting*)

- $\mathbb{P}(\hat{\mathcal{S}} = \mathcal{S}_m^*) \geq 1 - \delta$ (the algorithm is δ-PAC)
- The sample complexity $\mathbb{E}[\tau]$ is small

Challenges for m best arm identification

The regret minimization problem is 'solved' in some sense:

- An (asymptotic) lower bound on the regret of any good algorithm

$$\liminf_{n\to\infty} \frac{R_n}{\log(n)} \geq \sum_{a=2}^{K} \frac{\mu_1 - \mu_a}{\mathrm{KL}(\mathcal{B}(\mu_a), \mathcal{B}(\mu_1))}$$

- Algorithms matching this lower bound: KL-UCB, Thompson Sampling

Challenges for m best arm identification

The regret minimization problem is 'solved' in some sense:

- An (asymptotic) lower bound on the regret of any good algorithm

$$\liminf_{n\to\infty} \frac{R_n}{\log(n)} \geq \sum_{a=2}^{K} \frac{\mu_1 - \mu_a}{\mathrm{KL}(\mathcal{B}(\mu_a), \mathcal{B}(\mu_1))}$$

- Algorithms matching this lower bound: KL-UCB, Thompson Sampling

For m best arm identification, we would want to give:

- A lower bound on the sample complexity $\mathbb{E}[\tau]$ of any δ-PAC algorithm
- δ-PAC algorithms whose sample complexity matches this lower bound

A lower bound

Theorem [K.,Cappé, Garivier (14)]
Any algorithm that is δ-PAC on every bandit model such that $\mu_m > \mu_{m+1}$ satisfies, for $\delta \leq 0.15$,

$$\mathbb{E}[\tau] \geq \left(\sum_{t=1}^{m} \frac{1}{d(\mu_a, \mu_{m+1})} + \sum_{t=m+1}^{K} \frac{1}{d(\mu_a, \mu_m)} \right) \log \frac{1}{2\delta}$$

An algorithm: KL-LUCB

Generic notation:

- confidence interval (C.I.) on the mean of arm a at round t:

$$\mathcal{I}_a(t) = [L_a(t), U_a(t)]$$

- $J(t)$ the set of estimated m best arms at round t (m empirical best)

Our contribution: Introduce KL-based confidence intervals

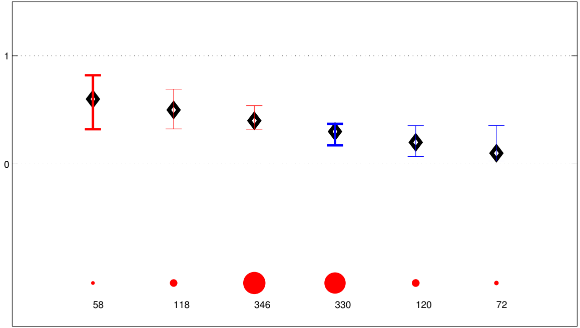
$$\begin{aligned} U_a(t) &= \max\left\{q \geq \hat{\mu}_a(t) : N_a(t) d(\hat{\mu}_a(t), q) \leq \beta(t,\delta)\right\} \\ L_a(t) &= \min\left\{q \leq \hat{\mu}_a(t) : N_a(t) d(\hat{\mu}_a(t), q) \leq \beta(t,\delta)\right\} \end{aligned}$$

for $\beta(t,\delta)$ some exploration rate.

An algorithm: KL-LUCB

At round t, the algorithm:

- draws only two well-chosen arms: u_t and l_t (in bold)
- stops when C.I. for arms in $J(t)$ and $J(t)^c$ are separated



$$m = 3, K = 6$$

Set $J(t)$, arm l_t in bold Set $J(t)^c$, arm u_t in bold

Theoretical guarantees

Theorem [K.,Kalyanakrishnan 2013]
KL-LUCB using the exploration rate

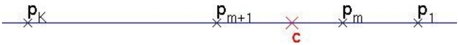
$$\beta(t,\delta) = \log\left(\frac{k_1 K t^\alpha}{\delta}\right),$$

with $\alpha > 1$ and $k_1 > 1 + \frac{1}{\alpha-1}$ satisfies $\mathbb{P}(\hat{\mathcal{S}} = \mathcal{S}_m^*) \geq 1 - \delta$.
For $\alpha > 2$,

$$\mathbb{E}[\tau] \leq 4\alpha H^* \left[\log\left(\frac{k_1 K (H^*)^\alpha}{\delta}\right) + \log\log\left(\frac{k_1 K (H^*)^\alpha}{\delta}\right)\right] + C_\alpha,$$

with

$$H^* = \min_{c\in[\mu_{m+1};\mu_m]} \sum_{a=1}^{K} \frac{1}{d^*(\mu_a, c)}.$$



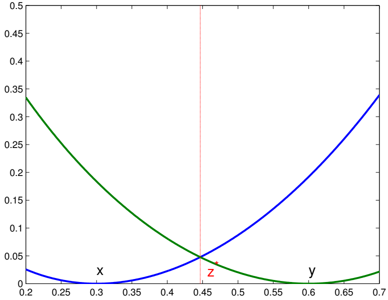
Theoretical guarantees

- **An alternative informational quantity: Chernoff information**

$$d^*(x,y) := d(z^*,x) = d(z^*,y),$$

where z^* is defined by the equality

$$d(z^*,x) = d(z^*,y).$$



Summary

- Lower bound:

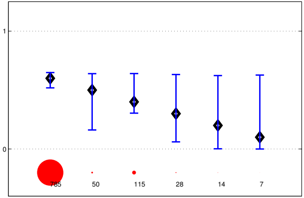
$$\limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau]}{\log\frac{1}{\delta}} \geq \sum_{t=1}^{m} \frac{1}{d(\mu_a,\mu_{m+1})} + \sum_{t=m+1}^{K} \frac{1}{d(\mu_a,\mu_m)}$$

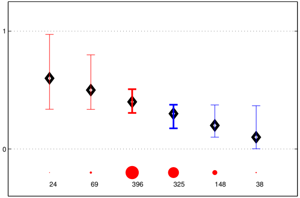
- Upper bound (for KL-LUCB):

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau]}{\log\frac{1}{\delta}} \leq 4 \min_{c\in[\mu_{m+1};\mu_m]} \sum_{a=1}^{K} \frac{1}{d^*(\mu_a,c)}$$

Regret minimization versus Best arms identification

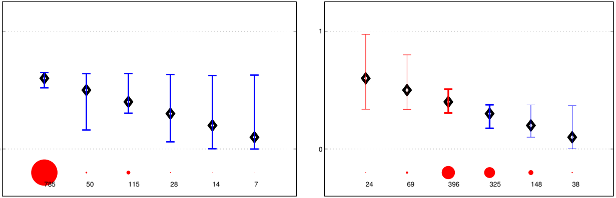
- KL-based confidence intervals are useful in both settings, although KL-UCB and KL-LUCB draw the arms in a different fashion





Regret minimization versus Best arms identification

- KL-based confidence intervals are useful in both settings, although KL-UCB and KL-LUCB draw the arms in a different fashion



- Do the complexity of these two problems feature the same information-theoretic quantities?

$$\inf_{\substack{\text{consistent}\\\text{algorithms}}} \limsup_{T\to\infty} \frac{R_T}{\log T} = \sum_{a=2}^{K} \frac{\mu_1 - \mu_a}{d(\mu_a, \mu_1)}$$

$$\inf_{\substack{\delta-PAC\\\text{algorithms}}} \limsup_{\delta\to\infty} \frac{\mathbb{E}[\tau]}{\log(1/\delta)} \geq \sum_{a=1}^{K} \frac{1}{d(\mu_a, \mu_{m+1})} + \sum_{a=m+1}^{K} \frac{1}{d(\mu_a, \mu_m)}$$

Conclusion

The use of KL-based confidence intervals is useful in bandits models:

- KL-UCB is asymptotically optimal in the regret setting
- KL-LUCB is provably very efficient in the pure-exploration setting

Conclusion

The use of KL-based confidence intervals is useful in bandits models:

- KL-UCB is asymptotically optimal in the regret setting
- KL-LUCB is provably very efficient in the pure-exploration setting

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 practice)

Conclusion

The use of KL-based confidence intervals is useful in bandits models:

- KL-UCB is asymptotically optimal in the regret setting
- KL-LUCB is provably very efficient in the pure-exploration setting

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 practice)

Natural open question:

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