# Stratégies bayésiennes et fréquentistes dans un modèle de bandit

Emilie Kaufmann

thèse effectuée à Telecom ParisTech, co-dirigée par Olivier Cappé, Aurélien Garivier et Rémi Munos

Journées MAS, Grenoble, 30 août 2016

# The multi-armed bandit model

$K$ arms $=$ $K$ probability distributions ($\nu_a$ has mean $\mu_a$)


$\nu_1$ $\nu_2$ $\nu_3$ $\nu_4$ $\nu_5$

At round $t$, an agent:

- chooses an arm $A_t$
- observes a sample $X_t \sim \nu_{A_t}$

using a sequential sampling strategy $(A_t)$:

$$A_{t+1} = F_t(A_1, X_1, \ldots, A_t, X_t).$$

**Generic goal:** learn the best arm, $a^* = \text{argmax}_a \, \mu_a$

# Regret minimization in a bandit model

Samples = **rewards**, $(A_t)$ is adjusted to

- maximize the (expected) sum of rewards,

$$\mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$$

- or equivalently minimize the *regret*:

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

$$\mu^* = \mu_{a^*} = \max_a \mu_a$$

$\Rightarrow$ **Exploration/Exploitation tradeoff**

# Modern motivation: recommendation tasks


$\nu_1$ $\nu_2$ $\nu_3$ $\nu_4$ $\nu_5$

For the $t$-th visitor of a website,

- recommend a movie $A_t$
- observe a rating $X_t \sim \nu_{A_t}$ (e.g. $X_t \in \{1, \dots, 5\}$)

**Goal:** maximize the sum of ratings

# Back to the initial motivation: clinical trials

$\mathcal{B}(\mu_1)$

$\mathcal{B}(\mu_2)$

$\mathcal{B}(\mu_3)$

$\mathcal{B}(\mu_4)$


$\mathcal{B}(\mu_5)$

For the $t$-th patient in a clinical study,

- chooses a treatment $A_t$
- observes a response $X_t \in \{0, 1\}$: $\mathbb{P}(X_t = 1) = \mu_{A_t}$

**Goal:** maximize the number of patient healed during the study

# Back to the initial motivation: clinical trials





$\mathcal{B}(\mu_1)$ $\mathcal{B}(\mu_2)$ $\mathcal{B}(\mu_3)$ $\mathcal{B}(\mu_4)$ $\mathcal{B}(\mu_5)$

For the $t$-th patient in a clinical study,

- chooses a treatment $A_t$
- observes a response $X_t \in \{0, 1\}$: $\mathbb{P}(X_t = 1) = \mu_{A_t}$

**Goal:** maximize the number of patient healed during the study

**Alternative goal:** allocate the treatments so as to identify as quickly as possible the best treatment
(no focus on curing patients during the study)

# Two bandit problems

| | **Regret minimization** | **Best arm identification** |
|---|---|---|
| Bandit algorithm | sampling rule $(A_t)$ | sampling rule $(A_t)$<br>stopping rule $\tau$<br>recommendation rule $\hat{a}_\tau$ |
| Input | horizon $T$ | risk parameter $\delta$ |
| Objective | minimize<br>$R_T = \mathbb{E}\left[\mu^* T - \sum_{t=1}^T X_t\right]$ | ensure $\mathbb{P}(\hat{a}_\tau = a^*) \geq 1 - \delta$<br>and minimize $\mathbb{E}[\tau]$ |
| | **Exploration/Exploitation** | **pure Exploration** |

➜ Goal: find efficient, optimal algorithms for both objectives

In the presentation, we focus on Bernoulli bandit models

$$\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$$

# Outline

1. Asymptotically optimal algorithms for regret minimization
2. A Bayesian approach for regret minimization
   - Bayes-UCB
   - Thompson Sampling
3. Optimal algorithms for Best Arm Identification

# Outline

1. Asymptotically optimal algorithms for regret minimization

2. A Bayesian approach for regret minimization
   - Bayes-UCB
   - Thompson Sampling

3. Optimal algorithms for Best Arm Identification

# Optimal algorithms for regret minimization

$\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$. $N_a(t)$ : number of draws of arm $a$ up to time $t$

$$\mathrm{R}_{\boldsymbol{\mu}}(\mathcal{A}, T) = \mu^* T - \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{T} X_t\right] = \sum_{a=1}^{K} (\mu^* - \mu_a)\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]$$

**Notation: Kullback-Leibler divergence**

$$\begin{aligned} d(\mu, \mu') &:= \mathrm{KL}\left(\mathcal{B}(\mu), \mathcal{B}(\mu')\right) \\ &= \mu \log(\mu/\mu') + (1-\mu)\log((1-\mu)/(1-\mu')) \end{aligned}$$

- [Lai and Robbins, 1985]: for uniformly efficient algorithms,

$$\mu_a < \mu^* \Rightarrow \liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]}{\log T} \geq \frac{1}{d(\mu_a, \mu^*)}$$

A bandit algorithm is **asymptotically optimal** if, for every $\boldsymbol{\mu}$,

$$\mu_a < \mu^* \Rightarrow \limsup_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]}{\log T} \leq \frac{1}{d(\mu_a, \mu^*)}$$

# Algorithms: naive ideas

- **Idea 1 :** Choose each arm $T/K$ times

$\Rightarrow$ EXPLORATION

- **Idea 2 :** Always choose the best arm so far

$$A_{t+1} = \underset{a}{\text{argmax}}\ \hat{\mu}_a(t)$$

$\Rightarrow$ EXPLOITATION

**...Linear regret**

# Algorithms: naive ideas

- **Idea 1 :** Choose each arm $T/K$ times

$\Rightarrow$ EXPLORATION

- **Idea 2 :** Always choose the best arm so far

$$A_{t+1} = \underset{a}{\text{argmax}} \; \hat{\mu}_a(t)$$

$\Rightarrow$ EXPLOITATION

**...Linear regret**

- **Idea 3 :** First explore the arms uniformly, then commit to the empirical best until the end

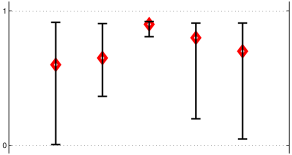
$\Rightarrow$ EXPLORATION followed by EXPLOITATION

**...Still sub-optimal**

# Mixing Exploration and Exploitation: the UCB approach

$$\mathcal{I}_a(t) = [\mathrm{LCB}_a(t), \mathrm{UCB}_a(t)]$$

a confidence interval on $\mu_a$, based on observation up to round $t$.



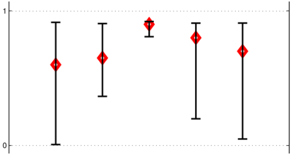

- A UCB-type (or *optimistic*) algorithm chooses at round $t$

$$A_{t+1} = \underset{a=1\dots K}{\text{argmax}}\ \mathrm{UCB}_a(t).$$

# Mixing Exploration and Exploitation: the UCB approach

$$\mathcal{I}_a(t) = [\mathrm{LCB}_a(t), \mathrm{UCB}_a(t)]$$

a confidence interval on $\mu_a$, based on observation up to round $t$.




- A UCB-type (or *optimistic*) algorithm chooses at round $t$

$$A_{t+1} = \underset{a=1\dots K}{\text{argmax}}\ \mathrm{UCB}_a(t).$$

**How to choose the Upper Confidence Bounds ?**

- use appropriate deviation inequalities...
- ... in order to guarantee that

$$\mathbb{P}(\mu_a \leq \mathrm{UCB}_a(t)) \gtrsim 1 - t^{-1}$$

# The KL-UCB algorithm

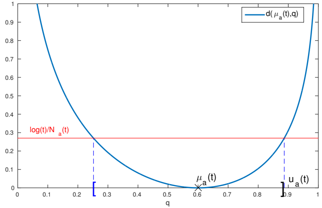
$\hat{\mu}_a(t)$: empirical mean of rewards from arm $a$ up to time $t$.

- A deviation inequality involving the KL-divergence:

$$\mathbb{P}\left(N_a(t) d(\hat{\mu}_a(t), \mu_a) \geq \gamma\right) \leq 2e(\log(t)+1)\gamma e^{-\gamma}$$

- The KL-UCB algorithm: $A_{t+1} = \arg\max_a \ u_a(t)$ with

$$u_a(t) := \max\left\{q : d\left(\hat{\mu}_a(t), q\right) \leq \frac{\log(t) + c\log\log(t)}{N_a(t)}\right\},$$




[Cappé et al. 13]: KL-UCB satisfies, for $c \geq 5$,

$$\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] \leq \frac{1}{d(\mu_a, \mu^*)}\log T + O(\sqrt{\log(T)}).$$

# Outline

# A frequentist or a Bayesian model?

$\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$.

- Two probabilistic modelings

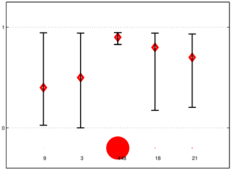
| **Frequentist model** | **Bayesian model** |
|---|---|
| $\mu_1, \dots, \mu_K$<br>unknown parameters | $\mu_1, \dots, \mu_K$ drawn from a<br>prior distribution : $\mu_a \sim \pi_a$ |
| arm $a$: $(Y_{a,s})_s \overset{\text{i.i.d.}}{\sim} \mathcal{B}(\mu_a)$ | arm $a$: $(Y_{a,s})_s \vert \boldsymbol{\mu} \overset{\text{i.i.d.}}{\sim} \mathcal{B}(\mu_a)$ |

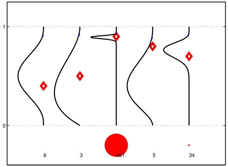
- The regret can be computed in each case

| Frequentist regret<br>(regret) | Bayesian regret<br>(Bayes risk) |
|---|---|
| $R_T(\mathcal{A}, \boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{T} (\mu^* - \mu_{A_t})\right]$ | $\mathcal{R}_T(\mathcal{A}, \pi) = \mathbb{E}_{\boldsymbol{\mu} \sim \pi}\left[\sum_{t=1}^{T} (\mu^* - \mu_{A_t})\right]$<br>$= \int R_T(\mathcal{A}, \boldsymbol{\mu}) d\pi(\boldsymbol{\mu})$ |

- Two types of tools to build bandit algorithms:

| Frequentist tools | Bayesian tools |
|---|---|
| MLE estimators of the means<br>Confidence Intervals | Posterior distributions<br>$\pi_a^t = \mathcal{L}(\mu_a \| X_{a,1}, \dots, X_{a,N_a(t)})$ |




- One can separate tools and objective:

We present efficient Bayesian algorithms for regret minimization

# Outline

# The Bayes-UCB algorithm

$\pi_a^t$ the posterior distribution over $\mu_a$ at the end of round $t$.

**Algorithm: Bayes-UCB [K., Cappé, Garivier 2012]**

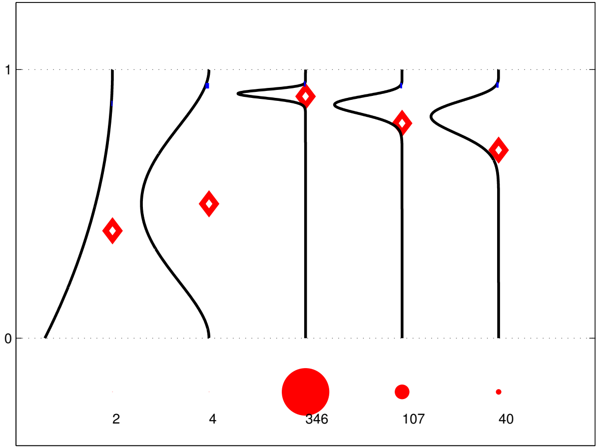
$$A_{t+1} = \underset{a}{\text{argmax}} \; Q\left(1 - \frac{1}{t(\log t)^c}, \pi_a^t\right)$$

where $Q(\alpha, p)$ is the quantile of order $\alpha$ of the distribution $p$.

Bernoulli reward with uniform prior:

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

# Bayes-UCB in practice



- **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 parameter $c \geq 5$ satisfies

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

# Links to a frequentist algorithm

Bayes-UCB index is close to KL-UCB indices:

**Lemma**

$$\tilde{u}_a(t) \leq q_a(t) \leq u_a(t)$$

with:

$$u_a(t) = \max\left\{q : d\left(\hat{\mu}_a(t), q\right) \leq \frac{\log(t) + c\log\log(t)}{N_a(t)}\right\}$$

$$\tilde{u}_a(t) = \max\left\{q : d\left(\frac{N_a(t)\hat{\mu}_a(t)}{N_a(t)+1}, q\right) \leq \frac{\log\left(\frac{t}{N_a(t)+2}\right) + c\log\log(t)}{(N_a(t)+1)}\right\}$$

**Bayes-UCB automatically builds confidence intervals based on the Kullback-Leibler divergence !**

# 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 inequalities

**Lemma**

For $k > nx$,

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

# Outline

# Thompson Sampling

$(\pi_1^t, .., \pi_K^t)$ posterior distribution on $(\mu_1, .., \mu_K)$ at round $t$.

**Algorithm: Thompson Sampling**

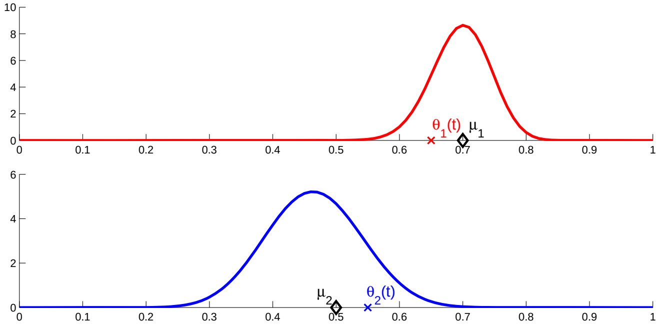
**Thompson Sampling** is a randomized Bayesian algorithm:

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

"Draw each arm according to its posterior probability of being optimal"

- the first bandit algorithm, introduced by [Thompson 1933]

# Illustration of the algorithm




*Posterior distributions of the mean of arm 1 (top) and arm 2 (bottom) in a two-armed bandit model*

# Thompson Sampling is asymptotically optimal

- good empirical performance in complex models
- first logarithmic regret bound by [Agrawal and Goyal 2012]

**Theorem** [K.,Korda,Munos 2012]

For all $\epsilon > 0$,

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

# Outline

1. Asymptotically optimal algorithms for regret minimization
2. A Bayesian approach for regret minimization
   - Bayes-UCB
   - Thompson Sampling
3. Optimal algorithms for Best Arm Identification

# A sample complexity lower bound

A Best Arm Identification algorithm $(A_t, \tau, \hat{a}_\tau)$ is $\delta$-PAC if

$$\forall \boldsymbol{\mu}, \ \mathbb{P}_{\boldsymbol{\mu}}(\hat{a}_\tau = a^*(\boldsymbol{\mu})) \geq 1 - \delta.$$

**Theorem** [Garivier and K. 2016]

For any $\delta$-PAC algorithm,

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geq T^*(\boldsymbol{\mu}) \log\left(\frac{1}{2.4\delta}\right),$$

where

$$T^*(\boldsymbol{\mu})^{-1} = \sup_{w \in \Sigma_K} \inf_{\{\boldsymbol{\lambda} : a^*(\boldsymbol{\lambda}) \neq a^*(\boldsymbol{\mu})\}} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a)$$

with $\Sigma_K = \{w \in [0,1]^K : \sum_{i=1}^K w_i = 1\}$.

# Optimal proportion of draws

The vector

$$w^*(\boldsymbol{\mu}) = \underset{w \in \Sigma_K}{\text{argmax}} \inf_{\{\boldsymbol{\lambda} : a^*(\boldsymbol{\lambda}) \neq a^*(\boldsymbol{\mu})\}} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a)$$

contains the optimal proportions of draws of the arms, i.e. an algorithm matching the lower bound should satisfy

$$\forall a \in \{1, \dots, K\}, \quad \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau)]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau]} \simeq w_a^*(\boldsymbol{\mu}).$$

- Building on this notion of optimal proportions, one can exhibit an asymptotically optimal algorithm:

$$\lim_{\delta \to \infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} = T^*(\boldsymbol{\mu}).$$

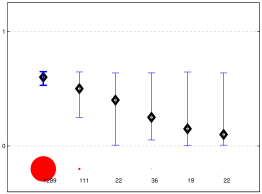
# Conclusion

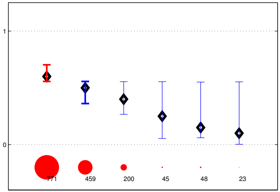
We saw two Bayesian algorithms that are good alternative to KL-UCB for regret minimization because:

- they are also asymptotically optimal in simple models
- they display better empirical performance
- … they can be easily generalized to more complex models

Algorithms for regret minimization and BAI are very different!

- playing mostly the best arm vs. optimal proportions




- different "complexity terms" (featuring KL-divergence)

$$R_T \simeq \left( \sum_{a \neq a^*} \frac{\mu^* - \mu_a}{d(\mu_a, \mu^*)} \right) \log(T) \qquad \mathbb{E}_{\boldsymbol{\mu}}[\tau] \simeq T^*(\boldsymbol{\mu}) \log(1/\delta)$$

# Merci de votre attention.

# Bayesian algorithms in contextual linear bandit models

At time $t$, a set of 'contexts' $\mathcal{D}_t \subset \mathbb{R}^d$ is revealed.

$= \text{characteristics of the items to recommend}$

**The model**:

- if the context $x_t \in \mathcal{D}_t$ is selected
- a reward $r_t = x_t^T \theta + \epsilon_t$ is received

$$\theta \in \mathbb{R}^d = \text{underlying preference vector}$$

**A Bayesian model**: (with Gaussian prior)

$$r_t = x_t^T \theta + \epsilon_t, \qquad \theta \sim \mathcal{N}\left(0, \kappa^2 \mathrm{I}_d\right), \qquad \epsilon_t \sim \mathcal{N}\left(0, \sigma^2\right).$$

Explicit posterior: $p(\theta | x_1, r_1, \dots, x_t, r_t) = \mathcal{N}\left(\hat{\theta}(t), \Sigma_t\right)$.

**Thompson Sampling**:

$$\tilde{\theta}(t) \sim \mathcal{N}\left(\hat{\theta}(t), \Sigma_t\right), \quad \text{and} \quad x_{t+1} = \underset{x \in \mathcal{D}_{t+1}}{\text{argmax}}\ x^T \tilde{\theta}(t).$$