Minimisation du regret vs. Exploration pure: Deux critères de performance pour des algorithmes de bandit

Emilie Kaufmann (Telecom ParisTech)  
joint work with Olivier Cappé, Aurélien Garivier  
and Shivaram Kalyanakrishnan (Yahoo Labs)

ANR Spadro, 9 avril 2014
1. Two bandit problems

2. Regret minimization: a well solved problem

3. Algorithms for pure-exploration

4. The complexity of $m$ best arms identification
1. Two bandit problems

2. Regret minimization: a well solved problem

3. Algorithms for pure-exploration

4. The complexity of $m$ best arms identification
A **multi-armed bandit model** is a set of $K$ arms where
- Arm $a$ is an unknown probability distribution $\nu_a$ with mean $\mu_a$
- Drawing arm $a$ is observing a realization of $\nu_a$
- Arms are assumed to be independent

In a **bandit game**, at round $t$, an agent
- 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**

$$a^* = \arg\max_a \mu_a$$
A **multi-armed bandit model** is a set of $K$ arms where
- Arm $a$ is a Bernoulli distribution $\mathcal{B}(\mu_a)$ (with unknown mean $\mu_a$)
- Drawing arm $a$ is observing a realization of $\mathcal{B}(\mu_a)$ (0 or 1)
- Arms are assumed to be independent

In a **bandit game**, at round $t$, an agent
- chooses arm $A_t$ to draw based on past observations, according to its **sampling strategy** (or **bandit algorithm**)
- observes a sample $X_t \sim \mathcal{B}(\mu_{A_t})$

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

$$a^* = \arg\max_a \mu_a$$
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**:

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

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

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

- **optimally explores** the environment,

...together with a stopping criterion, and then recommend a set $S$ of $m$ arms such that

$$
P(S \text{ is the set of } m \text{ best arms}) \geq 1 - \delta.
$$
A doctor can choose between \( K \) different treatments for a given symptom.

- treatment number \( a \) has unknown probability of success \( \mu_a \)
- **Unknown** best treatment \( a^* = \arg\max_a \mu_a \)
- If treatment \( a \) is given to patient \( t \), he is cured with probability \( p_a \)

The doctor:

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

Regret minimization versus best arm identification

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 $\mu_a$
- **Unknown** best treatment $a^* = \arg\max_a \mu_a$
- If treatment $a$ is given to patient $t$, he is cured with probability $p_a$

The doctor:

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

The doctor can adjust his strategy $(A_t)$ so as to

<table>
<thead>
<tr>
<th>Regret minimization</th>
<th>Pure exploration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximize the number of patient healed during a study involving $n$ patients</td>
<td>Identify the best treatment with probability at least $1 - \delta$ (and always give this one later)</td>
</tr>
</tbody>
</table>
1 Two bandit problems

2 Regret minimization: a well solved problem

3 Algorithms for pure-exploration

4 The complexity of $m$ best arms identification
Asymptotically optimal algorithms

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

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

- [Lai and Robbins, 1985]: every consistent policy satisfies

$$\mu_a < \mu^* \Rightarrow \liminf_{T \to \infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \geq \frac{1}{\text{KL}(\mathcal{B}(\mu_a), \mathcal{B}(\mu^*_a))}$$

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

$$\mu_a < \mu^* \Rightarrow \limsup_{n \to \infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \leq \frac{1}{\text{KL}(\mathcal{B}(\mu_a), \mathcal{B}(\mu^*_a))}$$
Algorithms: a family of optimistic index policies

- For each arm $a$, compute a confidence interval on $\mu_a$:
  \[ \mu_a \leq UCB_a(t) \text{ w.h.p} \]

- Act as if the best possible model was the true model (optimism-in-face-of-uncertainty):
  \[ A_t = \arg \max_a UCB_a(t) \]
Algorithms: a family of optimistic index policies

- For each arm $a$, compute a confidence interval on $\mu_a$:
  \[ \mu_a \leq UCB_a(t) \text{ w.h.p} \]

- Act as if the best possible model was the true model (optimism-in-face-of-uncertainty):
  \[ A_t = \arg \max_a UCB_a(t) \]

**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 is not asymptotically optimal, but one can show that

\[
E[N_a(T)] \leq \frac{K_1}{2(\mu_a - \mu^*)^2} \ln T + K_2, \quad \text{with } K_1 > 1.
\]
KL-UCB: and asymptotically optimal frequentist algorithm

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

\[
\alpha_a(t) = \text{argmax}_{x > \frac{S_a(t)}{N_a(t)}} \left\{ d\left( \frac{S_a(t)}{N_a(t)}, x \right) \leq \ln(t) + c \ln \ln(t) \right\}
\]

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

\[
\mathbb{E}[N_a(T)] \leq \frac{1}{d(\mu_a, \mu^*)} \ln T + C
\]
Regret minimization: Summary

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

\[
\liminf_{T \to \infty} \frac{R_T}{\log T} \geq \sum_{a: \mu_a < \mu} \frac{\mu^* - \mu_a}{\text{KL}(B(\mu_a), B(\mu^*))}
\]

- An algorithm based on confidence intervals matching this lower bound: KL-UCB
Regret minimization: Summary

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

\[
\liminf_{T \to \infty} \frac{R_T}{\log T} \geq \sum_{a : \mu_a < \mu^*} \frac{\mu^* - \mu_a}{\text{KL}(B(\mu_a), B(\mu^*))}
\]

- An algorithm based on confidence intervals matching this lower bound: KL-UCB

- A Bayesian approach of the MAB problem can also lead to asymptotically optimal algorithms (Thompson Sampling, Bayes-UCB)
1. Two bandit problems

2. Regret minimization: a well solved problem

3. Algorithms for pure-exploration

4. The complexity of $m$ best arms identification
Assume $\mu_1 \geq \cdots \geq \mu_m > \mu_{m+1} \geq \cdots \mu_K$.

**Parameters and notations**

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

Assume $\mu_1 \geq \cdots \geq \mu_m > \mu_{m+1} \geq \cdots \mu_K$.

Parameters and notations

- $m$ the number of arms to find
- $\delta \in ]0, 1[$ a risk parameter
- $S^*_m = \{1, \ldots, m\}$ 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 $\tau$
- recommends a set $S$ of $m$ arms
Assume $\mu_1 \geq \cdots \geq \mu_m > \mu_{m+1} \geq \cdots \geq \mu_K$.

**Parameters and notations**

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

**His goal**

- $\mathbb{P}(S = S^*_m) \geq 1 - \delta$, and $\mathbb{E}[\tau]$ is small (fixed-confidence setting)
**Generic algorithms based on confidence intervals**

**Generic notations:**

- confidence interval 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)

- $u_t \in J(t)^c$ and $l_t \in J(t)$ two 'critical' arms (likely to be misclassified)

  \[
  u_t = \arg\max_{a \notin J(t)} U_a(t) \quad \text{and} \quad l_t = \arg\min_{a \in J(t)} L_a(t).
  \]
(KL)-Racing: uniform sampling and eliminations

The algorithm maintains a set of remaining arms $\mathcal{R}$ and at round $t$:
- draw all the arms in $\mathcal{R}$ (uniform sampling)
- possibly accept the empirical best or discard the empirical worst

$$\mu = [0.6, 0.5, 0.4, 0.3, 0.2, 0.1] \quad m = 3 \quad \delta = 0.1$$

*In this situation, arm 1 is selected*
At round $t$, the algorithm:

- draw only two well-chosen arms: $u_t$ and $l_t$ (adaptive sampling)
- stops when CI for arms in $J(t)$ and $J(t)^c$ are separated

Set $J(t)$, arm $l_t$ in bold  Set $J(t)^c$, arm $u_t$ in bold
Two $\delta$-PAC algorithms

\[ L_a(t) = \min \{ q \in [0, \hat{\mu}_a(t)] : N_a(t)d(\hat{\mu}_a(t), q) \leq \beta(t, \delta) \} , \]
\[ U_a(t) = \max \{ q \in [\hat{\mu}_a(t), 1] : N_a(t)d(\hat{\mu}_a(t), q) \leq \beta(t, \delta) \} . \]

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

**Theorem**

The $KL$-Racing algorithm and $KL$-LUCB algorithm using

\[ \beta(t, \delta) = \log \left( \frac{k_1 K t^\alpha}{\delta} \right), \tag{1} \]

with $\alpha > 1$ and $k_1 > 1 + \frac{1}{\alpha - 1}$ satisfy $\mathbb{P}(S = S^*_m) \geq 1 - \delta$. 
Confidence intervals based on KL are always better

\[ B_1 : K = 15; \mu_1 = \frac{1}{2}; \mu_a = \frac{1}{2} - \frac{a}{40} \quad \text{for } a = 2, 3, \ldots, K. \quad B_2 = \frac{1}{2} B_1 \]

Adaptive Sampling seems to do better than Uniform Sampling
A new informational quantity: Chernoff information

\[ d^*(x, y) := d(z^*, x) = d(z^*, y), \]

where \( z^* \) is defined by the equality

\[ d(z^*, x) = d(z^*, y). \]
KL-LUCB with $\beta(t, \delta) = \log \left( \frac{k_1 K t^\alpha}{\delta} \right)$ is $\delta$-PAC and satisfies, 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

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

\[ \begin{array}{cccc}
\bullet_{p_1} & \times_{p_m} & \times_{c} & \bullet_{p_{m+1}} \end{array} \]
1. Two bandit problems

2. Regret minimization: a well solved problem

3. Algorithms for pure-exploration

4. The complexity of $m$ best arms identification
Lower bound on the number of sample used complexity

For KL-LUCB, $\mathbb{E}[\tau] = O\left( H^* \log \frac{1}{\delta} \right)$.

**Theorem**

*Any algorithm that is $\delta$-PAC on every bandit model such that $\mu_m > \mu_{m+1}$ satisfies, for $\delta \leq 0.15$,*

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

The informational complexity of \( m \) best arm identification

For a bandit model \( \nu \), one can introduce the complexity term

\[
\kappa_C(\nu) = \inf_{\text{A \& PAC algorithm}} \limsup_{\delta \to 0} \frac{\mathbb{E}_\nu[\tau]}{\log \frac{1}{\delta}}.
\]

Our results rewrite

\[
\sum_{t=1}^{m} \frac{1}{d(\mu_a, \mu_{m+1})} + \sum_{t=m+1}^{K} \frac{1}{d(\mu_a, \mu_m)} \leq \kappa_C(\nu) \leq 4 \min_{c \in [\mu_{m+1}; \mu_m]} \sum_{a=1}^{K} \frac{1}{d^*(\mu_a, c)}
\]
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?

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

\[
\inf_{\delta-PAC \text{ algorithms}} \limsup_{\delta \to 0} \frac{\mathbb{E}[\tau]}{\log(1/\delta)} \geq \sum_{a=1}^{K} \frac{1}{d(\mu_a, \mu_{m+1})} + \sum_{a=m+1}^{K} \frac{1}{d(\mu_a, \mu_m)}
\]