On the Complexity of Best Arm Identification in Multi-Armed Bandit Models

Aurélien Garivier

Institut de Mathématiques de Toulouse

Information Theory, Learning and Big Data
Simons Institute, Berkeley, March 2015
Roadmap

1. Simple Multi-Armed Bandit Model

2. Complexity of Best Arm Identification
   - Lower bounds on the complexities
   - Gaussian Feedback
   - Binary Feedback
The (stochastic) Multi-Armed Bandit Model

Environment \( K \) arms with parameters \( \theta = (\theta_1, \ldots, \theta_K) \) such that for any possible choice of arm \( a_t \in \{1, \ldots, K\} \) at time \( t \), one receives the reward

\[
X_t = X_{a_t,t}
\]

where, for any \( 1 \leq a \leq K \) and \( s \geq 1 \), \( X_{a,s} \sim \nu_a \), and the \((X_{a,s})_{a,s}\) are independent.

Reward distributions \( \nu_a \in \mathcal{F}_a \) parametric family, or not: canonical exponential family, general bounded rewards

Example Bernoulli rewards: \( \theta \in [0, 1]^K \), \( \nu_a = \mathcal{B}(\theta_a) \)

Strategy The agent’s actions follow a dynamical strategy

\( \pi = (\pi_1, \pi_2, \ldots) \) such that

\[
A_t = \pi_t(X_1, \ldots, X_{t-1})
\]
Real challenges

- Randomized clinical trials
  - original motivation since the 1930’s
  - dynamic strategies can save resources

- Recommender systems:
  - advertisement
  - website optimization
  - news, blog posts, …

- Computer experiments
  - large systems can be simulated in order to optimize some criterion over a set of parameters
  - but the simulation cost may be high, so that only few choices are possible for the parameters

- Games and planning (tree-structured options)
Performance Evaluation: Cumulated Regret

Cumulated Reward: \( S_T = \sum_{t=1}^{T} X_t \)

Goal: Choose \( \pi \) so as to maximize

\[
E[S_T] = \sum_{t=1}^{T} \sum_{a=1}^{K} E[X_t \mathbb{1}\{A_t = a\} \mid X_1, \ldots, X_{t-1}]
\]

\[
= \sum_{a=1}^{K} \mu_a E[N_{\pi a}(T)]
\]

where \( N_{\pi a}(T) = \sum_{t \leq T} \mathbb{1}\{A_t = a\} \) is the number of draws of arm \( a \) up to time \( T \), and \( \mu_a = E(\nu_a) \).

Regret Minimization: maximizing \( E[S_T] \iff \) minimizing

\[
R_T = T\mu^* - E[S_T] = \sum_{a : \mu_a < \mu^*} (\mu^* - \mu_a) E[N_{\pi a}(T)]
\]

where \( \mu^* \in \max\{\mu_a : 1 \leq a \leq K\} \)
Construct an upper confidence bound for the expected reward of each arm:

$$\frac{S_a(t)}{N_a(t)} + \sqrt{\frac{\log(t)}{2N_a(t)}}$$

Choose the arm with the highest UCB

It is an *index strategy* [Gittins ’79]

Its behavior is easily interpretable and intuitively appealing

Listen to Robert Nowak’s talk tomorrow!
Generalization of [Lai\&Robbins ’85]

**Theorem [Burnetas and Katehakis, ’96]**

If $\pi$ is a uniformly efficient strategy, then for any $\theta \in [0, 1]^K$,

$$
\liminf_{T \to \infty} \frac{\mathbb{E}[N_a(T)]}{\log(T)} \geq \frac{1}{K_{\inf}(\nu_a, \mu^*)}
$$

where

$$
K_{\inf}(\nu_a, \mu^*) = \inf \{ K(\nu_a, \nu') : \nu' \in \mathcal{F}_a, E(\nu') \geq \mu^* \}
$$

Idea: change of distribution
The KL-UCB Algorithm, AoS 2013
joint work with O. Cappé, O-A. Maillard, R. Munos, G. Stoltz

Parameters: An operator \( \Pi_F : \mathcal{M}_1(S) \rightarrow F \); a non-decreasing function \( f : \mathbb{N} \rightarrow \mathbb{R} \)

Initialization: Pull each arm of \( \{1, \ldots, K\} \) once

for \( t = K \) to \( T - 1 \) do

compute for each arm \( a \) the quantity

\[
U_a(t) = \sup \left\{ E(\nu) : \nu \in F \text{ and } KL \left( \Pi_F(\hat{\nu}_a(t)), \nu \right) \leq \frac{f(t)}{N_a(t)} \right\}
\]

pick an arm \( A_{t+1} \in \arg\max_{a \in \{1, \ldots, K\}} U_a(t) \)

end for
**Theorem:** Assume that \( \mathcal{F} \) is the set of finitely supported probability distributions over \( S = [0, 1] \), that \( \mu_a > 0 \) for all arms \( a \) and that \( \mu^* < 1 \). There exists a constant \( M(\nu_a, \mu^*) > 0 \) only depending on \( \nu_a \) and \( \mu^* \) such that, with the choice \( f(t) = \log(t) + \log(\log(t)) \) for \( t \geq 2 \), for all \( T \geq 3 \):

\[
\mathbb{E}[N_a(T)] \leq \frac{\log(T)}{K_{inf}(\nu_a, \mu^*)} + \frac{36}{(\mu^*)^4} (\log(T))^{4/5} \log(\log(T)) + \left( \frac{72}{(\mu^*)^4} + \frac{2\mu^*}{(1 - \mu^*)K_{inf}(\nu_a, \mu^*)^2} \right) (\log(T))^{4/5} + \frac{(1 - \mu^*)^2 M(\nu_a, \mu^*)}{2(\mu^*)^2} (\log(T))^{2/5} + \frac{\log(\log(T))}{K_{inf}(\nu_a, \mu^*)} + \frac{2\mu^*}{(1 - \mu^*)K_{inf}(\nu_a, \mu^*)^2} + 4.
\]
**Theorem:** Assume that $\mathcal{F}$ is the set of finitely supported probability distributions over $S = [0, 1]$, that $\mu_a > 0$ for all arms $a$ and that $\mu^* < 1$. There exists a constant $M(\nu_a, \mu^*) > 0$ only depending on $\nu_a$ and $\mu^*$ such that, with the choice $f(t) = \log(t) + \log(\log(t))$ for $t \geq 2$, for all $T \geq 3$:

$$
\mathbb{E}[N_a(T)] \leq \frac{\log(T)}{K_{inf}(\nu_a, \mu^*)} + \frac{36}{(\mu^*)^4} (\log(T))^{4/5} \log(\log(T)) \\
+ \left( \frac{72}{(\mu^*)^4} + \frac{2\mu^*}{(1 - \mu^*) K_{inf}(\nu_a, \mu^*)^2} \right) (\log(T))^{4/5} \\
+ \frac{(1 - \mu^*)^2 M(\nu_a, \mu^*)}{2(\mu^*)^2} (\log(T))^{2/5} \\
+ \frac{\log(\log(T))}{K_{inf}(\nu_a, \mu^*)} + \frac{2\mu^*}{(1 - \mu^*) K_{inf}(\nu_a, \mu^*)^2} + 4.
$$
Roadmap

1. Simple Multi-Armed Bandit Model

2. Complexity of Best Arm Identification
   - Lower bounds on the complexities
   - Gaussian Feedback
   - Binary Feedback
Best Arm Identification Strategies

A two-armed bandit model is

- a pair $\nu = (\nu_1, \nu_2)$ of probability distributions ('arms') with respective means $\mu_1$ and $\mu_2$
- $a^* = \arg\max_a \mu_a$ is the (unknown) best arm

Strategy =

- a **sampling rule** $(A_t)_{t \in \mathbb{N}}$ where $A_t \in \{1, 2\}$ is the arm chosen at time $t$ (based on past observations) a sample $Z_t \sim \nu_{A_t}$ is observed
- a **stopping rule** $\tau$ indicating when he stops sampling the arms
- a **recommendation rule** $\hat{a}_\tau \in \{1, 2\}$ indicating which arm he thinks is best (at the end of the interaction)

In classical A/B Testing, the sampling rule $A_t$ is uniform on $\{1, 2\}$ and the stopping rule $\tau = t$ is fixed in advance.
Best Arm Identification

Joint work with Emilie Kaufmann and Olivier Cappé (Telecom ParisTech)

Goal: design a strategy $\mathcal{A} = ((A_t), \tau, \hat{a}_\tau)$ such that:

<table>
<thead>
<tr>
<th>Fixed-budget setting</th>
<th>Fixed-confidence setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau = t$</td>
<td>$\mathbb{P}<em>\nu(\hat{a}</em>\tau \neq a^*) \leq \delta$</td>
</tr>
<tr>
<td>$p_t(\nu) := \mathbb{P}_\nu(\hat{a}_t \neq a^*)$ as small as possible</td>
<td>$\mathbb{E}_\nu[\tau]$ as small as possible</td>
</tr>
</tbody>
</table>

See also: [Mannor&Tsitsiklis ’04], [Even-Dar&al. ’06], [Audibert&al. ’10], [Bubeck&al. ’11,’13], [Kalyanakrishnan&al. ’12], [Karnin&al. ’13], [Jamieson&al. ’14]...
Two possible goals

Goal: design a strategy $\mathcal{A} = ((A_t), \tau, \hat{a}_\tau)$ such that:

<table>
<thead>
<tr>
<th>Fixed-budget setting</th>
<th>Fixed-confidence setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau = t$</td>
<td>$\mathbb{P}<em>\nu(\hat{a}</em>\tau \neq a^*) \leq \delta$</td>
</tr>
<tr>
<td>$p_t(\nu) := \mathbb{P}_\nu(\hat{a}_t \neq a^*)$ as small as possible</td>
<td>$\mathbb{E}_\nu[\tau]$ as small as possible</td>
</tr>
</tbody>
</table>

In the particular case of uniform sampling:

<table>
<thead>
<tr>
<th>Fixed-budget setting</th>
<th>Fixed-confidence setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>classical test of $(\mu_1 &gt; \mu_2)$ against $(\mu_1 &lt; \mu_2)$ based on $t$ samples</td>
<td>sequential test of $(\mu_1 &gt; \mu_2)$ against $(\mu_1 &lt; \mu_2)$ with probability of error uniformly bounded by $\delta$</td>
</tr>
</tbody>
</table>

[Siegmund 85]: sequential tests can save samples!
The complexities of best-arm identification

For a class $\mathcal{M}$ bandit models, algorithm $\mathcal{A} = ((A_t), \tau, \hat{a}_\tau)$ is...

<table>
<thead>
<tr>
<th>Fixed-budget setting</th>
<th>Fixed-confidence setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>consistent on $\mathcal{M}$ if</td>
<td>$\delta$-PAC on $\mathcal{M}$ if</td>
</tr>
<tr>
<td>$\forall \nu \in \mathcal{M}, p_t(\nu) = \mathbb{P}_\nu(\hat{a}_t \neq a^*) \xrightarrow{t \to \infty} 0$</td>
<td>$\forall \nu \in \mathcal{M}, \mathbb{P}<em>\nu(\hat{a}</em>\tau \neq a^*) \leq \delta$</td>
</tr>
</tbody>
</table>

From the literature

$$p_t(\nu) \simeq \exp \left( -\frac{t}{C_H(\nu)} \right)$$

[Audibert et al.’10], [Bubeck et al.’11],
[Bubeck et al.’13], ...

$$\mathbb{E}_\nu[\tau] \simeq C' H'(\nu) \log(1/\delta)$$

[Mannor & Tsitsiklis ’04], [Even-Dar et al. ’06],
[Kalanakrishnan et al.’12], ...

$\implies$ two complexities

$$\kappa_B(\nu) = \inf_{\mathcal{A} \text{ cons.}} \left( \limsup_{t \to \infty} -\frac{1}{t} \log p_t(\nu) \right)^{-1}$$

for a probability of error $\leq \delta$,
budget $t \simeq \kappa_B(\nu) \log(1/\delta)$

$$\kappa_C(\nu) = \inf_{\mathcal{A} \delta\text{-PAC}} \limsup_{\delta \to 0} \frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)}$$

for a probability of error $\leq \delta$,
$$\mathbb{E}_\nu[\tau] \simeq \kappa_C(\nu) \log(1/\delta)$$
Theorem: how to use (and hide) the change of distribution

Let $\nu$ and $\nu'$ be two bandit models with $K$ arms such that for all $a$, the distributions $\nu_a$ and $\nu'_a$ are mutually absolutely continuous. For any almost-surely finite stopping time $\sigma$ with respect to $(\mathcal{F}_t)$,

$$\sum_{a=1}^{K} \mathbb{E}_\nu [N_a(\sigma)] \text{KL}(\nu_a, \nu'_a) \geq \sup_{\mathcal{E} \in \mathcal{F}_\sigma} \text{kl}(\mathbb{P}_\nu(\mathcal{E}), \mathbb{P}_{\nu'}(\mathcal{E})),$$

where $\text{kl}(x, y) = x \log(x/y) + (1 - x) \log((1 - x)/(1 - y))$.

Useful remark:

$$\forall \delta \in [0, 1], \quad \text{kl}(\delta, 1 - \delta) \geq \log \frac{1}{2.4 \delta},$$
General lower bounds

**Theorem 1**

Let $\mathcal{M}$ be a class of two armed bandit models that are continuously parametrized by their means. Let $\nu = (\nu_1, \nu_2) \in \mathcal{M}$.

<table>
<thead>
<tr>
<th>Fixed-budget setting</th>
<th>Fixed-confidence setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>any consistent algorithm satisfies</td>
<td>any $\delta$-PAC algorithm satisfies</td>
</tr>
<tr>
<td>$\limsup_{t \to \infty} -\frac{1}{t} \log p_t(\nu) \leq K^*(\nu_1, \nu_2)$</td>
<td>$\mathbb{E}<em>\nu[\tau] \geq \frac{1}{K^*</em>*(\nu_1, \nu_2)} \log \left( \frac{1}{2.4\delta} \right)$</td>
</tr>
<tr>
<td>with $K^<em>(\nu_1, \nu_2) = KL(\nu^</em>, \nu_1) = KL(\nu^*, \nu_2)$</td>
<td>with $K^<em>_</em>(\nu_1, \nu_2) = KL(\nu_1, \nu^<em>) = KL(\nu_2, \nu^</em>)$</td>
</tr>
<tr>
<td>Thus, $\kappa_B(\nu) \geq \frac{1}{K^*(\nu_1, \nu_2)}$</td>
<td>Thus, $\kappa_C(\nu) \geq \frac{1}{K^<em>_</em>(\nu_1, \nu_2)}$</td>
</tr>
</tbody>
</table>
Gaussian Rewards: Fixed-Budget Setting

For fixed (known) values $\sigma_1, \sigma_2$, we consider Gaussian bandit models

$$\mathcal{M} = \{ \nu = (\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)) : (\mu_1, \mu_2) \in \mathbb{R}^2, \mu_1 \neq \mu_2 \}$$

- Theorem 1:

$$\kappa_B(\nu) \geq \frac{2(\sigma_1 + \sigma_2)^2}{(\mu_1 - \mu_2)^2}$$

- A strategy allocating $t_1 = \left\lceil \frac{\sigma_1}{\sigma_1 + \sigma_2} t \right\rceil$ samples to arm 1 and $t_2 = t - t_1$ samples to arm 1, and recommending the empirical best satisfies

$$\liminf_{t \to \infty} -\frac{1}{t} \log p_t(\nu) \geq \frac{(\mu_1 - \mu_2)^2}{2(\sigma_1 + \sigma_2)^2}$$

$$\kappa_B(\nu) = \frac{2(\sigma_1 + \sigma_2)^2}{(\mu_1 - \mu_2)^2}$$
Gaussian Rewards: Fixed-confidence setting

The $\alpha$-Elimination algorithm with exploration rate $\beta(t, \delta)$

- chooses $A_t$ in order to keep a proportion $N_1(t)/t \simeq \alpha$
- if $\hat{\mu}_a(t)$ is the empirical mean of rewards obtained from $a$ up to time $t$, $\sigma_t^2(\alpha) = \sigma_1^2/\lceil \alpha t \rceil + \sigma_2^2/(t - \lceil \alpha t \rceil)$,

$$\tau = \inf \left\{ t \in \mathbb{N} : |\hat{\mu}_1(t) - \hat{\mu}_2(t)| > \sqrt{2\sigma_t^2(\alpha)\beta(t, \delta)} \right\}$$

- recommends the empirical best arm $\hat{a}_\tau = \arg\max_a \hat{\mu}_a(\tau)$
Gaussian Rewards: Fixed-confidence setting

- From Theorem 1:
  \[ \mathbb{E}_\nu[\tau] \geq \frac{2(\sigma_1 + \sigma_2)^2}{(\mu_1 - \mu_2)^2} \log \left( \frac{1}{2.4\delta} \right) \]

- \( \frac{\sigma_1}{\sigma_1 + \sigma_2} \)-Elimination with \( \beta(t, \delta) = \log \frac{t}{\delta} + 2 \log \log (6t) \) is \( \delta \)-PAC
  and

\[ \forall \epsilon > 0, \quad \mathbb{E}_\nu[\tau] \leq (1 + \epsilon) \frac{2(\sigma_1 + \sigma_2)^2}{(\mu_1 - \mu_2)^2} \log \left( \frac{1}{2.4\delta} \right) + o_{\delta \to 0} \left( \log \frac{1}{\delta} \right) \]

\[ \kappa_C(\nu) = \frac{2(\sigma_1 + \sigma_2)^2}{(\mu_1 - \mu_2)^2} \]
Gaussian Rewards: Conclusion

For any two fixed values of $\sigma_1$ and $\sigma_2$,

$$\kappa_B(\nu) = \kappa_C(\nu) = \frac{2(\sigma_1 + \sigma_2)^2}{(\mu_1 - \mu_2)^2}$$

If the variances are equal, $\sigma_1 = \sigma_2 = \sigma$,

$$\kappa_B(\nu) = \kappa_C(\nu) = \frac{8\sigma^2}{(\mu_1 - \mu_2)^2}$$

- uniform sampling is optimal only when $\sigma_1 = \sigma_2$
- $1/2$-Elimination is $\delta$-PAC for a smaller exploration rate $\beta(t, \delta) \simeq \log(\log(t)/\delta)$
\[ \mathcal{M} = \{ \nu = (\mathcal{B}(\mu_1), \mathcal{B}(\mu_2)) : (\mu_1, \mu_2) \in ]0; 1[^2, \mu_1 \neq \mu_2 \}, \]

shorthand: \( K(\mu, \mu') = \text{KL} (\mathcal{B}(\mu), \mathcal{B}(\mu')) \).

<table>
<thead>
<tr>
<th>Fixed-budget setting</th>
<th>Fixed-confidence setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>any consistent algorithm satisfies</td>
<td>any ( \delta )-PAC algorithm satisfies</td>
</tr>
<tr>
<td>( \limsup_{t \to \infty} -\frac{1}{t} \log p_t(\nu) \leq K^*(\mu_1, \mu_2) )</td>
<td>( \mathbb{E}_\nu[\tau] \geq \frac{1}{K^*(\mu_1, \mu_2)} \log \left( \frac{1}{2\delta} \right) )</td>
</tr>
<tr>
<td>(Chernoff information)</td>
<td></td>
</tr>
</tbody>
</table>

\( K^*(\mu_1, \mu_2) > K^*(\mu_1, \mu_2) \)
**Binary Rewards: Uniform Sampling**

<table>
<thead>
<tr>
<th>... algorithm</th>
<th>For any consistent...</th>
<th>For any ( \delta )-PAC...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( p_t(\nu) \gtrsim e^{-K^*(\mu_1,\mu_2)t} )</td>
<td>( \mathbb{E}_\nu[\tau] \log(1/\delta) \gtrsim \frac{1}{K^*(\mu_1,\mu_2)} )</td>
</tr>
<tr>
<td>... algorithm using uniform sampling</td>
<td>( p_t(\nu) \gtrsim e^{-\frac{K(\overline{\mu},\mu_1)+K(\overline{\mu},\mu_2)}{2}t} ) with ( \overline{\mu} = f(\mu_1,\mu_2) )</td>
<td>( \mathbb{E}_\nu[\tau] \log(1/\delta) \gtrsim \frac{2}{K(\mu_1,\mu)+K(\mu_2,\mu)} ) with ( \mu = \frac{\mu_1+\mu_2}{2} )</td>
</tr>
</tbody>
</table>

**Remark:** Quantities in the same column appear to be close from one another

\( \Rightarrow \) **Binary rewards: uniform sampling close to optimal**
### Binary Rewards: Uniform Sampling

<table>
<thead>
<tr>
<th>... algorithm</th>
<th>For any consistent...</th>
<th>For any $\delta$-PAC...</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_t(\nu) \sim e^{\eta K^*(\mu_1, \mu_2) t}$</td>
<td>$\mathbb{E}_\nu [\tau] / \log(1/\delta) \gtrsim \frac{1}{K^*(\mu_1, \mu_2)}$</td>
<td></td>
</tr>
</tbody>
</table>

| ... algorithm using uniform sampling | $p_t(\nu) \sim e^{-\frac{K(\bar{\mu}, \mu_1) + K(\bar{\mu}, \mu_2)}{2} t}$ with $\bar{\mu} = f(\mu_1, \mu_2)$ | $\mathbb{E}_\nu [\tau] / \log(1/\delta) \gtrsim \frac{2}{K(\mu_1, \underline{\mu}) + K(\mu_2, \underline{\mu})}$ with $\underline{\mu} = \frac{\mu_1 + \mu_2}{2}$ |

**Remark:** Quantities in the same column appear to be close from one another

$\Rightarrow$ **Binary rewards: uniform sampling close to optimal**
In fact, 

$$\kappa_B(\nu) = \frac{1}{K^*(\mu_1, \mu_2)}$$

The algorithm using uniform sampling and recommending the empirical best arm is very close to optimal
Binary Rewards: Fixed-Confidence Setting

\( \delta \)-PAC algorithms using uniform sampling satisfy

\[
\frac{\mathbb{E}_\nu [\tau]}{\log(1/\delta)} \geq \frac{1}{I_*(\nu)} \quad \text{with} \quad I_*(\nu) = \frac{K(\mu_1, \frac{\mu_1 + \mu_2}{2}) + K(\mu_2, \frac{\mu_1 + \mu_2}{2})}{2}.
\]

The algorithm using uniform sampling and

\[
\tau = \inf \left\{ t \in 2\mathbb{N}^* : |\hat{\mu}_1(t) - \hat{\mu}_2(t)| > \log \frac{\log(t) + 1}{\delta} \right\}
\]

is \( \delta \)-PAC but not optimal:

\[
\frac{\mathbb{E}[\tau]}{\log(1/\delta)} \approx \frac{2}{(\mu_1 - \mu_2)^2} > \frac{1}{I_*(\nu)}.
\]

A better stopping rule NOT based on the difference of empirical means

\[
\tau = \inf \left\{ t \in 2\mathbb{N}^* : t I_*(\hat{\mu}_1(t), \hat{\mu}_2(t)) > \log \frac{\log(t) + 1}{\delta} \right\}
\]
Regarding the complexities:

- $\kappa_B(\nu) = \frac{1}{K^*(\mu_1, \mu_2)}$
- $\kappa_C(\nu) \geq \frac{1}{K^*(\mu_1, \mu_2)} > \frac{1}{K^*(\mu_1, \mu_2)}$

Thus

$$\kappa_C(\nu) > \kappa_B(\nu)$$

Regarding the algorithms

- There is not much to gain by departing from uniform sampling
- In the fixed-confidence setting, a sequential test based on the difference of the empirical means is no longer optimal
Conclusion

- The complexities $\kappa_B(\nu)$ and $\kappa_C(\nu)$ are not always equal (and feature some different informational quantities).

- Strategies using random stopping do not necessarily lead to a saving in terms of the number of samples used.

- For Bernoulli distributions and Gaussian with similar variances, strategies using uniform sampling are (almost) optimal.

- Generalization to $m$ best arms identification among $K$ arms.
## Elements of Bibliography (see references therein!)


