Bandits for Exploration: Best Arm Identification and Discovery with Probabilistic Experts

Aurélien Garivier

Institut de Mathématiques de Toulouse

Imperial Probability Centre Multi-armed Bandits Session
October 31, 2014
Roadmap

1. Classical Bandits

2. Best arm identification in two-armed bandits
   - Lower bounds on the complexities
   - The complexity of A/B Testing with Gaussian feedback
   - The complexity of A/B Testing with binary feedback

3. Optimal Exploration with Probabilistic Expert Advice
   - Missing mass and Good-UCB
   - Analysis: Classical and Macroscopic Optimality
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. Examples: 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, Regret

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

Our goal Choose \( \pi \) so as to maximize

\[
\mathbb{E} [S_T] = \sum_{t=1}^{T} \sum_{a=1}^{K} \mathbb{E} \left[ \mathbb{E} [X_t \mathbb{1} \{A_t = a\}|X_1, \ldots, X_{t-1}] \right] \\
= \sum_{a=1}^{K} \mu_a \mathbb{E} \left[ N_{\pi}^a (T) \right]
\]

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 \( \mu_a = \mathbb{E} (\nu_a) \).

Regret Minimization equivalent to minimizing

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

where \( \mu^* \in \max\{\mu_a : 1 \leq a \leq K\} \)
A strategy $\pi$ is said to be **consistent** if, for any $(\nu_a)_a \in \mathcal{F}^K$, 

$$\frac{1}{T} \mathbb{E}[S_T] \to \mu^*$$

The strategy is efficient if for all $\theta \in [0, 1]^K$ and all $\alpha > 0$, 

$$R_T = o(T^\alpha)$$

There are efficient strategies and we consider the best achievable asymptotic performance among efficient strategies.
The Bound of Lai and Robbins

One-parameter reward distribution $\nu_a = \nu_{\theta_a}$, $\theta_a \in \Theta \subset \mathbb{R}$.

Theorem [Lai and Robbins, ’85]

If $\pi$ is an efficient strategy, then, for any $\theta \in \Theta^K$,

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

where $\text{KL}(\nu, \nu')$ denotes the Kullback-Leibler divergence.

For example, in the Bernoulli case:

$$\text{KL}(\tilde{B}(p), \tilde{B}(q)) = d_{\text{BER}}(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}$$
The Bound of Burnetas and Katehakis

More general reward distributions $\nu_a \in \mathcal{F}_a$

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

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

$$
\lim_{T \to \infty} \inf \frac{R_T}{\log(T)} \geq \sum_{a: \mu_a < \mu^*} \frac{\mu^* - \mu_a}{K_{\inf}(\nu_a, \mu^*)}
$$

where

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

Performance of UCB

For rewards in $[0, 1]$, the regret of UCB is upper-bounded as

$$E[R_T] = O(\log(T))$$

(finite-time regret bound) and

$$\limsup_{T \to \infty} \frac{E[R_T]}{\log(T)} \leq \sum_{a: \mu_a < \mu^*} \frac{1}{2(\mu^* - \mu_a)}$$

Yet, in the case of Bernoulli variables, the rhs. is greater than suggested by the bound by Lai & Robbins.

Many variants have been suggested to incorporate an estimate of the variance in the exploration bonus (e.g., [Audibert&al ’07])
The KL-UCB algorithm

joint work with O. Cappé, O-A. Maillard, R. Munos, G. Stoltz

**Parameters:** An operator $\Pi_F : M_1(S) \rightarrow \mathcal{F}$; 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

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

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

end for
```
Assume that $\mathcal{F}_a = \mathcal{F} = \text{canonical exponential family}$, i.e. such that the pdf of the rewards is given by

$$p_{\theta_a}(x) = \exp \left( x\theta_a - b(\theta_a) + c(x) \right), \quad 1 \leq a \leq K$$

for a parameter $\theta \in \mathbb{R}^K$, expectation $\mu_a = \dot{b}(\theta_a)$

$$U_a(t) = \sup \left\{ \mu \in \bar{I} : d(\hat{\mu}_a(t), \mu) \leq \frac{f(t)}{N_a(t)} \right\}$$

For instance,

- for Bernoulli rewards:

  $$d_{\text{BER}}(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}$$

- for exponential rewards $p_{\theta_a}(x) = \theta_a e^{-\theta_a x}$:

  $$d_{\text{exp}}(u, v) = u - v + u \log \frac{u}{v}$$

The analysis is generic and yields a non-asymptotic regret bound optimal in the sense of Lai and Robbins.
**Theorem:** Assume that all arms belong to a canonical, regular, exponential family \( \mathcal{F} = \{ \nu_\theta : \theta \in \Theta \} \) of probability distributions indexed by its natural parameter space \( \Theta \subseteq \mathbb{R} \). Then, with the choice \( f(t) = \log(t) + 3 \log \log(t) \) for \( t \geq 3 \), the number of draws of any suboptimal arm \( a \) is upper bounded for any horizon \( T \geq 3 \) as

\[
\mathbb{E} [N_a(T)] \leq \frac{\log(T)}{d(\mu_a, \mu^*)} + 2 \sqrt{\frac{2\pi \sigma^2_{a,*} \left( d'(\mu_a, \mu^*) \right)^2}{\left( d(\mu_a, \mu^*) \right)^3} \sqrt{\log(T) + 3 \log(\log(T))}} \\
+ \left( 4e + \frac{3}{d(\mu_a, \mu^*)} \right) \log(\log(T)) + 8\sigma^2_{a,*} \left( \frac{d'(\mu_a, \mu^*)}{d(\mu_a, \mu^*)} \right)^2 + 6,
\]

where \( \sigma^2_{a,*} = \max \{ \text{Var}(\nu_\theta) : \mu_a \leq E(\nu_\theta) \leq \mu^* \} \) and where \( d'(\cdot, \mu^*) \) denotes the derivative of \( d(\cdot, \mu^*) \).
Results: Two-Arm Scenario

Figure: Performance of various algorithms when $\theta = (0.9, 0.8)$. Left: average number of draws of the sub-optimal arm as a function of time. Right: box-and-whiskers plot for the number of draws of the sub-optimal arm at time $T = 5,000$. Results based on 50,000 independent replications.
Classical Bandits

Non-parametric setting

- Rewards are only assumed to be bounded (say in $[0, 1]$)

- Need for an estimation procedure
  - with non-asymptotic guarantees
  - efficient in the sense of Stein / Bahadur

$\Rightarrow$ Idea 1: use $d_{BER}$ (Hoeffding)

$\Rightarrow$ Idea 2: Empirical Likelihood [Owen ’01]

- Bad idea: use Bernstein / Bennett
First idea: use $d_{\text{BER}}$

Idea: rescale to $[0, 1]$, and take the divergence $d_{\text{BER}}$.

→ because Bernoulli distributions maximize deviations among bounded variables with given expectation:

**Lemma (Hoeffding ’63)**

Let $X$ denote a random variable such that $0 \leq X \leq 1$ and denote by $\mu = \mathbb{E}[X]$ its mean. Then, for any $\lambda \in \mathbb{R}$,

$$E[\exp(\lambda X)] \leq 1 - \mu + \mu \exp(\lambda) .$$

This fact is well-known for the variance, but also true for all exponential moments and thus for Cramer-type deviation bounds.
Regret Bound for kl-UCB

**Theorem**

With the divergence $d_{\text{BER}}$, for all $T > 3$,

$$
\mathbb{E}[N_a(T)] \leq \frac{\log(T)}{d_{\text{BER}}(\mu_a, \mu^*)} + \frac{\sqrt{2\pi} \log\left(\frac{\mu^*(1-\mu_a)}{\mu_a(1-\mu^*)}\right)}{(d_{\text{BER}}(\mu_a, \mu^*))^{3/2}} \sqrt{\log(T) + 3 \log(\log(T))} + \frac{2\left(\log\left(\frac{\mu^*(1-\mu_a)}{\mu_a(1-\mu^*)}\right)\right)^2}{(d_{\text{BER}}(\mu_a, \mu^*))^2} + 6.
$$

- kl-UCB satisfies an **improved logarithmic finite-time regret bound**
- Besides, it is **asymptotically optimal in the Bernoulli case**
Classical Bandits

Comparison to UCB

KL-UCB addresses exactly the same problem as UCB, with the same generality, but it has always a smaller regret as can be seen from Pinsker’s inequality

\[ d_{BER}(\mu_1, \mu_2) \geq 2(\mu_1 - \mu_2)^2 \]
Idea 2: Empirical Likelihood

\[ U(\hat{\nu}_n, \epsilon) = \sup \left\{ E(\nu') : \nu' \in \mathcal{M}_1(\text{Supp}(\hat{\nu}_n)) \text{ and } KL(\hat{\nu}_n, \nu') \leq \epsilon \right\} \]

or, rather, modified Empirical Likelihood:

\[ U(\hat{\nu}_n, \epsilon) = \sup \left\{ E(\nu') : \nu' \in \mathcal{M}_1(\text{Supp}(\hat{\nu}_n) \cup \{1\}) \text{ and } KL(\hat{\nu}_n, \nu') \leq \epsilon \right\} \]
Coverage properties of the modified EL confidence bound

**Proposition:** Let $\nu_0 \in \mathcal{M}_1([0, 1])$ with $E(\nu_0) \in (0, 1)$ and let $X_1, \ldots, X_n$ be independent random variables with common distribution $\nu_0 \in \mathcal{M}_1([0, 1])$, not necessarily with finite support. Then, for all $\epsilon > 0$,

$$\mathbb{P}\{U(\hat{\nu}_n, \epsilon) \leq E(\nu_0)\} \leq \mathbb{P}\{K_{inf}(\hat{\nu}_n, E(\nu_0)) \geq \epsilon\} \leq e(n + 2) \exp(-n\epsilon).$$

**Remark:** For $\{0, 1\}$—valued observations, it is readily seen that $U(\hat{\nu}_n, \epsilon)$ boils down to the upper-confidence bound above. This proposition is at least not always optimal: the presence of the factor $n$ in front of the exponential $\exp(-n\epsilon)$ term is questionable.
Regret bound

**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 $T \geq 3$:

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

- for each arm $1 \leq a \leq 6$ is associated with $\nu_a$, a Poisson distribution with expectation $(2 + a)/4$, truncated at 10.
- $N = 10,000$ Monte-Carlo replications on an horizon of $T = 20,000$ steps.
Take-home message on classical bandit algorithms

1. Use kl-UCB rather than UCB-1 or UCB-2

2. Use KL-UCB if speed is not a problem

3. todo: improve on the deviation bounds, address general non-parametric families of distributions

4. Alternative: Bayesian-flavored methods:
   - Bayes-UCB [Kaufmann, Cappé, G.]
   - Thompson sampling [Kaufmann & al.]
Roadmap

1. Classical Bandits

2. Best arm identification in two-armed bandits
   - Lower bounds on the complexities
   - The complexity of A/B Testing with Gaussian feedback
   - The complexity of A/B Testing with binary feedback

3. Optimal Exploration with Probabilistic Expert Advice
   - Missing mass and Good-UCB
   - Analysis: Classical and Macroscopic Optimality
Motivation

On the Complexity of Best Arm Identification in Multi-Armed Bandit Models, ArXiv (COLT 2014)
joint work with O. Cappé and E. Kaufmann
Our goal

Improve performance:

- fixed number of test users $\rightarrow$ smaller probability of error
- fixed probability of error $\rightarrow$ fewer test users

Tools: sequential allocation and stopping
Best arm identification in two-armed bandits

The model

A two-armed bandit model is

- a set \( \nu = (\nu_1, \nu_2) \) of two probability distributions (’arms’) with respective means \( \mu_1 \) and \( \mu_2 \)
- \( \alpha^* = \arg\max_{\alpha} \mu_\alpha \) is the (unknown) best arm

To find the best arm, an agent interacts with the bandit model with

- 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) \( \rightarrow \) a sample \( Z_t \sim \nu_{A_t} \) is observed
- a **stopping rule** \( \tau \) indicating when he stops sampling the arms
- a **recommendation rule** \( \hat{\alpha}_\tau \in \{1, 2\} \) 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.
Two possible goals

The agent’s goal is to design a strategy \( \mathcal{A} = ((A_t), \tau, \hat{a}_\tau) \) satisfying

<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>

An algorithm using **uniform sampling** is

<table>
<thead>
<tr>
<th>Fixed-budget setting</th>
<th>Fixed-confidence setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>a classical test of ((\mu_1 &gt; \mu_2)) against ((\mu_1 &lt; \mu_2)) based on (t) samples</td>
<td>a 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 \( A = ((A_t), \tau, \hat{a}_\tau) \) is...

<table>
<thead>
<tr>
<th><strong>Fixed-budget setting</strong></th>
<th><strong>Fixed-confidence setting</strong></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

\[
p_t(\nu) \approx \exp\left(-\frac{t}{C_H(\nu)}\right)
\]

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

\[
\mathbb{E}_\nu[\tau] \approx C' H'(\nu) \log \frac{1}{\delta}
\]

[Mannor Tsitsilis 04],[Even-Dar al. 06]
[Kalanakrishnan et al.12],...

Two complexities

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

for a probability of error \( \leq \delta \),

budget \( t \approx \kappa_B(\nu) \log \frac{1}{\delta} \)

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

for a probability of error \( \leq \delta \)

\[
\mathbb{E}_\nu[\tau] \approx \kappa_C(\nu) \log \frac{1}{\delta}
\]
Changes of distribution

New formulation for a change of distribution

Let \( \nu \) and \( \nu' \) be two bandit models. Let \( N_1 \) (resp. \( N_2 \)) denote the total number of draws of arm 1 (resp. arm 2) by algorithm \( A \). For any \( A \in \mathcal{F}_\tau \) such that \( 0 < P_\nu(A) < 1 \)

\[
\mathbb{E}_\nu[N_1] \text{KL}(\nu_1, \nu'_1) + \mathbb{E}_\nu[N_2] \text{KL}(\nu_2, \nu'_2) \geq d_{\text{BER}}(P_\nu(A), P_{\nu'}(A)),
\]

where \( d_{\text{BER}}(x, y) = x \log(x/y) + (1 - x) \log((1 - x)/(1 - y)) \).
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>$\mathbb{E}<em>{\nu}[\tau] \geq \frac{1}{K^*</em>*(\nu_1, \nu_2)} \log \left( \frac{1}{2\delta} \right)$</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>
Fixed-budget setting

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

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

- Theorem 1:

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

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

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

$$\kappa_B(\nu) = \frac{2(\sigma_1 + \sigma_2)^2}{(\mu_1 - \mu_2)^2}$$
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 $t$, $\sigma_t^2(\alpha) = \sigma_1^2/[\alpha t] + \sigma_2^2/(t - [\alpha t])$,

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

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

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

- Elimination with \( \beta(t, \delta) = \log \frac{t}{\delta} + 2\log \log(6t) \) is \( \delta \)-PAC and
  \[ \forall \epsilon > 0, \quad \mathbb{E}_\nu[\tau] \leq \left( 1 + \epsilon \right) \frac{2(\sigma_1 + \sigma_2)^2}{(\mu_1 - \mu_2)^2} \log \left( \frac{1}{2\delta} \right) + o_{\epsilon, \delta} \log \left( \frac{1}{\delta} \right) \]

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

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 $\beta(t, \delta) \approx \log(\log(t)/\delta)$
Lower bounds for Bernoulli bandit models

\[ \mathcal{M} = \{ \nu = (B(\mu_1), B(\mu_2)) : (\mu_1, \mu_2) \in [0; 1]^2, \mu_1 \neq \mu_2 \}, \]

shorthand: \( K(\mu, \mu') = KL(B(\mu), B(\mu')) \).

<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>

\[ K^*(\mu_1, \mu_2) > K^*(\mu_1, \mu_2) \]
## Algorithms using uniform sampling

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

⇒ **Binary rewards:** uniform sampling close to optimal
**Algorithms using uniform sampling**

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

We show that

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

(matching algorithm not implementable in practice)

The algorithm using uniform sampling and recommending the empirical best arm is preferable (and very close to optimal)
Fixed-confidence setting

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

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

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

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

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 on Best Arm Identification

- The complexities $\kappa_B(\nu)$ and $\kappa_C(\nu)$ are not always equal (and feature some different informational quantities).
- For Bernoulli distributions and Gaussian with similar variances, strategies using uniform sampling are (almost) optimal.
- Strategies using random stopping do not necessarily lead to a saving in terms of the number of samples used.
- Generalization to $m$ best arms identification among $K$ arms.
Optimal Exploration with Probabilistic Expert Advice

Roadmap

1. Classical Bandits

2. Best arm identification in two-armed bandits
   - Lower bounds on the complexities
   - The complexity of A/B Testing with Gaussian feedback
   - The complexity of A/B Testing with binary feedback

3. Optimal Exploration with Probabilistic Expert Advice
   - Missing mass and Good-UCB
   - Analysis: Classical and Macroscopic Optimality
The model


joint work with S. Bubeck and D. Ernst

- Subset $A \subset \mathcal{X}$ of important items
- $|\mathcal{X}| \gg 1$, $|A| \ll |\mathcal{X}|$
- Access to $\mathcal{X}$ only by probabilistic experts $(P_i)_{1 \leq i \leq K}$: sequential independent draws

**Goal:** discover rapidly the elements of $A$
The model

**Optimal Discovery with Probabilistic Expert Advice**: Finite Time Analysis and Macroscopic Optimality, JMLR 2013
joint work with S. Bubeck and D. Ernst

- Subset $A \subset \mathcal{X}$ of important items
- $|\mathcal{X}| \gg 1$, $|A| \ll |\mathcal{X}|$
- Access to $\mathcal{X}$ only by probabilistic experts $(P_i)_{1 \leq i \leq K}$: sequential independent draws

Goal: discover rapidly the elements of $A$
The model

*Optimal Discovery with Probabilistic Expert Advice: Finite Time Analysis and Macroscopic Optimality*, JMLR 2013
joint work with S. Bubeck and D. Ernst

- Subset $A \subset X$ of important items
- $|X| \gg 1$, $|A| \ll |X|$  
- Access to $X$ only by probabilistic experts $(P_i)_{1 \leq i \leq K}$: sequential independent draws

**Goal:** discover rapidly the elements of $A$
The model

Optimal Discovery with Probabilistic Expert Advice: Finite Time Analysis and Macroscopic Optimality, JMLR 2013
joint work with S. Bubeck and D. Ernst

- Subset $A \subset \mathcal{X}$ of important items
- $|\mathcal{X}| \gg 1$, $|A| \ll |\mathcal{X}|$
- Access to $\mathcal{X}$ only by probabilistic experts $(P_i)_{1 \leq i \leq K}$: sequential independent draws

Goal: discover rapidly the elements of $A$
Search space: \( A \subset \Omega \) discrete set

Probabilistic experts: \( P_i \in \mathcal{M}_1(\Omega) \) for \( i \in \{1, \ldots, K\} \)

Requests: at time \( t \), calling expert \( I_t \) yields a realization of \( X_t = X_{I_t,t} \) independent with law \( P_a \)

Goal: find as many distinct elements of \( A \) as possible with few requests:

\[
F_n = \text{Card} \ (A \cap \{X_1, \ldots, X_n\})
\]
Goal

At each time step $t = 1, 2, \ldots$:

- pick an index $I_t = \pi_t(I_1, Y_1, \ldots, I_{s-1}, Y_{s-1}) \in \{1, \ldots, K\}$ according to past observations
- observe $Y_t = X_{I_t, n_{I_t,t}} \sim P_{I_t}$, where

$$n_{i,t} = \sum_{s \leq t} \mathbb{1}\{I_s = i\}$$

**Goal:** design the strategy $\pi = (\pi_t)_t$ so as to maximize the number of important items found after $t$ requests

$$F^{\pi}(t) = |A \cap \{Y_1, \ldots, Y_t\}|$$

**Assumption:** non-intersecting supports

$$A \cap \text{supp}(P_i) \cap \text{supp}(P_j) = \emptyset \text{ for } i \neq j$$
Is it a Bandit Problem?

It looks like a bandit problem...

- sequential choices among K options
- want to maximize cumulative rewards
- exploration vs exploitation dilemma

... but it is **not a bandit problem**!

- rewards are not i.i.d.
- **destructive rewards**: no interest to observe twice the same important item
- all strategies eventually equivalent
The oracle strategy

**Proposition:** Under the non-intersecting support hypothesis, the greedy oracle strategy selecting the expert with highest ‘missing mass’

\[ I_t^* \in \arg \max_{1 \leq i \leq K} P_i (A \setminus \{Y_1, \ldots, Y_t\}) \]

is optimal: for every possible strategy \( \pi \), \( \mathbb{E}[F^\pi(t)] \leq \mathbb{E}[F^*(t)] \).

**Remark:** the proposition if false if the supports may intersect

\[ \implies \text{estimate the “missing mass of important items”!} \]
Missing mass estimation

Let us first focus on one expert \( i \): \( P = P_i, X_n = X_{i,n} \)

\( X_1, \ldots, X_n \) independent draws of \( P \)

\[
O_n(x) = \sum_{m=1}^{n} \mathbb{1}\{X_m = x\}
\]

How to ’estimate’ the total mass of the unseen important items

\[
R_n = \sum_{x \in A} P(x) \mathbb{1}\{O_n(x) = 0\}
\]
The Good-Turing Estimator

Idea: use the **hapax**es = items seen only once (linguistic)

\[ \hat{R}_n = \frac{U_n}{n}, \quad \text{where} \quad U_n = \sum_{x \in A} 1\{O_n(x) = 1\} \]

**Lemma [Good ’53]:** For every distribution \( P \),

\[ 0 \leq \mathbb{E}[\hat{R}_n] - \mathbb{E}[R_n] \leq \frac{1}{n} \]

**Proposition:** With probability at least \( 1 - \delta \) for every \( P \),

\[ \hat{R}_n - \frac{1}{n} - (1 + \sqrt{2}) \sqrt{\log(4/\delta) \over n} \leq R_n \leq \hat{R}_n + (1 + \sqrt{2}) \sqrt{\log(4/\delta) \over n} \]

See [McAllester and Schapire ’00, McAllester and Ortiz ’03]:
- deviations of \( \hat{R}_n \): McDiarmid’s inequality
- deviations of \( R_n \): negative association
The Good-UCB algorithm [Bubeck, Ernst & G.]

Optimistic algorithm based on Good-Turing’s estimator:

\[
I_{t+1} = \arg \max_{i \in \{1, \ldots, K\}} \left\{ \frac{H_i(t)}{N_i(t)} + c \sqrt{\frac{\log(t)}{N_i(t)}} \right\}
\]

- \( N_i(t) = \) number of draws of \( P_i \) up to time \( t \)
- \( H_i(t) = \) number of elements of \( A \) seen exactly once thanks to \( P_i \)
- \( c = \) tuning parameter
**Theorem:** For any \( t \geq 1 \), under the non-intersecting support assumption, Good-UCB (with constant \( C = (1 + \sqrt{2})\sqrt{3} \)) satisfies

\[
\mathbb{E}[ F^*(t) - F^{UCB}(t) ] \leq 17\sqrt{Kt\log(t)} + 20\sqrt{Kt} + K + K\log(t/K)
\]

**Remark:** Usual result for bandit problem, but not-so-simple analysis
A Typical Run of Good-UCB
The macroscopic limit

- **Restricted framework:** $P_i = \mathcal{U}\{1, \ldots, N\}$
- $N \to \infty$
- $|A \cap \text{supp}(P_i)|/N \to q_i \in (0, 1)$, $q = \sum_i q_i$
The macroscopic limit

- **Restricted framework:** $P_i = \mathcal{U}\{1, \ldots, N\}$
- $N \to \infty$
- $|A \cap \text{supp}(P_i)|/N \to q_i \in (0, 1), \quad q = \sum_i q_i$
The macroscopic limit

- **Restricted framework:** \( P_i = \mathcal{U}\{1, \ldots, N\} \)
- \( N \to \infty \)
- \( |A \cap \text{supp}(P_i)|/N \to q_i \in (0, 1), \ q = \sum_i q_i \)
The limiting discovery process of the Oracle strategy is \textit{deterministic}.

**Proposition:** For every $\lambda \in (0, q_1)$, for every sequence $(\lambda^N)_N$ converging to $\lambda$ as $N$ goes to infinity, almost surely

$$\lim_{N \to \infty} \frac{T^*_N(\lambda^N)}{N} = \sum_i \left( \log \frac{q_i}{\lambda} \right)_+$$
Oracle vs. uniform sampling

**Oracle:** The proportion of important items not found after $Nt$ draws tends to

$$q - F^*(t) = I(t)q_{I(t)} \exp(-t/I(t)) \leq K \bar{q}_K \exp(-t/K)$$

with $\bar{q}_K = \left(\prod_{i=1}^{K} q_i\right)^{1/K}$ the geometric mean of the $(q_i)_i$.

**Uniform:** The proportion of important items not found after $Nt$ draws tends to $K \bar{q}_K \exp(-t/K)$

$\implies$ Asymptotic ratio of efficiency

$$\rho(q) = \frac{\bar{q}_K}{q_{\underline{K}}} = \frac{1}{K} \frac{\sum_{i=1}^{k} q_i}{\left(\prod_{i=1}^{k} q_i\right)^{1/K}} \geq 1$$

larger if the $(q_i)_i$ are unbalanced
Macroscopic optimality

**Theorem:** Take $C = (1 + \sqrt{2}) \sqrt{c + 2}$ with $c > 3/2$ in the Good-UCB algorithm.

- For every sequence $(\lambda^N)_N$ converging to $\lambda$ as $N$ goes to infinity, almost surely

  $$\limsup_{N \to +\infty} \frac{T_{UCB}^N(\lambda^N)}{N} \leq \sum_i \left( \log \frac{q_i}{\lambda} \right)_+$$

- The proportion of items found after $Nt$ steps $F^{GUCB}$ converges *uniformly* to $F^*$ as $N$ goes to infinity.
Simulation

Number of items found by Good-UCB (line), the oracle (bold dashed), and by uniform sampling (light dotted) as a function of time, for sample sizes $N = 128, N = 500$, $N = 1000$ et $N = 10000$, in an environment with 7 experts.