Revisiting the Exploration-Exploitation Tradeoff in Bandit Models

Emilie Kaufmann

joint work with Aurélien Garivier (IMT, Toulouse) and Tor Lattimore (University of Alberta)

Workshop on Optimization and Decision-Making in Uncertainty, Simons Institute, Berkeley, September 21st, 2016
The multi-armed bandit model

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

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, \( a^* = \arg \max_a \mu_a \)
of mean \( \mu^* = \max_a \mu_a \)
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:

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

\( N_a(T) \): number of draws of arm \( a \) up to time \( T \)

\( \Rightarrow \) Exploration/Exploitation tradeoff
Algorithms: naive ideas

- **Idea 1**: Choose each arm $T/K$ times
  \[ \Rightarrow \text{EXPLORATION} \]

- **Idea 2**: Always choose the best arm so far
  \[ A_{t+1} = \arg\max_a \hat{\mu}_a(t) \]
  \[ \Rightarrow \text{EXPLOITATION} \]

...Linear regret
Idea 1: Choose each arm $T/K$ times

$\Rightarrow$ EXPLORATION

Idea 2: Always choose the best arm so far

$$A_{t+1} = \arg\max_a \hat{\mu}_a(t)$$

$\Rightarrow$ EXPLOITATION

...Linear regret

A better idea:
First explore the arms uniformly,
then commit to the empirical best until the end

$\Rightarrow$ EXPLORATION followed by EXPLOITATION

...Still sub-optimal
A motivation: should we minimize regret?

\[ B(\mu_1) \quad B(\mu_2) \quad B(\mu_3) \quad B(\mu_4) \quad B(\mu_5) \]

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
A motivation: should we minimize regret?

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 different objectives

<table>
<thead>
<tr>
<th>Bandit algorithm</th>
<th>Regret minimization</th>
<th>Best arm identification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sampling rule ((A_t))</td>
<td>sampling rule ((A_t))</td>
</tr>
<tr>
<td>Input</td>
<td>horizon (T)</td>
<td>stopping rule (\tau)</td>
</tr>
<tr>
<td>Objective</td>
<td>minimize (R_T = \mu^*T - \mathbb{E}\left[\sum_{t=1}^{T}X_t\right])</td>
<td>recommendation rule (\hat{a}_\tau)</td>
</tr>
<tr>
<td>Exploration/Exploitation</td>
<td>and minimize (\mathbb{E}[\tau])</td>
<td>pure Exploration</td>
</tr>
</tbody>
</table>

This talk:
- (distribution-dependent) optimal algorithm for both objectives
- best performance of an Explore-Then-Commit strategy?

We focus on distributions parameterized by their means
\[\mu = (\mu_1, \ldots, \mu_K)\]
(Bernoulli, Gaussian)
Outline

1. Optimal algorithms for Regret Minimization

2. Optimal algorithms for Best Arm Identification

3. Explore-Then-Commit strategies
Optimal algorithms for regret minimization

\( \mu = (\mu_1, \ldots, \mu_K) \). \( N_a(t) \): number of draws of arm \( a \) up to time \( t \)

\[
R_\mu(A, T) = \sum_{a=1}^{K} (\mu^* - \mu_a) \mathbb{E}_\mu[N_a(T)]
\]

Notation: Kullback-Leibler divergence

\[
d(\mu, \mu') := \text{KL}(\nu_\mu, \nu_{\mu'})
\]

(Gaussian): \( d(\mu, \mu') = \frac{(\mu - \mu')^2}{2\sigma^2} \)

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

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

A bandit algorithm is **asymptotically optimal** if, for every \( \mu \),

\[
\mu_a < \mu^* \Rightarrow \lim_{T \to \infty} \frac{\mathbb{E}_\mu[N_a(T)]}{\log T} \leq \frac{1}{d(\mu_a, \mu^*)}
\]
Optimal algorithms for regret minimization

\[ \mu = (\mu_1, \ldots, \mu_K). \]  
\[ N_a(t) : \text{number of draws of arm } a \text{ up to time } t \]

\[ R_\mu(A, T) = \sum_{a=1}^{K} (\mu^* - \mu_a) \mathbb{E}_\mu[N_a(T)] \]

**Notation:** Kullback-Leibler divergence

\[ d(\mu, \mu') := \text{KL}(\nu_\mu, \nu_{\mu'}) \]

(Bernoulli): \[ d(\mu, \mu') = \mu \log \frac{\mu}{\mu'} + (1 - \mu) \log \frac{1 - \mu}{1 - \mu'} \]

[Laï and Robbins, 1985]: for uniformly efficient algorithms,

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

A bandit algorithm is **asymptotically optimal** if, for every \( \mu \),

\[ \mu_a < \mu^* \Rightarrow \lim_{T \to \infty} \sup \frac{\mathbb{E}_\mu[N_a(T)]}{\log T} \leq \frac{1}{d(\mu_a, \mu^*)} \]
A UCB-type (or optimistic) algorithm chooses at round $t$

$$A_{t+1} = \arg\max_{a=1...K} UCB_a(t).$$

where $UCB_a(t)$ is an Upper Confidence Bound on $\mu_a$.

The KL-UCB index

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

satisfies $\mathbb{P}(\mu_a \leq UCB_a(t)) \gtrsim 1 - t^{-1}$. 
A UCB-type (or optimistic) algorithm chooses at round $t$

$$A_{t+1} = \arg\max_{a=1\ldots K} \text{UCB}_a(t).$$

where $\text{UCB}_a(t)$ is an Upper Confidence Bound on $\mu_a$.

The KL-UCB index [Cappé et al. 13]: KL-UCB satisfies

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

2. Optimal algorithms for Best Arm Identification

3. Explore-Then-Commit strategies
A sample complexity lower bound

A Best Arm Identification algorithm \((A_t, \tau, \hat{a}_\tau)\) is \(\delta\)-PAC if
\[
\forall \mu, \quad P_\mu(\hat{a}_\tau = a^*(\mu)) \geq 1 - \delta.
\]

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

For any \(\delta\)-PAC algorithm,
\[
E_\mu[\tau] \geq T^*(\mu) \log \left(\frac{1}{2.4\delta}\right),
\]
where
\[
T^*(\mu)^{-1} = \sup_{w \in \Sigma_K} \inf_{\lambda \in \text{Alt}(\mu)} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a)
\]
\[
\Sigma_K = \{w \in [0, 1]^K : \sum_{i=1}^{K} w_i = 1\}, \quad \text{Alt}(\mu) = \{\lambda : a^*(\lambda) \neq a^*(\mu)\}
\]

Moreover, the vector of optimal proportions,
\[
\left(\frac{E_\mu[N_a(\tau)]}{E_\mu[\tau]} \simeq w^*_a(\mu)\right)
\]

\[
w^*(\mu) = \arg\max_{w \in \Sigma_K} \inf_{\lambda \in \text{Alt}(\mu)} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a)
\]
is well-defined, and we propose an efficient way to compute it.
Sampling rule: Tracking the optimal proportions

\[ \hat{\mu}(t) = (\hat{\mu}_1(t), \ldots, \hat{\mu}_K(t)) : \text{vector of empirical means} \]

- Introducing

\[ U_t = \{a : N_a(t) < \sqrt{t}\}, \]

the arm sampled at round \( t + 1 \) is

\[ A_{t+1} \in \begin{cases} 
\text{argmin}_{a \in U_t} N_a(t) \text{ if } U_t \neq \emptyset & \text{(forced exploration)} \\
\text{argmax}_{1 \leq a \leq K} \left[ t w^*_a(\hat{\mu}(t)) - N_a(t) \right] & \text{(tracking)}
\end{cases} \]

**Lemma**

Under the Tracking sampling rule,

\[ \mathbb{P}_\mu \left( \lim_{t \to \infty} \frac{N_a(t)}{t} = w^*_a(\mu) \right) = 1. \]
The Track-and-Stop strategy, that uses

- the **Tracking** sampling rule
- a stopping rule based on GLRT tests:
  \[ \tau_\delta = \inf \{ t \in \mathbb{N} : Z(t) > \log \frac{2Kt}{\delta} \}, \]
  with
  \[ Z(t) := t \times \sup_{\lambda \in \text{Alt}(\hat{\mu}(t))} \left[ \sum_{a=1}^{K} \frac{N_a(t)}{t} d(\hat{\mu}_a(t), \lambda_a) \right] \]
  and recommends \( \hat{a}_\tau = \arg\max_{a=1...K} \hat{\mu}_a(\tau) \)

is \( \delta \)-PAC for every \( \delta \in ]0, 1[ \) and satisfies
\[
\limsup_{\delta \to 0} \frac{\mathbb{E}_{\mu}[\tau_\delta]}{\log(1/\delta)} = T^*(\mu).
\]
Regret minimization versus Best Arm Identification

Algorithms for regret minimization and BAI are very different!

- playing mostly the best arm vs. optimal proportions

- different “complexity terms” (featuring KL-divergence)

\[ R_T(\mu) \approx \left( \sum_{a \neq a^*} \frac{\mu^* - \mu_a}{d(\mu_a, \mu^*)} \right) \log(T) \]

\[ \mathbb{E}_\mu[\tau] \approx T^*(\mu) \log \left( \frac{1}{\delta} \right) \]
Outline

1. Optimal algorithms for Regret Minimization
2. Optimal algorithms for Best Arm Identification
3. Explore-Then-Commit strategies
Gaussian two-armed bandits

\[ \nu_1 = \mathcal{N}(\mu_1, 1) \] and \[ \nu_2 = \mathcal{N}(\mu_2, 1) \]. \[ \mu = (\mu_1, \mu_2) \].

Let \[ \Delta = |\mu_1 - \mu_2| \]

**Regret minimization**

For any uniformly efficient algorithm \( \mathcal{A} = (A_t) \),

\[
\lim_{T \to \infty} \inf \frac{R_\mu(T, A)}{\log(T)} \geq \frac{2}{\Delta}
\]

u.e.:
\[
\forall \mu, \forall \alpha \in ]0, 1[, R_\mu(T, A) = o(T^\alpha)
\]

**Best Arm Identification**

For any \( \delta \)-PAC algorithm \( \mathcal{A} = (A_t, \tau, \hat{a}_\tau) \),

\[
\lim_{\delta \to 0} \inf \frac{\mathbb{E}_\mu[\tau_\delta]}{\log(1/\delta)} \geq \frac{8}{\Delta^2}
\]

(optimal algorithms use uniform sampling)
ETC strategies: given a stopping rule $\tau$ and a commit rule $\hat{a}$,

$$A_t = \begin{cases} 
1 & \text{if } t \leq \tau \text{ and } t \text{ is odd}, \\
2 & \text{if } t \leq \tau \text{ and } t \text{ is even}, \\
\hat{a} & \text{otherwise}.
\end{cases}$$

Assume $\mu_1 > \mu_2$.

$$R_{\mu} (T, A^{ETC}) = \Delta \mathbb{E}_\mu [N_2(T)]$$

$$= \Delta \mathbb{E}_\mu \left[ \frac{\tau \wedge T}{2} + (T - \tau) + 1(\hat{a} = 2) \right]$$

$$\leq \frac{\Delta}{2} \mathbb{E}_\mu [\tau] + T \Delta \mathbb{P}_\mu (\hat{a} = 2).$$
Explore-Then-Commit (ETC) strategies

ETC strategies: given a stopping rule \( \tau \) and a commit rule \( \hat{a} \),
\[
A_t = \begin{cases} 
1 & \text{if } t \leq \tau \text{ and } t \text{ is odd} , \\
2 & \text{if } t \leq \tau \text{ and } t \text{ is even} , \\
\hat{a} & \text{otherwise}. 
\end{cases}
\]

Assume \( \mu_1 > \mu_2 \).

For \( A = (\tau, \hat{a}) \) as in an optimal BAI algorithm with \( \delta = \frac{1}{T} \)
\[
R_\mu (T, A^{\text{ETC}}) = \Delta \mathbb{E}_\mu [N_2 (T)] \\
= \Delta \mathbb{E}_\mu \left[ \frac{\tau \land T}{2} + (T - \tau)_+ 1_{(\hat{a} = 2)} \right] \\
\leq \frac{\Delta}{2} \mathbb{E}_\mu [\tau] + T \Delta \mathbb{P}_\mu (\hat{a} = 2).
\]

Hence
\[
\lim \sup \frac{R_\mu (T, A)}{\log T} \leq \frac{4}{\Delta}.
\]
Lemma

Let $\mu, \lambda : a^*(\mu) \neq a^*(\lambda)$

Let $\sigma$ s.t. $N_2(T)$ is $\mathcal{F}_\sigma$-measurable. For any u.e. algorithm,

$$\liminf_{T \to \infty} \frac{\mathbb{E}_\mu[N_1(\sigma)](\lambda_1 - \mu_1)^2}{2} + \frac{\mathbb{E}_\mu[N_2(\sigma)](\lambda_2 - \mu_2)^2}{2} \log(T) \geq 1.$$

Proof. Introducing the log-likelihood ratio

$$L_t(\mu, \lambda) = \log \frac{p_\mu(X_1, \ldots, X_t)}{p_\lambda(X_1, \ldots, X_t)},$$

one needs to prove that $\liminf_{T \to \infty} \frac{\mathbb{E}_\mu[L_\sigma(\mu, \lambda)]}{\log(T)} \geq 1$.

$$\mathbb{E}_\mu[L_\sigma(\mu, \lambda)] = \text{KL} \left( \mathcal{L}_\mu(X_1, \ldots, X_\sigma), \mathcal{L}_\lambda(X_1, \ldots, X_\sigma) \right) \geq \text{kl}(\mathbb{E}_\mu[Z], \mathbb{E}_\lambda[Z]) \text{ for any } Z \in [0, 1], \mathcal{F}_\sigma\text{-mesurable}$$

[Garivier et al. 16]

$$\mathbb{E}_\mu[L_\sigma(\mu, \lambda)] \geq \text{kl}(\mathbb{E}_\mu[N_2(T)/T], \mathbb{E}_\lambda[N_2(T)/T]) \sim \log(T) \quad (u.e.)$$
Is this the best we can do? Lower bounds.

**Lemma**

Let \( \mu, \lambda : a^*(\mu) \neq a^*(\lambda) \).
Let \( \sigma \) s.t. \( N_2(T) \) is \( \mathcal{F}_\sigma \)-measurable. For any u.e. algorithm,

\[
\liminf_{T \to \infty} \frac{\mathbb{E}_\mu[N_1(\sigma)] \frac{(\lambda_1 - \mu_1)^2}{2} + \mathbb{E}_\mu[N_2(\sigma)] \frac{(\lambda_2 - \mu_2)^2}{2}}{\log(T)} \geq 1.
\]

Assume \( \mu_1 > \mu_2 \):

- **Lai and Robbins’ bound:**
  \[
  \lambda_1 = \mu_1, \quad \lambda_2 = \mu_1 + \epsilon \\
  \sigma = T \\
  \Rightarrow \quad \liminf_{T \to \infty} \frac{\mathbb{E}_\mu[N_2(T)] \frac{(\Delta + \epsilon)^2}{2}}{\log(T)} \geq 1.
  \]

- **For ETC strategies:**
  \[
  \lambda_1 = \frac{\mu_1 + \mu_2 - \epsilon}{2}, \quad \lambda_2 = \frac{\mu_1 + \mu_2 + \epsilon}{2} \\
  \sigma = \tau \wedge T \\
  \Rightarrow \quad \liminf_{T \to \infty} \frac{\mathbb{E}_\mu[\tau \wedge T] \frac{(\Delta + \epsilon)^2}{8}}{\log(T)} \geq 1
  \]
  \[
  \Rightarrow \quad \liminf_{T \to \infty} \frac{R_{\mu}(T, A)}{\log(T)} \geq \frac{4}{\Delta}.
  \]
An interesting matching algorithm

Theorem

Any uniformly efficient ETC strategy satisfies

\[
\lim_{T \to \infty} \inf \frac{R_\mu(T, A)}{\log(T)} \geq \frac{4}{\Delta}.
\]

The ETC strategy based on the stopping rule

\[
\tau = \inf \left\{ t = 2n : |\hat{\mu}_{1,n} - \hat{\mu}_{2,n}| > \sqrt{\frac{4 \log \left( \frac{T}{2n} \right)}{n}} \right\}
\]

satisfies, for \( T \Delta^2 > 4e^2 \),

\[
R_\mu(T, A) \leq 4 \log \left( \frac{T \Delta^2}{4} \right) + \frac{334}{\Delta} \sqrt{\log \left( \frac{T \Delta^2}{4} \right)} + \frac{178}{\Delta} + \Delta,
\]

\[
R_\mu(T, A) \leq 32 \sqrt{T} + \Delta.
\]
Conclusion

In Gaussian two-armed bandits, ETC strategies are sub-optimal by a factor two compared to UCB strategies

⇒ rather than A/B Test + always showing the best product, dynamically present products to customers all day long!

On-going work:

- how does Optimal BAI + Commit behave in general?

\[ T^*(\mu) \left( \sum_{a=2}^{K} w_a^*(\mu)(\mu_1 - \mu_a) \right) \quad \text{v.s} \quad \sum_{a=2}^{K} \frac{\mu_1 - \mu_a}{d(\mu_a, \mu_1)}. \]