Optimal Best Arm Identification
with Fixed Confidence

Emilie Kaufmann,

joint work with Aurélien Garivier
(Institut de Mathématiques de Toulouse)

Workshop on Computational and Statistical Trade-offs in Learning,
March 22nd, 2016
The stochastic multi-armed bandit model (MAB)

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

At round \( t \), an agent:

- chooses an arm \( A_t \in A := \{1, \ldots, K\} \)
- observes a sample \( X_t \sim \nu_{A_t} \)

using a sequential sampling strategy (\( A_t \)):

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

aimed for a prescribed objective, e.g. related to learning

\[ a^* = \arg\max_a \mu_a \] and \[ \mu^* = \max_a \mu_a. \]
A possible objective: Regret minimization

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 regret:

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

⇒ exploration/exploitation tradeoff

**Motivation:** clinical trials [1933]

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

**Goal:** maximize the number of patients healed during the trial
A possible objective: Regret minimization

Samples = \textit{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 \textit{regret}:

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

\( \Rightarrow \) exploration/exploitation tradeoff

**Motivation**: clinical trials [1933]

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

Goal: maximize the number of patients healed during the trial

Alternative goal: identify as quickly as possible the best treatment
Our objective: Best-arm identification

Goal: identify the best arm, $a^*$, as fast/accurately as possible.
No incentive to draw arms with high means!

⇒ optimal exploration

The agent’s strategy is made of:
- a sequential sampling strategy ($A_t$)
- a stopping rule $\tau$ (stopping time)
- a recommendation rule $\hat{a}_\tau$

Possible goals:

<table>
<thead>
<tr>
<th>Fixed-budget setting</th>
<th>Fixed-confidence setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau = T$</td>
<td>minimize $\mathbb{E}[\tau]$</td>
</tr>
<tr>
<td>minimize $\mathbb{P}(\hat{a}_\tau \neq a^*)$</td>
<td>$\mathbb{P}(\hat{a}_\tau \neq a^*) \leq \delta$</td>
</tr>
</tbody>
</table>

Motivation: Market research, A/B Testing, clinical trials...
Our objective: Best-arm identification

**Goal**: identify the best arm, $a^*$, as fast/accurately as possible. No incentive to draw arms with high means!

⇒ **optimal exploration**

The agent’s strategy is made of:

- a sequential **sampling strategy** ($A_t$)
- a **stopping rule** $\tau$ (stopping time)
- a **recommendation rule** $\hat{a}_\tau$

Possible goals:

<table>
<thead>
<tr>
<th>Fixed-budget setting</th>
<th>Fixed-confidence setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau = T$</td>
<td>minimize $\mathbb{E}[\tau]$</td>
</tr>
<tr>
<td>minimize $\mathbb{P}(\hat{a}_\tau \neq a^*)$</td>
<td>$\mathbb{P}(\hat{a}_\tau \neq a^*) \leq \delta$</td>
</tr>
</tbody>
</table>

**Motivation**: Market research, A/B Testing, clinical trials...
A class of bandit models $\nu = (\nu_1, \ldots, \nu_K)$.

A strategy is $\delta$-PAC on $S$ is $\forall \nu \in S, \mathbb{P}_\nu(\hat{\tau} = a^*) \geq 1 - \delta$.

Goal: for some classes $S$, and $\nu \in S$, find

- a lower bound on $\mathbb{E}_\nu[\tau]$ for any $\delta$-PAC strategy
- a $\delta$-PAC strategy such that $\mathbb{E}_\nu[\tau]$ matches this bound

(distribution-dependent bounds)
Exponential family bandit models

\( \nu_1, \ldots, \nu_K \) belong to a one-dimensional exponential family:

\[ \mathcal{P}_{\lambda, \Theta, \theta} = \{ \nu_\theta, \theta \in \Theta : \nu_\theta \text{ has density } f_\theta(x) = \exp(\theta x - b(\theta)) \text{ w.r.t. } \lambda \} \]

**Example:** Gaussian, Bernoulli, Poisson distributions...

\( \nu_\theta \) can be parametrized by its mean \( \mu = \omega b(\theta) : \nu^\mu := \nu_{b^{-1}(\mu)} \)

**Notation:** Kullback-Leibler divergence

For a given exponential family \( \mathcal{P} \),

\[ d_\mathcal{P}(\mu, \mu') := KL(\nu^\mu, \nu^{\mu'}) = \mathbb{E}_{X \sim \nu^\mu} \left[ \log \frac{d
u^\mu}{d
u^{\mu'}(X)} \right] \]

is the KL-divergence between the distributions of mean \( \mu \) and \( \mu' \).

**Example:** Bernoulli distributions

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

We identify \( \nu = (\nu^{\mu_1}, \ldots, \nu^{\mu_K}) \) and \( \mu = (\mu_1, \ldots, \mu_K) \) and consider

\[ S = \left\{ \mu \in (\omega b(\Theta))^K : \exists a \in A : \mu_a > \max_{i \neq a} \mu_i \right\} \]
Outline

1. Regret minimization

2. Sample complexity lower bounds
   - Tools and a first lower bound
   - Characteristic time and optimal proportions of draws

3. The Track-and-Stop Strategy
   - The Tracking Sampling rule
   - The Chernoff Stopping Rule
   - Asymptotic optimality

4. Practical performance
Outline

1. Regret minimization

2. Sample complexity lower bounds
   - Tools and a first lower bound
   - Characteristic time and optimal proportions of draws

3. The Track-and-Stop Strategy
   - The Tracking Sampling rule
   - The Chernoff Stopping Rule
   - Asymptotic optimality

4. Practical performance
Optimal algorithms for regret minimization

\( \mu = (\mu_1, \ldots, \mu_K) \in S. \)

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

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

- consistent algorithm: \( \forall \nu \in S, \forall \alpha \in [0, 1], R_T(\mu) = o(T^\alpha) \)
- [Lai and Robbins 1985]: every consistent algorithm satisfies

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

**Definition**

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

\[
\mu_a < \mu^* \Rightarrow \limsup_{T \to \infty} \frac{\mathbb{E}_\mu[N_a(T)]}{\log T} \leq \frac{1}{d(\mu_a, \mu^*)}
\]
KL-UCB: an asymptotically optimal algorithm

KL-UCB [Cappé et al. 2013] \( A_{t+1} = \arg \max_a u_a(t) \), with

\[
u_a(t) = \arg\max_x \left\{ d(\hat{\mu}_a(t), x) \leq \frac{\log(t)}{N_a(t)} \right\},
\]

where \( d(\mu, \mu') = \text{KL} \left( \nu^\mu, \nu^{\mu'} \right) \).

\[
\mathbb{E}_\mu \left[ N_a(T) \right] \leq \frac{1}{d(\mu_a, \mu^*)} \log T + O\left( \sqrt{\log(T)} \right).
\]
We showed that

\[
\inf_{\mathcal{A} \text{ consistent}} \limsup_{T \to \infty} \frac{R_T(\mu)}{\log(T)} = \sum_{a=1}^{K} \frac{(\mu^* - \mu_a)}{d(\mu_a, \mu^*)}.
\]

The history of this result:

- Asymptotic lower bound [Lai and Robbins 85]
- First asymptotically optimal algorithms [Lai and Robbins 85, Agarwal et al. 95]
- Finite-time analysis of simple and explicit asymptotically optimal algorithms: KL-UCB, Bayesian algorithms...
The best arm identification problem

Assume $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$.

Given $\delta \in ]0, 1[$, we want to design a strategy, that is

- a sampling rule $(A_t)$
- a stopping rule $\tau (= \tau_\delta)$
- a recommendation rule $\hat{a}_\tau$

such that, for all $\mu \in S$,

$$\mathbb{P}_\mu (\hat{a}_\tau = a^*(\mu)) \geq 1 - \delta \quad \text{(the strategy is $\delta$-PAC)}$$

and the sample complexity, $\mathbb{E}_\mu[\tau]$ is as small as possible.

**State-of-the-art:** $\delta$-PAC algorithms for which

$$\mathbb{E}_\mu[\tau] = O \left( H(\mu) \log \frac{1}{\delta} \right), \quad H(\mu) = \frac{1}{(\mu_2 - \mu_1)^2} + \sum_{a=2}^{K} \frac{1}{(\mu_a - \mu_1)^2}$$

[Even Dar et al. 2006, Kalyanakrishnan et al. 2012]

$\rightarrow$ the optimal sample complexity is not identified...
Outline

1. Regret minimization

2. Sample complexity lower bounds
   - Tools and a first lower bound
   - Characteristic time and optimal proportions of draws

3. The Track-and-Stop Strategy
   - The Tracking Sampling rule
   - The Chernoff Stopping Rule
   - Asymptotic optimality

4. Practical performance
A first lower bound

\( \mu = (\mu_1, \ldots, \mu_K) \) and \( \lambda = (\lambda_1, \ldots, \lambda_K) \) be two bandit models.

**Change of distribution lemma** [K., Cappé, Garivier 15]

If \( a^*(\mu) \neq a^*(\lambda) \), any \( \delta \)-PAC algorithm satisfies

\[
\sum_{a=1}^{K} \mathbb{E}_\mu [N_a(\tau)] d(\mu_a, \lambda_a) \geq \text{kl}(\delta, 1 - \delta),
\]

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

- For any \( a \in \{2, \ldots, K\} \), introducing \( \lambda \):

\[
\lambda_a = \mu_1 + \epsilon, \\
\lambda_i = \mu_i, \quad \text{if } i \neq a
\]

\[
\mathbb{E}_\mu [N_a(\tau)] d(\mu_a, \mu_1 + \epsilon) \geq \text{kl}(\delta, 1 - \delta)
\]

\[
\mathbb{E}_\mu [N_a(\tau)] \geq \frac{1}{d(\mu_a, \mu_1)} \text{kl}(\delta, 1 - \delta).
\]
A first lower bound

\( \mu = (\mu_1, \ldots, \mu_K) \) and \( \lambda = (\lambda_1, \ldots, \lambda_K) \) be two bandit models.

Change of distribution lemma [K., Cappé, Garivier 15]

If \( a^*(\mu) \neq a^*(\lambda) \), any \( \delta \)-PAC algorithm satisfies

\[
\sum_{a=1}^{K} \mathbb{E}_\mu[N_a(\tau)]d(\mu_a, \lambda_a) \geq \text{kl}(\delta, 1 - \delta),
\]

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

One obtains:

**Theorem**

For any \( \delta \)-PAC algorithm,

\[
\mathbb{E}_\mu[\tau] \geq \left( \frac{1}{d(\mu_1, \mu_2)} + \sum_{a=2}^{K} \frac{1}{d(\mu_a, \mu_1)} \right) \text{kl}(\delta, 1 - \delta)
\]

**Remark:** \( \text{kl}(\delta, 1 - \delta) \sim \log \left( \frac{1}{\delta} \right) \) and \( \text{kl}(\delta, 1 - \delta) \geq \log \left( \frac{1}{2.4\delta} \right) \).
1. Regret minimization

2. Sample complexity lower bounds
   - Tools and a first lower bound
   - Characteristic time and optimal proportions of draws

3. The Track-and-Stop Strategy
   - The Tracking Sampling rule
   - The Chernoff Stopping Rule
   - Asymptotic optimality

4. Practical performance
The best possible lower bound

\[ \mu = (\mu_1, \ldots, \mu_K) \text{ and } \lambda = (\lambda_1, \ldots, \lambda_K) \] be two bandit models.

**Change of distribution lemma** [K., Cappé, Garivier 15]

If \( a^∗(\mu) \neq a^∗(\lambda) \), any \( \delta \)-PAC algorithm satisfies

\[
\sum_{a=1}^{K} \mathbb{E}_{\mu}[N_a(\tau)]d(\mu_a, \lambda_a) \geq \text{kl}(\delta, 1 - \delta).
\]

Let \( \text{Alt}(\mu) = \{ \lambda : a^∗(\lambda) \neq a^∗(\mu) \} \).

\[
\inf_{\lambda \in \text{Alt}(\mu)} \sum_{a=1}^{K} \mathbb{E}_{\mu}[N_a(\tau)]d(\mu_a, \lambda_a) \geq \text{kl}(\delta, 1 - \delta)
\]

\[
\mathbb{E}_{\mu}[\tau] \times \inf_{\lambda \in \text{Alt}(\mu)} \sum_{a=1}^{K} \frac{\mathbb{E}_{\mu}[N_a(\tau)]}{\mathbb{E}_{\mu}[\tau]}d(\mu_a, \lambda_a) \geq \text{kl}(\delta, 1 - \delta)
\]

\[
\mathbb{E}_{\mu}[\tau] \times \left( \sup_{w \in \Sigma_K} \inf_{\lambda \in \text{Alt}(\mu)} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) \right) \geq \text{kl}(\delta, 1 - \delta)
\]
The best possible lower bound

Theorem

For any $\delta$-PAC algorithm,

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

where

$$
T^*(\mu)^{-1} = \sup_{w \in \Sigma_K} \inf_{\lambda \in \text{Alt}(\mu)} \left( \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) \right).
$$

other non-explicit lower bounds:


Moreover, the vector

$$
w^*(\mu) = \arg\max_{w \in \Sigma_K} \inf_{\lambda \in \text{Alt}(\mu)} \left( \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) \right)
$$

contains the optimal proportions of draws of the arms.
Computing the optimal proportions

$$w^* \in \arg\max_{w \in \Sigma_K} \inf_{\lambda \in \text{Alt}(\mu)} \left( \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) \right).$$

An explicit calculation yields

\begin{align*}
(\ast) &= \min_{a \neq 1} \left[ w_1 d \left( \mu_1, \frac{w_1 \mu_1 + w_a \mu_a}{w_1 + w_a} \right) + w_a d \left( \mu_a, \frac{w_1 \mu_1 + w_a \mu_a}{w_1 + w_a} \right) \right] \\
&= w_1 \min_{a \neq 1} g_a \left( \frac{w_a}{w_1} \right) \quad (w_1 \neq 0)
\end{align*}

where $g_a(x) = d \left( \mu_1, \frac{\mu_1 + x \mu_a}{1 + x} \right) + xd \left( \mu_a, \frac{\mu_1 + x \mu_a}{1 + x} \right)$.

g_a is a one-to-one mapping from $[0, +\infty[$ onto $[0, d(\mu_1, \mu_a)[$. 

Emilie Kaufmann

Optimal Best Arm Identification
Computing the optimal proportions

\[ w^* \in \arg\max_{w \in \Sigma_K} \inf_{\lambda \in \text{Alt}(\mu)} \left( \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) \right). \]

An explicit calculation yields

\[ (*) = \min_{a \neq 1} \left[ w_1 d \left( \mu_1, \frac{w_1 \mu_1 + w_a \mu_a}{w_1 + w_a} \right) + w_a d \left( \mu_a, \frac{w_1 \mu_1 + w_a \mu_a}{w_1 + w_a} \right) \right] \]

\[ = w_1 \min_{a \neq 1} g_a \left( \frac{w_a}{w_1} \right) \quad (w_1 \neq 0) \]

where \( g_a(x) = d \left( \mu_1, \frac{\mu_1 + x \mu_a}{1 + x} \right) + xd \left( \mu_a, \frac{\mu_1 + x \mu_a}{1 + x} \right). \)

\( g_a \) is a one-to-one mapping from \([0, +\infty[\) onto \([0, d(\mu_1, \mu_a)][\).

\[ x_1^* = 1 \quad x_2^* = \frac{w_2^*}{w_1^*} \quad \ldots \quad x_K^* = \frac{w_K^*}{w_1^*} \]
Computing the optimal proportions

Letting $x_a^* = w_a^*/w_1^*$ for all $a \geq 2$,

$$x_2^*, \ldots, x_K^* \in \arg\max_{x_2, \ldots, x_K \geq 0} \frac{\min_{a \neq 1} g_a(x_a)}{1 + x_2 + x_K}.$$ 

It is easy to check that there exists $y^* \in [0, d(\mu_1, \mu_2)]$ such that

$$\forall a \in \{2, \ldots, K\}, g_a(x_a^*) = y^*.$$

Letting $x_a(y) = g_a^{-1}(y)$, one has $x_a^* = x_a(y^*)$ where

$$y^* \in \arg\max_{y \in [0, d(\mu_1, \mu_2)]} \frac{y}{1 + x_2(y) + x_K(y)}.$$
Theorem

For every $a \in A$, 

$$w_a^*(\mu) = \frac{x_a(y^*)}{\sum_{a=1}^{K} x_a(y^*)} ,$$

where $y^*$ is the unique solution of the equation $F_{\mu}(y) = 1$, where

$$F_{\mu} : y \mapsto \sum_{a=2}^{K} d \left( \mu_1, \frac{\mu_1 + x_a(y)\mu_a}{1 + x_a(y)} \right)$$

is a continuous, increasing function on $[0, d(\mu_1, \mu_2)]$ such that

$F_{\mu}(0) = 0$ and $F_{\mu}(y) \to \infty$ when $y \to d(\mu_1, \mu_2)$.

→ an efficient way to compute the vector of proportions $w^*(\mu)$.
Outline

1. Regret minimization

2. Sample complexity lower bounds
   - Tools and a first lower bound
   - Characteristic time and optimal proportions of draws

3. The Track-and-Stop Strategy
   - The Tracking Sampling rule
   - The Chernoff Stopping Rule
   - Asymptotic optimality

4. Practical performance
Sampling rule: Tracking the optimal proportions

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

- Introducing

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

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

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

\[
\text{(tracking)}
\]

Lemma

Under the Tracking sampling rule,

\[
P_{\mu} \left( \lim_{t \to \infty} \frac{N_a(t)}{t} = w^*_a(\mu) \right) = 1.
\]
Outline

1. Regret minimization

2. Sample complexity lower bounds
   - Tools and a first lower bound
   - Characteristic time and optimal proportions of draws

3. The Track-and-Stop Strategy
   - The Tracking Sampling rule
   - The Chernoff Stopping Rule
   - Asymptotic optimality

4. Practical performance
Stopping rule: performing statistical tests

High values of the Generalized Likelihood Ratio

\[ Z_{a,b}(t) := \log \frac{\max\{\lambda: \lambda_a \geq \lambda_b\} \ell(X_1, \ldots, X_t; \lambda)}{\max\{\lambda: \lambda_a \leq \lambda_b\} \ell(X_1, \ldots, X_t; \lambda)}, \]

reject the hypothesis that \((\mu_a < \mu_b)\).

We stop when one arm is accessed to be significantly larger than all other arms, according to a GLR Test:

\[ \tau_\delta = \inf \{ t \in \mathbb{N} : \exists a \in \{1, \ldots, K\}, \forall b \neq a, Z_{a,b}(t) > \beta(t, \delta) \} \]

\[ = \inf \left\{ t \in \mathbb{N} : \max_{a \in A} \min_{b \neq a} Z_{a,b}(t) > \beta(t, \delta) \right\} \]

Chernoff stopping rule [Chernoff 59]
Stopping rule: alternative interpretations

One has $Z_{a,b}(t) = -Z_{b,a}(t)$ and, if $\hat{\mu}_a(t) \geq \hat{\mu}_b(t)$,

$$Z_{a,b}(t) = N_a(t) d(\hat{\mu}_a(t), \hat{\mu}_{a,b}(t)) + N_b(t) d(\hat{\mu}_b(t), \hat{\mu}_{a,b}(t)),$$

where $\hat{\mu}_{a,b}(t) := \frac{N_a(t)}{N_a(t)+N_b(t)} \hat{\mu}_a(t) + \frac{N_b(t)}{N_a(t)+N_b(t)} \hat{\mu}_b(t)$.

A link with the lower bound

$$\max_{a} \min_{b \neq a} Z_{a,b}(t) = t \times \inf_{\lambda \in \text{Alt}(\hat{\mu}(t))} \sum_{a=1}^{K} \frac{N_a(t)}{t} d(\hat{\mu}_a(t), \lambda_a) \approx \frac{t}{T^*(\mu)}$$

under a “good” sampling strategy (for $t$ large)
Stopping rule: alternative interpretations

One has $Z_{a,b}(t) = -Z_{b,a}(t)$ and, if $\hat{\mu}_a(t) \geq \hat{\mu}_b(t)$,

$$Z_{a,b}(t) = N_a(t) d(\hat{\mu}_a(t), \hat{\mu}_{a,b}(t)) + N_b(t) d(\hat{\mu}_b(t), \hat{\mu}_{a,b}(t)),$$

where $\hat{\mu}_{a,b}(t) := \frac{N_a(t)}{N_a(t)+N_b(t)} \hat{\mu}_a(t) + \frac{N_b(t)}{N_a(t)+N_b(t)} \hat{\mu}_b(t)$.

A Minimum Description Length interpretation

If $H(\mu) = \mathbb{E}_{X \sim \nu^\mu} [-\log p_{\mu}(X)]$ is the Shannon entropy,

$$Z_{a,b}(t) = \underbrace{(N_a(t) + N_b(t)) H(\hat{\mu}_{a,b}(t))}_{\text{average #bits to encode the samples of a and b together}}$$

$$- \underbrace{[N_a(t) H(\hat{\mu}_a(t)) + N_b(t) H(\hat{\mu}_b(t))]}_{\text{average #bits to encode the sample of a and b separately}},$$
Stopping rule: $\delta$-PAC property

The Chernoff rule is $\delta$-PAC for $\beta(t, \delta) = \log \left( \frac{2(K-1)t}{\delta} \right)$.

**Lemma**

If $\mu_a < \mu_b$, whatever the sampling rule,

$$
P_\mu \left( \exists t \in \mathbb{N} : Z_{a,b}(t) > \log(2t/\delta) \right) \leq \delta.
$$

i.e., $P(T_{a,b} < \infty) \leq \delta$, for $T_{a,b} = \inf\{ t \in \mathbb{N} : Z_{a,b}(t) > \log(2t/\delta) \}$.

Using that

$$(T_{a,b} = t) \subseteq \left( \frac{\max_{\mu'_a \leq \mu'_b} p_{\mu'_a}(X^a_t)p_{\mu'_b}(X^b_t)}{\max_{\mu'_a \geq \mu'_b} p_{\mu'_a}(X^a_t)p_{\mu'_b}(X^b_t)} \geq \frac{2t}{\delta} \right),$$

one has

$$
P_\mu(T_{a,b} < \infty) = \sum_{t=1}^{\infty} E_\mu \left[ 1(T_{a,b}=t) \right]$$

$$
\leq \sum_{t=1}^{\infty} \frac{\delta}{2t} E_\mu \left[ 1(T_{a,b}=t) \max_{\mu'_a \geq \mu'_b} \frac{p_{\mu'_a}(X^a_t)p_{\mu'_b}(X^b_t)}{\max_{\mu'_a \leq \mu'_b} p_{\mu'_a}(X^a_t)p_{\mu'_b}(X^b_t)} \right].$$
Stopping rule: $\delta$-PAC property

$$\Pr_{\mu}(T_{a,b} < \infty) \leq \sum_{t=1}^{\infty} \frac{\delta}{2t} \mathbb{E}_{\mu} \left[ 1(T_{a,b}=t) \frac{\max_{\mu'_a \geq \mu'_b} p_{\mu'_a}(X^a_t)p_{\mu'_b}(X^b_t)}{p_{\mu_a}(X^a_t)p_{\mu_b}(X^b_t)} \right]$$

$$= \sum_{t=1}^{\infty} \frac{\delta}{2t} \sum_{x_t \in \{0,1\}^t} 1(T_{a,b}=t)(x_t) \max_{\mu'_a \geq \mu'_b} p_{\mu'_a}(X^a_t)p_{\mu'_b}(X^b_t) \prod_{i \in A \setminus \{a,b\}} p_{\mu_i}(x^i_t)$$

not a probability density...

Lemma [Willems et al. 95]

The Krichevsky-Trofimov distribution

$$kt(x) = \int_0^1 \frac{1}{\pi \sqrt{u(1-u)}} p_u(x) du$$

is a probability law on $\{0,1\}^n$ that satisfies

$$\sup_{x \in \{0,1\}^n} \sup_{u \in [0,1]} \frac{p_u(x)}{kt(x)} \leq 2\sqrt{n}.$$
Stopping rule: $\delta$-PAC property

$$\mathbb{P}_\mu(T_{a,b} < \infty) \leq \sum_{t=1}^{\infty} \frac{\delta}{2t} \mathbb{E}_\mu \left[ 1(T_{a,b}=t) \frac{\max_{\mu_a' \geq \mu_b'} p_{\mu_a'}(X^a_t)p_{\mu_b'}(X^b_t)}{p_{\mu_a}(X^a_t)p_{\mu_b}(X^b_t)} \right]$$

$$= \sum_{t=1}^{\infty} \frac{\delta}{2t} \sum_{x_t \in \{0,1\}^t} 1(T_{a,b}=t)(x_t) \max_{\mu_a' \geq \mu_b'} p_{\mu_a'}(x^a_t)p_{\mu_b'}(x^b_t) \prod_{i \in A \setminus \{a,b\}} p_{\mu_i}(x^i_t)$$

$$\leq \sum_{t=1}^{\infty} \frac{\delta}{2t} \sum_{x_t \in \{0,1\}^t} 1(T_{a,b}=t)(x_t) 4 \sqrt{n^a_t n^b_t} k(t)(x^a_t) k(t)(x^b_t) \prod_{i \in A \setminus \{a,b\}} p_{\mu_i}(x^i_t)$$

$$\leq \sum_{t=1}^{\infty} \delta \sum_{x_t \in \{0,1\}^t} 1(T_{a,b}=t)(x_t) I(x_t)$$

$$= \delta \sum_{t=1}^{\infty} \mathbb{E}[1(T_{a,b}=t)] = \delta \tilde{\mathbb{P}}(T_{a,b} < \infty) \leq \delta.$$
Outline

1. Regret minimization

2. Sample complexity lower bounds
   - Tools and a first lower bound
   - Characteristic time and optimal proportions of draws

3. The Track-and-Stop Strategy
   - The Tracking Sampling rule
   - The Chernoff Stopping Rule
   - Asymptotic optimality

4. Practical performance
Theorem

The Track-and-Stop strategy, that uses
- the Tracking sampling rule
- the Chernoff stopping rule with \( \beta(t, \delta) = \log \left( \frac{2(K-1)t}{\delta} \right) \)
- and recommends \( \hat{a}_\tau = \arg\max_{a=1 \ldots K} \hat{\mu}_a(\tau) \)

is \( \delta \)-PAC for every \( \delta \in ]0, 1[ \) and satisfies

\[
\limsup_{\delta \to 0} \frac{\mathbb{E}_{\mu} [\tau_\delta]}{\log(1/\delta)} = T^*(\mu).
\]
1. Regret minimization

2. Sample complexity lower bounds
   - Tools and a first lower bound
   - Characteristic time and optimal proportions of draws

3. The Track-and-Stop Strategy
   - The Tracking Sampling rule
   - The Chernoff Stopping Rule
   - Asymptotic optimality

4. Practical performance
State-of-the-art algorithms

An algorithm based on confidence intervals: **KL-LUCB**

[K., Kalyanakrishnan 13]

\[
u_a(t) = \max \{ q : N_a(t) d(\hat{\mu}_a(t), q) \leq \beta(t, \delta) \} \\
\ell_a(t) = \min \{ q : N_a(t) d(\hat{\mu}_a(t), q) \leq \beta(t, \delta) \}
\]

- **sampling rule:** \( A_{t+1} = \arg\max_a \hat{\mu}_a(t), \ B_{t+1} = \arg\max_{b \neq A_{t+1}} u_b(t) \)
- **stopping rule:** \( \tau = \inf \{ t \in \mathbb{N} : \ell_{A_t}(t) > u_{B_t}(t) \} \)

Emilie Kaufmann
Optimal Best Arm Identification
State-of-the-art algorithms

A Racing-type algorithm: **KL-Racing** [K., Kalyanakrishnan 13]

\[ \mathcal{R} = \{1, \ldots, K\} \text{ set of remaining arms.} \]

\( r = 0 \) current round

**while** \( |\mathcal{R}| > 1 \)

- \( r = r + 1 \)
- draw each \( a \in \mathcal{R} \), compute \( \hat{\mu}_{a,r} \), the empirical mean of the \( r \) samples observed so far
- compute the **empirical best and empirical worst arms**:

\[
\begin{align*}
    b_r &= \arg\max_{a \in \mathcal{R}} \hat{\mu}_{a,r} \\
    w_r &= \arg\min_{a \in \mathcal{R}} \hat{\mu}_{a,r}
\end{align*}
\]

- **Elimination step**: if

\[
    \ell_{b_r}(r) > u_{w_r}(r),
\]

eliminate \( w_r \) : \( \mathcal{R} = \mathcal{R} \setminus \{w_r\} \)

**Outpout**: \( \hat{a} \) the single element in \( \mathcal{R} \).
The Chernoff-Racing algorithm

\[ \mathcal{R} = \{1, \ldots, K\} \] set of remaining arms.

\( r = 0 \) current round

while \( |\mathcal{R}| > 1 \)

- \( r=r+1 \)
- draw each \( a \in \mathcal{R} \), compute \( \hat{\mu}_{a,r} \), the empirical mean of the \( r \) samples observed sofar
- compute the empirical best and empirical worst arms:
  \[ b_r = \arg\max_{a \in \mathcal{R}} \hat{\mu}_{a,r} \quad w_r = \arg\min_{a \in \mathcal{R}} \hat{\mu}_{a,r} \]

- Elimination step: if \((Z_{b_r,w_r}(r) > \beta(r, \delta))\), or
  \[ rd \left( \frac{\hat{\mu}_{a,r} + \hat{\mu}_{b,r}}{2} \right) + rd \left( \frac{\hat{\mu}_{b,r} + \hat{\mu}_{a,r}}{2} \right) > \beta(r, \delta), \]
  eliminate \( w_r : \mathcal{R} = \mathcal{R} \setminus \{w_r\} \)

end

Output: \( \hat{a} \) the single element in \( \mathcal{R} \).
Numerical experiments

Experiments on two Bernoulli bandit models:

- $\mu_1 = [0.5 \ 0.45 \ 0.43 \ 0.4]$, such that
  $w^*(\mu_1) = [0.417 \ 0.390 \ 0.136 \ 0.057]$

- $\mu_2 = [0.3 \ 0.21 \ 0.2 \ 0.19 \ 0.18]$, such that
  $w^*(\mu_2) = [0.336 \ 0.251 \ 0.177 \ 0.132 \ 0.104]$

In practice, set the threshold to $\beta(t, \delta) = \log \left( \frac{\log(t)+1}{\delta} \right)$.

<table>
<thead>
<tr>
<th></th>
<th>Track-and-Stop</th>
<th>Chernoff-Racing</th>
<th>KL-LUCB</th>
<th>KL-Racing</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_1$</td>
<td>4052</td>
<td>4516</td>
<td>8437</td>
<td>9590</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>1406</td>
<td>3078</td>
<td>2716</td>
<td>3334</td>
</tr>
</tbody>
</table>

Table: Expected number of draws $E_\mu[\tau_\delta]$ for $\delta = 0.1$, averaged over $N = 3000$ experiments.
Conclusion

For best arm identification, we showed that

$$\inf_{\text{PAC algorithm}} \limsup_{\delta \to 0} \frac{\mathbb{E}_\mu[\tau_\delta]}{\log(1/\delta)} = \sup_{w \in \Sigma_K} \inf_{\lambda \in \text{Alt}(\mu)} \left( \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) \right)$$

and provided an efficient strategy matching this bound.

Future work:

- a finite-time analysis
- combine the knowledge of $w^*(\mu)$ with other successful heuristics (UCB, Thompson Sampling)
References

- E. Kaufmann, S. Kalyanakrishnan. The information complexity of best arm identification, COLT 2013