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 (ν_a has mean μ_a)

ν_1

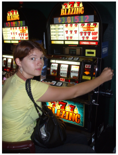
ν_2

ν_3

ν_4

ν_5

At round t, an agent:



- chooses an arm $A_t \in \mathcal{A} := \{1, \ldots, K\}$
- 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

$$a^* = \text{argmax}_a \, \mu_a \quad \text{and} \quad \mu^* = \max_a \mu_a.$$

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*:

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

$\Rightarrow$ **exploration/exploitation tradeoff**

Motivation: clinical trials [1933]





$\mathcal{B}(\mu_1)$ $\mathcal{B}(\mu_2)$ $\mathcal{B}(\mu_3)$ $\mathcal{B}(\mu_4)$ $\mathcal{B}(\mu_5)$

<u>Goal</u>: maximize the number of patients healed during the trial

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*:

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

$\Rightarrow$ **exploration/exploitation tradeoff**

Motivation: clinical trials [1933]

$\mathcal{B}(\mu_1)$

$\mathcal{B}(\mu_2)$

$\mathcal{B}(\mu_3)$

$\mathcal{B}(\mu_4)$

$\mathcal{B}(\mu_5)$

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 !

$\Rightarrow$ **optimal exploration**

The agent's strategy is made of:

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

Possible goals:

Fixed-budget setting	Fixed-confidence setting
$\tau = T$ minimize $\mathbb{P}(\hat{a}_\tau \neq a^*)$	minimize $\mathbb{E}[\tau]$ $\mathbb{P}(\hat{a}_\tau \neq a^*) \leq \delta$

Motivation: Market research, A/B Testing, clinical trials...

Our objective: Best-arm identification

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

$\Rightarrow$ **optimal exploration**

The agent's strategy is made of:

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

Possible goals:

Fixed-budget setting	Fixed-confidence setting
$\tau = T$ minimize $\mathbb{P}(\hat{a}_\tau \neq a^*)$	minimize $\mathbb{E}[\tau]$ $\mathbb{P}(\hat{a}_\tau \neq a^*) \leq \delta$

Motivation: Market research, A/B Testing, clinical trials...

Wanted: Optimal algorithms in the PAC formulation

$\mathcal{S}$ a class of bandit models $\nu = (\nu_1, \ldots, \nu_K)$.
A strategy is δ-PAC on $\mathcal{S}$ is $\forall \nu \in \mathcal{S}, \mathbb{P}_\nu(\hat{a}_\tau = a^*) \geq 1 - \delta$.

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

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

(distribution-dependent bounds)

Exponential family bandit models

$\nu_1, \dots, \nu_K$ belong to a one-dimensional exponential family:

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

Example: Gaussian, Bernoulli, Poisson distributions...

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

Notation: Kullback-Leibler divergence

For a given exponential family $\mathcal{P}$,

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

is the KL-divergence between the distributions of mean μ and μ'.

Example: Bernoulli distributions

$$d(\mu, \mu') = \mathrm{KL}(\mathcal{B}(\mu), \mathcal{B}(\mu')) = \mu \log \frac{\mu}{\mu'} + (1-\mu) \log \frac{1-\mu}{1-\mu'}.$$

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

$$\mathcal{S} = \left\{\boldsymbol{\mu} \in (\dot{b}(\Theta))^K : \exists a \in \mathcal{A} : \mu_a > \max_{i \neq a} \mu_i\right\}$$

Outline

Outline

Optimal algorithms for regret minimization

$\boldsymbol{\mu} = (\mu_1, \dots, \mu_K) \in \mathcal{S}$.

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

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

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

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

Definition

A bandit algorithm is **asymptotically optimal** if, for every $\boldsymbol{\mu} \in \mathcal{S}$,

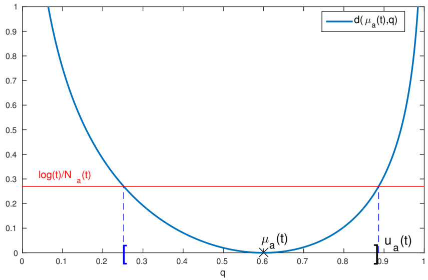
$$\mu_a < \mu^* \Rightarrow \limsup_{T \to \infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]}{\log T} \leq \frac{1}{d(\mu_a, \mu^*)}$$

KL-UCB: an asymptotically optimal algorithm

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

$$u_a(t) = \underset{x}{\text{argmax}} \left\{ d\left(\hat{\mu}_a(t), x\right) \leq \frac{\log(t)}{N_a(t)} \right\},$$

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



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

The information complexity of regret minimization

We showed that

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

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 $\mu_1 > \mu_2 \geq \dots \geq \mu_K$.

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 $\boldsymbol{\mu} \in \mathcal{S}$,

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

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

State-of-the-art: δ-PAC algorithms for which

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

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

➜ the optimal sample complexity is not identified...

Outline

A first lower bound

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

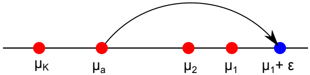
Change of distribution lemma [K., Cappé, Garivier 15]

If $a^*(\boldsymbol{\mu}) \neq a^*(\boldsymbol{\lambda})$, any δ-PAC algorithm satisfies

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

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

- For any $a \in \{2, \ldots, K\}$, introducing λ:



$$\begin{cases} \lambda_a = \mu_1 + \epsilon \\ \lambda_i = \mu_i, \quad \text{if } i \neq a \end{cases}$$

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau)] d(\mu_a, \mu_1 + \epsilon) &\geq \mathrm{kl}(\delta, 1-\delta) \\ \mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau)] &\geq \frac{1}{d(\mu_a, \mu_1)} \mathrm{kl}(\delta, 1-\delta). \end{aligned}$$

A first lower bound

$\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ and $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)$ be two bandit models.

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

If $a^*(\boldsymbol{\mu}) \neq a^*(\boldsymbol{\lambda})$, any δ-PAC algorithm satisfies

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

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

- One obtains:

Theorem

For any δ-PAC algorithm,

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

Remark: $\mathrm{kl}(\delta, 1-\delta) \underset{\delta \to 0}{\sim} \log\left(\frac{1}{\delta}\right)$ and $\mathrm{kl}(\delta, 1-\delta) \geq \log\left(\frac{1}{2.4\delta}\right)$.

Outline

The best possible lower bound

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

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

If $a^*(\boldsymbol{\mu}) \neq a^*(\boldsymbol{\lambda})$, any δ-PAC algorithm satisfies

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

- Let $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} : a^*(\boldsymbol{\lambda}) \neq a^*(\boldsymbol{\mu})\}$.

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

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

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

The best possible lower bound

Theorem

For any δ-PAC algorithm,

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

where

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

other non-explicit lower bounds:
[Graves and Lai 1997, Vaidhyan and Sundaresan, 2015]

Moreover, the vector

$$w^*(\boldsymbol{\mu}) = \underset{w \in \Sigma_K}{\mathrm{argmax}} \inf_{\lambda \in \mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^{K} w_a d(\mu_a, \lambda_a)\right)$$

contains the optimal proportions of draws of the arms.

Computing the optimal proportions

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

An explicit calculation yields

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

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 $[0, d(\mu_1, \mu_a)[$.

Computing the optimal proportions

$$w^* \in \underset{w \in \Sigma_K}{\operatorname{argmax}} \underbrace{\inf_{\lambda \in \mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) \right)}_{(*)}.$$

An explicit calculation yields

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

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

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

$$x_1^* = 1 \qquad x_2^* = w_2^*/w_1^* \qquad \dots \qquad x_K^* = w_K^*/w_1^*$$

Computing the optimal proportions

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

$$x_2^*, \dots, x_K^* \in \underset{x_2,\dots,x_K \geq 0}{\text{argmax}} \frac{\min_{a\neq 1} g_a(x_a)}{1 + x_2 + x_K}.$$

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

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

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

$$y^* \in \underset{y\in[0,d(\mu_1,\mu_2)[}{\text{argmax}} \frac{y}{1 + x_2(y) + x_K(y)}.$$

Computing the optimal proportions

Theorem

For every $a \in \mathcal{A}$,

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

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

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

is a continuous, increasing function on $[0, d(\mu_1, \mu_2)[$ such that $F_{\boldsymbol{\mu}}(0) = 0$ and $F_{\boldsymbol{\mu}}(y) \to \infty$ when $y \to d(\mu_1, \mu_2)$.

➜ an efficient way to compute the vector of proportions $w^*(\boldsymbol{\mu})$

Outline

Sampling rule: Tracking the optimal proportions

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

- Introducing

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

the arm sampled at round $t+1$ is

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

Lemma

Under the Tracking sampling rule,

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\lim_{t\to\infty} \frac{N_a(t)}{t} = w_a^*(\boldsymbol{\mu})\right) = 1.$$

Outline

Stopping rule: performing statistical tests

High values of the Generalized Likelihood Ratio

$$Z_{a,b}(t) := \log \frac{\max_{\{\boldsymbol{\lambda}:\lambda_a \geq \lambda_b\}} \ell(X_1, \dots, X_t; \boldsymbol{\lambda})}{\max_{\{\boldsymbol{\lambda}:\lambda_a \leq \lambda_b\}} \ell(X_1, \dots, X_t; \boldsymbol{\lambda})},$$

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:

$$\begin{aligned} \tau_\delta &= \inf\{t \in \mathbb{N} : \exists a \in \{1, \dots, K\}, \forall b \neq a, Z_{a,b}(t) > \beta(t, \delta)\} \\ &= \inf\left\{t \in \mathbb{N} : \max_{a \in \mathcal{A}} \min_{b \neq a} Z_{a,b}(t) > \beta(t, \delta)\right\} \end{aligned}$$

Chernoff stopping rule [Chernoff 59]

Stopping rule: alternative interpretations

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

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

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

$$\begin{aligned} \max_a \min_{b \neq a} Z_{a,b}(t) &= t \times \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\mu}(t))} \sum_{a=1}^{K} \frac{N_a(t)}{t} d(\hat{\mu}_a(t), \lambda_a) \\ &\simeq \frac{t}{T^*(\boldsymbol{\mu})} \end{aligned}$$

under a "good" sampling strategy (for t large)

Stopping rule: alternative interpretations

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

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

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,

$$\begin{aligned} Z_{a,b}(t) &= \underbrace{(N_a(t)+N_b(t))H(\hat{\mu}_{a,b}(t))}_{\text{average \#bits to encode the samples of a and b together}} \\ &- \underbrace{[N_a(t)H(\hat{\mu}_a(t)) + N_b(t)H(\hat{\mu}_b(t))]}_{\text{average \#bits to encode the sample of a and b separately}}, \end{aligned}$$

Stopping rule: δ-PAC property

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

Lemma

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

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

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

Using that

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

one has

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\mu}}(T_{a,b} < \infty) &= \sum_{t=1}^{\infty} \mathbb{E}_{\boldsymbol{\mu}}\left[\mathbb{1}_{(T_{a,b}=t)}\right] \\ &\leq \sum_{t=1}^{\infty} \frac{\delta}{2t} \mathbb{E}_{\boldsymbol{\mu}}\left[\mathbb{1}_{(T_{a,b}=t)} \frac{\max_{\mu'_a \geq \mu'_b} p_{\mu'_a}(\underline{X}^a_t) p_{\mu'_b}(\underline{X}^b_t)}{\max_{\mu'_a \leq \mu'_b} p_{\mu'_a}(\underline{X}^a_t) p_{\mu'_b}(\underline{X}^b_t)}\right].\end{aligned}$$

Stopping rule: δ-PAC property

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

$$= \sum_{t=1}^{\infty} \frac{\delta}{2t} \sum_{\underline{x}_t \in \{0,1\}^t} \mathbb{1}_{(T_{a,b}=t)}(\underline{x}_t) \underbrace{\max_{\mu'_a \geq \mu'_b} p_{\mu'_a}(\underline{x}_t^a) p_{\mu'_b}(\underline{x}_t^b) \prod_{i \in \mathcal{A} \setminus \{a,b\}} p_{\mu_i}(\underline{x}_t^i)}_{\text{not a probability density...}}$$

Lemma [Willems et al. 95]

The Krichevsky-Trofimov distribution

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

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

$$\sup_{x \in \{0,1\}^n} \sup_{u \in [0,1]} \frac{p_u(x)}{\mathrm{kt}(x)} \leq 2\sqrt{n} \, .$$

Stopping rule: δ-PAC property

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

$$= \sum_{t=1}^{\infty} \frac{\delta}{2t} \sum_{\underline{x}_t \in \{0,1\}^t} \mathbb{1}_{(T_{a,b}=t)}(\underline{x}_t) \max_{\mu'_a \geq \mu'_b} p_{\mu'_a}(\underline{x}_t^a) p_{\mu'_b}(\underline{x}_t^b) \prod_{i \in \mathcal{A} \backslash \{a,b\}} p_{\mu_i}(\underline{x}_t^i)$$

$$\leq \sum_{t=1}^{\infty} \frac{\delta}{2t} \sum_{\underline{x}_t \in \{0,1\}^t} \mathbb{1}_{(T_{a,b}=t)}(\underline{x}_t) 4\sqrt{n_t^a n_t^b}\, \underbrace{\mathrm{kt}(\underline{x}_t^a) \mathrm{kt}(\underline{x}_t^b) \prod_{i \in \mathcal{A} \backslash \{a,b\}} p_{\mu_i}(\underline{x}_t^i)}_{I(\underline{x}_t)}$$

$$\leq \sum_{t=1}^{\infty} \delta \sum_{\underline{x}_t \in \{0,1\}^t} \mathbb{1}_{(T_{a,b}=t)}(\underline{x}_t) I(\underline{x}_t)$$

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

Outline

An asymptotically optimal algorithm

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 $\hat{a}_\tau = \underset{a=1\dots K}{\operatorname{argmax}}\ \hat{\mu}_a(\tau)$

is δ-PAC for every $\delta \in]0,1[$ and satisfies

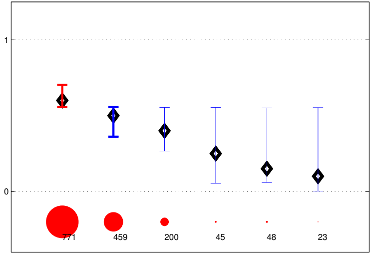
$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} = T^*(\boldsymbol{\mu}).$$

Outline

State-of-the-art algorithms

An algorithm based on confidence intervals : **KL-LUCB**
[K., Kalyanakrishnan 13]

$$
\begin{aligned}
u_a(t) &= \max\left\{q : N_a(t) d(\hat{\mu}_a(t), q) \leq \beta(t,\delta)\right\} \\
\ell_a(t) &= \min\left\{q : N_a(t) d(\hat{\mu}_a(t), q) \leq \beta(t,\delta)\right\}
\end{aligned}
$$



- sampling rule: $A_{t+1} = \underset{a}{\operatorname{argmax}}\ \hat{\mu}_a(t)$, $B_{t+1} = \underset{b \neq A_{t+1}}{\operatorname{argmax}}\ u_b(t)$
- stopping rule: $\tau = \inf\{t \in \mathbb{N} : \ell_{A_t}(t) > u_{B_t}(t)\}$

State-of-the-art algorithms

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

$\mathcal{R} = \{1, \ldots, K\}$ 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:

$$b_r = \underset{a \in \mathcal{R}}{\text{argmax}} \ \hat{\mu}_{a,r} \qquad w_r = \underset{a \in \mathcal{R}}{\text{argmin}} \ \hat{\mu}_{a,r}$$

- Elimination step: if

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

eliminate w_r : $\mathcal{R} = \mathcal{R} \backslash \{w_r\}$

end

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:

$$b_r = \underset{a \in \mathcal{R}}{\text{argmax}}\ \hat{\mu}_{a,r} \qquad w_r = \underset{a \in \mathcal{R}}{\text{argmin}}\ \hat{\mu}_{a,r}$$

- Elimination step: if $(Z_{b_r,w_r}(r) > \beta(r,\delta))$, or

$$rd\left(\hat{\mu}_{a,r}, \frac{\hat{\mu}_{a,r} + \hat{\mu}_{b,r}}{2}\right) + rd\left(\hat{\mu}_{b,r}, \frac{\hat{\mu}_{a,r} + \hat{\mu}_{b,r}}{2}\right) > \beta(r,\delta),$$

  eliminate w_r : $\mathcal{R} = \mathcal{R} \backslash \{w_r\}$

end

Outpout: $\hat{a}$ the single element in $\mathcal{R}$.

Numerical experiments

Experiments on two Bernoulli bandit models:

- $\boldsymbol{\mu}_1 = [0.5\ 0.45\ 0.43\ 0.4]$, such that
$$w^*(\boldsymbol{\mu}_1) = [0.417\ 0.390\ 0.136\ 0.057]$$
- $\boldsymbol{\mu}_2 = [0.3\ 0.21\ 0.2\ 0.19\ 0.18]$, such that
$$w^*(\boldsymbol{\mu}_2) = [0.336\ 0.251\ 0.177\ 0.132\ 0.104]$$

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

	Track-and-Stop	Chernoff-Racing	KL-LUCB	KL-Racing
$\boldsymbol{\mu}_1$	4052	4516	8437	9590
$\boldsymbol{\mu}_2$	1406	3078	2716	3334

Table : Expected number of draws $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$ for $\delta = 0.1$, averaged over $N = 3000$ experiments.

For best arm identification, we showed that

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

and provided an efficient strategy matching this bound.

Future work:

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

References

- O. Cappé, A. Garivier, O-A. Maillard, R. Munos, and G. Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 2013.
- H. Chernoff. Sequential design of Experiments. The Annals of Mathematical Statistics, 1959.
- E. Even-Dar, S. Mannor, Y. Mansour, Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. *JMLR*, 2006.
- T.L. Graves and T.L. Lai. Asymptotically Efficient adaptive choice of control laws in controlled markov chains. SIAM Journal on Control and Optimization, 35(3):715–743, 1997.
- S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. PAC subset selection in stochastic multi- armed bandits. ICML, 2012.
- E. Kaufmann, O. Cappé, A. Garivier. On the Complexity of Best Arm Identification in Multi-Armed Bandit Models. *JMLR*, 2015
- A. Garivier, E. Kaufmann. Optimal Best Arm Identification with Fixed Confidence, arXiv:1602.04589, 2016
- E. Kaufmann, S. Kalyanakrishnan. The information complexity of best arm identification, COLT 2013
- T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 1985.
- N.K. Vaidhyan and R. Sundaresan. Learning to detect an oddball target. arXiv:1508.05572, 2015.