On the Complexity of Best Arm Identification with Fixed Confidence

Discrete Optimization with Noise

Aurélien Garivier, joint work with Emilie Kaufmann (CNRS, CRIStAL)

to be presented at COLT'16, New York

June 6^{th}, 2016

Institut de Mathématiques de Toulouse
LabeX CIMI
Université Paul Sabatier

Table of contents

Preliminaries: Basics of Large Deviation Bounds

Chernoff Bound for Bernoulli variables

Let $\mu \in (0,1)$. Let $X_1, X_2, \dots X_n \sim \mathcal{B}(\mu)$, and let $\bar{X}_n = (X_1 + \cdots + X_n)/n$.

Theorem

For all $x > \mu$,

$$P_\mu\left(\bar{X}_n \geq x\right) \leq e^{-n\,\mathrm{kl}(x,\mu)}$$

where $\mathrm{kl}(x,y) = x \log \frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$ is the binary relative entropy

Corollary

For every $\delta > 0$,

$$\mathbb{P}_\mu\left(n\,\mathrm{kl}\left(\bar{X}_n, \mu\right) \geq \log\frac{1}{\delta}\right) \leq 2\delta$$

Proof: Fenchel-Legendre transform of log-Laplace

For every $\lambda > 0$,

$$\begin{aligned}\mathbb{P}_\mu\left(\bar{X}_n \geq x\right) &= \mathbb{P}_\mu\left(e^{\lambda(X_1+\cdots+X_n)} \geq e^{\lambda n x}\right) \\ &\leq \frac{\mathbb{E}_\mu\left[e^{\lambda(X_1+\cdots+X_n)}\right]}{e^{\lambda n x}} \\ &= e^{-n\left(\lambda x - \log \mathbb{E}_\mu[\exp \lambda X_1]\right)} .\end{aligned}$$

Thus,

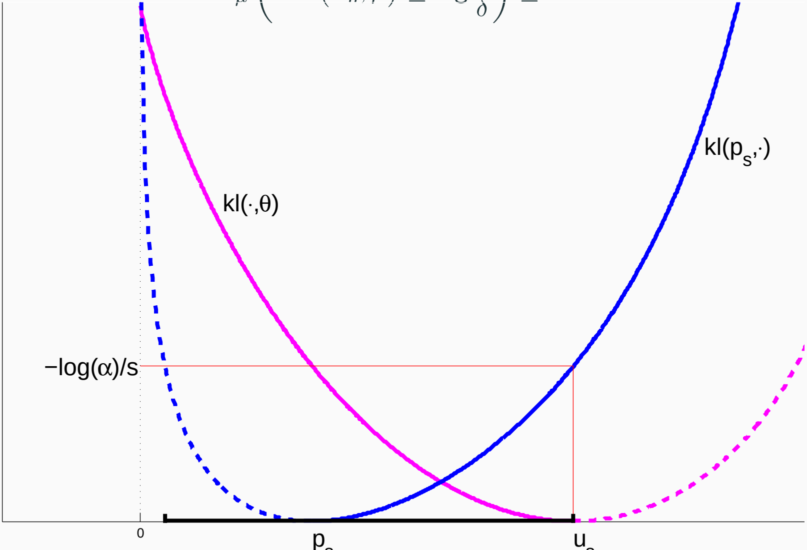
$$\begin{aligned}-\frac{1}{n}\log \mathbb{P}_\mu\left(\bar{X}_n \geq x\right) &\geq \sup_{\lambda>0}\left\{\lambda x - \log \mathbb{E}_\mu[\exp \lambda X_1]\right)\right\} \\ &= \sup_{\lambda>0}\left\{\lambda x - \log\left(1-\mu+\mu e^{\lambda}\right)\right\} \\ &= \mathrm{kl}(x,\mu) .\end{aligned}$$

kl = binary Kullback-Leibler divergence: more generally

$$\mathrm{KL}(P,Q) = \mathbb{E}_{X\sim P}\left[\log \frac{dP}{dQ}(X)\right]$$

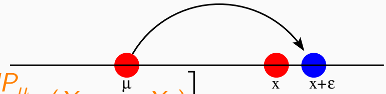
A Divergence on the Set of Possible Means

$$\mathbb{P}_\mu\left(n\,\mathrm{kl}\left(\bar{X}_n, \mu\right) \geq \log\frac{1}{\delta}\right) \leq 2\delta$$



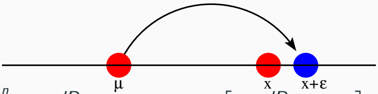
Lower Bound: Change of Measure

For all $\epsilon > 0$ and all $\alpha > 0$,



$$
\begin{aligned}
\mathbb{P}_\mu\left(\bar{X}_n \geq x\right) &= \mathbb{E}_\mu\left[\mathbb{1}\{\bar{X}_n \geq x\}\right] \\
&= \mathbb{E}_{x+\epsilon}\left[\mathbb{1}\{\bar{X}_n \geq x\} \times \frac{dP_\mu}{dP_{x+\epsilon}}\left(X_1, \ldots, X_n\right)\right] \\
&= \mathbb{E}_{x+\epsilon}\left[\mathbb{1}\{\bar{X}_n \geq x\} \times e^{-\sum_{i=1}^n \log \frac{dP_{x+\epsilon}}{dP_\mu}(X_i)}\right] \\
&\geq \mathbb{E}_{x+\epsilon}\left[\mathbb{1}\{\bar{X}_n \geq x\}\ \mathbb{1}\left\{\frac{1}{n}\sum_{i=1}^n \log \frac{dP_{x+\epsilon}}{dP_\mu}\left(X_i\right) \leq \mathbb{E}_{x+\epsilon}\left[\log \frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right] + \alpha\right\}\right. \\
&\qquad\qquad \left. \times\, e^{-\sum_{i=1}^n \log \frac{dP_{x+\epsilon}}{dP_\mu}(X_i)}\right] \\
&\geq e^{-n\left\{\mathbb{E}_{x+\epsilon}\left[\log \frac{dP_{x+\epsilon}}{dP_\mu}(X_i)\right]+\alpha\right\}}\left[1 - \mathbb{P}_{x+\epsilon}\left(\bar{X}_n < x\right)\right. \\
&\qquad \left. - \mathbb{P}_{x+\epsilon}\left(\frac{1}{n}\sum_{i=1}^n \log \frac{dP_{x+\epsilon}}{dP_\mu}\left(X_i\right) > \mathbb{E}_{x+\epsilon}\left[\log \frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right] + \alpha\right)\right] \\
&= e^{-n\left\{\mathrm{kl}(x+\epsilon,\mu)+\alpha\right\}}\left(1 - o_n(1)\right).
\end{aligned}
$$

For all $\epsilon > 0$ and all $\alpha > 0$,



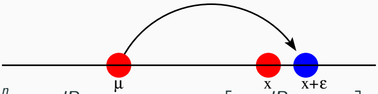
$$
\begin{aligned}
\mathbb{P}_\mu\left(\bar{X}_n \geq x\right) &= \mathbb{E}_\mu\left[\mathbb{1}\{\bar{X}_n \geq x\}\right] \\
&\geq \mathbb{E}_{x+\epsilon}\Bigg[\mathbb{1}\{\bar{X}_n \geq x\}\ \mathbb{1}\left\{\frac{1}{n}\sum_{i=1}^n \log\frac{dP_{x+\epsilon}}{dP_\mu}(X_i) \leq \mathbb{E}_{x+\epsilon}\left[\log\frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right] + \alpha\right\} \\
&\qquad\qquad \times e^{-\sum_{i=1}^n \log\frac{dP_{x+\epsilon}}{dP_\mu}(X_i)}\Bigg] \\
&\geq e^{-n\left\{\mathbb{E}_{x+\epsilon}\left[\log\frac{dP_{x+\epsilon}}{dP_\mu}(X_i)\right]+\alpha\right\}}\Bigg[1 - \mathbb{P}_{x+\epsilon}\left(\bar{X}_n < x\right) \\
&\qquad - \mathbb{P}_{x+\epsilon}\left(\frac{1}{n}\sum_{i=1}^n \log\frac{dP_{x+\epsilon}}{dP_\mu}(X_i) > \mathbb{E}_{x+\epsilon}\left[\log\frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right] + \alpha\right)\Bigg] \\
&= e^{-n\left\{\mathrm{kl}(x+\epsilon,\mu)+\alpha\right\}}\left(1 - o_n(1)\right).
\end{aligned}
$$

Asymptotic Optimality (Large Deviation Lower Bound)

$$
\liminf_n \frac{1}{n}\log\mathbb{P}_\mu\left(\bar{X}_n \geq x\right) \geq -\mathrm{kl}(x,\mu)
$$

Lower Bound: Change of Measure

For all $\epsilon > 0$ and all $\alpha > 0$,



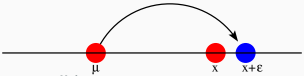
$$
\begin{aligned}
\mathbb{P}_\mu\left(\bar{X}_n \geq x\right) &= \mathbb{E}_\mu\left[\mathbb{1}\{\bar{X}_n \geq x\}\right] \\
&\geq \mathbb{E}_{x+\epsilon}\Bigg[\mathbb{1}\{\bar{X}_n \geq x\}\ \mathbb{1}\left\{\frac{1}{n}\sum_{i=1}^n \log\frac{dP_{x+\epsilon}}{dP_\mu}(X_i) \leq \mathbb{E}_{x+\epsilon}\left[\log\frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right] + \alpha\right\} \\
&\qquad\qquad \times e^{-\sum_{i=1}^n \log\frac{dP_{x+\epsilon}}{dP_\mu}(X_i)}\Bigg] \\
&\geq e^{-n\left\{\mathbb{E}_{x+\epsilon}\left[\log\frac{dP_{x+\epsilon}}{dP_\mu}(X_i)\right]+\alpha\right\}}\Bigg[1 - \mathbb{P}_{x+\epsilon}\left(\bar{X}_n < x\right) \\
&\qquad - \mathbb{P}_{x+\epsilon}\left(\frac{1}{n}\sum_{i=1}^n \log\frac{dP_{x+\epsilon}}{dP_\mu}(X_i) > \mathbb{E}_{x+\epsilon}\left[\log\frac{dP_{x+\epsilon}}{dP_\mu}(X_1)\right] + \alpha\right)\Bigg] \\
&= e^{-n\left\{\mathrm{kl}(x+\epsilon,\mu)+\alpha\right\}}\left(1 - o_n(1)\right) .
\end{aligned}
$$

Asymptotic Optimality (Large Deviation Lower Bound)

$$
\frac{1}{n}\log\mathbb{P}_\mu\left(\bar{X}_n \geq x\right) \underset{n\to\infty}{\longrightarrow} -\mathrm{kl}(x,\mu)
$$

Lower Bound: the Entropy Way

Notation: $\mathcal{KL}(Y,Z) = \mathrm{KL}\big(\mathcal{L}(Y),\mathcal{L}(Z)\big)$.



For all $\epsilon > 0$, if $X_1, \dots, X_n' \overset{iid}{\sim} \mathcal{B}(\mu)$ and $X_1', \dots, X_n' \overset{iid}{\sim} \mathcal{B}(x+\epsilon)$:

$$
\begin{aligned}
n\,\mathrm{kl}(x+\epsilon,\mu) &= \mathrm{KL}\left(\mathcal{B}(x+\epsilon)^{\otimes n}, \mathcal{B}(\mu)^{\otimes n}\right) && \mathrm{KL}(P\otimes P', Q\otimes Q') = \mathrm{KL}(P,Q) + \mathrm{KL}(P',Q') \\
&= \mathcal{KL}\big((X_1',\dots,X_n'),(X_1,\dots,X_n)\big) \\
&\geq \mathcal{KL}\Big(\mathbb{1}\left\{\bar{X}_n' \geq x\right\}, \mathbb{1}\left\{\bar{X}_n \geq x\right\}\Big) && \text{contraction of entropy} = \text{data-processing inequality} \\
&= \mathrm{kl}\Big(\mathbb{P}_{x+\epsilon}\left(\bar{X}_n' \geq x\right), \mathbb{P}_{\mu}\left(\bar{X}_n \geq x\right)\Big) \\
&\geq \mathbb{P}_{x+\epsilon}\left(\bar{X}_n' \geq x\right) \log \frac{1}{\mathbb{P}_{\mu}\left(\bar{X}_n \geq x\right)} - \log(2) && \mathrm{kl}(p,q) \geq p \log \frac{1}{q} - \log 2
\end{aligned}
$$

A non-asymptotic lower bound

$$
\mathbb{P}_{\mu}\left(\bar{X}_n \geq x\right) \geq e^{-\frac{n\,\mathrm{kl}(x+\epsilon,\mu)+\log(2)}{1-e^{-2n\epsilon^2}}}
$$

Identifying the Best Arm with Fixed Confidence

The Stochastic Multi-Armed Bandit Model (MAB)

K arms $= K$ probability distributions (ν_a has mean μ_a, here: $\nu_a = \mathcal{B}(\mu_a)$)

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 $X_t \sim \mathcal{B}(\mu_{A_t})$

using a sequential sampling strategy (A_t):

$$A_{t+1} = \phi_t(A_1, X_1, \ldots, A_t, X_t),$$

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

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

Usual 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]


$\mathcal{B}(\mu_1)$


$\mathcal{B}(\mu_2)$


$\mathcal{B}(\mu_3)$


$\mathcal{B}(\mu_4)$


$\mathcal{B}(\mu_5)$

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

Usual 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 and 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
given $\tau = T$ minimize $\mathbb{P}(\hat{a}_\tau \neq a^*)$	minimize $\mathbb{E}[\tau]$ under constraint $\mathbb{P}(\hat{a}_\tau \neq a^*) \leq \delta$

Motivation: clinical trials, market research, A/B testing...

Wanted: Optimal Algorithms for PAC-BAI

$\mathcal{S}$ a class of bandit models $\nu = (\nu_1, \dots, \nu_K)$.

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

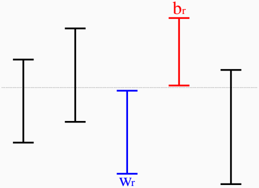
Goal: for some classes $\mathcal{S}$, find

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

(distribution-dependent bounds)

best achievable $\mathbb{E}_\nu[\tau]$ = sample complexity of model ν

Racing Strategy see [Kaufmann & 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

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

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

end

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

Lower Bound on the Sample Complexity

Key Inequality for Lower Bounds in Bandit Models

Let $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ and $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)$ be two bandit models with KL-divergence d ($= \mathrm{kl}$ for Bernoulli models).

Change of distribution lemma [G., Ménard, Stoltz '16]

For every stopping time τ and every $\mathcal{F}_\tau$-measurable variable Z almost surely bounded in $[0, 1]$,

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

- cf lower bound $n\, d(x+\epsilon, \mu) \geq \mathrm{kl}\Big(\mathbb{P}_{x+\epsilon}\left(\bar{X}'_n \geq x\right), \mathbb{P}_{\mu}\left(\bar{X}_n \geq x\right)\Big)$
- Useful if the behaviour of the algorithm (and of Z) is supposed to be very different under $\boldsymbol{\mu}$ and under $\boldsymbol{\lambda}$.
- Permits to prove the famous Lai&Robbins lower bound on regret clearly in a few lines.

Key Inequality for PAC-BAI

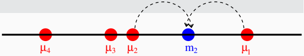
Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$ be two bandit models.

Change of distribution lemma **[Kaufmann, Cappé, G.'15]**

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

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

Using it for each arm separately, 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 $\mathrm{kl}(\delta, 1-\delta) \geq \log\left(\frac{1}{2.4\delta}\right)$

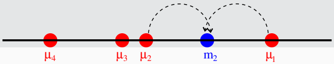
Combining the Inequalities

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

Uniform δ-PAC Constraint **[Kaufmann, Cappé, G. '15]**

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

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



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

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

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

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

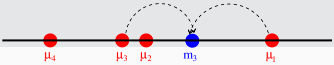
Combining the Inequalities

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

Uniform δ-PAC Constraint **[Kaufmann, Cappé, G. '15]**

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

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



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

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

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

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

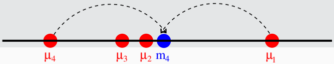
Combining the Inequalities

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

Uniform δ-PAC Constraint [Kaufmann, Cappé, G. '15]

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

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



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

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

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

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

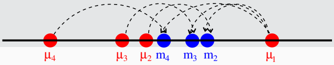
Combining the Inequalities

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

Uniform δ-PAC Constraint [Kaufmann, Cappé, G. '15]

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

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



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

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

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

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

Lower Bound: the Complexity of BAI

Theorem

For any δ-PAC algorithm,

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geq T^*(\boldsymbol{\mu}) \, \mathrm{kl}(\delta, 1-\delta),$$

where

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

- Cf. [Graves and Lai 1997, Vaidhyan and Sundaresan, 2015]
- A kind of game : you choose the proportions of draws $(w_a)_a$, the opponent chooses the alternative

➜ the optimal proportions of arm draws are

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

PAC-BAI as a Game

Given a parameter $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ such that $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$:

- the statistician chooses proportions of arm draws $\boldsymbol{w} = (w_a)_a$
- the opponent chooses an alternative model $\boldsymbol{\lambda}$
- the payoff is the minimal number $T = T(\boldsymbol{w}, \boldsymbol{\lambda})$ of draws necessary to ensure that he does not violate the δ-PAC constraint

$$\sum_{a=1}^{K} T w_a \, d(\mu_a, \lambda_a) \geq \mathrm{kl}(\delta, 1-\delta)$$

- $T^*(\boldsymbol{\mu})$ = value of the game
  $\boldsymbol{w}^*$ = optimal action for the statistician

PAC-BAI as a Game

Given a parameter $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ such that $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$:

- the statistician chooses proportions of arm draws $\boldsymbol{w} = (w_a)_a$
- the opponent chooses an arm $a \in \{2, \ldots, K\}$ and $\lambda_a = \arg\min_\lambda w_1\, d(\mu_1, \lambda) + w_a\, d(\mu_a, \lambda)$
- the payoff is the minimal number $T = T(\boldsymbol{w}, a)$ of draws necessary to ensure that

$$\mathbb{P}_{\boldsymbol{\mu}}\big(\hat{\mu}_{1,Tw_1} \leq \hat{\mu}_{a,Tw_a}\big) \approx \mathbb{P}_{\boldsymbol{\mu}}\big(\hat{\mu}_{1,Tw_1} < \lambda_a \text{ and } \hat{\mu}_{a,Tw_a} \geq \lambda_a\big)$$
$$\leq \exp\Big(- T\big(w_1 \mathrm{kl}(\mu_1, \lambda_a)) + w_a \mathrm{kl}(\mu_a, \lambda_a)\big)\Big) \leq \delta$$

  that is $T(\boldsymbol{w}, a) = \dfrac{\log(1/\delta)}{w_1\, \mathrm{kl}(\mu_1, \lambda_a - \epsilon)) + w_a\, \mathrm{kl}(\mu_a, \lambda)}$

- $T^*(\boldsymbol{\mu})$ = value of the game
  $\boldsymbol{w}^*$ = optimal action for the statistician

Computing the optimal proportions

Computing $\boldsymbol{w}^*$

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

An explicit calculation yields

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

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

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

Computing the optimal proportions

Computing $\boldsymbol{w}^*$

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

An explicit calculation yields

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

where $g_a(x) = d\left(\mu_1, \frac{\mu_1 + x\mu_a}{1+x}\right) + x d\left(\mu_a, \frac{\mu_1 + x\mu_a}{1+x}\right)$ (Jensen-Shannon divergence)

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

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

Computing the optimal proportions

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

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

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

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

Computing the optimal proportions

Theorem

For every $a \in \{1, \dots, K\}$,

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

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 $w^*(\boldsymbol{\mu})$

Properties of $T^*(\mu)$ and $w^*(\mu)$

1. For all $\boldsymbol{\mu} \in \mathcal{S}$, for all a, $w_a^*(\boldsymbol{\mu}) > 0$
2. w^* is continuous in every $\boldsymbol{\mu} \in \mathcal{S}$
3. If $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$, one has $w_2^*(\boldsymbol{\mu}) \geq \cdots \geq w_K^*(\boldsymbol{\mu})$
   (one may have $w_1^*(\boldsymbol{\mu}) < w_2^*(\boldsymbol{\mu})$)
4. Case of two arms [Kaufmann, Cappé, G. '14]:

   $$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq \frac{\mathrm{kl}(\delta, 1-\delta)}{d_*(\mu_1, \mu_2)} .$$

   where d_* is the 'reversed' Chernoff information

   $$d_*(\mu_1, \mu_2) := d(\mu_1, \mu_*) = d(\mu_2, \mu_*).$$

5. Gaussian arms : algebraic equation but no simple formula when $K \geq 3$, only:

   $$\sum_{a=1}^{K} \frac{2\sigma^2}{\Delta_a^2} \leq T^*(\boldsymbol{\mu}) \leq 2 \sum_{a=1}^{K} \frac{2\sigma^2}{\Delta_a^2}.$$

The Track-and-Stop Strategy

Sampling rule: Tracking the optimal proportions

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

Introducing

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

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,

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

Sequential Generalized Likelihood Test

High values of the Generalized Likelihood Ratio

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

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:

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

Chernoff stopping rule [Chernoff '59]

Stopping Rule: Alternative Formulations

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

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

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 Formulations

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,

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

Calibration

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,

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

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

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

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

Stopping rule: δ-PAC property

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

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

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

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

Stopping rule: δ-PAC property

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

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

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

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

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

Asymptotic Optimality of the T&S strategy

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

is δ-PAC for every $\delta \in (0,1)$ and satisfies

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

Sketch of proof (almost-sure convergence only)

- forced exploration $\implies$ $N_a(t) \to \infty$ a.s. for all $a \in \{1, \dots, K\}$

➜ $\boldsymbol{\mu}(t) \to \boldsymbol{\mu}$ a.s.

➜ $\boldsymbol{w}^*\left(\hat{\mu}(t)\right) \to \boldsymbol{w}^*$ a.s.

➜ tracking rule: $\dfrac{N_a(t)}{t} \underset{t\to\infty}{\to} w_a^*$ a.s.

- but the mapping $F : (\boldsymbol{\mu}', w) \mapsto \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\mu}')} \sum_{a=1}^{K} w_a d(\mu'_a, \lambda_a)$ is continuous at $(\boldsymbol{\mu}, w^*(\boldsymbol{\mu}))$:

➜ as $\max_a \min_{b\neq a} Z_{a,b}(t) = t\, F\left(\hat{\mu}(t), (N_a(t)/t)_{a=1}^{K}\right)$, for every $\epsilon > 0$ there exists t_0 such that

$$t \geq t_0 \;\Rightarrow\; \max_a \min_{b\neq a} Z_{a,b}(t) \geq (1+\epsilon)^{-1} T^*(\boldsymbol{\mu})^{-1} t$$

$\implies$ Thus $\tau \leq t_0 \wedge \inf\left\{t \in \mathbb{N} : (1+\epsilon)^{-1} T^*(\boldsymbol{\mu})^{-1} t \geq \log(2(K-1)t/\delta)\right\}$

and $\limsup_{\delta\to 0} \dfrac{\tau}{\log(1/\delta)} \leq (1+\epsilon) T^*(\boldsymbol{\mu})$.

Numerical Experiments

- $\boldsymbol{\mu}_1 = [0.5\ 0.45\ 0.43\ 0.4]$ ➜ $w^*(\boldsymbol{\mu}_1) = [0.42\ 0.39\ 0.14\ 0.06]$
- $\boldsymbol{\mu}_2 = [0.3\ 0.21\ 0.2\ 0.19\ 0.18]$ ➜ $w^*(\boldsymbol{\mu}_2) = [0.34\ 0.25\ 0.18\ 0.13\ 0.10]$

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

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

Table 1: Expected number of draws $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$ for $\delta = 0.1$, averaged over $N = 3000$ experiments.

➜ Empirically good even for large values of the risk δ

➜ Racing is sub-optimal in general, because it plays $w_1 = w_2$

Perspectives

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:

- (easy) find an ϵ-optimal arm
- give a simple algorithm with a finite-time analysis
- extend to structured and continuous settings

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):715743, 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, COLT'16, New York, arXiv:1602.04589
- A. Garivier, P. Ménard, G. Stoltz. Explore First, Exploit Next: The True Shape of Regret in Bandit Problems.
- 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.