The information complexity of best-arm identification

Emilie Kaufmann,

joint work with Olivier Cappé and Aurélien Garivier

MAB workshop, Lancaster, January 11th, 2016

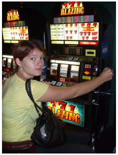
Context: the multi-armed bandit model (MAB)

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
- 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

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 $\mathbb{E}[\tau]$
minimize $\mathbb{P}(\hat{a}_\tau \neq a^*)$	$\mathbb{P}(\hat{a}_\tau \neq a^*) \leq \delta$

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 !

$\Rightarrow$ **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

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

Goal: for some classes $\mathcal{M}$, and $\nu \in \mathcal{M}$, 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)

Outline

1. Regret minimization
2. Lower bound on the sample complexity
3. The complexity of A/B Testing
4. Algorithms for the general case

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 $\mathcal{P}$,

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

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

Example: Bernoulli distributions

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

Outline

Optimal algorithms for regret minimization

$\nu = (\nu^{\mu_1}, \dots, \nu^{\mu_K}) \in \mathcal{M} = (\mathcal{P})^K$.

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

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

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

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

Definition

A bandit algorithm is **asymptotically optimal** if, for every $\nu \in \mathcal{M}$,

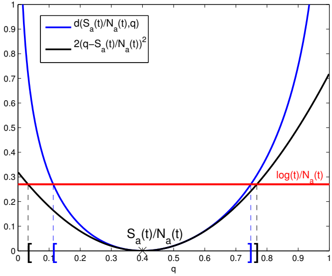
$$\mu_a < \mu^* \Rightarrow \limsup_{T \to \infty} \frac{\mathbb{E}_\nu[N_a(T)]}{\log T} \leq \frac{1}{d(\mu_a, \mu^*)}$$

KL-UCB: an asymptotically optimal algorithm

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

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

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



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

The information complexity of regret minimization

Letting

$$\kappa_R(\nu) := \inf_{\mathcal{A} \text{ consistent}} \liminf_{T\to\infty} \frac{R_T(\nu)}{\log(T)},$$

we showed that

$$\kappa_R(\nu) = \sum_{a=1}^{K} \frac{(\mu^* - \mu_a)}{\color{red}{d(\mu_a, \mu^*)}}.$$

Outline

A general lower bound

$\mathcal{M}$ a class of exponential family bandit models

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

Theorem [K.,Cappé, Garivier 15]

Let $\nu = (\nu^{\mu_1}, \dots, \nu^{\mu_K})$ be such that $\mu_1 > \mu_2 \geq \dots \geq \mu_K$.
Let $\delta \in]0,1[$. Any algorithm that is δ-PAC on $\mathcal{M}$ satisfies

$$\mathbb{E}_\nu[\tau] \geq \left(\frac{1}{d(\mu_1, \mu_2)} + \sum_{a=2}^{K} \frac{1}{d(\mu_a, \mu_1)} \right) \log \left(\frac{1}{2.4\delta} \right).$$

$$d(\mu, \mu') = \mathrm{KL}(\nu^{\mu}, \nu^{\mu'})$$

Behind the lower bound: changes of distribution

Lemma [K., Cappé, Garivier 2015]

$\nu = (\nu_1, \nu_2, \dots, \nu_K)$, $\nu' = (\nu'_1, \nu'_2, \dots, \nu'_K)$ two bandit models.

$$\sum_{a=1}^{K} \mathbb{E}_\nu[N_a(\tau)]\mathrm{KL}(\nu_a, \nu'_a) \geq \sup_{\mathcal{E} \in \mathcal{F}_\tau} \mathrm{kl}(\mathbb{P}_\nu(\mathcal{E}), \mathbb{P}_{\nu'}(\mathcal{E})).$$

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

Behind the lower bound: changes of distribution

Lemma [K., Cappé, Garivier 2015]

$\nu = (\nu_1, \nu_2, \ldots, \nu_K)$, $\nu' = (\nu'_1, \nu'_2, \ldots, \nu'_K)$ two bandit models.

$$\sum_{a=1}^{K} \mathbb{E}_\nu[N_a(\tau)]\mathrm{KL}(\nu_a, \nu'_a) \geq \sup_{\mathcal{E}\in\mathcal{F}_\tau} \mathrm{kl}(\mathbb{P}_\nu(\mathcal{E}), \mathbb{P}_{\nu'}(\mathcal{E})).$$

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

L_t the log-likelihood ratio of past observations under ν and ν':

- Wald's equality: $\mathbb{E}_\nu[L_\tau] = \sum_{a=1}^{K} \mathbb{E}_\nu[N_a(\tau)]\mathrm{KL}(\nu_a, \nu'_a)$
- change of distribution: $\forall \mathcal{E} \in \mathcal{F}_\tau,\ \mathbb{P}_{\nu'}(\mathcal{E}) = \mathbb{E}_\nu\left[\exp(-L_\tau)\mathbb{1}_{\mathcal{E}}\right]$

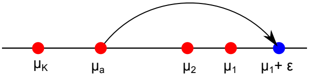
Behind the lower bound: changes of distribution

Exponential bandits: $\nu = (\mu_1, \mu_2, \ldots, \mu_K)$, $\nu' = (\mu'_1, \mu'_2, \ldots, \mu'_K)$

$$\forall \mathcal{E} \in \mathcal{F}_\tau, \quad \sum_{a=1}^{K} \mathbb{E}_\nu[N_a(\tau)] d(\mu_a, \mu'_a) \geq \mathrm{kl}(\mathbb{P}_\nu(\mathcal{E}), \mathbb{P}_{\nu'}(\mathcal{E})).$$

$\mathbb{E}_\nu[\tau] = \sum_{a=1}^{K} \mathbb{E}_\nu[N_a(\tau)]$. Then, for $a \neq 1$,

① choose ν' such that arm 1 is no longer the best :



$$\begin{cases} \mu'_a &= \mu_1 + \epsilon \\ \mu'_i &= \mu_i, \quad \text{if } i \neq a \end{cases}$$

② $\mathcal{E} = (\hat{a}_\tau = 1)$: $\mathbb{P}_\nu(\mathcal{E}) \geq 1 - \delta$ and $\mathbb{P}_{\nu'}(\mathcal{E}) \leq \delta$.

$$\begin{aligned} \mathbb{E}_\nu[N_a(\tau)] d(\mu_a, \mu_1 + \epsilon) &\geq kl(\delta, 1 - \delta) \\ \mathbb{E}_\nu[N_a(\tau)] &\geq \frac{1}{d(\mu_a, \mu_1)} \log\left(\frac{1}{2.4\delta}\right). \end{aligned}$$

The complexity of best arm identification

$\mathcal{M}$ a class of exponential family bandit models

Theorem [K.,Cappé, Garivier 15]

Let $\nu = (\nu^{\mu_1}, \dots, \nu^{\mu_K})$ be such that $\mu_1 > \mu_2 \geq \dots \geq \mu_K$.
Let $\delta \in]0,1[$. Any algorithm that is δ-PAC on $\mathcal{M}$ satisfies

$$\mathbb{E}_\nu[\tau] \geq \left(\frac{1}{d(\mu_1, \mu_2)} + \sum_{a=2}^{K} \frac{1}{d(\mu_a, \mu_1)} \right) \log\left(\frac{1}{2.4\delta}\right).$$

The complexity of best arm identification

$\mathcal{M}$ a class of exponential family bandit models

Theorem [K.,Cappé, Garivier 15]

Let $\nu = (\nu^{\mu_1}, \dots, \nu^{\mu_K})$ be such that $\mu_1 > \mu_2 \geq \dots \geq \mu_K$.
Let $\delta \in]0,1[$. Any algorithm that is δ-PAC on $\mathcal{M}$ satisfies

$$\mathbb{E}_\nu[\tau] \geq \left(\frac{1}{d(\mu_1,\mu_2)} + \sum_{a=2}^{K} \frac{1}{d(\mu_a,\mu_1)}\right) \log\left(\frac{1}{2.4\delta}\right).$$

- For any class $\mathcal{M}$, the complexity term of $\nu \in \mathcal{M}$ is defined as

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

$\mathcal{A} = (\mathcal{A}(\delta))$ is PAC if for all $\delta \in]0,1[$, $\mathcal{A}(\delta)$ is δ-PAC on $\mathcal{M}$.

Outline

Computing the complexity term

$\mathcal{M}$ a class of two-armed bandit models. For $\nu = (\nu_1, \nu_2)$ recall that

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

We now compute $\kappa_C(\nu)$ for two types of classes $\mathcal{M}$:

- Exponential family bandit models:

$$\mathcal{M} = \{\nu = (\nu^{\mu_1}, \nu^{\mu_2}) : \nu^\mu \in \mathcal{P}, \mu_1 \neq \mu_2\}$$

- Gaussian with known variances σ_1^2 and σ_2^2:

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

From our previous lower bound (or a similar method)

- Exponential family bandit models:

$$\kappa_C(\nu) \geq \frac{1}{d(\mu_1, \mu_2)} + \frac{1}{d(\mu_2, \mu_1)}.$$

- Gaussian with known variances σ_1^2, σ_2^2:

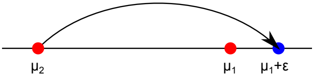
$$\kappa_C(\nu) \geq \frac{2\sigma_1^2}{(\mu_1 - \mu_2)^2} + \frac{2\sigma_2^2}{(\mu_2 - \mu_1)^2}.$$

Towards tighter lower bounds

Exponential bandits: $\nu = (\mu_1, \mu_2)$, $\nu' = (\mu'_1, \mu'_2)$: $\mu'_1 < \mu'_2$

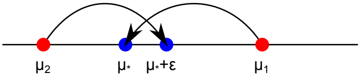
$$\mathbb{E}_\nu[N_1(\tau)]d(\mu_1, \mu'_1) + \mathbb{E}_\nu[N_2(\tau)]d(\mu_2, \mu'_2) \geq \log\left(\frac{1}{2.4\delta}\right).$$

previously,



$$\begin{aligned} \mu'_1 &= \mu_1 \\ \mu'_2 &= \mu_1 + \epsilon \end{aligned}$$

a new change of distribution:



$$\begin{aligned} \mu'_1 &= \mu_* \\ \mu'_2 &= \mu_* + \epsilon \end{aligned}$$

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

$$\begin{aligned} d_*(\mu_1, \mu_2)\mathbb{E}_\nu[\tau] &\geq \log\left(\frac{1}{2.4\delta}\right) \\ \mathbb{E}_\nu[\tau] &\geq \frac{1}{d_*(\mu_1, \mu_2)} \log\left(\frac{1}{2.4\delta}\right) \end{aligned}$$

Tighter lower bounds in the two-armed case

- New lower bounds (tighter!)

Exponential families	Gaussian with known σ_1^2, σ_2^2
$\kappa_C(\nu) \geq \frac{1}{d_*(\mu_1,\mu_2)}$	$\kappa_C(\nu) \geq \frac{2(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2}$

$d_*(\mu_1, \mu_2) := d(\mu_1, \mu_*) = d(\mu_2, \mu_*)$ is a Chernoff information.

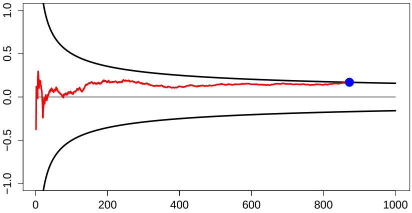
- Previous lower bounds

Exponential families	Gaussian with known σ_1^2, σ_2^2
$\kappa_C(\nu) \geq \frac{1}{d(\mu_1,\mu_2)} + \frac{1}{d(\mu_2,\mu_1)}$	$\kappa_C(\nu) \geq \frac{2(\sigma_1^2+\sigma_2^2)}{(\mu_1-\mu_2)^2}$

Upper bounds on the complexity: algorithms

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

The α-Elimination algorithm with exploration rate $\beta(t, \delta)$:



➜ chooses A_t in order to keep a proportion $N_1(t)/t \simeq \alpha$
i.e. $A_t = 2$ if and only if $\lceil \alpha t \rceil = \lceil \alpha(t+1) \rceil$

➜ if $\hat{\mu}_a(t)$ is the empirical mean of rewards obtained from a up to time t, $\sigma_t^2(\alpha) = \sigma_1^2/\lceil \alpha t \rceil + \sigma_2^2/(t - \lceil \alpha t \rceil)$,

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

Gaussian case: matching algorithm

Theorem [K., Cappé, Garivier 14]

With $\alpha = \frac{\sigma_1}{\sigma_1 + \sigma_2}$ and $\beta(t,\delta) = \log \frac{t}{\delta} + 2\log\log(6t)$,

α-Elimination is δ-PAC and

$$\forall \epsilon > 0, \quad \mathbb{E}_\nu[\tau] \leq (1+\epsilon)\frac{2(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2}\log\left(\frac{1}{\delta}\right) + \underset{\delta\to 0}{o_\epsilon}\left(\log\frac{1}{\delta}\right)$$

In the Gaussian case,

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

and finally

$$\boxed{\kappa_C(\nu) = \frac{2(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2}}.$$

$$\mathcal{M} = \{\nu = (\nu^{\mu_1}, \nu^{\mu_2}) : \nu^{\mu} \in \mathcal{P}, \mu_1 \neq \mu_2\}$$

Another lower bound...

A δ-PAC algorithm using uniform sampling ($A_t = t[2]$) satisfy

$$\mathbb{E}_\nu[\tau] \geq \frac{1}{I_*(\mu_1, \mu_2)} \log\left(\frac{1}{2.4\delta}\right)$$

with

$$I_*(\mu_1, \mu_2) = \frac{d\left(\mu_1, \frac{\mu_1+\mu_2}{2}\right) + d\left(\mu_2, \frac{\mu_1+\mu_2}{2}\right)}{2}.$$

Remark: $I_*(\mu_1, \mu_2)$ is very close to $d_*(\mu_1, \mu_2)$...

$\Rightarrow$ find a good strategy with a uniform sampling strategy !

Exponential families: uniform sampling

- For Bernoulli bandit models, uniform sampling and

$$\tau = \inf \left\{ t \in \mathbb{N}^* : |\hat{\mu}_1(t) - \hat{\mu}_2(t)| > \log\left(\frac{t}{\delta}\right) \right\}$$

is δ-PAC but *not* optimal: $\frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)} \simeq \frac{2}{(\mu_1 - \mu_2)^2} > \frac{1}{I_*(\mu_1,\mu_2)}$.

SGLRT algorithm (Sequential Generalized Likelihood Ratio Test)

Let $\alpha > 0$. There exists $C = C_\alpha$ such that the algorithm using a uniform sampling strategy and the stopping rule

$$\tau = \inf \left\{ t \in \mathbb{N}^* : t I_*(\hat{\mu}_1(t), \hat{\mu}_2(t)) > \beta(t, \delta) \right\}$$

with $\beta(t,\delta) = \log\left(\frac{Ct^{1+\alpha}}{\delta}\right)$ is δ-PAC and

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)} \leq \frac{1+\alpha}{I_*(\mu_1, \mu_2)}.$$

$$\kappa_C(\nu) \leq \frac{1}{I_*(\mu_1, \mu_2)}$$

The complexity of A/B Testing

- For Gaussian bandit models with known variances σ_1^2 and σ_2^2, if $\nu = (\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2))$,

$$\boxed{\kappa_C(\nu) = \frac{(\sigma_1 + \sigma_2)^2}{2(\mu_1 + \mu_2)^2}}$$

  and the optimal strategy draws the arms proportionally to their standard deviation.

- For exponential bandit models, if $\nu = (\nu^{\mu_1}, \nu^{\mu_2})$,

$$\boxed{\frac{1}{d_*(\mu_1, \mu_2)} \leq \kappa_C(\nu) \leq \frac{1}{I_*(\mu_1, \mu_2)}}$$

  and uniform sampling is close to optimal.

Outline

Racing (or elimination) algorithms

$\mathcal{S} = \{1, \dots, K\}$ set of remaining arms
$r = 0$ current round
while $|\mathcal{S}| > 1$

- r=r+1
- draw each $a \in \mathcal{S}$, compute $\hat{\mu}_{a,r}$, the empirical mean of the r samples observed sofar
- compute the empirical best and empirical worst arms:

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

- if $\text{EliminationRule}(r, b_r, w_r)$, eliminate w_r : $\mathcal{S} = \mathcal{S} \backslash \{w_r\}$

end

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

Elimination rules

In the literature:

- Successive Elimination for Bernoulli bandits

$$\text{Elimination}(r, a, b) = \left(\hat{\mu}_{a,r} - \hat{\mu}_{b,r} > \sqrt{\frac{\log\left(cKt^2/\delta\right)}{r}}\right)$$

  [Even Dar et al. 06]

- KL-Racing for exponential family bandits

$$\text{Elimination}(r, a, b) = (l_{a,r} > u_{b,r})$$

  with

$$\left\{\begin{array}{lcl} l_{a,r} & = & \min\{x : rd(\hat{\mu}_{a,r}, x) \leq \beta(r,\delta)\} \\ u_{b,r} & = & \max\{x : rd(\hat{\mu}_{b,r}, x) \leq \beta(r,\delta)\} \end{array}\right.$$

  [K. and Kalyanakrishnan 13]

The Racing-SGLRT algorithm

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

Analysis of Racing-SGLRT

Let $\alpha > 0$. For an exploration rate of the form

$$\beta(r,\delta) = \log\left(\frac{Ct^{1+\alpha}}{\delta}\right),$$

Racing-SGLRT is δ-PAC and satifies

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)} \leq (1+\alpha)\left(\sum_{a=2}^{K} \frac{1}{I_*(\mu_a, \mu_1)}\right).$$

Conclusion

We presented:

- a simple methodology to derive lower bounds on the sample complexity of a δ-PAC strategy
- a characterization of the complexity of best arm identification among two-arm, involving alternative information-theoretic quantities (e.g. Chernoff information)

To be continued...

- A/B Testing: for which classes of distributions is uniform sampling a good idea?
- the complexity of best arm identification is still to be understood in the general case...

$$\frac{1}{d_*(\mu_1,\mu_2)} + \sum_{a=2}^{K} \frac{1}{d(\mu_a,\mu_1)} \leq \kappa_C(\nu) \leq \sum_{a=2}^{K} \frac{1}{I_*(\mu_a,\mu_1)}$$

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.
- E. Even-Dar, S. Mannor, Y. Mansour, Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. *JMLR*, 2006.
- E. Kaufmann, S. Kalyanakrishnan, Information Complexity in Bandit Subset Selection. In *Proceedings of the 26th Conference On Learning Theory* (COLT), 2013.
- E. Kaufmann, O. Cappé, A. Garivier. On the Complexity of A/B Testing. In *Proceedings of the 27th Conference On Learning Theory* (COLT), 2014.
- E. Kaufmann, O. Cappé, A. Garivier. On the Complexity of Best Arm Identification in Multi-Armed Bandit Models. *JMLR*, 2015
- T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 1985.