Quelques outils statistiques pour la prise de décision séquentielle

Emilie Kaufmann (CRIStAL)

GRETSI, Lille, 27 août 2019

The multi-armed bandit model

K arms $\leftrightarrow$ 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
- receives a a sample $X_t \sim \nu_{A_t}$

Sequential sampling strategy (**bandit algorithm**) :

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

A reinforcement learning problem ?

K arms $\leftrightarrow$ 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
- receives a a reward $X_t \sim \nu_{A_t}$

Sequential sampling strategy (**bandit algorithm**) :

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

Possible goal : maximize the sum of collected rewards $\mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$.

Clinical trials

Historical motivation [Thompson, 1933]

$\mathcal{B}(\mu_1)$

$\mathcal{B}(\mu_2)$

$\mathcal{B}(\mu_3)$

$\mathcal{B}(\mu_4)$

$\mathcal{B}(\mu_5)$

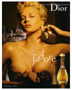
For the t-th patient in a clinical study,

- chooses a treatment A_t
- observes a response $X_t \in \{0, 1\} : \mathbb{P}(X_t = 1 | A_t = a) = \mu_a$

Goal : Maximize the expected number of patients healed

Online content optimization

Modern motivation ($$) [Li et al., 2010]
(recommender systems, online advertisement)




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

For the t-th visitor of a website,

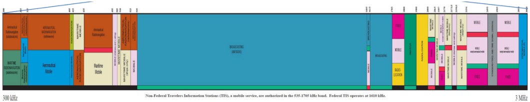
- display an advertisement A_t
- observe a possible click $X_t \sim \mathcal{B}(\mu_{A_t})$

Goal : Maximize the total number of clicks

Cognitive radios

Opportunistic spectrum access

[Jouini et al., 2009, Anandkumar et al., 2010]



streams indicating channel quality :

Channel 1	$X_{1,1}$	$X_{1,2}$	$\dots$	$X_{1,t}$	$\dots$	$X_{1,T}$	$\sim \nu_1$
Channel 2	$X_{2,1}$	$X_{2,2}$	$\dots$	$X_{2,t}$	$\dots$	$X_{2,T}$	$\sim \nu_2$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	
Channel K	$X_{K,1}$	$X_{K,2}$	$\dots$	$X_{K,t}$	$\dots$	$X_{K,T}$	$\sim \nu_K$

At round t, the device :

- selects a channel A_t
- observes the quality of its communication $X_t = X_{A_t,t} \in [0,1]$

Goal : Maximize the overall quality of communications

A performance measure : Regret

$$\mu_\star = \max_{a \in \{1,\dots,K\}} \mu_a \qquad a_\star = \underset{a \in \{1,\dots,K\}}{\text{argmax}}\ \mu_a.$$

Maximizing rewards $\leftrightarrow$ selecting $a_\star$ as much as possible
$\leftrightarrow$ minimizing the regret [Robbins, 52]

$$\mathcal{R}_\nu(\mathcal{A}, T) := \underbrace{T\mu_\star}_{\substack{\text{sum of rewards of} \\ \text{an oracle strategy} \\ \text{always selecting } a_\star}} - \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} X_t\right]}_{\substack{\text{sum of rewards of} \\ \text{the strategy } \mathcal{A}}}$$

Regret decomposition

$$\mathcal{R}_\nu(\mathcal{A}, T) = \sum_{a=1}^{K} \mathbb{E}_\nu[N_a(T)](\mu_\star - \mu_a)$$

$N_a(T)$: number of selections of arm a up to round T.

➜ Wanted : $\mathcal{R}_\nu(\mathcal{A}, T) = o(T)$

A performance measure : Regret

$$\mu_\star = \max_{a \in \{1,\dots,K\}} \mu_a \qquad a_\star = \underset{a \in \{1,\dots,K\}}{\text{argmax}}\ \mu_a.$$

Maximizing rewards $\leftrightarrow$ selecting $a_\star$ as much as possible
$\leftrightarrow$ minimizing the regret [Robbins, 52]

$$\mathcal{R}_\nu(\mathcal{A}, T) := \underbrace{T\mu_\star}_{\substack{\text{sum of rewards of} \\ \text{an oracle strategy} \\ \text{always selecting } a_\star}} - \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} X_t\right]}_{\substack{\text{sum of rewards of} \\ \text{the strategy} \mathcal{A}}}$$

Regret decomposition

$$\mathcal{R}_\nu(\mathcal{A}, T) = \sum_{a=1}^{K} \mathbb{E}_\nu[N_a(T)](\mu_\star - \mu_a)$$

$N_a(T)$: number of selections of arm a up to round T.

➜ sub-linear regret requires an exploration/exploitation trade-off

How to minimize regret ?

- **Idea 1 :**

Draw each arm T/K times

$\Rightarrow$ EXPLORATION

- **Idea 2 :** Always trust the empirical best arm

$$A_{t+1} = \underset{a \in \{1,\dots,K\}}{\text{argmax}} \ \hat{\mu}_a(t)$$

where

$$\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} X_s \mathbb{1}_{(A_s = a)}$$

is an estimate of the unknown mean μ_a.

$\Rightarrow$ EXPLOITATION

Linear regret...

How to minimize regret ?

- **Idea 1 :**

Draw each arm T/K times

$\Rightarrow$ EXPLORATION

- **Idea 2 :** Always trust the empirical best arm

$$A_{t+1} = \underset{a \in \{1,\dots,K\}}{\text{argmax}} \ \hat{\mu}_a(t)$$

where

$$\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} X_s \mathbb{1}_{(A_s = a)}$$

is an estimate of the unknown mean μ_a.

$\Rightarrow$ EXPLOITATION

Linear regret...

- **A Better Idea :** Mix Exploration and Exploitation

The optimism principle

Step 1 : construct a set of statistically plausible models

- For each arm a, build a confidence interval on the mean μ_a :

$$\mathcal{I}_a(t) = [\mathrm{LCB}_a(t), \mathrm{UCB}_a(t)]$$

LCB = Lower Confidence Bound
UCB = Upper Confidence Bound

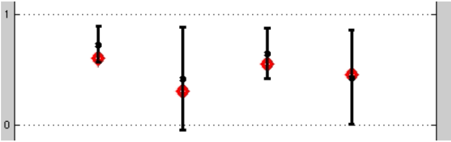
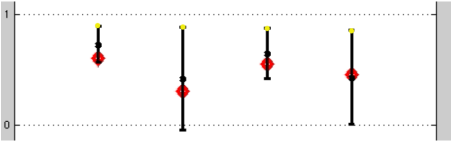


FIGURE – Confidence intervals on the means after t rounds

The optimism principle

Step 2 : act as if the best possible model were the true model

(optimism in face of uncertainty)



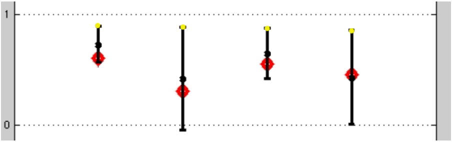
- That is, select

$$A_{t+1} = \underset{a=1,\dots,K}{\text{argmax}}\ \text{UCB}_a(t).$$

[Agrawal, 1995, Katehakis and Robbins, 1995, Auer, 2002, Audibert et al., 2009, Cappé et al., 2013] and others

The optimism principle

Step 2 : act as if the best possible model were the true model
(optimism in face of uncertainty)



- That is, select

$$A_{t+1} = \underset{a=1,\dots,K}{\text{argmax}} \ \text{UCB}_a(t).$$

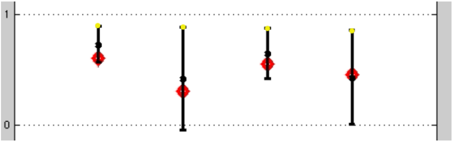
[Agrawal, 1995, Katehakis and Robbins, 1995, Auer, 2002, Audibert et al., 2009, Cappé et al., 2013] and others

$$\mathbb{P}\left(\text{UCB}_a(t) > \mu_a\right) \gtrsim 1 - \frac{1}{t}$$

The optimism principle

Step 2 : act as if the best possible model were the true model

(optimism in face of uncertainty)



- That is, select

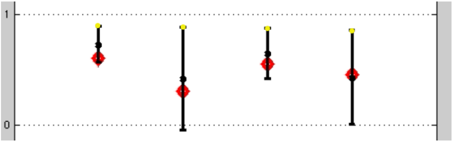
$$A_{t+1} = \underset{a=1,\dots,K}{\text{argmax}} \ \text{UCB}_a(t).$$

[Agrawal, 1995, Katehakis and Robbins, 1995, Auer, 2002, Audibert et al., 2009, Cappé et al., 2013] and others

Example : $\text{UCB}_a(t) = \hat{\mu}_a(t) + \sqrt{\frac{\ln(t)}{2N_a(t)}}$ [Auer, 2002]

The optimism principle

Step 2 : act as if the best possible model were the true model
(optimism in face of uncertainty)



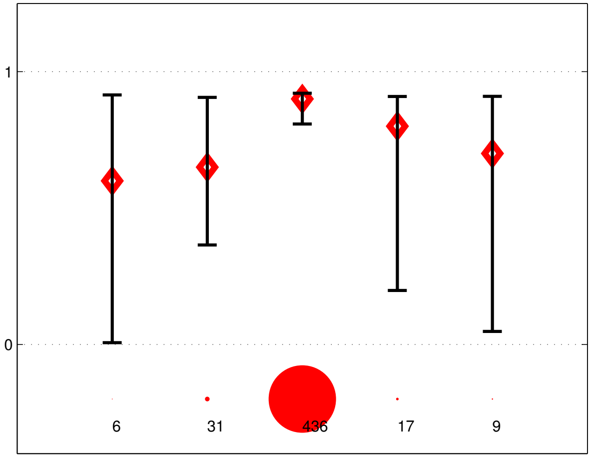
- That is, select

$$A_{t+1} = \underset{a=1,\dots,K}{\mathrm{argmax}}\ \mathrm{UCB}_a(t).$$

[Agrawal, 1995, Katehakis and Robbins, 1995, Auer, 2002, Audibert et al., 2009, Cappé et al., 2013] and others

Example : $\mathrm{UCB}_a(t) = \max\left\{q : N_a(t)\mathrm{kl}(\hat{\mu}_a(t), q) \leq \ln(t)\right\}$

A UCB algorithm in action



A Bayesian algorithm : Thompson Sampling

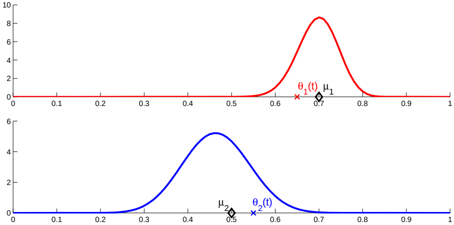
Two equivalent interpretations :

- "randomize the arm selection so that the probability to select an arm is equal to its posterior probability of being the best arm" [Thompson, 1933]
- "sample a possible bandit model from the posterior distribution and act optimally in this sampled model"

$\neq$ optimistic

Thompson Sampling : a randomized Bayesian algorithm

$$\left\{ \begin{array}{l} \forall a \in \{1..K\}, \quad \theta_a(t) \sim \pi_a(t) \\ A_{t+1} = \underset{a=1\dots K}{\text{argmax}} \; \theta_a(t). \end{array} \right.$$



Regret minimization is "solved" (in simple cases)

Example : Bernoulli bandit model $\nu = (\mathcal{B}(\mu_1), \dots, \mathcal{B}(\mu_K))$

A regret lower bound

[Lai and Robbins, 1985] : any uniformly efficient bandit algorithm satisfies

$$\mu_a < \mu_\star \Rightarrow \liminf_{T \to \infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]}{\ln T} \geq \frac{1}{\mathrm{kl}(\mu_a, \mu_\star)},$$

where

$$\mathrm{kl}(\mu, \mu') = \mathrm{KL}(\mathcal{B}(\mu), \mathcal{B}(\mu')) = \mu \ln\left(\frac{\mu}{\mu'}\right) + (1-\mu)\ln\left(\frac{1-\mu}{1-\mu'}\right).$$

Matching upper bounds

kl-UCB and Thompson Sampling satisfy, for any sub-optimal arm a,

$$\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] \leq \frac{\ln(T)}{\mathrm{kl}(\mu_a, \mu_\star)} + o(\ln(T)).$$

[Cappé et al., 2013, Kaufmann et al., 2012, Agrawal and Goyal, 2013]

But... should we maximize rewards ?

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

Best treatment : $a_\star = \underset{a=1,\dots,K}{\mathrm{argmax}}\ \mu_a$

Sequential protocol : for the t-th patient,

- choose a treatment A_t
- observe a response $X_t \in \{0,1\}$: $\mathbb{P}(X_t = 1) = \mu_{A_t}$

Maximize rewards $\leftrightarrow$ cure as many patients as possible

But... should we maximize rewards ?





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

Best treatment : $a_\star = \underset{a=1,\dots,K}{\text{argmax}}\ \mu_a$

Sequential protocol : for the t-th patient,

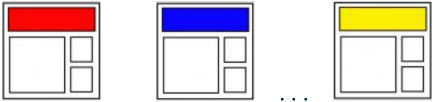
- choose a treatment A_t
- observe a response $X_t \in \{0,1\}$: $\mathbb{P}(X_t = 1) = \mu_{A_t}$

Maximize rewards $\leftrightarrow$ cure as many patients as possible

Alternative goal : identify as quickly as possible the best treatment
(without trying to cure patients during the study)

But... should we maximize rewards ?

Probability that some version of a website generates a conversion :



μ_1 μ_2 μ_K

Best version : $a_\star = \underset{a=1,\dots,K}{\text{argmax}}\ \mu_a$

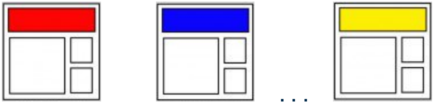
Sequential protocol : for the t-th visitor :

- display version A_t
- observe conversion indicator $X_t \sim \mathcal{B}(\mu_{A_t})$.

Maximize rewards $\leftrightarrow$ maximize the number of conversions

But... should we maximize rewards ?

Probability that some version of a website generates a conversion :



μ_1 μ_2 μ_K

Best version : $a_\star = \underset{a=1,\dots,K}{\text{argmax}}\ \mu_a$

Sequential protocol : for the t-th visitor :

- display version A_t
- observe conversion indicator $X_t \sim \mathcal{B}(\mu_{A_t})$.

Maximize rewards $\leftrightarrow$ maximize the number of conversions

Alternative goal : identify the best version
(without trying to maximize conversions during the test)

Outline

1. Optimal Best Arm Identification
2. Active Identification in a Bandit Model
3. A Particular Case : Murphy Sampling

based on joint works with Aurélien Garivier & Wouter Koolen

Outline

based on joint works with Aurélien Garivier & Wouter Koolen

Best Arm Identification

Assumption : Bernoulli bandit model (can be extended to any one-dimensional exponential family)

$$\boldsymbol{\mu} = (\mu_1, \dots, \mu_K) \qquad a_\star(\boldsymbol{\mu}) = \underset{a=1,\dots,K}{\operatorname{argmax}} \ \mu_a$$

A **best arm identification algorithm** is made of

- a sampling rule A_t : which arm is sampled at round t ?
- a stopping rule τ : when can we stop sampling the arms ?
- a recommendation rule $\hat{a}_\tau$: a guess for $a_\star(\boldsymbol{\mu})$ when we stop

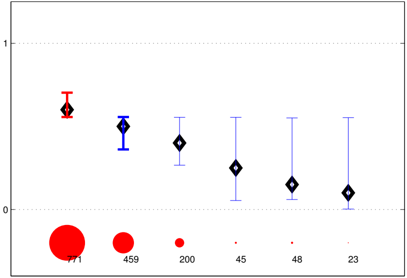
BAI in the fixed-confidence setting

The objective is to build [Even-Dar et al., 2006]

- a δ-correct algorithm : $\forall \boldsymbol{\mu}, \mathbb{P}_{\boldsymbol{\mu}} \left(\hat{a}_\tau = a_\star(\boldsymbol{\mu}) \right) \geq 1 - \delta.$
- with a small sample complexity $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$

The LUCB algorithm [Kalyanakrishnan et al., 2012]

$$\mathcal{I}_a(t) = [\mathrm{LCB}_a(t), \mathrm{UCB}_a(t)].$$



- At round t, draw
$$B_t = \underset{b}{\mathrm{argmax}}\ \hat{\mu}_b(t)$$
$$C_t = \underset{c \neq B_t}{\mathrm{argmax}}\ \mathrm{UCB}_c(t)$$
- Stop at round t if
$$\mathrm{LCB}_{B_t}(t) > \mathrm{UCB}_{C_t}(t)$$
- Recommend $\hat{a}_\tau = B_\tau$

Theorem [Kalyanakrishnan et al., 2012]

For well-chosen confidence intervals, $\mathbb{P}_{\boldsymbol{\mu}}(\hat{a}_\tau = a_\star(\boldsymbol{\mu})) \geq 1 - \delta$ and

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

The best we can do ? Lower bound.

- a change-of-measure lemma

Lemma (e.g., [Garivier et al., 2019])

$\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ two different bandit instances.
τ a stopping time and $\mathcal{E}$ an event in $\sigma(X_1, \dots, X_\tau)$.

$$\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{(X_1,\dots,X_\tau)}; \mathbb{P}_{\boldsymbol{\lambda}}^{(X_1,\dots,X_\tau)}\right) \geq \mathrm{kl}(\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}), \mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E})),$$

where KL is the Kullback-Leibler divergence and

$$\mathrm{kl}(x, y) = \mathrm{KL}\left(\mathcal{B}(x), \mathcal{B}(y)\right) = x \ln\left(\frac{x}{y}\right) + (1-x)\ln\left(\frac{1-x}{1-y}\right)$$

The best we can do ? Lower bound.

- a change-of-measure lemma

Lemma (e.g., [Garivier et al., 2019])

$\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ two different bandit instances.
τ a stopping time and $\mathcal{E}$ an event in $\sigma(X_1, \dots, X_\tau)$.

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

where KL is the Kullback-Leibler divergence and

$$\mathrm{kl}(x, y) = \mathrm{KL}\left(\mathcal{B}(x), \mathcal{B}(y)\right) = x \ln\left(\frac{x}{y}\right) + (1-x)\ln\left(\frac{1-x}{1-y}\right)$$

The best we can do ? Lower bound.

- a change-of-measure lemma

Lemma (e.g., [Garivier et al., 2019])

$\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ two different bandit instances.
τ a stopping time and $\mathcal{E}$ an event in $\sigma(X_1, \dots, X_\tau)$.

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

where KL is the Kullback-Leibler divergence and

$$\mathrm{kl}(x, y) = \mathrm{KL}\left(\mathcal{B}(x), \mathcal{B}(y)\right) = x \ln\left(\frac{x}{y}\right) + (1 - x) \ln\left(\frac{1 - x}{1 - y}\right)$$

Under a δ-correct algorithm,

$$\left.\begin{array}{r} \boldsymbol{\lambda} \text{ such that } a_\star(\boldsymbol{\lambda}) \neq a_\star(\boldsymbol{\mu}) \\ \mathcal{E} = (\hat{a}_\tau = a_\star(\boldsymbol{\lambda})) \end{array}\right\} \Rightarrow \left\{\begin{array}{l} \mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}) \leq \delta \\ \mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}) \geq 1 - \delta \end{array}\right.$$

The best we can do ? Lower bound.

Lemma

$\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ be such that $a_\star(\boldsymbol{\mu}) \neq a_\star(\boldsymbol{\lambda})$. For any δ-correct algorithm,

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

The best we can do ? Lower bound.

Lemma

$\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ be such that $a_\star(\boldsymbol{\mu}) \neq a_\star(\boldsymbol{\lambda})$. For any δ-correct algorithm,

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

- Let $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} : a_\star(\boldsymbol{\lambda}) \neq a_\star(\boldsymbol{\mu})\}$.

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

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \times \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^{K} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau)]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}\mathrm{kl}(\mu_a, \lambda_a) \geq \ln\left(\frac{1}{3\delta}\right)$$

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \times \left(\sup_{w \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a \mathrm{kl}(\mu_a, \lambda_a)\right) \geq \ln\left(\frac{1}{3\delta}\right)$$

The best we can do ? Lower bound.

Theorem [Garivier and Kaufmann, 2016]

For any δ-correct algorithm,

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geq T_\star(\boldsymbol{\mu}) \ln\left(\frac{1}{3\delta}\right),$$

where

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

Moreover, the vector of optimal proportions,

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

is well-defined, and can be computed efficiently.

The best we can do ? Lower bound.

Theorem [Garivier and Kaufmann, 2016]

For any δ-correct algorithm,

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geq T_\star(\boldsymbol{\mu}) \ln\left(\frac{1}{3\delta}\right),$$

where

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

Moreover, the vector of optimal proportions,

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

is well-defined, and can be computed efficiently.

➜ **inspires (optimal) algorithms !**

How to match the lower bound ? Sampling rule.

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

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

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

Lemma

Under the Tracking sampling rule,

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\lim_{t \to \infty} \frac{N_a(t)}{t} = (w_\star(\boldsymbol{\mu}))_a \right) = 1.$$

How to match the lower bound ? Stopping rule.

Idea : perform statistical tests

Individual Generalized Likelihood Ratio test : fix $a \in \{1, \dots, K\}$

$$\mathcal{H}_0 : (a_\star(\boldsymbol{\mu}) \neq a) \quad \text{against} \quad \mathcal{H}_1 : (a_\star(\boldsymbol{\mu}) = a)$$

High values of the GLR statistic tend to reject $\mathcal{H}_0$:

$$\hat{Z}_a(t) = \ln \frac{\sup_{\{\boldsymbol{\lambda} \in [0,1]^K\}} \ell(X_1, \dots, X_t; \boldsymbol{\lambda})}{\sup_{\{\boldsymbol{\lambda} : a_\star(\boldsymbol{\lambda}) \neq a\}} \ell(X_1, \dots, X_t; \boldsymbol{\lambda})}.$$

GLRT stopping rule for BAI : run the K GLR tests in parallel, and stop when one of them rejects $\mathcal{H}_0$:

$$\tau = \inf \left\{ t \in \mathbb{N} : \underbrace{\max_{a=1,\dots,K} \hat{Z}_a(t)}_{:=\hat{Z}(t)} > \beta(t, \delta) \right\}$$

[Chernoff, 1959]

Rewriting the stopping statistic

$$\hat{Z}(t) = \max_{a=1,\dots,K} \hat{Z}_a(t)$$

Using that $\hat{Z}_a(t) = 0$ for $a \neq B_t$, $\hat{Z}(t) = \hat{Z}_{B_t}(t)$ and

$$\hat{Z}(t) = \ln \frac{\ell\left(X_1,\dots,X_t;\hat{\boldsymbol{\mu}}(t)\right)}{\max_{\boldsymbol{\lambda}\in\mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} \ell(X_1,\dots,X_t;\boldsymbol{\lambda})} = \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} \sum_{a=1}^{K} N_a(t)\mathrm{kl}(\hat{\mu}_a(t),\lambda_a)$$

➜ reminiscent of the lower bound

Rewriting the stopping statistic

$$\hat{Z}(t) = \max_{a=1,\dots,K} \hat{Z}_a(t)$$

Using that $\hat{Z}_a(t) = 0$ for $a \neq B_t$, $\hat{Z}(t) = \hat{Z}_{B_t}(t)$ and

$$\hat{Z}(t) = \ln \frac{\ell(X_1,\dots,X_t;\hat{\boldsymbol{\mu}}(t))}{\max_{\boldsymbol{\lambda}\in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} \ell(X_1,\dots,X_t;\boldsymbol{\lambda})} = \inf_{\boldsymbol{\lambda}\in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} \sum_{a=1}^{K} N_a(t)\mathrm{kl}(\hat{\mu}_a(t),\lambda_a)$$

➜ reminiscent of the lower bound

Stopping and recommendation rule

$$\tau_\delta = \inf\left\{ t \in \mathbb{N} : \inf_{\boldsymbol{\lambda}\in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} \sum_{a=1}^{K} N_a(t)\mathrm{kl}(\hat{\mu}_a(t),\lambda_a) > \beta(t,\delta) \right\}$$

$$\hat{a}_{\tau_\delta} = B_{\tau_\delta} = \underset{a=1,\dots,K}{\mathrm{argmax}}\ \hat{\mu}_a(\tau).$$

- How to choose the threshold to ensure a δ-correct algorithm ?

An asymptotically optimal algorithm

Theorem [Garivier and Kaufmann, 2016]

The Track-and-Stop strategy, that uses

- the Tracking sampling rule
- the GLRT stopping rule with

$$\beta(t,\delta) = \ln\left(\frac{2(K-1)t}{\delta}\right)$$

- and recommends $\hat{a}_{\tau_\delta} = \underset{a=1\dots K}{\text{argmax}}\ \hat{\mu}_a(\tau)$

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

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\ln(1/\delta)} = T_\star(\boldsymbol{\mu}).$$

Why ?

$$\tau_\delta = \inf\left\{ t \in \mathbb{N}_\star : \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} \sum_{a=1}^{K} N_a(t)\mathrm{kl}\left(\hat{\mu}_a(t),\lambda_a\right) > \beta(t,\delta)\right\}$$

An asymptotically optimal algorithm

Theorem [Garivier and Kaufmann, 2016]

The Track-and-Stop strategy, that uses

- the Tracking sampling rule
- the GLRT stopping rule with

$$\beta(t,\delta) = \ln\left(\frac{2(K-1)t}{\delta}\right)$$

- and recommends $\hat{a}_{\tau_\delta} = \underset{a=1\dots K}{\text{argmax}}\ \hat{\mu}_a(\tau)$

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

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\ln(1/\delta)} = T_\star(\boldsymbol{\mu}).$$

Why ?

$$\tau_\delta = \inf\left\{ t \in \mathbb{N}_\star : t \times \inf_{\boldsymbol{\lambda}\in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} \sum_{a=1}^{K} \frac{N_a(t)}{t} \mathrm{kl}\left(\hat{\mu}_a(t), \lambda_a\right) > \beta(t,\delta) \right\}$$

An asymptotically optimal algorithm

Theorem [Garivier and Kaufmann, 2016]

The Track-and-Stop strategy, that uses

- the Tracking sampling rule
- the GLRT stopping rule with

$$\beta(t,\delta) = \ln\left(\frac{2(K-1)t}{\delta}\right)$$

- and recommends $\hat{a}_{\tau_\delta} = \underset{a=1\dots K}{\text{argmax}}\ \hat{\mu}_a(\tau)$

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

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\ln(1/\delta)} = T_\star(\boldsymbol{\mu}).$$

Why ?

$$\tau_\delta \simeq \inf\left\{ t \in \mathbb{N}_\star : t \times \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^{K} (w_\star(\boldsymbol{\mu}))_a \mathrm{kl}\left(\mu_a, \lambda_a\right) > \beta(t,\delta) \right\}$$

An asymptotically optimal algorithm

Theorem [Garivier and Kaufmann, 2016]

The Track-and-Stop strategy, that uses

- the Tracking sampling rule
- the GLRT stopping rule with

$$\beta(t, \delta) = \ln\left(\frac{2(K-1)t}{\delta}\right)$$

- and recommends $\hat{a}_{\tau_\delta} = \underset{a=1\dots K}{\operatorname{argmax}}\ \hat{\mu}_a(\tau)$

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

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\ln(1/\delta)} = T_\star(\boldsymbol{\mu}).$$

Why ?

$$\tau_\delta \simeq \inf\left\{ t \in \mathbb{N}_\star : t \times T_\star^{-1}(\boldsymbol{\mu}) > \beta(t, \delta) \right\}$$

Numerical experiments

Experiments on two Bernoulli bandit models :

- $\boldsymbol{\mu}_1 = [0.5\ 0.45\ 0.43\ 0.4]$, such that

$$w_\star(\boldsymbol{\mu}_1) = [0.417\ 0.390\ 0.136\ 0.057]$$

- $\boldsymbol{\mu}_2 = [0.3\ 0.21\ 0.2\ 0.19\ 0.18]$, such that

$$w_\star(\boldsymbol{\mu}_2) = [0.336\ 0.251\ 0.177\ 0.132\ 0.104]$$

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

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

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

Outline

A more general objective

$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$
$\mathcal{R}_1, \ldots, \mathcal{R}_M$ be M regions of possible parameters ($\mathcal{R}_i \subseteq [0,1]^K$).
$\mathcal{R} = \bigcup_{i=1}^M \mathcal{R}_i$.

Active identification : identify *one* region to which $\boldsymbol{\mu}$ belongs.
⚠the regions may be *overlapping*

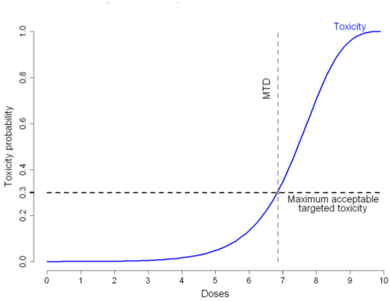
Formalization : build a

- sampling rule (A_t)
- stopping rule τ
- recommendation rule $\hat{\imath}_\tau \in \{1, \ldots, M\}$

such that, for some risk parameter δ, for all $\boldsymbol{\mu} \in \mathcal{R}$

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\boldsymbol{\mu} \notin \mathcal{R}_{\hat{\imath}_\tau}\right) \leq \delta \quad \text{and} \quad \mathbb{E}_{\boldsymbol{\mu}}[\tau] \text{ is small.}$$

Example : Dose Finding in Clinical Trials



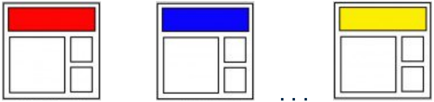
Goal : identify the arm whose mean (= toxicity probability) is closest to a threshold θ

$$\mathcal{R}_i = \left\{ \boldsymbol{\mu} : \mu_1 \leq \cdots \leq \mu_K, i = \underset{k}{\operatorname{argmin}} \, |\mu_k - \theta| \right\}$$

[Garivier et al., 2017]

Example : Back to A/B Testing

Conversion probabilities :



μ_1 μ_2 μ_K

There may be several near-optimal versions.

ϵ-Best arm identification :

$$\mathcal{R}_i = \left\{ \boldsymbol{\mu} \in [0,1]^K : \mu_i > \max_{a \neq i} \mu_a - \epsilon \right\}$$

Goal :

- small error probability : $\forall \boldsymbol{\mu}, \mathbb{P}_{\boldsymbol{\mu}} \left(\mu_{\hat{\imath}_\tau} < \mu_{i_\star} - \epsilon \right) \leq \delta$
- test as short as possible : $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$ small

[Even-Dar et al., 2006]

A GLRT stopping rule

→ the stopping rule introduced for best arm identification can be generalized to any active identification problem !

Individual Generalized Likelihood Ratio test : fix $i \in \{1, \dots, M\}$

$$\mathcal{H}_0 : (\boldsymbol{\mu} \in \mathcal{R} \backslash \mathcal{R}_i) \quad \text{against} \quad \mathcal{H}_1 : (\boldsymbol{\mu} \in \mathcal{R}_i)$$

High values of the GLR statistic tend to reject $\mathcal{H}_0$:

$$\hat{Z}_i(t) = \ln \frac{\sup_{\{\boldsymbol{\lambda} \in \mathcal{R}\}} \ell(X_1, \dots, X_t; \boldsymbol{\lambda})}{\sup_{\{\boldsymbol{\lambda} \in \mathcal{R} \backslash \mathcal{R}_i\}} \ell(X_1, \dots, X_t; \boldsymbol{\lambda})}.$$

GLRT stopping rule for Active Identification : run the M GLR tests in parallel, and stop when one of them rejects $\mathcal{H}_0$:

$$\tau = \inf \left\{ t \in \mathbb{N} : \underbrace{\max_{i=1,\dots,M} \hat{Z}_i(t)}_{:=\hat{Z}(t)} > \beta(t, \delta) \right\}$$

A GLRT stopping rule

→ the stopping rule introduced for best arm identification can be generalized to any active identification problem !

Individual Generalized Likelihood Ratio test : fix $i \in \{1, \dots, M\}$

$$\mathcal{H}_0 : (\boldsymbol{\mu} \in \mathcal{R} \backslash \mathcal{R}_i) \quad \text{against} \quad \mathcal{H}_1 : (\boldsymbol{\mu} \in \mathcal{R}_i)$$

High values of the GLR statistic tend to reject $\mathcal{H}_0$:

$$\hat{Z}_i(t) = \inf_{\boldsymbol{\lambda} \in \mathcal{R} \backslash \mathcal{R}_i} \sum_{a=1}^{K} N_a(t) \mathrm{kl}(\hat{\mu}_a(t), \lambda_a).$$

GLRT stopping rule for Active Identification : run the M GLR tests in parallel, and stop when one of them rejects $\mathcal{H}_0$:

$$\tau = \inf \left\{ t \in \mathbb{N} : \underbrace{\max_{i=1,\dots,M} \hat{Z}_i(t)}_{:= \hat{Z}(t)} > \beta(t, \delta) \right\}$$

A δ-correct stopping rule

$$
\begin{aligned}
\tau_\delta &= \inf\left\{ t \in \mathbb{N} : \max_{i=1,\dots,M} \inf_{\boldsymbol{\lambda} \in \mathcal{R} \setminus \mathcal{R}_i} \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \lambda_a) > \beta(t,\delta) \right\} \\
\hat{\imath}_{\tau_\delta} &\in \operatorname*{argmax}_{i=1,\dots,M} \inf_{\boldsymbol{\lambda} \in \mathcal{R} \setminus \mathcal{R}_i} \sum_{a=1}^{K} N_a(t) \mathrm{kl}(\hat{\mu}_a(t), \lambda_a).
\end{aligned}
$$

Theorem

We can propose a threshold $\beta(t,\delta)$ such that

$$\beta(t,\delta) \simeq \ln(1/\delta) + K \ln\ln(1/\delta) + 3K \ln(1 + \ln t)$$

and for all $\boldsymbol{\mu} \in \mathcal{R}$, $\mathbb{P}_{\boldsymbol{\mu}}\left(\tau_\delta < \infty, \boldsymbol{\mu} \notin \mathcal{R}_{\hat{\imath}_{\tau_\delta}}\right) \leq \delta$.

Proof (1/2)

$$\begin{aligned}
& \mathbb{P}_{\boldsymbol{\mu}}\left(\tau_\delta < \infty, \boldsymbol{\mu} \notin \mathcal{R}_{\hat{\imath}_{\tau_\delta}}\right) \\
\leq \;& \mathbb{P}\left(\exists t \in \mathbb{N}^*, \exists i : \boldsymbol{\mu} \notin \mathcal{R}_i, \inf_{\boldsymbol{\lambda} \in \mathcal{R} \setminus \mathcal{R}_i} \sum_{a=1}^{K} N_a(t) \mathrm{kl}(\hat{\mu}_a(t), \lambda_i) > \beta(t,\delta)\right) \\
\leq \;& \mathbb{P}\left(\exists t \in \mathbb{N}^*, \exists i : \boldsymbol{\mu} \in \mathcal{R} \setminus \mathcal{R}_i, \sum_{a=1}^{K} N_a(t) \mathrm{kl}(\hat{\mu}_a(t), \mu_a) > \beta(t,\delta)\right) \\
\leq \;& \mathbb{P}\left(\exists t \in \mathbb{N}^*, \sum_{a=1}^{K} N_a(t) \mathrm{kl}(\hat{\mu}_a(t), \mu_a) > \beta(t,\delta)\right)
\end{aligned}$$

Proof (1/2)

$$
\begin{aligned}
& \mathbb{P}_{\boldsymbol{\mu}}\left(\tau_\delta < \infty, \boldsymbol{\mu} \notin \mathcal{R}_{\hat{\imath}_{\tau_\delta}}\right) \\
\leq \quad & \mathbb{P}\left(\exists t \in \mathbb{N}^*, \exists i : \boldsymbol{\mu} \notin \mathcal{R}_i, \inf_{\boldsymbol{\lambda} \in \mathcal{R} \setminus \mathcal{R}_i} \sum_{a=1}^{K} N_a(t) \mathrm{kl}(\hat{\mu}_a(t), \lambda_i) > \beta(t,\delta)\right) \\
\leq \quad & \mathbb{P}\left(\exists t \in \mathbb{N}^*, \exists i : \boldsymbol{\mu} \in \mathcal{R} \setminus \mathcal{R}_i, \sum_{a=1}^{K} N_a(t) \mathrm{kl}(\hat{\mu}_a(t), \mu_a) > \beta(t,\delta)\right) \\
\leq \quad & \mathbb{P}\left(\exists t \in \mathbb{N}^*, \sum_{a=1}^{K} N_a(t) \mathrm{kl}(\hat{\mu}_a(t), \mu_a) > \beta(t,\delta)\right)
\end{aligned}
$$

Need for a deviation inequality with the following properties :

➜ deviations are measured with KL-divergence

Proof (1/2)

$$\begin{aligned}
& \mathbb{P}_{\boldsymbol{\mu}}\left(\tau_\delta < \infty, \boldsymbol{\mu} \notin \mathcal{R}_{\hat{\imath}_{\tau_\delta}}\right) \\
\leq \;& \mathbb{P}\left(\exists t \in \mathbb{N}^*, \exists i : \boldsymbol{\mu} \notin \mathcal{R}_i, \inf_{\boldsymbol{\lambda} \in \mathcal{R} \backslash \mathcal{R}_i} \sum_{a=1}^{K} N_a(t) \mathrm{kl}(\hat{\mu}_a(t), \lambda_i) > \beta(t,\delta)\right) \\
\leq \;& \mathbb{P}\left(\exists t \in \mathbb{N}^*, \exists i : \boldsymbol{\mu} \in \mathcal{R} \backslash \mathcal{R}_i, \sum_{a=1}^{K} N_a(t) \mathrm{kl}(\hat{\mu}_a(t), \mu_a) > \beta(t,\delta)\right) \\
\leq \;& \mathbb{P}\left(\exists t \in \mathbb{N}^*, \sum_{a=1}^{K} N_a(t) \mathrm{kl}(\hat{\mu}_a(t), \mu_a) > \beta(t,\delta)\right)
\end{aligned}$$

Need for a deviation inequality with the following properties :

- ➜ deviations are measured with KL-divergence
- ➜ deviations are uniform over time

Proof (1/2)

$$\begin{aligned}
& \mathbb{P}_{\boldsymbol{\mu}}\left(\tau_\delta < \infty, \boldsymbol{\mu} \notin \mathcal{R}_{\hat{\imath}_{\tau_\delta}}\right) \\
\leq \;& \mathbb{P}\left(\exists t \in \mathbb{N}^*, \exists i : \boldsymbol{\mu} \notin \mathcal{R}_i, \inf_{\boldsymbol{\lambda} \in \mathcal{R} \setminus \mathcal{R}_i} \sum_{a=1}^{K} N_a(t) \mathrm{kl}(\hat{\mu}_a(t), \lambda_i) > \beta(t, \delta)\right) \\
\leq \;& \mathbb{P}\left(\exists t \in \mathbb{N}^*, \exists i : \boldsymbol{\mu} \in \mathcal{R} \setminus \mathcal{R}_i, \sum_{a=1}^{K} N_a(t) \mathrm{kl}(\hat{\mu}_a(t), \mu_a) > \beta(t, \delta)\right) \\
\leq \;& \mathbb{P}\left(\exists t \in \mathbb{N}^*, \sum_{a=1}^{K} N_a(t) \mathrm{kl}(\hat{\mu}_a(t), \mu_a) > \beta(t, \delta)\right)
\end{aligned}$$

Need for a deviation inequality with the following properties :

➜ deviations are measured with KL-divergence

➜ deviations are uniform over time

➜ deviations that take into account multiple arms

Proof (2/2)

Theorem [Kaufmann and Koolen, 2018]

There exists $\mathcal{T} : \mathbb{R}^+ \to \mathbb{R}^+$ a threshold function such that

$$\mathcal{T}(x) \simeq x + \ln(x)$$

one has

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \sum_{a=1}^{K} N_a(t)\mathrm{kl}(\hat{\mu}_a(t), \mu_a) \geq 3\sum_{a=1}^{K} \ln(1 + \ln(N_a(t))) + K\mathcal{T}\left(\frac{x}{K}\right)\right) \leq e^{-x}.$$

Consequence :

$$\mathbb{P}\left(\exists t : \sum_{a=1}^{K} N_a(t)\mathrm{kl}(\hat{\mu}_a(t), \mu_a) \geq 3\ln(1 + \ln(t)) + K\mathcal{T}\left(\frac{\ln(1/\delta)}{K}\right)\right) \leq \delta.$$

Optimal Active Identification ?

Non-Overlapping case : Same lower bound

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geq T_\star(\boldsymbol{\mu}) \ln\left(\frac{1}{3\delta}\right),$$

with

$$T_\star(\boldsymbol{\mu})^{-1} = \sup_{w \in \boldsymbol{\Sigma}_K} \inf_{\boldsymbol{\lambda} \in \mathcal{R} \setminus \mathcal{R}_{i_\star(\boldsymbol{\mu})}} \left(\sum_{a=1}^{K} w_a \mathrm{kl}(\mu_a, \lambda_a) \right).$$

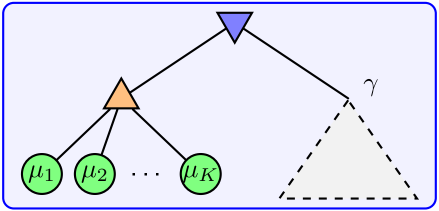
- Tracking + GLRT is asymptotically optimal provided that the optimal weights can easily be computed...

Overlapping case : can be slightly harder

[Degenne and Koolen, 2019, Garivier and Kaufmann, 2019]

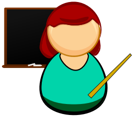
Outline

Comparing the Smallest Mean to a Threshold



Fix threshold γ.

$\mu_{\min} := \min_i \mu_i \lessgtr \gamma$?



For $t = 1, \ldots, \tau$

- pick a leaf A_t
- observe $X_t \sim \mathcal{B}(\mu_{A_t})$

After stopping, recommend $\hat{m} \in \{<, >\}$

Goal : controlled error $\mathbb{P}_{\boldsymbol{\mu}}(\hat{m} \neq m_\star) \leq \delta$
and small sample complexity $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$

[Kaufmann et al., 2018]

Lower Bound and Oracle Allocation

Lower bound : for *any* δ-correct algorithm,

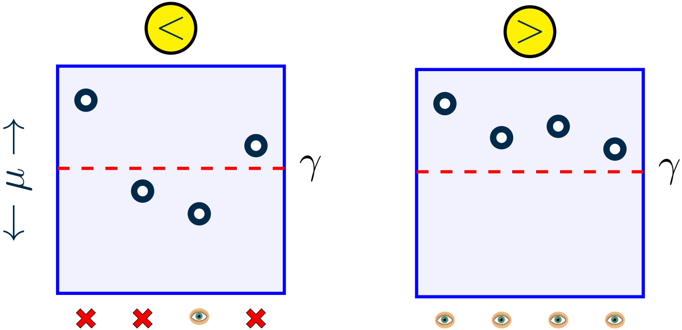
$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geq T_\star(\boldsymbol{\mu}) \ln\left(\frac{1}{3\delta}\right).$$

For our problem the *characteristic time* and *oracle weights* are

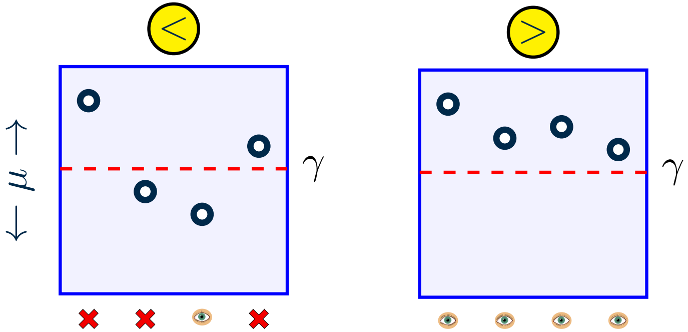
$$T_\star(\boldsymbol{\mu}) = \begin{cases} \dfrac{1}{\mathrm{kl}(\mu_{\min}, \gamma)} & \mu_{\min} < \gamma, \\ \displaystyle\sum_a \frac{1}{\mathrm{kl}(\mu_a, \gamma)} & \mu_{\min} > \gamma, \end{cases} \qquad (w_\star(\boldsymbol{\mu}))_a = \begin{cases} \mathbb{1}_{(a = a_\star)} & \mu_{\min} < \gamma, \\ \dfrac{\frac{1}{\mathrm{kl}(\mu_a,\gamma)}}{\sum_j \frac{1}{\mathrm{kl}(\mu_j,\gamma)}} & \mu_{\min} > \gamma. \end{cases}$$

$(w_\star(\boldsymbol{\mu}))_a$: fraction of selections of the leaf a under a strategy that would match the lower bound

Dichotomous Oracle Behaviour !



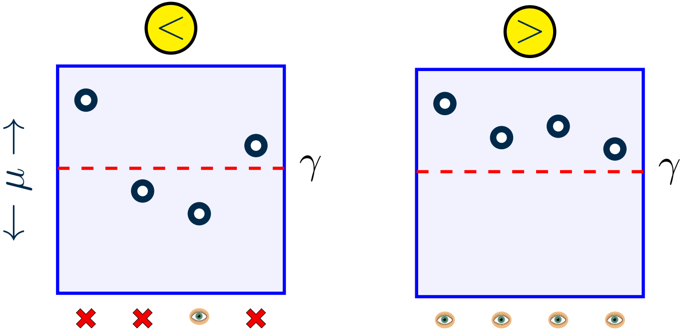
Dichotomous Oracle Behaviour !



Two different ideas to get those sampling profiles :

- **Thompson Sampling** (Π_{t-1} is posterior after $t-1$ rounds)
  Sample $\theta \sim \Pi_{t-1}$, then play $A_t = \mathrm{argmin}_a\ \theta_a$.
- **a Lower Confidence Bound algorithm**
  Play $A_t = \mathrm{argmin}_a\ \mathrm{LCB}_a(t)$

A Solution : Murphy Sampling !



A more flexible idea :

- **Murphy Sampling** condition on *low* minimum mean
  Sample $\boldsymbol{\theta} \sim \Pi_{t-1}\left(\cdot \middle| \min_a \theta_a < \gamma\right)$, then play $A_t = \arg\min_a \ \theta_a$.

➜ converges to the optimal allocation in both cases !

Properties of Murphy Sampling

Theorem

For all $\boldsymbol{\mu}$, Murphy Sampling satisfies, for all a,

$$\frac{N_a(t)}{t} \to (w_\star(\boldsymbol{\mu}))_a$$

Sampling rule	$<$	$>$
Thompson Sampling	✓	✗
Lower Confidence Bounds	✗	✓
Murphy Sampling	✓	✓

Corollary

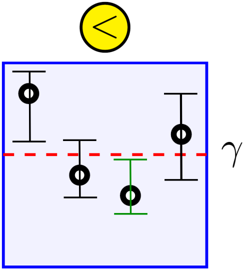
Murphy Sampling combined with a "good" stopping rule satisfies

$$\limsup_{\delta \to 0} \frac{\tau_\delta}{\ln \frac{1}{\delta}} \leq T_\star(\boldsymbol{\mu}), \; a.s.$$

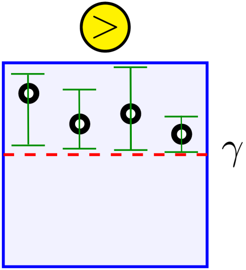
A good stopping rule

Sufficient for asymptotic guarantees : a simple stopping rule based on individual confidence intervals $\tau^{\text{Box}} := \min(\tau_<; \tau_>)$ where

$$\begin{aligned}\tau_< &= \inf\{t \in \mathbb{N} : \exists a : \mathrm{UCB}_a(t) < \gamma\} \\ \tau_> &= \inf\{t \in \mathbb{N} : \forall a, \mathrm{LCB}_a(t) > \gamma\}\end{aligned}$$



$\tau = \tau_<$



$\tau = \tau_>$

Better stopping rules

The GLRT stopping rule

Improved test for rejecting $\mathcal{H}_>$: (summing evidence)

$$\tau_<^{\text{GLRT}} = \inf\left\{ t \in \mathbb{N} : \sum_{a:\hat{\mu}_a(t)\leq\gamma} N_a(t)\mathrm{kl}(\hat{\mu}_a(t),\gamma) > \beta(t,\delta) \right\}$$

- **Beyond the GLRT** : aggregating evidence

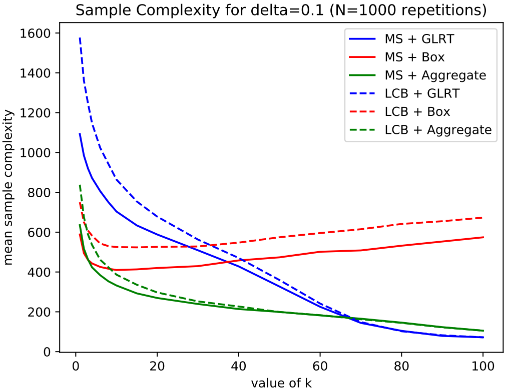
$$\tau_<^{\text{Aggr}} = \inf\left\{ t \in \mathbb{N} : \exists \mathcal{S} : N_{\mathcal{S}}(t)\mathrm{kl}^+(\hat{\mu}_{\mathcal{S}}(t),\gamma) > \beta_{\mathcal{S}}(t,\delta) \right\}$$

where $N_{\mathcal{S}}(t)$ and $\hat{\mu}_{\mathcal{S}}(t)$ are computed based on all the samples gathered from all arms in $\mathcal{S}$.

→ new concentration inequality showing this rule is δ-correct for

$$\beta_{\mathcal{S}}(t,\delta) \simeq \ln\left(\frac{1}{\delta\pi(\mathcal{S})}\right) + 3\ln(1+\ln(t)), \quad \text{where} \quad \sum_{\mathcal{S}} \pi(\mathcal{S}) = 1.$$

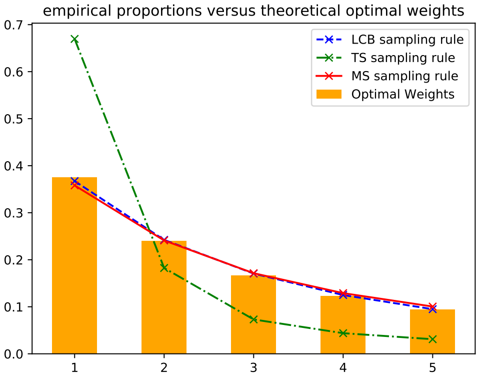
Sample complexity results



Agg *beats* Box and GLRT in adapting to the number k of low arms.
Here $\mu_a \in \{-1, 0\}$ and $\gamma = 0$ (Gaussian arms).

Sampling rule : $\mu \in \mathcal{H}_>$

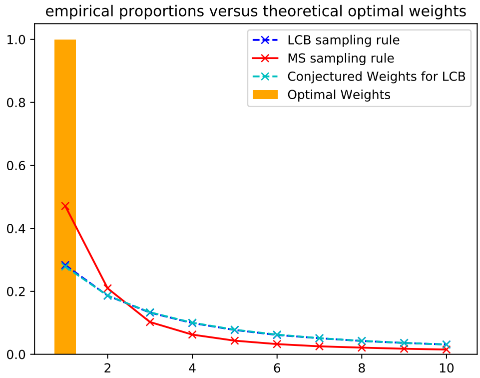
$\boldsymbol{\mu} = \text{linspace}(1/2, 1, 5) \in \mathcal{H}_>$



Sampling proportions vs oracle, $\delta = e^{-7}$.

Sampling rule : $\mu \in \mathcal{H}_<$

$\boldsymbol{\mu} = \text{linspace}(-1, 1, 10) \in \mathcal{H}_<$



Sampling proportions vs oracle, $\delta = e^{-23}$.

Conclusion

- Many interesting bandit problems beyond rewards maximization !
- Generalized Likelihood Ratios are powerful for general active identification in a bandit model :
  - ➜ they can guarantee δ-correct identification
  - ➜ they reach the optimal sample complexity when coupled with an appropriate sampling rule
- Murphy Sampling : a first step beyond lower bound inspired (Tracking) sampling rules

Merci !

Agrawal, R. (1995).
Sample mean based index policies with O(log n) regret for the multi-armed bandit problem.
Advances in Applied Probability, 27(4) :1054–1078.

Agrawal, S. and Goyal, N. (2013).
Further Optimal Regret Bounds for Thompson Sampling.
In *Proceedings of the 16th Conference on Artificial Intelligence and Statistics*.

Anandkumar, A., Michael, N., and Tang, A. K. (2010).
Opportunistic Spectrum Access with multiple users : Learning under competition.
In *IEEE INFOCOM*.

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2009).
Exploration-exploitation trade-off using variance estimates in multi-armed bandits.
Theoretical Computer Science, 410(19).

Auer (2002).
Using Confidence bounds for Exploration Exploitation trade-offs.
Journal of Machine Learning Research, 3 :397–422.

Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., and Stoltz, G. (2013).
Kullback-Leibler upper confidence bounds for optimal sequential allocation.
Annals of Statistics, 41(3) :1516–1541.

Chernoff, H. (1959).
Sequential design of Experiments.
The Annals of Mathematical Statistics, 30(3) :755–770.

Degenne, R. and Koolen, W. M. (2019).
Pure exploration with multiple correct answers.
arXiv :1902.03475.

Even-Dar, E., Mannor, S., and Mansour, Y. (2006).
Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems.
Journal of Machine Learning Research, 7 :1079–1105.

Garivier, A. and Kaufmann, E. (2016).
Optimal best arm identification with fixed confidence.
In *Proceedings of the 29th Conference On Learning Theory*.

Garivier, A. and Kaufmann, E. (2019).

Non-asymptotic sequential tests for overlapping hypotheses and application to near optimal arm identification in bandit models.
arXiv :1905.03495.

Garivier, A., Ménard, P., and Rossi, L. (2017).
Thresholding bandit for dose-ranging : The impact of monotonicity.
arXiv :1711.04454.

Garivier, A., Ménard, P., and Stoltz, G. (2019).
Explore first, exploit next : The true shape of regret in bandit problems.
Math. Oper. Res., 44(2) :377–399.

Jouini, W., Ernst, D., Moy, C., and Palicot, J. (2009).
Multi-armed bandit based policies for cognitive radio's decision making issues.
In *International Conference Signals, Circuits and Systems (IEEE).*

Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. (2012).
PAC subset selection in stochastic multi-armed bandits.
In *International Conference on Machine Learning (ICML).*

Katehakis, M. and Robbins, H. (1995).
Sequential choice from several populations.
Proceedings of the National Academy of Science, 92 :8584–8585.

Kaufmann, E. and Koolen, W. (2018).
Mixture martingales revisited with applications to sequential tests and confidence intervals.
arXiv :1811.11419.

Kaufmann, E., Koolen, W., and Garivier, A. (2018).
Sequential test for the lowest mean : From thompson to murphy sampling.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kaufmann, E., Korda, N., and Munos, R. (2012).
Thompson Sampling : an Asymptotically Optimal Finite-Time Analysis.
In *Proceedings of the 23rd conference on Algorithmic Learning Theory*.

Lai, T. and Robbins, H. (1985).
Asymptotically efficient adaptive allocation rules.
Advances in Applied Mathematics, 6(1) :4–22.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010).
A contextual-bandit approach to personalized news article recommendation.
In *WWW*.

Thompson, W. (1933).

On the likelihood that one unknown probability exceeds another in view of the evidence of two samples.
Biometrika, 25 :285–294.