Generalized Likelihood Ratios Tests applied to Sequential Decision Making

Emilie Kaufmann,

based on joint works with Lilian Besson (CentraleSupélec), Aurélien Garivier (ENS Lyon), Wouter Koolen (CWI)

Centre de Recherche en Informatique, Signal et Automatique de Lille

Machine Learning and Statistics for Structures
Leiden, May 3rd, 2019

Outline

1. The bandit framework for sequential decision making
2. Active identification in a bandit model
   - A generic δ-correct stopping rule
   - Towards optimal sample complexity
   - the Best Arm Identification example
3. Rewards maximization in a non-stationary bandit model
   - The kl-UCB algorithm in the stationary case
   - A non-parametric sequential change point detector
   - kl-UCB meets the Bernoulli-GLRT

Outline

1. The bandit framework for sequential decision making
2. Active identification in a bandit model
   - A generic δ-correct stopping rule
   - Towards optimal sample complexity
   - the Best Arm Identification example
3. Rewards maximization in a non-stationary bandit model
   - The kl-UCB algorithm in the stationary case
   - A non-parametric sequential change point detector
   - kl-UCB meets the Bernoulli-GLRT

The multi-armed bandit model

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):

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

Generic goal: learn *something* about the means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$

Bernoulli bandit model

K arms $=$ K probability distributions (ν_a has mean μ_a)





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

At round t, an agent:

- chooses an arm A_t
- observes a sample $X_t \sim \mathcal{B}(\mu_{A_t})$

using a sequential sampling strategy (A_t):

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

Generic goal: learn *something* about the means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$

Bernoulli bandit model

K arms $=$ K probability distributions (ν_a has mean μ_a)

$\mathcal{B}(\mu_1)$

$\mathcal{B}(\mu_2)$

$\mathcal{B}(\mu_3)$

$\mathcal{B}(\mu_4)$

$\mathcal{B}(\mu_5)$

For the t-th patient in a clinical study,

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

using a sequential sampling strategy (A_t):

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

Possible goals:

- identify the best treatment, i.e. $a^* = \text{argmax}_a \ \mu_a$
- maximize the number of healed patients, $\sum_{t=1}^{K} X_t$

Outline

Active Identification in a Bandit Model

Assumption: arms belong to a one-dimensional exponential family
$\rightarrow$ each arm is parameterized by its mean $\mu_a \in \mathcal{I}$
(Bernoulli, Gaussian with known variance, Poisson...)

Active identification: $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$
Given M regions of $\mathcal{I}^K$, $\mathcal{R}_1, \dots, \mathcal{R}_M$, the goal is to identify one region to which $\boldsymbol{\mu}$ belongs.

Formalization: build a

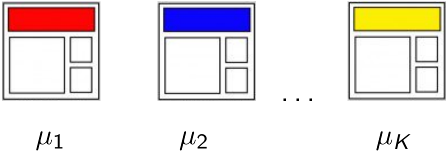
- sampling rule (A_t)
- stopping rule τ
- recommendation rule $\hat{\imath}_\tau \in \{1, \dots, M\}$

such that, for some risk parameter δ,

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

Example: A/B/C Testing

Probability that some version of a website generates a conversion:



Best version: $i^* = \underset{a}{\text{argmax}}\ \mu_a$

Active identification of the best version:

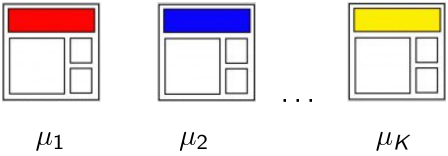
- which version A_t should be displayed to the t-th visitor?
- when to stop the test (after τ visitors)?
- which version should be recommend as the best one ($\hat{\imath}_\tau$)?

Goal:

- small error probability: $\mathbb{P}\left(\hat{\imath}_\tau \neq i^*\right) \leq 0.05$
- test as short as possible: $\mathbb{E}[\tau]$ small

Example: A/B/C Testing

Mean of each arm:



Best arm: $i^* = \underset{a}{\text{argmax}}\ \mu_a$

Best arm identification: $\mathcal{R}_i = \{\boldsymbol{\mu} : \mu_i > \max_{a \neq i} \mu_a\}$

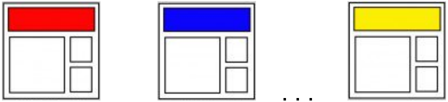
- sampling rule A_t
- stopping rule τ
- recommendation rule $\hat{\imath}_\tau$

Goal:

- small error probability: $\mathbb{P}\left(\hat{\imath}_{\tau_\delta} \neq i^*\right) \leq \delta$
- test as short as possible: $\mathbb{E}[\tau]$ small

Example: A/B/C Testing

Mean of each arm:



μ_1 μ_2 μ_K

Best arm: $i^* = \underset{a}{\text{argmax}}\ \mu_a$

ϵ-**Best arm identification**: $\mathcal{R}_i = \{\boldsymbol{\mu} : \mu_i > \max_{a \neq i} \mu_a - \epsilon\}$

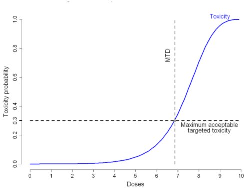
- sampling rule A_t
- stopping rule τ
- recommendation rule $\hat{\imath}_\tau$

Goal:

- small error probability: $\mathbb{P}\left(\mu_{\hat{\imath}_\tau} \geq \mu_{i^*} - \epsilon\right) \leq \delta$
- test as short as possible: $\mathbb{E}[\tau]$ small

Beyond Best Arm Identification

- Dose finding in Phase I Clinical Trials



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

$$\mathcal{R}_i = \left\{ \boldsymbol{\mu} : i = \underset{k}{\text{argmin}} \, |\mu_k - \theta| \right\}$$

- Anomaly detection: $\mathcal{R}_1 = \{\boldsymbol{\mu} : \min_i \mu_i \leq \gamma\}$, $\mathcal{R}_2 = \mathcal{R}_1^c$

K., Koolen, Garivier, Sequential Test for the Lowest Mean: From Thompson to Murphy Sampling, NeurIPS 2018

Outline

Objective

For a given sampling rule, we want to build stopping and recommendation rules $(\tau_\delta, \hat{\imath}_{\tau_\delta})$ for the test

$$\mathcal{H}_1 : (\boldsymbol{\mu} \in \mathcal{R}_1) \quad \mathcal{H}_2 : (\boldsymbol{\mu} \in \mathcal{R}_2) \quad \dots \quad \mathcal{H}_M : (\boldsymbol{\mu} \in \mathcal{R}_M)$$

(possibly with overlapping hypotheses!)

Assumption: $\mathcal{R} := \bigcup_{i=1}^{M} \mathcal{R}_i$, $\overline{\mathcal{R}} = \mathcal{I}^K$ (all possible means).

Definition

A **δ-correct sequential test** is a pair $(\tau_\delta, \hat{\imath}_{\tau_\delta})$ where

- τ_δ is a stopping time with respect to $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$
- $\hat{\imath}_{\tau_\delta}$ is $\mathcal{F}_{\tau_\delta}$-measurable

such that

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

The parallel GLRT

Idea: run M statistical tests of

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

in parallel until one of them rejects $\tilde{\mathcal{H}}_0$.

Individual test: a Generalized Likelihood Ratio rejects $\tilde{\mathcal{H}}_0$ for large values of the Generalized Likelihood Ratio

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

where $\ell(X_1, \dots, X_t; \boldsymbol{\lambda})$ is the likelihood of the observations under a bandit model with means $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)$.

The parallel GLRT

$$\hat{\mathrm{GLR}}(t) = \frac{\sup_{\boldsymbol{\lambda}\in\mathcal{R}} \ell(X_1,\dots,X_t;\boldsymbol{\lambda})}{\sup_{\boldsymbol{\lambda}\in\mathcal{R}\setminus\mathcal{R}_i} \ell(X_1,\dots,X_t;\boldsymbol{\lambda})} = \inf_{\boldsymbol{\lambda}\in\mathcal{R}\setminus\mathcal{R}_i} \frac{\ell(X_1,\dots,X_t;\hat{\boldsymbol{\mu}}(t))}{\ell(X_1,\dots,X_t;\boldsymbol{\lambda})}$$

where $\hat{\boldsymbol{\mu}}(t) = (\hat{\mu}_1(t),\dots,\hat{\mu}_K(t))$ is the MLE.

- With arms in a one-dimensional exponential family,

$$\ln \frac{\ell(X_1,\dots,X_\tau;\hat{\boldsymbol{\mu}}(t))}{\ell(X_1,\dots,X_t;\boldsymbol{\lambda})} = \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t),\lambda_a)$$

with the Kullback-Leibler divergence

$$d(\mu,\lambda) = \mathrm{KL}(\nu_\mu,\nu_\lambda) = \mathbb{E}_{X\sim\nu_\mu}\left[\ln\frac{f_\mu(X)}{f_\lambda(X)}\right]$$

and

- f_μ is the density of an arm with mean μ
- $N_a(t)$: number of selections of arm a up to time t
- $\hat{\mu}_a(t)$: empirical mean of the observation received from arm a

The parallel GLRT

$$\hat{\text{GLR}}(t) = \frac{\sup_{\boldsymbol{\lambda}\in\mathcal{R}} \ell(X_1,\dots,X_t;\boldsymbol{\lambda})}{\sup_{\boldsymbol{\lambda}\in\mathcal{R}\setminus\mathcal{R}_i} \ell(X_1,\dots,X_t;\boldsymbol{\lambda})} = \inf_{\boldsymbol{\lambda}\in\mathcal{R}\setminus\mathcal{R}_i} \frac{\ell(X_1,\dots,X_t;\hat{\boldsymbol{\mu}}(t))}{\ell(X_1,\dots,X_t;\boldsymbol{\lambda})}$$

where $\hat{\boldsymbol{\mu}}(t) = (\hat{\mu}_1(t),\dots,\hat{\mu}_K(t))$ is the MLE.

- With arms in a one-dimensional exponential family,

$$\ln \frac{\ell(X_1,\dots,X_\tau;\hat{\boldsymbol{\mu}}(t))}{\ell(X_1,\dots,X_t;\boldsymbol{\lambda})} = \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t),\lambda_a)$$

with the Kullback-Leibler divergence

$$d(\mu,\lambda) = \frac{(\mu-\lambda)^2}{2\sigma^2} \quad \text{(Gaussian distributions)}$$

and

- f_μ is the density of an arm with mean μ
- $N_a(t)$: number of selections of arm a up to time t
- $\hat{\mu}_a(t)$: empirical mean of the observation received from arm a

The parallel GLRT

$$\hat{\text{GLR}}(t) = \frac{\sup_{\boldsymbol{\lambda} \in \mathcal{R}} \ell(X_1, \dots, X_t; \boldsymbol{\lambda})}{\sup_{\boldsymbol{\lambda} \in \mathcal{R} \setminus \mathcal{R}_i} \ell(X_1, \dots, X_t; \boldsymbol{\lambda})} = \inf_{\boldsymbol{\lambda} \in \mathcal{R} \setminus \mathcal{R}_i} \frac{\ell(X_1, \dots, X_t; \hat{\boldsymbol{\mu}}(t))}{\ell(X_1, \dots, X_t; \boldsymbol{\lambda})}$$

where $\hat{\boldsymbol{\mu}}(t) = (\hat{\mu}_1(t), \dots, \hat{\mu}_K(t))$ is the MLE.

- With arms in a one-dimensional exponential family,

$$\ln \frac{\ell(X_1, \dots, X_\tau; \hat{\boldsymbol{\mu}}(t))}{\ell(X_1, \dots, X_t; \boldsymbol{\lambda})} = \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \lambda_a)$$

with the Kullback-Leibler divergence

$$d(\mu, \lambda) = \mu \ln \frac{\mu}{\lambda} + (1 - \mu) \ln \frac{1 - \mu}{1 - \lambda} \text{ (Bernoulli distributions)}$$

and

- f_μ is the density of an arm with mean μ
- $N_a(t)$: number of selections of arm a up to time t
- $\hat{\mu}_a(t)$: empirical mean of the observation received from arm a

The parallel GLRT

Idea: run M statistical tests of

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

in parallel until one of them rejects $\tilde{\mathcal{H}}_0$.

Individual test: a Generalized Likelihood Ratio rejects $\tilde{\mathcal{H}}_0$ for large values of the Generalized Likelihood Ratio

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

with

- $N_a(t)$: number of selections of arm a up to time t
- $\hat{\mu}_a(t)$: empirical mean of the observation received from arm a

The parallel GLRT

Idea: run M GLR tests of

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

in parallel until one of them rejects $\tilde{\mathcal{H}}_0$.

Global test:

$$\tau_\delta = \inf \left\{ t \in \mathbb{N} : \max_{i=1,\dots,M} \inf_{\boldsymbol{\lambda} \in \mathcal{R} \backslash \mathcal{R}_i} \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \lambda_a) > \beta(t, \delta) \right\}$$

$$\hat{\imath}_{\tau_\delta} \in \underset{i=1,\dots,M}{\text{argmax}} \inf_{\boldsymbol{\lambda} \in \mathcal{R} \backslash \mathcal{R}_i} \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \lambda_a).$$

depends on a threshold function $\beta(t, \delta)$.

A closer look at the stopping rule

$$\tau_\delta = \inf \left\{ t \in \mathbb{N} : \max_{i=1,\dots,M} \inf_{\lambda \in \mathcal{R} \backslash \mathcal{R}_i} \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \lambda_a) > \beta(t, \delta) \right\}$$

Interpretation: $\sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \lambda_a)$ measures a distance between $\hat{\boldsymbol{\mu}}(t)$ and $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)$.

➜ we stop when there exists a region $\mathcal{R}_i$ such that $\hat{\boldsymbol{\mu}}(t) \in \mathcal{R}_i$ and $\hat{\boldsymbol{\mu}}(t)$ is "far enough" from all instances $\boldsymbol{\lambda} \in \mathcal{R} \backslash \mathcal{R}_i$.

Example: ϵ-BAI, Gaussian case

$$\max_{a \in \hat{A}_\epsilon(t)} \min_{b \neq a} \frac{N_a(t) N_b(t)}{2\sigma^2 (N_a(t) + N_b(t))} \Big(|\hat{\mu}_a(t) - \hat{\mu}_b(t)| + \epsilon \Big)^2 > \beta(t, \delta)$$

A δ-correct parallel GLRT

$$\tau_\delta = \inf\left\{t \in \mathbb{N} : \max_{i=1,\dots,M} \inf_{\lambda \in \mathcal{R}\setminus\mathcal{R}_i} \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \lambda_a) > \beta(t,\delta)\right\}$$

$$\hat{\imath}_{\tau_\delta} \in \operatorname*{argmax}_{i=1,\dots,M} \inf_{\lambda \in \mathcal{R}\setminus\mathcal{R}_i} \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \lambda_a).$$

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)

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

Proof (1/2)

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

Need for a deviation inequality with the following properties:

➜ deviations are measured with KL-divergence

Proof (1/2)

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

Need for a deviation inequality with the following properties:

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

Proof (1/2)

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

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 [K. 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

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

Consequence:

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

Optimal Active Identification?

So far we proved, that the parallel GLRT $(\hat{\tau}_\delta, \hat{\imath}_{\tau_\delta})$ can be made δ-correct for active identification for any sampling rule (A_t).

Question: what about the expected duration of the test $\mathbb{E}_\mu[\tau_\delta]$?

- requires a not too crazy sampling rule
- can we find a sampling rule that attains the smallest possible sample complexity when combined with a parallel GLRT?

Outline

1. The bandit framework for sequential decision making

2. Active identification in a bandit model
   - A generic δ-correct stopping rule
   - Towards optimal sample complexity
   - the Best Arm Identification example

3. Rewards maximization in a non-stationary bandit model
   - The kl-UCB algorithm in the stationary case
   - A non-parametric sequential change point detector
   - kl-UCB meets the Bernoulli-GLRT

Sample complexity lower bound

Change of distribution argument: pick an alternative $\boldsymbol{\lambda}$ close enough to $\boldsymbol{\mu}$ such that the behaviour of the algorithm needs to be different under $\boldsymbol{\lambda}$ and under $\boldsymbol{\mu}$.

➜ some event C will be very likely under $\boldsymbol{\mu}$, very unlikely under $\boldsymbol{\lambda}$, which gives constraints on the observed samples

Elementary change of distribution: Introducing

$$L_t(\boldsymbol{\mu}, \boldsymbol{\lambda}) := \ln \frac{\ell(X_1, \dots, X_t; \boldsymbol{\mu})}{\ell(X_1, \dots, X_t; \boldsymbol{\lambda})},$$

for every event $C \in \mathcal{F}_n$,

$$\mathbb{P}_{\boldsymbol{\lambda}}(C) = \mathbb{E}_{\boldsymbol{\mu}}\left[\mathbb{1}_C \exp\left(-L_n(\boldsymbol{\mu}, \boldsymbol{\lambda})\right)\right]$$

Sample complexity lower bound

More sophisticated change of distribution [Garivier et al. 2016]

Let $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ be two bandit models. For any event $C \in \mathcal{F}_\tau$,

$$\mathbb{E}_{\boldsymbol{\mu}}[L_\tau(\boldsymbol{\mu}, \boldsymbol{\lambda})] \geq \mathrm{kl}\big(\mathbb{P}_{\boldsymbol{\mu}}(C), \mathbb{P}_{\boldsymbol{\lambda}}(C)\big).$$

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

Sample complexity lower bound

More sophisticated change of distribution [Garivier et al. 2016]

Let $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ be two bandit models. For any event $C \in \mathcal{F}_\tau$,

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

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

Sample complexity lower bound

More sophisticated change of distribution [Garivier et al. 2016]

Let $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ be two bandit models. For any event $C \in \mathcal{F}_\tau$,

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

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

If $\boldsymbol{\mu}$ belongs to a unique region $\mathcal{R}_{i^*(\boldsymbol{\mu})}$, then for all $\boldsymbol{\lambda} \in \mathcal{R} \backslash \mathcal{R}_{i^*(\boldsymbol{\mu})}$, under a δ-correct strategy,

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\hat{\imath}_{\tau_\delta} = i^*(\boldsymbol{\mu})\right) \geq 1 - \delta \quad \text{and} \quad \mathbb{P}_{\boldsymbol{\lambda}}\left(\hat{\imath}_{\tau_\delta} = i^*(\boldsymbol{\mu})\right) \leq \delta$$

For any $\boldsymbol{\lambda} \in \mathcal{R} \backslash \mathcal{R}_{i^*(\boldsymbol{\mu})}$,

$$\sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau_\delta)] d(\mu_a, \lambda_a) \geq (1 - 2\delta) \ln\left(\frac{1-\delta}{\delta}\right)$$

Sample Complexity Lower Bound

Assumption: the regions form a partition $\mathcal{R} = \bigcup_{i=1}^{M} \mathcal{R}_i$.

Theorem

Any δ-correct algorithm satisfies

$$\mathbb{E}[\tau_\delta] \geq T^*(\boldsymbol{\mu}) \ln\left(\frac{1}{3\delta}\right)$$

where

$$T^*(\boldsymbol{\mu})^{-1} = \sup_{w \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathcal{R} \setminus \mathcal{R}_{i^*(\boldsymbol{\mu})}} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a)$$

$\Sigma_K = \{w \in [0,1]^K : \sum_{i=1}^K w_i = 1\}$

Proof.

$$\inf_{\boldsymbol{\lambda} \in \mathcal{R} \setminus \mathcal{R}_{i^*(\boldsymbol{\mu})}} \sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau)] d(\mu_a, \lambda_a) \geq (1 - 2\delta) \ln\left(\frac{1-\delta}{\delta}\right)$$

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \times \inf_{\boldsymbol{\lambda} \in \mathcal{R} \setminus \mathcal{R}_{i^*(\boldsymbol{\mu})}} \sum_{a=1}^{K} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau)]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau]} d(\mu_a, \lambda_a) \geq \ln(1/(3\delta))$$

Sample Complexity Lower Bound

Assumption: the regions form a partition $\mathcal{R} = \bigcup_{i=1}^{M} \mathcal{R}_i$.

Theorem

Any δ-correct algorithm satisfies

$$\mathbb{E}[\tau_\delta] \geq T^*(\boldsymbol{\mu}) \ln\left(\frac{1}{3\delta}\right)$$

where

$$T^*(\boldsymbol{\mu})^{-1} = \sup_{w \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathcal{R} \setminus \mathcal{R}_{i^*(\boldsymbol{\mu})}} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a)$$

$\Sigma_K = \{w \in [0,1]^K : \sum_{i=1}^K w_i = 1\}$

Proof.

$$\inf_{\boldsymbol{\lambda} \in \mathcal{R} \setminus \mathcal{R}_{i^*(\boldsymbol{\mu})}} \sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau)] d(\mu_a, \lambda_a) \geq (1-2\delta) \ln\left(\frac{1-\delta}{\delta}\right)$$

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \times \left(\sup_{w \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathcal{R} \setminus \mathcal{R}_{i^*(\boldsymbol{\mu})}} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) \right) \geq \ln(1/(3\delta))$$

Sample Complexity Lower Bound

An algorithm matching the lower bound should satisfy

$$\forall a \in \{1,\ldots,K\}, \ \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau_\delta)]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau]} \simeq w_a^*(\boldsymbol{\mu})$$

for a vector of optimal proportions

$$\boldsymbol{w}^*(\boldsymbol{\mu}) \in \underset{w \in \Sigma_K}{\text{argmax}} \inf_{\boldsymbol{\lambda} \in \mathcal{R} \setminus \mathcal{R}_{i^*(\boldsymbol{\mu})}} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a).$$

Remark: in general $\boldsymbol{w}^*(\boldsymbol{\mu})$

- ➜ may be non unique
- ➜ may be hard to compute

Parallel GLRT can match the lower bound

If $\mathcal{R} = \bigcup_{i=1}^{M} \mathcal{R}_i$ forms a partition,

$$\begin{aligned}
\tau_\delta &= \inf\left\{ t \in \mathbb{N} : \inf_{\boldsymbol{\lambda} \in \mathcal{R} \setminus \mathcal{R}_{\hat{\imath}(t)}} \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \lambda_a) > \beta(t,\delta) \right\} \\
&= \inf\left\{ t \in \mathbb{N} : t \times \inf_{\boldsymbol{\lambda} \in \mathcal{R} \setminus \mathcal{R}_{\hat{\imath}(t)}} \sum_{a=1}^{K} \frac{N_a(t)}{t} d(\hat{\mu}_a(t), \lambda_a) > \beta(t,\delta) \right\} \\
&\simeq \inf\left\{ t \in \mathbb{N} : t \times \underbrace{\inf_{\boldsymbol{\lambda} \in \mathcal{R} \setminus \mathcal{R}_{i^*(\boldsymbol{\mu})}} \sum_{a=1}^{K} w_a^*(\boldsymbol{\mu}) d(\mu_a, \lambda_a)}_{T^*(\boldsymbol{\mu})^{-1}} > \beta(t,\delta) \right\}
\end{aligned}$$

under a good sampling rule satisfying

$$\forall a, \ \lim_{t \to \infty} \frac{N_a(t)}{t} = w_a^*(\boldsymbol{\mu}) \quad a.s.$$

➜ $\tau_\delta \simeq \inf\{t \in \mathbb{N} : t > T^*(\boldsymbol{\mu})\beta(t,\delta)\} \simeq T^*(\boldsymbol{\mu}) \ln \frac{1}{\delta}$.

Outline

The Best Arm Identification problem

$$\mathcal{R}_1 : \left\{ \boldsymbol{\mu} : \mu_1 > \max_{a \neq 1} \mu_a \right\} \quad \dots \quad \mathcal{R}_K : \left\{ \boldsymbol{\mu} : \mu_K > \max_{a \neq K} \mu_a \right\}$$

A Best Arm Identification algorithm $(A_t, \tau, \hat{\imath}_{\tau_\delta})$ made of a

- sampling rule A_t
- stopping rule τ_δ and recommendation rule $\hat{\imath}_{\tau_\delta}$

is δ- correct if

$$\forall \boldsymbol{\mu} \in \mathcal{R}, \ \mathbb{P}_{\boldsymbol{\mu}} \left(\hat{\imath}_{\tau_\delta} = \arg\max_a \mu_a \right) \geq 1 - \delta.$$

Goal: A δ-correct algorithm with small sample complexity
[Even Dar et al. 06, Kalyanakrishanan et al. 12, Gabillon et al. 12]

A good sampling rule for BAI

Theorem [Garivier and K. 2016]

For any δ-correct algorithm,

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

where

$$T^*(\boldsymbol{\mu})^{-1} = \sup_{w \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathcal{R} \setminus \mathcal{R}_{i^*(\boldsymbol{\mu})}} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a)$$

$\Sigma_K = \{w \in [0,1]^K : \sum_{i=1}^K w_i = 1\}$.

Moreover, the vector of optimal proportions

$$w^*(\boldsymbol{\mu}) = \underset{w \in \Sigma_K}{\operatorname{argmax}} \inf_{\boldsymbol{\lambda} \in \mathcal{R} \setminus \mathcal{R}_{i^*(\boldsymbol{\mu})}} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a)$$

is well-defined, and we propose an efficient way to compute it.

The Tracking sampling rule

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

- Introducing

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

the arm sampled at round $t+1$ is

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

Lemma

Under the Tracking sampling rule,

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

The Parallel GLRT for BAI

Letting $\hat{a}(t) = \underset{a}{\text{argmax}}\ \hat{\mu}_a(t)$,

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

$$= \inf\left\{t \in \mathbb{N} : \min_{b \neq \hat{a}(t)} \inf_{\boldsymbol{\lambda}:\lambda_{\hat{a}} < \lambda_b} \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \lambda_a) > \beta(t,\delta)\right\}$$

$$= \inf\left\{t : \min_{b \neq \hat{a}(t)} \underbrace{\inf_{\lambda} \left[N_{\hat{a}(t)}(t) d(\hat{\mu}_{\hat{a}}(t), \lambda) + N_b(t) d(\hat{\mu}_b(t), \lambda)\right]}_{\lambda_{\min} = \frac{N_{\hat{a}}(t)\hat{\mu}_{\hat{a}}(t) + N_b(t)\hat{\mu}_b(t)}{N_{\hat{a}}(t) + N_b(t)}} > \beta(t,\delta)\right\}$$

➜ explicit expression featuring only pairs of arms

An asymptotically optimal algorithm for BAI

Theorem [Garivier and K., 2016]

The Track-and-Stop strategy, that uses

- the **Tracking sampling rule**
- the **Parallel GLRT stopping rule** with

$$\beta(t,\delta) \simeq \ln\left(\frac{K-1}{\delta}\right) + 2\ln\ln(1/\delta) + 6\ln(1+\ln t)$$

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

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

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

Outline

A different objective

$\mathcal{B}(\mu_1)$

$\mathcal{B}(\mu_2)$

$\mathcal{B}(\mu_3)$

$\mathcal{B}(\mu_4)$

$\mathcal{B}(\mu_5)$

At round t, an agent:

- chooses an arm A_t
- observes a **reward** $X_t \sim \mathcal{B}(\mu_{A_t})$

using a sequential sampling strategy (A_t):

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

Goal: maximize the expected sum of rewards $\mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{T} X_t\right]$.

Regret

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 the *regret*:

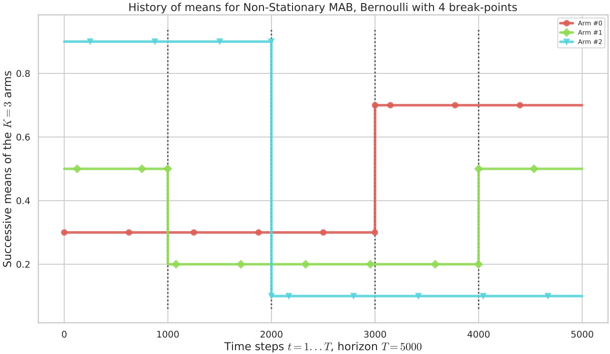
$$R_T = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right] = \sum_{a=1}^{K}(\mu^* - \mu_a)\mathbb{E}[N_a(T)]$$

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

$\Rightarrow$ **Exploration/Exploitation tradeoff**

Piecewise stationary bandit model

Sequence of means $(\mu_a(t))_t$ for each arm a
$a_t^* = \text{argmax}_a\ \mu_a(t)$: optimal arm at time t



few breakpoints: $\Upsilon_T = 4$

Goal: minimize the dynamic regret $R_T = \mathbb{E}\left[\sum_{t=1}^{T}(\mu_{a_t^*} - \mu_{A_T})\right]$

Assumption: bounded rewards, $X_t \in [0,1]$.

Positioning

(Quick) related work

- Existing guarantees for an adversarial bandit algorithm EXP3.S [Auer et al. 2002]
- Many recent attempts to adapt *stochastic bandit algorithms* to this problem: CUSUM-UCB [Liu et al, 2018], Monitored-UCB [Cao et al, 2019]
- Those attemps require the knowledge of

the number of breakpoints + a lower bound on the minimal magnitude of change

Positioning

(Quick) related work

- Existing guarantees for an adversarial bandit algorithm EXP3.S [Auer et al. 2002]
- Many recent attempts to adapt *stochastic bandit algorithms* to this problem: CUSUM-UCB [Liu et al, 2018], Monitored-UCB [Cao et al, 2019]
- Those attemps require the knowledge of

the number of breakpoints + a lower bound on the minimal magnitude of change

Our contributions:

- kl-UCB + un efficient adaptive sliding window
- no need to know anything about the size of a change

Outline

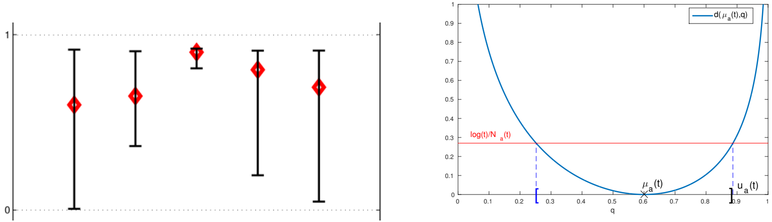
1. The bandit framework for sequential decision making
2. Active identification in a bandit model
   - A generic δ-correct stopping rule
   - Towards optimal sample complexity
   - the Best Arm Identification example
3. Rewards maximization in a non-stationary bandit model
   - The kl-UCB algorithm in the stationary case
   - A non-parametric sequential change point detector
   - kl-UCB meets the Bernoulli-GLRT

The kl-UCB algorithm

- A UCB-type (or *optimistic*) algorithm chooses at round t

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

where $\text{UCB}_a(t)$ is an Upper Confidence Bound on μ_a.



The kl-UCB index

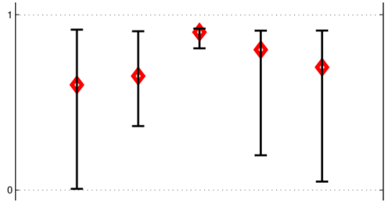
$$\text{UCB}_a(t) := \max\left\{q : d\left(\hat{\mu}_a(t), q\right) \leq \frac{\log(t)}{N_a(t)}\right\},$$

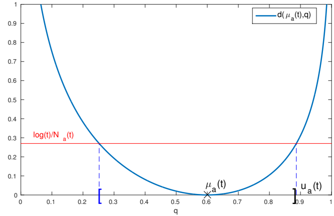
satisfies $\mathbb{P}(\mu_a \leq \text{UCB}_a(t)) \gtrsim 1 - t^{-1}$.

- A UCB-type (or *optimistic*) algorithm chooses at round t

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

where $\mathrm{UCB}_a(t)$ is an Upper Confidence Bound on μ_a.





The kl-UCB index [Cappé et al. 13]: kl-UCB satisfies

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

→ matching a lower bound by [Lai and Robbins 1985]

Outline

The Bernoulli GLRT

Question: How to detect a change in the mean of a stream of independent observations (X_t) bounded in $[0, 1]$?

Answer: a GLR test assuming a Bernoulli likelihood

$$\mathcal{H}_0 \; : \; \left(\exists \mu_0 : \forall i \in \mathbb{N}, X_i \overset{\text{i.i.d.}}{\sim} \mathcal{B}(\mu_0)\right)$$

$$\mathcal{H}_1 \; : \; \left(\exists \mu_0 \neq \mu_1, \tau \in \mathbb{N}^* : X_1, \dots, X_\tau \overset{\text{i.i.d.}}{\sim} \mathcal{B}(\mu_0) \text{ and } X_{\tau+1}, \dots \overset{\text{i.i.d.}}{\sim} \mathcal{B}(\mu_1)\right)$$

The Generalized Likelihood Ratio for this test is

$$\begin{aligned} \hat{\mathrm{GLR}}(t) &= \frac{\sup_{\mu_0, \mu_1, \tau \leq t} \ell(X_1, \dots, X_t; \mu_0, \mu_1, \tau)}{\sup_{\mu_0} \ell(X_1, \dots, X_t; \mu_0)} \\ &= \sup_{s \in [1, t]} \left[s \times \mathrm{kl}\left(\hat{\mu}_{1:s}, \hat{\mu}_{1:t}\right) + (t - s) \times \mathrm{kl}\left(\hat{\mu}_{s+1:t}, \hat{\mu}_{1:t}\right)\right] \end{aligned}$$

with $\hat{\mu}_{s:s'} = (\sum_{k=s}^{s'} X_s)/(s' - s + 1)$.

The Bernoulli GLR

Definition

Given a stream of samples $(X_s) \in [0,1]$, the Bernoulli-GLRT detects a change-point after n samples if

$$\sup_{s\in[1,n]} \left[s \times \mathrm{kl}\left(\hat{\mu}_{1:s}, \hat{\mu}_{1:n}\right) + (n-s) \times \mathrm{kl}\left(\hat{\mu}_{s+1:n}, \hat{\mu}_{1:n}\right) \right] \geq \beta(n,\delta)$$

We let T_δ be the first instant of detection.

- asymptotic study by [Lai and Xing, 2010] (for Bernoulli rewards)
- non-asymptotic properties established by [Maillard, 2018] for the Gaussian-GLR that can also be used for bounded rewards (sub-Gaussian)

Non-asymptotic properties of the Bernoulli-GLRT

- **Upper bound on the probability of false alarm**

Lemma

Assume that there exists $\mu_0 \in [0,1]$ *such that* $\mathbb{E}[X_t] = \mu_0$ *and that* $X_i \in [0,1]$ *for all* i*. Then the Bernoulli GLR test satisfies* $\mathbb{P}_{\mu_0}(T_\delta < \infty) \leq \delta$ *with the threshold function*

$$\beta(n,\delta) = 2\mathcal{T}\left(\frac{\ln(3n\sqrt{n}/\delta)}{2}\right) + 6\ln(1+\ln(n)).$$

Proof. require some modification of the martingale tools of [K. and Koolen 2018]

Non-asymptotic properties of the Bernoulli GLR

- **Upper bound on the detection delay**

Lemma

Let $\mathbb{P}_{\mu_0,\mu_1,\tau}$ be a model such that $\mathbb{E}[X_t] = \mu_0$ for $t \leq \tau$, and μ_1 for $t > \tau$, with $\mu_0 \neq \mu_1$. The Bernoulli-GLRT satisfies

$$\mathbb{P}_{\mu_0,\mu_1,\tau}(T_\delta \geq \tau + u)$$
$$\leq \exp\left(-\frac{2\tau u}{\tau + u}\left(\max\left[0, \Delta - \sqrt{\frac{\tau + u}{2\tau u}\beta(\tau + u, \delta)}\right]\right)^2\right)$$

with $\Delta = |\mu_1 - \mu_0|$.

Proof. Pinsker's inequality and similar technique as for the sub-Gaussian case [Maillard 2018].

Outline

1. The bandit framework for sequential decision making
2. Active identification in a bandit model
   - A generic δ-correct stopping rule
   - Towards optimal sample complexity
   - the Best Arm Identification example
3. Rewards maximization in a non-stationary bandit model
   - The kl-UCB algorithm in the stationary case
   - A non-parametric sequential change point detector
   - kl-UCB meets the Bernoulli-GLRT

The GLR-kl-UCB algorithm

Parameters: $\alpha \in (0,1)$, $\delta > 0$.
Arm selection: at round t,

- if $\alpha > 0$ and $t \mod \lfloor K/\alpha \rfloor \in \{1, \ldots, K\}$,

$$(\textit{forced exploration}) \quad A_t \leftarrow t \mod \lfloor K/\alpha \rfloor$$

- else, select

$$(\text{kl-}\textit{UCB}) \qquad A_t \leftarrow \arg\max_a \mathrm{UCB}_a(t)$$

$\tau_a(t)$: instant of the last **restart**
$n_a(t)$: number of selection of arm a since the last restart
$\hat{\mu}_a(t)$: empirical mean of samples from arm a since last restart

$$\mathrm{UCB}_a(t) := \max\big\{q \in [0,1] : n_a(t) \times \mathrm{kl}\left(\hat{\mu}_a(t), q\right) \leq f(t - \tau_a(t))\big\}.$$

Restarts: Local or Global after a change is detected by the Bernoulli-GLRT on the mean of the selected arm

Results

- a unified analysis of Local and Global changes
- a tuning of the algorithm that ensures $O(\Upsilon_T \sqrt{T})$ when Υ_T is unknown and $O(\sqrt{\Upsilon_T T})$ regret if Υ_T is known

Theorem

For piece-wise stationnary instances in which the breakpoints are "far enough"

1. Choosing $\alpha = \sqrt{\frac{\ln(T)}{T}}$, $\delta = \frac{1}{\sqrt{T}}$ gives
$$R_T = O\left(\frac{K}{(\Delta^{\text{change}})^2}\Upsilon_T\sqrt{T\ln(T)} + \frac{(K-1)}{\Delta^{\text{opt}}}\Upsilon_T \ln(T)\right),$$
2. Choosing $\alpha = \sqrt{\frac{\Upsilon_T \ln(T)}{T}}$, $\delta = \frac{1}{\sqrt{\Upsilon_T T}}$ gives
$$R_T = O\left(\frac{K}{(\Delta^{\text{change}})^2}\sqrt{\Upsilon_T T\ln(T)} + \frac{(K-1)}{\Delta^{\text{opt}}}\Upsilon_T \ln(T)\right).$$

Results

- Good practical performance!

Algorithmes \ Problèmes	Pb 1	Pb 2	Pb 3
Oracle-Restart kl-UCB	**37 ± 37**	**45 ± 34**	**257 ± 86**
kl-UCB	270 ± 76	162 ± 59	529 ± 148
Discounted- kl-UCB	1456 ± 214	1442 ± 440	1376 ± 37
SW- kl-UCB	177 ± 34	182 ± 34	1794 ± 71
M- kl-UCB	290 ± 29	534 ± 93	645 ± 141
CUSUM- kl-UCB	148 ± 32	152 ± 42	**490 ± 133**
GLR-kl-UCB (Local)	**74 ± 31**	**113 ± 34**	513 ± 97
GLR - kl-UCB (Global)	97 ± 32	134 ± 33	621 ± 103

Table: Mean regret for different algorithms at time T on three piecewise stationary bandit instances ($T = 5000$ for 1,2 and $T = 20000$ for 3).

Thanks!

References:

- A. Garivier, E. Kaufmann, Optimal Best Arm Identification with Fixed Confidence, COLT 2016
- E. Kaufmann, W. Koolen, Mixture Martingale Revisited and Applications to Sequential Tests and Confidence Intervals, arXiv 2018
- L. Besson, E. Kaufmann, The Generalized Likelihood Ratio Test meets klUCB: an Improved Algorithm for Piece-Wise Non-Stationary Bandits, arXiv 2019
- A. Garivier, E. Kaufmann, Non-Asymptotic Sequential Tests for Overlapping Hypotheses and application to near optimal arm identification in bandit models (soon on arXiv)