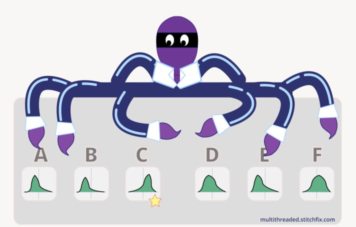
# Contributions to the Optimal Solution of Several Bandit Problems

Emilie Kaufmann

HDR defense
November 13th, 2020

# The stochastic MAB model

Arms = probability distributions an agent can choose from:




In each round $t$, the agent

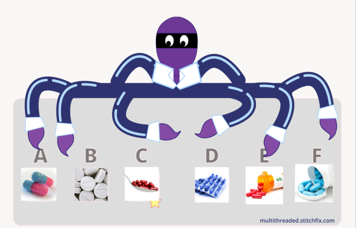
- selects arm $A_t \in \{1, \ldots, A\}$
- observes a sample $X_t \sim \nu_{A_t}$ independent from past data

sequential protocol:

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

# The stochastic MAB model

Arms = probability distributions an agent can choose from:




In each round $t$, the agent

- selects arm $A_t \in \{1, \dots, A\}$
- observes a sample $X_t \sim \mathcal{B}(\mu_{A_t})$ independent from past data

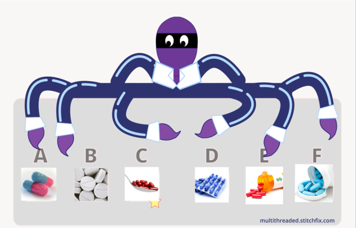
sequential protocol:

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

**Assumption** (in this work): arms are simple distribution parameterized by their means (e.g. Bernoulli , exponential families)

# The stochastic MAB model

Arms = probability distributions an agent can choose from:




In each round $t$, the agent

- selects arm $A_t \in [A]$
- observes a sample $X_t \sim \mathcal{B}(\mu_{A_t})$ independent from past data

sequential protocol:

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

**Assumption** (in this work): arms are simple distribution parameterized by their means (e.g. Bernoulli , exponential families)

**Notation**: $\nu_a = \nu_{\mu_a}$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_A) \in \mathcal{I}^A$.

# One bandit model, many bandit problems

**rewards maximization... with a twist**

- feedback $\neq$ reward [Ch. 1]
- structured bandits [Ch. 1]
- multi-player bandits [Ch. 2]

**pure exploration**

- a generic stopping rule for active identification [Ch. 3]
- the complexity of best arm identification [Ch. 4]
- two MCTS-related examples [Ch. 5]

Common emphasis on designing *optimal* algorithms

# One bandit model, many bandit problems

**rewards maximization...**
**with a twist**

- feedback $\neq$ reward 
- structured bandits 
- multi-player bandits 

**pure exploration**

- a generic stopping rule for active identification
- the complexity of best arm identification
- two MCTS-related examples

Common emphasis on designing *optimal* algorithms

Research *motivated by* some applications

# One bandit model, many bandit problems

**rewards maximization...**
**with a twist**

- feedback $\neq$ reward 
- structured bandits 
- multi-player bandits 

**pure exploration**

- a generic stopping rule for active identification
- the complexity of best arm identification
- two MCTS-related examples

Common emphasis on designing *optimal* algorithms

Research *motivated by* some applications

# Technical tools

- Lower bounds...
  and how they inspire algorithms
- Mixture martingales for new deviation inequalities
- Recent tools for the analysis of Thompson Sampling
  [Agrawal and Goyal, 2013, Russo, 2016]

# Outline

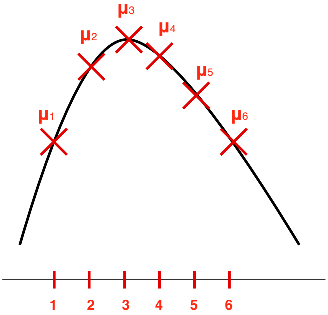
1. Thompson Sampling for a Structured Bandit Problem
2. The Complexity of Pure Exploration
3. Thompson Sampling for Pure Exploration?

# Outline

1. Thompson Sampling for a Structured Bandit Problem
2. The Complexity of Pure Exploration
3. Thompson Sampling for Pure Exploration?

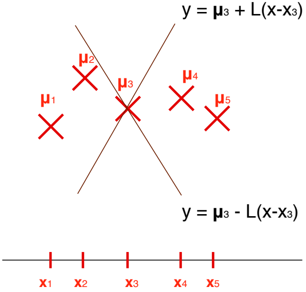
- Classical bandits: $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_A) \in \mathcal{I}^A$
- Structured bandits: $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_A) \in \mathcal{S} \subset \mathcal{I}^A$

➜ can we exploit the knowledge of $\mathcal{S}$ to gain more reward?




unimodal bandit

[Combes and Proutière, 2014]




Lipschitz bandit

[Magureanu et al., 2014]

# Lower Bounds can help

In each round $t$, the agent

- selects arm $A_t \in [A]$, observes a *reward* $X_t \sim \nu_{A_t}$

**Goal**: maximize the expected total reward $\leftrightarrow$ minimize the regret

$$\begin{aligned}\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}, T) &= \mu_\star T - \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{T} X_t\right] \\ &= \sum_{a\in[A]} (\mu_\star - \mu_a)\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]\end{aligned}$$

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

**Theorem** [Graves and Lai, 1997] (Theorem 1.8 in the HDR document)

Let $\mathcal{A}$ be such that $\forall \boldsymbol{\mu} \in \mathcal{S}, \forall \alpha \in (0,1]$, $\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}, T) = o(T^\alpha)$.

$$\forall \boldsymbol{\mu} \in \mathcal{S}, \quad \lim_{T\to\infty} \frac{\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}, T)}{\log(T)} \geq C_{\mathcal{S}}(\boldsymbol{\mu}).$$

➜ $\mathcal{A}$ is asymptotically optimal if $\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}, T) = C_{\mathcal{S}}(\boldsymbol{\mu})\log(T) + o(\log(T))$

# Lower Bounds can help

In each round $t$, the agent

- selects arm $A_t \in [A]$, observes a *reward* $X_t \sim \nu_{A_t}$

**Goal**: maximize the expected total reward $\leftrightarrow$ minimize the regret

$$\begin{aligned}\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}, T) &= \mu_{a_\star} T - \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{T} X_t\right] \\ &= \sum_{a\in[A]} (\mu_{a_\star} - \mu_a)\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]\end{aligned}$$

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

**Theorem** [Graves and Lai, 1997] (Theorem 1.8 in the HDR document)

Let $\mathcal{A}$ be such that $\forall \boldsymbol{\mu} \in \mathcal{S}, \forall \alpha \in (0,1]$, $\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}, T) = o(T^\alpha)$.

$$\forall \boldsymbol{\mu} \in \mathcal{S}, \quad \lim_{T\to\infty} \frac{\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}, T)}{\log(T)} \geq C_{\mathcal{S}}(\boldsymbol{\mu}).$$

➜ $\mathcal{A}$ is asymptotically optimal if $\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}, T) = C_{\mathcal{S}}(\boldsymbol{\mu}) \log(T) + o(\log(T))$

# Lower bounds can help

$C_{\mathcal{S}}(\boldsymbol{\mu})$ features the Kullback-Leibler divergence $d(\mu, \mu') := \mathrm{KL}(\nu_\mu, \nu_{\mu'})$

- $\mathcal{S} = \mathcal{I}^A$, $C_{\mathcal{S}}(\boldsymbol{\mu}) = \sum_{a=1}^{A} \frac{\mu_\star - \mu_a}{d(\mu_a, \mu_\star)}$ [Lai and Robbins, 1985]
- in general, $C_{\mathcal{S}}(\boldsymbol{\mu})$ has no closed-form expression (solution of a complex optimization problem)

## Special case [Combes and Proutière, 2014]

$\boldsymbol{\mu}$ is unimodal with respect to a graph $G = ([A], E)$: for all $a \in [A]$ there exists an increasing path to the optimal arm $a_\star$:

$$(a_1 = a, \dots, a_{m_a} = a_\star) : (a_i, a_{i+1}) \in E \text{ and } \mu_{a_i} < \mu_{a_{i+1}}.$$

For graphical unimodal bandits,

$$C_{\mathcal{S}}(\boldsymbol{\mu}) = \sum_{a \in \mathcal{N}_G(a_\star)} \frac{\mu_\star - \mu_a}{d(\mu_a, \mu_\star)} \quad \mathcal{N}_G(a^\star) = \{a : (a, a_\star) \in E\}$$

➜ an optimal algorithm focusses on neighbors of the optimal arm

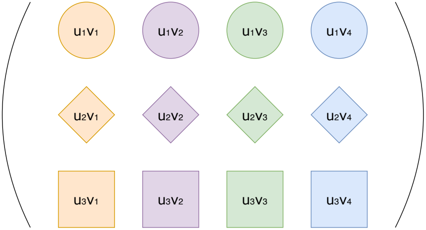
# Solving Rank-One Bandits

$$\mathcal{S}_{R1} = \left\{ \boldsymbol{\mu} = (\mu_{(k,\ell)})_{\substack{1\leq k\leq K \\ 1\leq \ell\leq L}} \;\middle|\; \exists \boldsymbol{u} \in [0,1]^K, \boldsymbol{v} \in [0,1]^L : \mu_{(k,\ell)} = u_k v_\ell \right\}$$

[Katariya et al., 2017]

**Example:** content optimization with two independent factors

**μ** =




clic probability $\mu_{(k,\ell)} = u_k \times v_\ell$

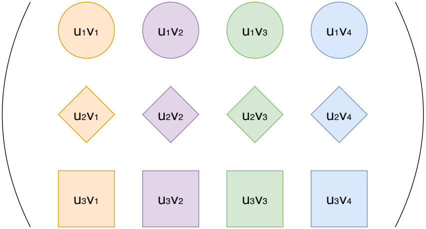
# Solving Rank-One Bandits

$$\mathcal{S}_{R1} = \left\{ \boldsymbol{\mu} = (\mu_{(k,\ell)})_{\substack{1\leq k\leq K \\ 1\leq \ell \leq L}} \;\middle|\; \exists \boldsymbol{u} \in [0,1]^K, \boldsymbol{v} \in [0,1]^L : \mu_{(k,\ell)} = u_k v_\ell \right\}$$

[Katariya et al., 2017]

**Example:** content optimization with two independent factors

$\boldsymbol{\mu}$ =




clic probability $\mu_{(k,\ell)} = u_k \times v_\ell$

## Key observation

$\boldsymbol{\mu}$ is unimodal with respect to the graph $G_1 = ([K]\times[L], E)$

$$\big((i,j),(k,\ell)\big) \in E \;\text{ if }\; (i = k \text{ (x)or } j = \ell)$$

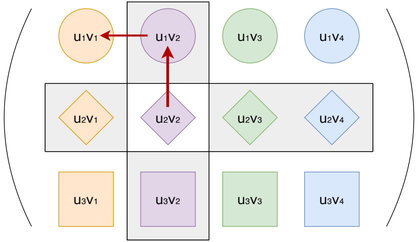
# Solving Rank-One Bandits

$$\mathcal{S}_{R1} = \left\{ \boldsymbol{\mu} = (\mu_{(k,\ell)})_{\substack{1\leq k\leq K\\ 1\leq \ell\leq L}} \;\middle|\; \exists \boldsymbol{u} \in [0,1]^K, \boldsymbol{v} \in [0,1]^L : \mu_{(k,\ell)} = u_k v_\ell \right\}$$

[Katariya et al., 2017]

**Example:** content optimization with two independent factors

**μ** =




clic probability $\mu_{(k,\ell)} = u_k \times v_\ell$

## Key observation

$\boldsymbol{\mu}$ is unimodal with respect to the graph $G_1 = ([K]\times[L], E)$

$$((i,j),(k,\ell)) \in E \;\text{ if }\; (i = k \text{ (x)or } j = \ell)$$

## Unimodal Thompson Sampling for Rank-One Bandits

**Idea:** use an optimal algorithm for graphical unimodal bandits

- Unimodal Thompson Sampling [Paladino et al., 2017]

**UTS with parameter $\gamma \in \{2, 3, \dots\}$ for Bernoulli bandits**

In each round $t+1$:

- compute the empirical *leader* $B_{t+1} = \underset{a\in[A]}{\text{argmax}}\ \hat{\mu}_a(t)$
- if $\ell_{B_{t+1}}(t+1) = 0[\gamma]$, select $A_{t+1} = B_{t+1}$ (leader exploration)
- else, draw **posterior samples** for arms in $\mathcal{N}_G(B_{t+1}) \cup \{B_{t+1}\}$:

$$\theta_a(t) \sim \mathrm{Beta}\big(S_a(t)+1, N_a(t)-S_a(t)+1\big)$$

and $A_{t+1} = \underset{a\in\mathcal{N}_G(B_{t+1})\cup\{B_{t+1}\}}{\text{argmax}}\ \theta_a(t)$ (TS around the leader)

$S_a(t) = \sum_{s=1}^{t} X_s \mathbb{1}(A_s = a)$: sum of rewards from arm $a$
$\hat{\mu}_a(t) = S_a(t)/N_a(t)$: empirical mean of arm $a$

$\ell_b(t) = \sum_{s=1}^{t} \mathbb{1}(B_s = b)$: number of times arm $b$ has been leader

# Unimodal Thompson Sampling for Rank-One Bandits

**Idea:** use an optimal algorithm for graphical unimodal bandits

- Unimodal Thompson Sampling [Paladino et al., 2017]

## UTS with parameter $\gamma \in \{2, 3, \dots\}$ for Bernoulli bandits

In each round $t+1$:

- compute the empirical *leader* $B_{t+1} = \underset{(k,\ell)\in[K]\times[L]}{\text{argmax}} \ \hat{\mu}_{(k,\ell)}(t)$
- if $\ell_{B_{t+1}}(t+1) = 0[\gamma]$, select $A_{t+1} = B_{t+1}$ (leader exploration)
- else, draw **posterior samples** for arms in $\mathcal{N}_G(B_{t+1}) \cup \{B_{t+1}\}$:

$$\theta_{(k,\ell)}(t) \sim \mathrm{Beta}\big(S_{(k,\ell)}(t)+1, N_{(k,\ell)}(t) - S_{(k,\ell)}(t)+1\big)$$

and $A_{t+1} = \underset{(k,\ell)\in\{(k',B_t^2)\}\cup\{(B_t^1,\ell')\}}{\text{argmax}} \ \theta_{(k,\ell)}(t)$ (TS around the leader)

$S_a(t) = \sum_{s=1}^t X_s \mathbb{1}(A_s = a)$: sum of rewards from arm $a$
$\hat{\mu}_a(t) = S_a(t)/N_a(t)$: empirical mean of arm $a$

$\ell_b(t) = \sum_{s=1}^t \mathbb{1}(B_s = b)$: number of times arm $b$ has been leader

# Unimodal Thompson Sampling for Rank-One Bandits

**Theorem** [Trinh, K., Vernade, Combes, ALT 2020]

Let $\boldsymbol{\mu}$ be a unimodal bandit instance with respect to a graph $G$, with Bernoulli rewards. For all $\gamma \geq 2$, UTS with parameter $\gamma$ satisfies, for every $\varepsilon > 0$,

$$\mathcal{R}_{\boldsymbol{\mu}}(\mathrm{UTS}(\gamma), T) \leq (1+\varepsilon) \sum_{a \in \mathcal{N}_G(a_\star)} \frac{(\mu_\star - \mu_a)}{d(\mu_a, \mu_\star)} \log(T) + C(\boldsymbol{\mu}, \gamma, \varepsilon).$$

- a novel analysis, valid for any leader exploration parameter $\gamma$, with $\gamma = 2$ being the best choice in practice
- $\mathrm{UTS}(\gamma)$ is asymptotically optimal for Rank-One bandits (matching the existing lower bound of [Katariya et al., 2017])
- ... and greatly outperforms the previous state-of-the-art

# Outline


1. Thompson Sampling for a Structured Bandit Problem
2. The Complexity of Pure Exploration
3. Thompson Sampling for Pure Exploration?

# Active Identification in a bandit model

**Goal:** answer *some question* about the unknown mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_A)$ by adaptively sampling the arms

**Input:**

- $\mathcal{R} \subseteq \mathcal{I}^A$ a subset that contains $\boldsymbol{\mu}$
- $I$ regions $\mathcal{R}_1, \ldots, \mathcal{R}_I$ such that $\mathcal{R} \subseteq \bigcup_{i=1}^{I} \mathcal{R}_i$

**Output:** one region $\mathcal{R}_i$ that contains $\boldsymbol{\mu}$.

## Active Identification with fixed-confidence

Given a risk parameter $\delta \in (0, 1)$, the goal is to build a

- sampling rule $(A_t)$
- stopping rule $\tau$
- recommendation rule $\hat{\imath}_\tau \in [I]$

such that $\mathbb{P}_{\boldsymbol{\mu}}\left(\boldsymbol{\mu} \notin \mathcal{R}_{\hat{\imath}_\tau}\right) \leq \delta$ and the sample complexity $\tau$ is small.

# Best Arm Identification

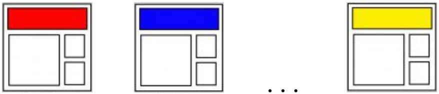
➜ Identify the arm with largest mean:

$$\mathcal{R} = \left\{ \boldsymbol{\mu} \in \mathcal{I}^A : \exists a \in [A] : \mu_a > \max_{b \neq a} \mu_b \right\}$$

$$\text{and} \quad \mathcal{R}_i = \left\{ \boldsymbol{\mu} \in \mathcal{I}^A : \mu_i > \max_{b \neq i} \mu_b \right\} \quad \text{for } i \in [A]$$

[Even-Dar et al., 2006]

**Example**: identify the version of a webpage with the largest conversion probability (A/B/C testing)



$\mu_1$ $\quad$ $\mu_2$ $\quad$ $\dots$ $\quad$ $\mu_A$

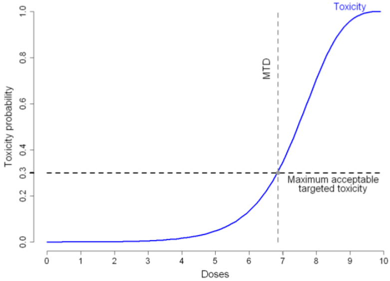
# Bandits and thresholds

➜ Identify the arm whose mean is the closest to some threshold:

$$\mathcal{R}_i = \left\{ \boldsymbol{\mu} \in \mathcal{R} : |\mu_i - \theta| = \min_a |\mu_a - \theta| \right\}$$

[Garivier et al., 2019a] [Aziz, K., Rivière, JMLR 2021]

**Motivation:** identify the Maximum Tolerated Dose in a dose-finding clinical trial



# Designing a good stopping rule

Let us fix some sampling rule $(A_t)_{t\in\mathbb{N}}$, giving a data stream

$$A_1, X_1, A_2, X_2, \dots, A_t, X_t, \dots \quad \text{where} \quad X_t \sim \nu_{\mu_{A_t}}$$

**Goal:** construct a *sequential test* $(\tau, \hat{\imath}_\tau)$ for the hypotheses

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

➜ multiple, composite hypotheses (possibly overlapping)

## Definition

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

- $\tau$ is a stopping time with respect to $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$
- $\hat{\imath}_\tau \in [I]$ is $\mathcal{F}_\tau$-measurable

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

# The Parallel GLRT rule

**Idea:** run $I$ statistical tests of

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

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

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

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

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

[Wilks, 1938]

$\hat{\boldsymbol{\mu}}(t) = (\hat{\mu}_1(t), \dots, \hat{\mu}_A(t))$, Maximum Likelihood Estimator.

# The Parallel GLRT rule

## Parallel GLRT

Given some threshold function $\beta(t,\delta)$,

$$\tau_\delta = \inf\left\{ t \in \mathbb{N} : \max_{i\in[I]} \inf_{\boldsymbol{\lambda}\in\mathcal{R}\setminus\mathcal{R}_i} \log \frac{\ell(X_1,\dots,X_t;\hat{\boldsymbol{\mu}}(t))}{\ell(X_1,\dots,X_t;\boldsymbol{\lambda})} > \beta(t,\delta) \right\}$$

$$\hat{\imath}_{\tau_\delta} \in \operatorname*{arg\,max}_{i\in[I]} \inf_{\boldsymbol{\lambda}\in\mathcal{R}\setminus\mathcal{R}_i} \log \frac{\ell(X_1,\dots,X_{\tau_\delta};\hat{\boldsymbol{\mu}}(\tau_\delta))}{\ell(X_1,\dots,X_{\tau_\delta};\boldsymbol{\lambda})}$$

In an exponential family bandit model,

$$\log \frac{\ell(X_1,\dots,X_t;\hat{\boldsymbol{\mu}}(t))}{\ell(X_1,\dots,X_t;\boldsymbol{\lambda})} = \sum_{a\in[A]} N_a(t) d(\hat{\mu}_a(t),\lambda_a)$$

with $d(\mu,\mu') = \mathrm{KL}\left(\nu_\mu,\nu_{\mu'}\right)$.

(rewards in a one-parameter exponential family: Bernoulli, Gaussian, Poisson...)

# The Parallel GLRT rule

## Parallel GLRT

Given some threshold function $\beta(t,\delta)$,

$$\begin{aligned}\tau_\delta &= \inf\left\{t \in \mathbb{N} : \max_{i\in[I]} \inf_{\boldsymbol{\lambda}\in\mathcal{R}\backslash\mathcal{R}_i} \sum_{a\in[A]} N_a(t) d(\hat{\mu}_a(t), \lambda_a) > \beta(t,\delta)\right\} \\ \hat{\imath}_{\tau_\delta} &\in \operatorname*{arg\,max}_{i\in[I]} \inf_{\boldsymbol{\lambda}\in\mathcal{R}\backslash\mathcal{R}_i} \sum_{a\in[A]} N_a(\tau_\delta) d(\hat{\mu}_a(\tau_\delta), \lambda_a)\end{aligned}$$

In an exponential family bandit model,

$$\log \frac{\ell(X_1,\dots,X_t;\hat{\boldsymbol{\mu}}(t))}{\ell(X_1,\dots,X_t;\boldsymbol{\lambda})} = \sum_{a\in[A]} N_a(t) d(\hat{\mu}_a(t), \lambda_a)$$

with $d(\mu,\mu') = \mathrm{KL}\left(\nu_\mu, \nu_{\mu'}\right)$.

(rewards in a one-parameter exponential family: Bernoulli, Gaussian, Poisson...)

# Upper bound on the error probability

For any sampling rule, under the GLRT stopping rule,

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

# Upper bound on the error probability

For any sampling rule, under the GLRT stopping rule,

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

**Wanted:** a deviation inequality in which

➜ deviations are measured with KL-divergence

# Upper bound on the error probability

For any sampling rule, under the GLRT stopping rule,

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

**Wanted:** a deviation inequality in which

- ➜ deviations are measured with KL-divergence
- ➜ deviations are uniform over time (*martingales*...)

# Upper bound on the error probability

For any sampling rule, under the GLRT stopping rule,

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

**Wanted:** a deviation inequality in which

- ➜ deviations are measured with KL-divergence
- ➜ deviations are uniform over time (*martingales...*)
- ➜ deviations take into account multiple arms (*..products*)

# A universal $\delta$-correct stopping rule

**Theorem** [K. and Koolen, 2018, under review]

Let $\boldsymbol{\mu}$ be an exponential family bandit model. There exists a threshold function $\mathcal{T}(x) \simeq x + \log(x)$ such that, for any subset $\mathcal{S} \subseteq [A]$, for all $x > 0$,

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

**Consequence**: the Parallel GLRT stopping rule with threshold

$$\beta(t, \delta) = 3A \log(1 + \log t) + A\mathcal{T}\left(\frac{\log(1/\delta)}{A}\right)$$

is $\delta$-correct

- ➜ for any active identification problem
- ➜ regardless of the sampling rule

# A universal $\delta$-correct stopping rule

**Theorem** [K. and Koolen, 2018, under review]

Let $\boldsymbol{\mu}$ be an exponential family bandit model. There exists a threshold function $\mathcal{T}(x) \simeq x + \log(x)$ such that, for any subset $\mathcal{S} \subseteq [A]$, for all $x > 0$,

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

**Consequence**: the Parallel GLRT stopping rule with threshold

$$\beta(t, \delta) \simeq \log(1/\delta) + A \log\log(1/\delta) + 3A \log\log(t)$$

is $\delta$-correct

- ➜ for any active identification problem
- ➜ regardless of the sampling rule

# A universal $\delta$-correct stopping rule

**Theorem** [K. and Koolen, 2018, under review]

Let $\boldsymbol{\mu}$ be an exponential family bandit model. There exists a threshold function $\mathcal{T}(x) \simeq x + \log(x)$ such that, for any subset $\mathcal{S} \subseteq [A]$, for all $x > 0$,

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

**Consequence**: the Parallel GLRT stopping rule with threshold

$$\beta(t, \delta) \simeq \log(1/\delta) + A \log\log(1/\delta) + 3A \log\log(t)$$

is $\delta$-correct

- ➜ for any active identification problem
- ➜ regardless of the sampling rule

The sample complexity $\tau_\delta$ crucially depends on the sampling rule!

# Best achievable sample complexity

$\mathcal{R} = \bigcup_{i=1}^{I} \mathcal{R}_i$ forms a partition
$i_\star(\boldsymbol{\mu})$: unique region that contains $\boldsymbol{\mu}$.

**Theorem** [K. and Garivier, COLT 2016]

Any $\delta$-correct algorithm satisfies, for all $\boldsymbol{\mu} \in \mathcal{R}$,

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq T^\star(\boldsymbol{\mu}) \log(1/(3\delta))$$

with

$$T^\star(\boldsymbol{\mu})^{-1} = \sup_{w \in \Sigma_A} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a \in [A]} w_a d(\mu_a, \lambda_a).$$

$\Sigma_A = \{w \in [0,1]^A : \sum_{a \in [A]} w_a = 1\}$ $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} : i_\star(\boldsymbol{\lambda}) \neq i_\star(\boldsymbol{\mu})\}$

**Proof.** change of distribution between $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ : $i_\star(\boldsymbol{\lambda}) \neq i_\star(\boldsymbol{\mu})$

$$\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{X_1,\dots,X_\tau}, \mathbb{P}_{\boldsymbol{\lambda}}^{X_1,\dots,X_\tau}\right) \geq \mathrm{kl}\left(\mathbb{P}_{\boldsymbol{\mu}}(\hat{\imath}_\tau = i_\star(\boldsymbol{\lambda})), \mathbb{P}_{\boldsymbol{\lambda}}(\hat{\imath}_\tau = i_\star(\boldsymbol{\lambda}))\right)$$

with $\mathrm{kl}(x,y) = \mathrm{KL}(\mathcal{B}(x), \mathcal{B}(y))$. [Garivier et al., 2019b]

# Best achievable sample complexity

$\mathcal{R} = \bigcup_{i=1}^{I} \mathcal{R}_i$ forms a partition
$i_\star(\boldsymbol{\mu})$: unique region that contains $\boldsymbol{\mu}$.

**Theorem** [K. and Garivier, COLT 2016]

Any $\delta$-correct algorithm satisfies, for all $\boldsymbol{\mu} \in \mathcal{R}$,

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq T^\star(\boldsymbol{\mu}) \log(1/(3\delta))$$

with

$$T^\star(\boldsymbol{\mu})^{-1} = \sup_{w \in \Sigma_A} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a \in [A]} w_a d(\mu_a, \lambda_a).$$

$\Sigma_A = \{w \in [0,1]^A : \sum_{a \in [A]} w_a = 1\}$ $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} : i_\star(\boldsymbol{\lambda}) \neq i_\star(\boldsymbol{\mu})\}$

**Proof.** change of distribution between $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ : $i_\star(\boldsymbol{\lambda}) \neq i_\star(\boldsymbol{\mu})$

$$\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{X_1,\dots,X_\tau}, \mathbb{P}_{\boldsymbol{\lambda}}^{X_1,\dots,X_\tau}\right) \geq \mathrm{kl}\big(\underbrace{\mathbb{P}_{\boldsymbol{\mu}}(\hat{\imath}_\tau = i_\star(\boldsymbol{\lambda}))}_{\leq \delta}, \underbrace{\mathbb{P}_{\boldsymbol{\lambda}}(\hat{\imath}_\tau = i_\star(\boldsymbol{\lambda}))}_{\geq 1-\delta}\big)$$

with $\mathrm{kl}(x,y) = \mathrm{KL}(\mathcal{B}(x), \mathcal{B}(y))$. [Garivier et al., 2019b]

# Best achievable sample complexity

$\mathcal{R} = \bigcup_{i=1}^{I} \mathcal{R}_i$ forms a partition
$i_\star(\boldsymbol{\mu})$: unique region that contains $\boldsymbol{\mu}$.

**Theorem** [K. and Garivier, COLT 2016]

Any $\delta$-correct algorithm satisfies, for all $\boldsymbol{\mu} \in \mathcal{R}$,

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq T^\star(\boldsymbol{\mu}) \log(1/(3\delta))$$

with

$$T^\star(\boldsymbol{\mu})^{-1} = \sup_{w \in \Sigma_A} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a \in [A]} w_a d(\mu_a, \lambda_a).$$

$\Sigma_A = \{w \in [0,1]^A : \sum_{a \in [A]} w_a = 1\}$ $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} : i_\star(\boldsymbol{\lambda}) \neq i_\star(\boldsymbol{\mu})\}$

**Proof.** change of distribution between $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ : $i_\star(\boldsymbol{\lambda}) \neq i_\star(\boldsymbol{\mu})$

$$\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{X_1,\dots,X_\tau}, \mathbb{P}_{\boldsymbol{\lambda}}^{X_1,\dots,X_\tau}\right) \geq \mathrm{kl}\left(\delta, 1-\delta\right)$$

with $\mathrm{kl}(x,y) = \mathrm{KL}(\mathcal{B}(x), \mathcal{B}(y))$. [Garivier et al., 2019b]

# Best achievable sample complexity

$\mathcal{R} = \bigcup_{i=1}^{I} \mathcal{R}_i$ forms a partition
$i_\star(\boldsymbol{\mu})$: unique region that contains $\boldsymbol{\mu}$.

**Theorem** [K. and Garivier, COLT 2016]

Any $\delta$-correct algorithm satisfies, for all $\boldsymbol{\mu} \in \mathcal{R}$,

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq T^\star(\boldsymbol{\mu}) \log(1/(3\delta))$$

with

$$T^\star(\boldsymbol{\mu})^{-1} = \sup_{w \in \Sigma_A} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a \in [A]} w_a d(\mu_a, \lambda_a).$$

$\Sigma_A = \{w \in [0,1]^A : \sum_{a \in [A]} w_a = 1\}$ $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} : i_\star(\boldsymbol{\lambda}) \neq i_\star(\boldsymbol{\mu})\}$

**Proof.** change of distribution between $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ : $i_\star(\boldsymbol{\lambda}) \neq i_\star(\boldsymbol{\mu})$

$$\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{X_1,\dots,X_\tau}, \mathbb{P}_{\boldsymbol{\lambda}}^{X_1,\dots,X_\tau}\right) \geq \log(1/(3\delta))$$

# Best achievable sample complexity

$\mathcal{R} = \bigcup_{i=1}^{I} \mathcal{R}_i$ forms a partition
$i_\star(\boldsymbol{\mu})$: unique region that contains $\boldsymbol{\mu}$.

**Theorem** [K. and Garivier, COLT 2016]

Any $\delta$-correct algorithm satisfies, for all $\boldsymbol{\mu} \in \mathcal{R}$,

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq T^\star(\boldsymbol{\mu}) \log(1/(3\delta))$$

with

$$T^\star(\boldsymbol{\mu})^{-1} = \sup_{w \in \Sigma_A} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a \in [A]} w_a d(\mu_a, \lambda_a).$$

$\Sigma_A = \{w \in [0,1]^A : \sum_{a \in [A]} w_a = 1\}$ $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} : i_\star(\boldsymbol{\lambda}) \neq i_\star(\boldsymbol{\mu})\}$

**Proof.** change of distribution between $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ : $i_\star(\boldsymbol{\lambda}) \neq i_\star(\boldsymbol{\mu})$

$$\sum_{a \in [A]} \mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau)] d(\mu_a, \lambda_a) \geq \log(1/(3\delta))$$

# Best achievable sample complexity

$\mathcal{R} = \bigcup_{i=1}^{I} \mathcal{R}_i$ forms a partition
$i_\star(\boldsymbol{\mu})$: unique region that contains $\boldsymbol{\mu}$.

**Theorem** [K. and Garivier, COLT 2016]

Any $\delta$-correct algorithm satisfies, for all $\boldsymbol{\mu} \in \mathcal{R}$,

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq T^\star(\boldsymbol{\mu}) \log(1/(3\delta))$$

with

$$T^\star(\boldsymbol{\mu})^{-1} = \sup_{w \in \Sigma_A} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a \in [A]} w_a d(\mu_a, \lambda_a).$$

$\Sigma_A = \{w \in [0,1]^A : \sum_{a \in [A]} w_a = 1\}$ $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} : i_\star(\boldsymbol{\lambda}) \neq i_\star(\boldsymbol{\mu})\}$

**Proof.** change of distribution between $\boldsymbol{\mu}$ and $\boldsymbol{\lambda} : i_\star(\boldsymbol{\lambda}) \neq i_\star(\boldsymbol{\mu})$

$$\inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a \in [A]} \mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau)] d(\mu_a, \lambda_a) \geq \log(1/(3\delta))$$

# Best achievable sample complexity

$\mathcal{R} = \bigcup_{i=1}^{I} \mathcal{R}_i$ forms a partition
$i_\star(\boldsymbol{\mu})$: unique region that contains $\boldsymbol{\mu}$.

**Theorem** [K. and Garivier, COLT 2016]

Any $\delta$-correct algorithm satisfies, for all $\boldsymbol{\mu} \in \mathcal{R}$,

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq T^\star(\boldsymbol{\mu}) \log(1/(3\delta))$$

with

$$T^\star(\boldsymbol{\mu})^{-1} = \sup_{w \in \Sigma_A} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a \in [A]} w_a d(\mu_a, \lambda_a).$$

$\Sigma_A = \{w \in [0,1]^A : \sum_{a \in [A]} w_a = 1\}$ $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} : i_\star(\boldsymbol{\lambda}) \neq i_\star(\boldsymbol{\mu})\}$

**Proof.** change of distribution between $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ : $i_\star(\boldsymbol{\lambda}) \neq i_\star(\boldsymbol{\mu})$

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \times \left[ \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a \in [A]} \underbrace{\frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau)]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}}_{w_a} d(\mu_a, \lambda_a) \right] \geq \log(1/(3\delta))$$

# Best achievable sample complexity

$\mathcal{R} = \bigcup_{i=1}^{I} \mathcal{R}_i$ forms a partition
$i_\star(\boldsymbol{\mu})$: unique region that contains $\boldsymbol{\mu}$.

**Theorem [K. and Garivier, COLT 2016]**

Any $\delta$-correct algorithm satisfies, for all $\boldsymbol{\mu} \in \mathcal{R}$,

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq T^\star(\boldsymbol{\mu}) \log(1/(3\delta))$$

with

$$T^\star(\boldsymbol{\mu})^{-1} = \sup_{w \in \Sigma_A} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a \in [A]} w_a d(\mu_a, \lambda_a).$$

$\Sigma_A = \{w \in [0,1]^A : \sum_{a \in [A]} w_a = 1\}$ $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} : i_\star(\boldsymbol{\lambda}) \neq i_\star(\boldsymbol{\mu})\}$

**Proof.** change of distribution between $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ : $i_\star(\boldsymbol{\lambda}) \neq i_\star(\boldsymbol{\mu})$

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \times \left[ \sup_{w \in \Sigma_A} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a \in [A]} w_a d(\mu_a, \lambda_a) \right] \geq \log(1/(3\delta))$$

# Insights from the lower bound

An algorithm matching the lower bound should satisfy

$$\forall a \in [A], \ \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau)]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau]} \simeq w_a^\star(\boldsymbol{\mu})$$

for a vector of **optimal proportions**

$$\boldsymbol{w}^\star(\boldsymbol{\mu}) \in \underset{w \in \Sigma_A}{\operatorname{argmax}} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a \in [A]} w_a d(\mu_a, \lambda_a).$$

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

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

# A lower-bound-inspired sampling rule for BAI

**Optimal proportions**

For the Best Arm Identification (BAI) problem, we propose an efficient algorithm to compute $w^\star(\boldsymbol{\mu})$ for any $\boldsymbol{\mu}$.

**The Tracking sampling rule:**

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

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

**Lemma**

Under the Tracking sampling rule,

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

# Optimal Best Arm Identification

**The Parallel GLRT for BAI**:

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

**Characteristic time**:

$$(T^\star(\boldsymbol{\mu}))^{-1} = \sup_{w\in\Sigma_A} \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\mu})} \sum_{a\in[A]} w_a d(\mu_a, \lambda_a)$$

**Theorem [K. and Garivier, COLT 2016]**

The Track-and-Stop algorithm which uses

- the Tracking sampling rule
- the Parallel GLRT stopping rule $\tau_\delta$
- recommends the empirical best arm $\hat{a}_{\tau_\delta} = \arg\max_a \hat{\mu}_a(\tau_\delta)$

satisfies $\mathbb{P}_{\boldsymbol{\mu}}(\hat{a}_{\tau_\delta} \neq a_\star(\boldsymbol{\mu})) \leq \delta$ and $\limsup_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \leq T^\star(\boldsymbol{\mu})$.

→ an asymptotically optimal algorithm for fixed-confidence BAI!

# Outline


1. Thompson Sampling for a Structured Bandit Problem
2. The Complexity of Pure Exploration
3. Thompson Sampling for Pure Exploration?

# Thompson Sampling for BAI

Track-and-Stop can be a bit computationally heavy due to the computation of $w^\star(\hat{\boldsymbol{\mu}}(t))$ in every round

➜ more efficient Thompson Sampling based alternatives?

**Top-Two Thompson Sampling [Russo, 2016]**

**Input:** parameter $\beta \in (0,1)$. In round $t+1$:

- draw a posterior sample $\boldsymbol{\theta} \sim \Pi_t$, $a_\star(\boldsymbol{\theta}) = \arg\max_a \theta_a$
- with probability $\beta$, select $A_{t+1} = a_\star(\boldsymbol{\theta})$
- with probability $1-\beta$, re-sample the posterior $\boldsymbol{\theta}' \sim \Pi_t$ until $a_\star(\boldsymbol{\theta}') \neq a_\star(\boldsymbol{\theta})$, select $A_{t+1} = a_\star(\boldsymbol{\theta}')$

[Russo, 2016] performs a Bayesian analysis of TTTS:

$$\Pi_t(\mathrm{Alt}(\boldsymbol{\mu})) \lesssim C \exp\left(-t/T^\star_\beta(\boldsymbol{\mu})\right) \quad \text{a.s.}$$

- New fixed-confidence guarantees for Gaussian bandits

**Theorem** [Shang, De Heide, K., Ménard, Valko, AISTATS 2020]

Using the TTTS sampling rule and the Parallel GLRT yields a $\delta$-correct BAI algorithm satisfying

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

where

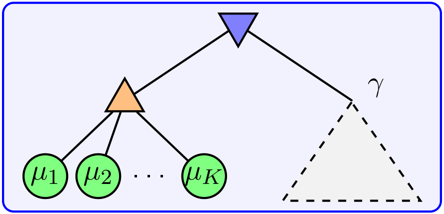
$$\left(T^\star_\beta(\boldsymbol{\mu})\right)^{-1} = \sup_{\substack{w \in \Sigma_A \\ w_{a_\star(\boldsymbol{\mu})} = \beta}} \inf_{\lambda \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a \in [A]} w_a d(\mu_a, \lambda_a)$$

➜ oracle tuning $\beta = w^\star_{a_\star}(\boldsymbol{\mu})$ needed for asymptotic optimality...

# Comparing the Smallest Mean to a Threshold

Fix threshold $\gamma$, let $\mu_{\min} = \min_a \mu_a$. Does $\boldsymbol{\mu}$ belong to

$$\begin{aligned} \mathcal{R}_< &= \{\boldsymbol{\mu} \in \mathcal{I}^A : \mu_{\min} < \gamma\} \\ \text{or to } \mathcal{R}_> &= \{\boldsymbol{\mu} \in \mathcal{I}^A : \mu_{\min} > \gamma\}? \end{aligned}$$




**Algorithm**:

- sampling rule $A_t$
- stopping rule $\tau$
- recommendation rule $\hat{m}_\tau \in \{<, >\}$.

**Goal**: $\mathbb{P}_{\boldsymbol{\mu}}(\hat{m}_\tau \neq m^\star(\boldsymbol{\mu})) \leq \delta$, small sample complexity $\tau$.

# Optimal allocation for this problem

For any $\delta$-correct strategy,

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geq T_\star(\boldsymbol{\mu}) \log(1/(3\delta))$$

Oracle allocation: $w^\star(\boldsymbol{\mu}) = \underset{w \in \Sigma_A}{\operatorname{argmax}} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^{A} w_a d(\mu_a, \lambda_a)$.

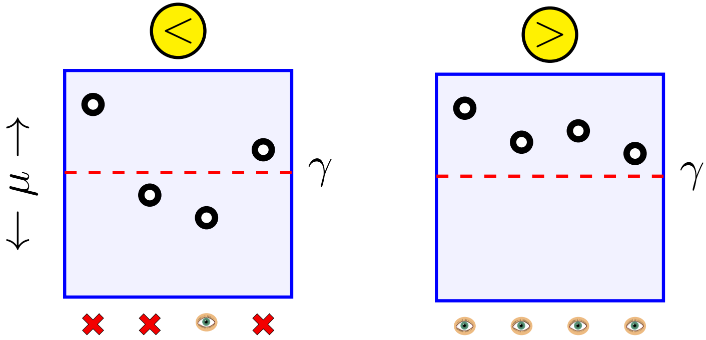
Closed-form expression for the optimal allocation :

$$w_a^\star(\boldsymbol{\mu}) = \begin{cases} \mathbb{1}_{(a = a_{\min})} & \text{if } \boldsymbol{\mu} \in \mathcal{R}_< \\ \frac{\frac{1}{d(\mu_a, \gamma)}}{\sum_j \frac{1}{d(\mu_j, \gamma)}} & \text{if } \boldsymbol{\mu} \in \mathcal{R}_> \end{cases}$$

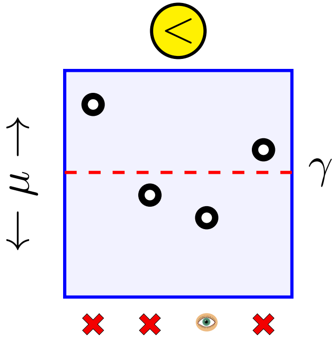
and the characteristic time

$$T_\star(\boldsymbol{\mu}) = \begin{cases} \frac{1}{d(\mu_{\min}, \gamma)} & \text{if } \boldsymbol{\mu} \in \mathcal{R}_< \\ \sum_a \frac{1}{d(\mu_a, \gamma)} & \text{if } \boldsymbol{\mu} \in \mathcal{R}_> \end{cases}$$

# Dichotomous Oracle Behaviour!



# Dichotomous Oracle Behaviour!




Two different ideas to converge to those sampling profiles:

- **Thompson Sampling**

Sample $\boldsymbol{\theta}(t) \sim \Pi_t$
Select $A_{t+1} = \arg\min_a \; \theta_a(t)$

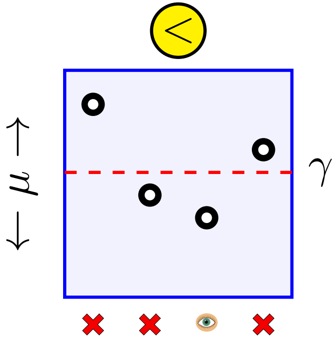
($\Pi_t$: posterior after $t$ rounds)

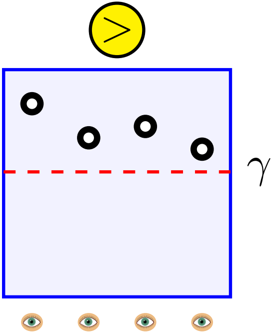
- **a LCB algorithm**

Compute a LCB on $\mu_a$
Select $A_{t+1} = \arg\min_a \; \mathrm{LCB}_a(t)$

(Lower Confidence Bound on $\mu_a$)

# A Solution: Murphy Sampling!




**Murphy Sampling**

Sample $\boldsymbol{\theta}(t) \sim \Pi_t \left(\cdot | \min_a \theta_a < \gamma\right)$
Select $A_{t+1} = \arg\min_a \ \theta_a(t)$.

**Idea:** condition on *low* minimum mean

# Properties of Murphy Sampling

**Theorem** [K., Koolen and Garivier, NeurIPS 2018]

For all exponential family bandit model $\boldsymbol{\mu}$, Murphy Sampling satisfies, for all $a$,

$$\frac{N_a(t)}{t} \to w_a^\star(\boldsymbol{\mu}).$$

| Sampling rule: | < | > |
|---|---|---|
| Thompson Sampling | ✓ | ✗ |
| Lower Confidence Bound | ✗ | ✓ |
| **Murphy Sampling** | ✓ | ✓ |

**Corollary** [K., Koolen and Garivier, NeurIPS 2018]

Murphy Sampling combined with a "good" stopping rule satisfies

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

# Conclusion

For both regret minimization and pure exploration:

- lower bounds are crucial to validate the (asymptotic) optimality of an algorithm
- ... and can also guide the design of optimal algorithms
- variants of Thompson Sampling provide efficient algorithms in different contexts

# Perspective

- Solving best arm identification in the fixed-budget setting
- Towards universal, optimal and efficient lower-bound inspired algorithms
- ... based on Thompson Sampling?

- Beyond "simple parameteric distributions":
  the power of re-sampling / sub-sampling based approaches?
- Beyond bandits:
  pure exploration done right in reinforcement learning

- Sequential methods for drug design?

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

Aziz, M., Kaufmann, E., and Riviere, M. (2018).
On multi-armed bandit designs for dose-finding clinical trials.
*arXiv:1903.07082*.

Combes, R. and Proutière, A. (2014).
Unimodal bandits: Regret lower bounds and optimal algorithms.
In *International Conference on Machine Learning (ICML)*.

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., Ménard, P., and Rossi, L. (2019a).
Thresholding bandit for dose-ranging: The impact of monotonicity.
In *International Conference on Machine Learning, Artificial Intelligence and Applications*.

Garivier, A., Ménard, P., and Stoltz, G. (2019b).
Explore first, exploit next: The true shape of regret in bandit problems.
*Mathemathics of Opereration Research*, 44(2):377–399.

Graves, T. and Lai, T. (1997).
Asymptotically Efficient adaptive choice of control laws in controlled markov chains.
*SIAM Journal on Control and Optimization*, 35(3):715–743.

Katariya, S., Kveton, B., Szepesvári, C., Vernade, C., and Wen, Z. (2017).
Bernoulli rank-1 bandits for click feedback.
In *IJCAI*.

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)*.

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

Magureanu, S., Combes, R., and Proutière, A. (2014).
Lipschitz Bandits: Regret lower bounds and optimal algorithms.
In *Proceedings on the 27th Conference On Learning Theory*.

Paladino, S., Trovò, F., Restelli, M., and Gatti, N. (2017).
Unimodal thompson sampling for graph-structured arms.
In *AAAI*.

Russo, D. (2016).
Simple Bayesian algorithms for best arm identification.
In *Proceedings of the 29th Conference on Learning Theory (COLT)*.

Shang, X., de Heide, R., Kaufmann, E., Ménard, P., and Valko, M. (2020).

Fixed-confidence guarantees for bayesian best-arm identification.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Trinh, C., Kaufmann, E., Vernade, C., and Combes, R. (2020).
Solving bernoulli rank-one bandits with unimodal thompson sampling.
In *Algorithmic Learning Theory (ALT)*.

Wilks, S. (1938).
The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses.
*The Annals of Mathematical Statistics*, 9(1):60–62.

# More explicit expression for BAI

**Characteristic time:** (for $a_\star(\boldsymbol{\mu}) = 1$)

$$
\begin{aligned}
(T^\star(\boldsymbol{\mu}))^{-1} &= \sup_{w \in \Sigma_A} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\mu)} \sum_{a \in [A]} w_a d(\mu_a, \lambda_a) \\
&= \sup_{w \in \Sigma_A} \min_{a \neq 1} \left[ w_1 d\left(\mu_1, \frac{w_1\mu_1 + w_a\mu_a}{w_1 + w_a}\right) + w_a d\left(\mu_a, \frac{w_1\mu_1 + w_a\mu_a}{w_1 + w_a}\right)\right]
\end{aligned}
$$

**Parallel GLRT:**

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

with

$$
\begin{aligned}
\hat{Z}(t) &= \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} \sum_{a \in [A]} w_a d(\hat{\mu}_a(t), \lambda_a) \\
&= \min_{a \neq \hat{a}_\star(t)} \left[ N_{\hat{a}_\star(t)}(t) d\left(\hat{\mu}_{\hat{a}_\star(t)}(t), \hat{\mu}_{\hat{a}_\star(t),a}(t)\right) + N_a(t) d\left(\hat{\mu}_a(t), \hat{\mu}_{\hat{a}_\star,a}(t)\right)\right],
\end{aligned}
$$

letting $\hat{\mu}_{a,b}(t) = \frac{N_a(t)\hat{\mu}_a(t) + N_b(t)\hat{\mu}_b(t)}{N_a(t) + N_b(t)}$.

# Practical impact of Track-and-Stop

Using the right stopping rule can make a big difference in practice!

- $\boldsymbol{\mu}_1 = [0.5\ 0.45\ 0.43\ 0.4]$, 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
$$w_\star(\boldsymbol{\mu}_2) = [0.336\ 0.251\ 0.177\ 0.132\ 0.104]$$

NB. GLRT with "stylized" threshold set to $\log\left(\frac{\log(t)+1}{\delta}\right)$.

| | Track-and-Stop | GLRT-SE* | KL-LUCB | KL-SE* |
|---|---|---|---|---|
| $\boldsymbol{\mu}_1$ | 4052 | 4516 | 8437 | 9590 |
| $\boldsymbol{\mu}_2$ | 1406 | 3078 | 2716 | 3334 |

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

* Succesive Elimination

# Mixture martingales

**How to prove**

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \in \mathbb{N}^* : \sum_{a\in\mathcal{S}} N_a(t) d(\hat{\mu}_a(t), \mu_a) \geq 3 \sum_{a\in\mathcal{S}} \log(1+\log N_a(t)) + |\mathcal{S}| \mathcal{T}\left(\frac{x}{|\mathcal{S}|}\right)\right) \leq e^{-x} \; ?$$

Letting $X_a(t) = N_a(t)d(\hat{\mu}_a(t), \mu_a) - 3\log(1 + \log N_a(t))$, find a martingale $M_a^\lambda(t)$ and a function $g : \Lambda \to \mathbb{R}$ such that

$$\forall \lambda \in \Lambda, \forall t \in \mathbb{N}, M_a^\lambda(t) \geq e^{\lambda X_a(t) - g(\lambda)}$$

and such that $\prod_{a\in\mathcal{S}} M_a^\lambda(t)$ is still a martingale.

➜ Cramer-Chernoff method + Doob inequality easily yields

$$\forall \lambda \in \Lambda, \quad \mathbb{P}\left(\exists t \in \mathbb{N} : \sum_{a\in\mathcal{S}} X_a(t) > \frac{|\mathcal{S}| g(\lambda) + x}{\lambda}\right) \leq e^{-x}$$

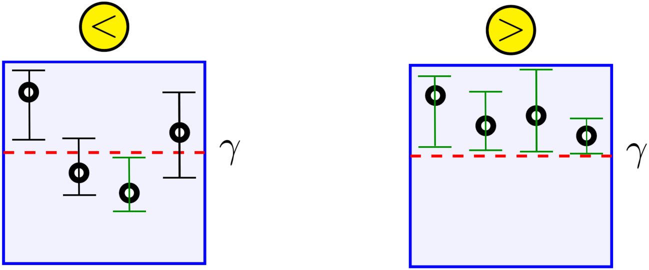
**Building the martingale(s)**:

$$\tilde{Z}_a^\pi(t) = \int \exp\left(\eta S_a(t) - \phi_{\mu_a}(\eta) N_a(t)\right) d\pi(\eta)$$

for a well chosen continuous mixture of discrete priors.

# Good stopping rules for the Smallest Minimum

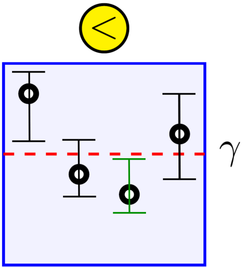
**Sufficient for asymptotic guarantees:** a simple stopping rule based on individual confidence intervals $\tau^{\text{Box}} := \min\left(\tau_<; \tau_>\right)$ where

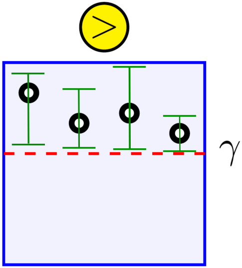



$\tau_< = \inf\{t : \exists a : \mathrm{UCB}_a(t) < \gamma\}$ $\tau_> = \inf\{t : \forall a, \mathrm{LCB}_a(t) > \gamma\}$

## Good stopping rules for the Smallest Minimum

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



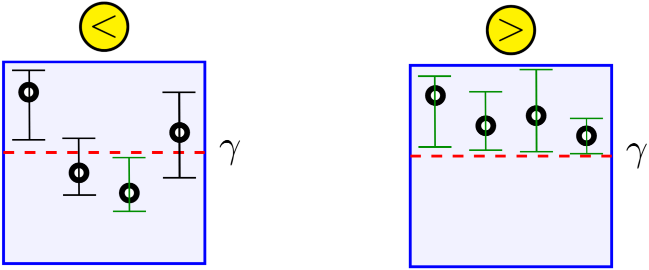

$\tau_< = \inf\{t : \exists a : \mathrm{UCB}_a(t) < \gamma\}$ $\quad$ $\tau_> = \inf\{t : \forall a, \mathrm{LCB}_a(t) > \gamma\}$

**The Parallel GLRT?**

$$\tau_>^{\text{GLRT}} = \inf\left\{ t \in \mathbb{N}^* : \min_{a\in[A]} N_a(t) d(\hat{\mu}_a(t), \gamma) \mathbb{1}(\hat{\mu}_a(t) \geq \gamma) > \beta(t,\delta) \right\}$$

# Good stopping rules for the Smallest Minimum

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



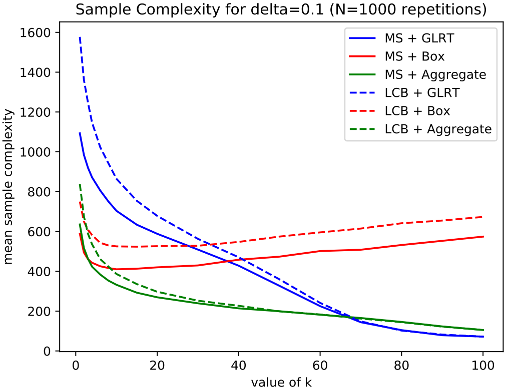

$\tau_< = \inf\{t : \exists a : \mathrm{UCB}_a(t) < \gamma\}$ $\quad$ $\tau_> = \inf\{t : \forall a, \mathrm{LCB}_a(t) > \gamma\}$

**The Parallel GLRT?**

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

# Practical performance of Murphy Sampling

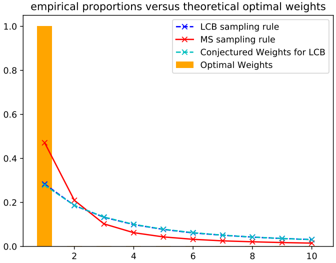
Empirical sample complexity for a Gaussian instance with $\mu_a \in \{-1, 0\}$ and $\gamma = 0$ as a function of the number $k$ of low arms




$$(\boldsymbol{\mu} \in \mathcal{R}_<)$$

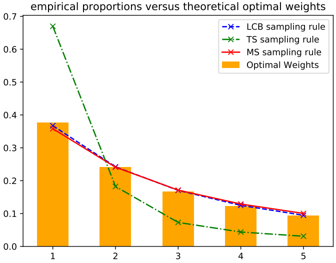
# Convergence of Murphy Sampling

$\boldsymbol{\mu} = \text{linspace}(-1, 1, 10) \in \mathcal{R}_<$
$\gamma = 0$




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

$\boldsymbol{\mu} = \text{linspace}(1/2, 1, 5) \in \mathcal{R}_>$
$\gamma = 0$




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