Contributions to the Optimal Solution of Several Bandit Problems

Emilie Kaufmann

HDR defense
November 13th, 2020
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 \( A_t \in \{1, \ldots, A\} \)
- observes a sample \( X_t \sim B(\mu_{A_t}) \) independent from past data

**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 $X_t \sim B(\mu_{A_t})$ independent from past data

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

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

Notation: $\nu_a = \nu_{\mu_a}$, $\mu = (\mu_1, \ldots, \mu_A) \in \mathcal{I}^A$. 
One bandit model, many bandit problems

rewards maximization...

with a twist

- feedback ≠ 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 ≠ 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?
Structured bandits

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

→ can we exploit the knowledge of $S$ to gain more reward?

\[ y = \mu_3 + L(x-x_3) \]
\[ y = \mu_3 - L(x-x_3) \]

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

$$R_\mu(A, T) = \mu^*_T - \mathbb{E}_\mu \left[ \sum_{t=1}^{T} X_t \right]$$

$$= \sum_{a \in [A]} (\mu^*_a - \mu_a) \mathbb{E}_\mu [N_a(T)]$$

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

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

Let $A$ be such that $\forall \mu \in S, \forall \alpha \in (0, 1], R_\mu(A, T) = o(T^\alpha)$.

$$\forall \mu \in S, \lim_{T \to \infty} \frac{R_\mu(A, T)}{\log(T)} \geq C_S(\mu).$$

$\Rightarrow A$ is asymptotically optimal if $R_\mu(A, T) = C_S(\mu) \log(T) + o(\log(T))$.
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

$$R_\mu(A, T) = \mu_{a^*} T - \mathbb{E}_\mu \left[ \sum_{t=1}^{T} X_t \right]$$

$$= \sum_{a \in [A]} (\mu_{a^*} - \mu_a) \mathbb{E}_\mu [N_a(T)]$$

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

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

Let $A$ be such that $\forall \mu \in S, \forall \alpha \in (0, 1], R_\mu(A, T) = o(T^\alpha)$.  

$$\forall \mu \in S, \lim_{T \to \infty} \frac{R_\mu(A, T)}{\log(T)} \geq C_S(\mu).$$

$\Rightarrow$ $A$ is asymptotically optimal if $R_\mu(A, T) = C_S(\mu) \log(T) + o(\log(T))$
Lower bounds can help

\[ C_S(\mu) \text{ features the Kullback-Leibler divergence } d(\mu, \mu') := KL(\nu_\mu, \nu_{\mu'}) \]

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

**Special case** [Combes and Proutière, 2014]

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

\[ (a_1 = a, \ldots, a_m = a_\star) : (a_i, a_{i+1}) \in E \text{ and } \mu_{a_i} < \mu_{a_{i+1}}. \]

For graphical unimodal bandits,

\[
C_S(\mu) = \sum_{a \in N_G(a_\star)} \frac{\mu_\star - \mu_a}{d(\mu_a, \mu_\star)} \quad \mathcal{N}_G(a_\star) = \{a : (a, a_\star) \in E\}
\]

an optimal algorithm focuses on neighbors of the optimal arm
Solving Rank-One Bandits

\[ S_{R1} = \left\{ \mu = (\mu(k,\ell))_{1 \leq k \leq K, 1 \leq \ell \leq L} \mid \exists u \in [0, 1]^K, v \in [0, 1]^L : \mu(k,\ell) = u_k v_\ell \right\} \]

Example: content optimization with two independent factors

\[ \mu = \begin{pmatrix} u_1 v_1 & u_1 v_2 & u_1 v_3 & u_1 v_4 \\ u_2 v_1 & u_2 v_2 & u_2 v_3 & u_2 v_4 \\ u_3 v_1 & u_3 v_2 & u_3 v_3 & u_3 v_4 \end{pmatrix} \]

clic probability \( \mu(k,\ell) = u_k \times v_\ell \)
Solving Rank-One Bandits

\[ S_{R1} = \left\{ \mu = (\mu(k,\ell))_{1 \leq k \leq K, 1 \leq \ell \leq L} \ \mid \exists \mathbf{u} \in [0,1]^K, \mathbf{v} \in [0,1]^L: \mu(k,\ell) = u_k v_\ell \right\} \]

[Katariya et al., 2017]

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

\[ \mu = \begin{pmatrix} U_1 V_1 & U_1 V_2 & U_1 V_3 & U_1 V_4 \\ U_2 V_1 & U_2 V_2 & U_2 V_3 & U_2 V_4 \\ U_3 V_1 & U_3 V_2 & U_3 V_3 & U_3 V_4 \end{pmatrix} \]

\[ \text{clic probability } \mu(k,\ell) = u_k \times v_\ell \]

**Key observation**

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

\[ ((i,j),(k,\ell)) \in E \text{ if } (i = k \text{ or } j = \ell) \]
Solving Rank-One Bandits

\[ S_{R1} = \left\{ \mu = (\mu(k,\ell))_{1 \leq k \leq K, 1 \leq \ell \leq L} \middle| \exists u \in [0,1]^K, v \in [0,1]^L : \mu(k,\ell) = u_k v_\ell \right\} \]

[Katariya et al., 2017]

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

\[
\mu = \begin{pmatrix}
U_1V_1 & U_1V_2 & U_1V_3 & U_1V_4 \\
U_2V_1 & U_2V_2 & U_2V_3 & U_2V_4 \\
U_3V_1 & U_3V_2 & U_3V_3 & U_3V_4
\end{pmatrix}
\]

clic probability \( \mu(k,\ell) = u_k \times v_\ell \)

**Key observation**

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

\( ((i,j), (k,\ell)) \in E \) if \( i = k \) (x)or \( j = \ell \)
**Idea:** use an optimal algorithm for graphical unimodal bandits

- Unimodal Thompson Sampling \cite{Paladino2017}

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

In each round $t + 1$:

- compute the empirical *leader* $B_{t+1} = \arg\max_{a \in [A]} \hat{\mu}_a(t)$

- 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 $\mathcal{N}_G(B_{t+1}) \cup \{B_{t+1}\}$:

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

and $A_{t+1} = \arg\max_{a \in \mathcal{N}_G(B_{t+1}) \cup \{B_{t+1}\}} \theta_a(t)$ (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** $\gamma \in \{2, 3, \ldots\}$ for Bernoulli bandits

In each round $t + 1$:

- compute the empirical *leader* $B_{t+1} = \arg\max_{(k,\ell)\in[K]\times[L]} \hat{\mu}(k,\ell)(t)$

- 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 $\mathcal{N}_G(B_{t+1}) \cup \{B_{t+1}\}$:

\[
\theta_{(k,\ell)}(t) \sim \text{Beta}(S_{(k,\ell)}(t) + 1, N_{(k,\ell)}(t) - S_{(k,\ell)}(t) + 1)
\]

and $A_{t+1} = \arg\max_{(k,\ell)\in\{(k',B_{t+1}^2)\}\cup\{(B_{t+1}^1,\ell')\}} \theta_{(k,\ell)}(t)$ (TS around the leader)

\[
S_a(t) = \sum_{s=1}^{t} X_s \mathbbm{1}(A_s = a): \text{sum of rewards from arm } a
\]

\[
\hat{\mu}_a(t) = S_a(t)/N_a(t): \text{empirical mean of arm } a
\]

\[
\ell_{b}(t) = \sum_{s=1}^{t} \mathbbm{1}(B_s = b): \text{number of times arm } b \text{ has been leader}
\]
Theorem [Trinh, K., Vernade, Combes, ALT 2020]

Let \( \mu \) 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 \( \varepsilon > 0 \),

\[
R_{\mu}(\text{UTS}(\gamma), T) \leq (1 + \varepsilon) \sum_{a \in N_G(a_*)} \frac{(\mu_\star - \mu_a)}{d(\mu_a, \mu_\star)} \log(T) + C(\mu, \gamma, \varepsilon).
\]

- a novel analysis, valid for any leader exploration parameter \( \gamma \), with \( \gamma = 2 \) being the best choice in practice
- 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
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 $\mu = (\mu_1, \ldots, \mu_A)$ by adaptively sampling the arms

**Input:**
- $\mathcal{R} \subseteq \mathcal{I}^A$ a subset that contains $\mu$
- $l$ regions $\mathcal{R}_1, \ldots, \mathcal{R}_l$ such that $\mathcal{R} \subseteq \bigcup_{i=1}^l \mathcal{R}_i$

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

**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 $\hat{i}_\tau \in [l]$

such that $\mathbb{P}_\mu (\mu \notin \mathcal{R}_{\hat{i}_\tau}) \leq \delta$ and the sample complexity $\tau$ is small.
Identify the arm with largest mean:

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

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

[Even-Dar et al., 2006]

**Example**: identify the version of a webpage with the largest conversion probability (A/B/C testing)
Identify the arm whose mean is the closest to some threshold:

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

[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

$$A_1, X_1, A_2, X_2, \ldots, A_t, X_t, \ldots$$

where $X_t \sim \nu_{\mu_A t}$

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

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

→ multiple, composite hypotheses (possibly overlapping)

**Definition**

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

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

such that $\forall \mu \in \mathcal{R}$, $\mathbb{P}_\mu (\tau < \infty, \mu \notin \mathcal{R}_{\hat{i}_\tau}) \leq \delta$. 
The Parallel GLRT rule

**Idea:** run \( l \) statistical tests of

\[
\tilde{H}_0 : (\mu \in \mathcal{R} \setminus \mathcal{R}_i) \quad \text{against} \quad \tilde{H}_1 : (\mu \in \mathcal{R}_i)
\]

in parallel until one of them rejects \( \tilde{H}_0 \).

**Individual test:** a GLR Test rejects \( \tilde{H}_0 \) for large values of the Generalized Likelihood Ratio

\[
\frac{\sup_{\lambda \in \mathcal{R}} \ell(X_1, \ldots, X_t; \lambda)}{\sup_{\lambda \in \mathcal{R} \setminus \mathcal{R}_i} \ell(X_1, \ldots, X_t; \lambda)} = \inf_{\lambda \in \mathcal{R} \setminus \mathcal{R}_i} \frac{\ell(X_1, \ldots, X_t; \hat{\mu}(t))}{\ell(X_1, \ldots, X_t; \lambda)}
\]

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

[Wilks, 1938]

\( \hat{\mu}(t) = (\hat{\mu}_1(t), \ldots, \hat{\mu}_A(t)) \), Maximum Likelihood Estimator.
The Parallel GLRT rule

Parallel GLRT

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

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

$$\hat{i}_{\tau_\delta} \in \arg \max_{i \in [I]} \inf_{\lambda \in \mathcal{R} \setminus \mathcal{R}_i} \log \frac{\ell(X_1, \ldots, X_{\tau_\delta}; \hat{\mu}(\tau_\delta))}{\ell(X_1, \ldots, X_{\tau_\delta}; \lambda)}$$

In an exponential family bandit model,

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

with $d(\mu, \mu') = \text{KL} (\nu_\mu, \nu_{\mu'})$.

(rewards in a one-parameter exponential family: Bernoulli, Gaussian, Poisson...
The Parallel GLRT rule

**Parallel GLRT**

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

\[
\tau_\delta = \inf \left\{ t \in \mathbb{N} : \max_{i \in [I]} \inf_{\lambda \in \mathcal{R} \setminus \mathcal{R}_i} \sum_{a \in [A]} N_a(t) d(\hat{\mu}_a(t), \lambda_a) > \beta(t, \delta) \right\}
\]

\[
\hat{\iota}_{\tau_\delta} \in \arg \max_{i \in [I]} \inf_{\lambda \in \mathcal{R} \setminus \mathcal{R}_i} \sum_{a \in [A]} N_a(\tau_\delta) d(\hat{\mu}_a(\tau_\delta), \lambda_a)
\]

In an exponential family bandit model,

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

with $d(\mu, \mu') = \text{KL} (\nu_\mu, \nu_{\mu'})$.

(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,

\[
P_{\mu} \left( \tau_\delta < \infty, \mu \notin R_{\hat{\tau}_\delta} \right)
\]

\[
\leq P \left( \exists t \in \mathbb{N}^*, \exists i : \mu \notin R_i, \inf_{\lambda \in R \setminus R_i} \sum_{a \in [A]} N_a(t)d(\hat{\mu}_a(t), \lambda_a) > \beta(t, \delta) \right)
\]

\[
\leq P \left( \exists t \in \mathbb{N}^*, \exists i : \mu \in R \setminus R_i, \sum_{a \in [A]} N_a(t)d(\hat{\mu}_a(t), \mu_a) > \beta(t, \delta) \right)
\]

\[
\leq P \left( \exists t \in \mathbb{N}^*, \sum_{a \in [A]} N_a(t)d(\hat{\mu}_a(t), \mu_a) > \beta(t, \delta) \right)
\]
Upper bound on the error probability

For any sampling rule, under the GLRT stopping rule,

\[
P_\mu \left( \tau_\delta < \infty, \mu \notin R_{\hat{\tau}_\delta} \right)
\leq P \left( \exists t \in \mathbb{N}^*, \exists i : \mu \notin R_i, \inf_{\lambda \in R \setminus R_i} \sum_{a \in [A]} N_a(t) d(\hat{\mu}_a(t), \lambda_a) > \beta(t, \delta) \right)
\leq P \left( \exists t \in \mathbb{N}^*, \exists i : \mu \in R \setminus R_i, \sum_{a \in [A]} N_a(t) d(\hat{\mu}_a(t), \mu_a) > \beta(t, \delta) \right)
\leq P \left( \exists t \in \mathbb{N}^*, \sum_{a \in [A]} N_a(t) d(\hat{\mu}_a(t), \mu_a) > \beta(t, \delta) \right)
\]

**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,

\[ P_{\mu} \left( \tau_{\delta} < \infty, \mu \notin \mathcal{R}_{\hat{\tau}_{\delta}} \right) \]

\[ \leq P \left( \exists t \in \mathbb{N}^*, \exists i : \mu \notin \mathcal{R}_i, \inf_{\lambda \in \mathcal{R} \setminus \mathcal{R}_i} \sum_{a \in [A]} N_a(t) d(\hat{\mu}_a(t), \lambda_a) > \beta(t, \delta) \right) \]

\[ \leq P \left( \exists t \in \mathbb{N}^*, \exists i : \mu \in \mathcal{R} \setminus \mathcal{R}_i, \sum_{a \in [A]} N_a(t) d(\hat{\mu}_a(t), \mu_a) > \beta(t, \delta) \right) \]

\[ \leq P \left( \exists t \in \mathbb{N}^*, \sum_{a \in [A]} N_a(t) d(\hat{\mu}_a(t), \mu_a) > \beta(t, \delta) \right) \]

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,

\[ P_{\mu} \left( \tau_\delta < \infty, \mu \notin \hat{R}_{t_{\tau_\delta}} \right) \]

\[ \leq P \left( \exists t \in \mathbb{N}^*, \exists i : \mu \notin \mathcal{R}_i, \inf_{\lambda \in \mathcal{R}\setminus \mathcal{R}_i} \sum_{a \in [A]} N_a(t) d(\hat{\mu}_a(t), \lambda_a) > \beta(t, \delta) \right) \]

\[ \leq P \left( \exists t \in \mathbb{N}^*, \exists i : \mu \in \mathcal{R}\setminus \mathcal{R}_i, \sum_{a \in [A]} N_a(t) d(\hat{\mu}_a(t), \mu_a) > \beta(t, \delta) \right) \]

\[ \leq P \left( \exists t \in \mathbb{N}^*, \sum_{a \in [A]} N_a(t) d(\hat{\mu}_a(t), \mu_a) > \beta(t, \delta) \right) \]

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

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

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

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

is $\delta$-correct

$\Rightarrow$ for any active identification problem

$\Rightarrow$ regardless of the sampling rule
A universal $\delta$-correct stopping rule

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

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

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

**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 δ-correct stopping rule

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

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

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

Consequence: the Parallel GLRT stopping rule with threshold

$$
\beta(t, \delta) \approx \log \left( \frac{1}{\delta} \right) + A \log \log \left( \frac{1}{\delta} \right) + 3A \log \log(t)
$$

is δ-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

\[ R = \bigcup_{i=1}^{l} R_i \text{ forms a partition} \]
\[ i_\star(\mu): \text{ unique region that contains } \mu. \]

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

Any \( \delta \)-correct algorithm satisfies, for all \( \mu \in R \),

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

with

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

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

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

\[ \text{KL} \left( \mathbb{P}_{X_1, \ldots, X_\tau}^{\mu}, \mathbb{P}_{X_1, \ldots, X_\tau}^{\lambda} \right) \geq \text{kl} \left( \mathbb{P}_{\mu} (\hat{i}_\tau = i_\star(\lambda)), \mathbb{P}_{\lambda} (\hat{i}_\tau = i_\star(\lambda)) \right) \]

with \( \text{kl}(x, y) = \text{KL}(\mathcal{B}(x), \mathcal{B}(y)) \).  

[Garivier et al., 2019b]
Best achievable sample complexity

\[ \mathcal{R} = \bigcup_{i=1}^{l} \mathcal{R}_i \] forms a partition
\[ i_*(\mu) \]: unique region that contains \( \mu \).

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

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

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

with

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

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

**Proof.** change of distribution between \( \mu \) and \( \lambda : i_*(\lambda) \neq i_*(\mu) \)

\[ \text{KL} \left( \mathbb{P}_{\mu}^{X_1, \ldots, X_\tau}, \mathbb{P}_{\lambda}^{X_1, \ldots, X_\tau} \right) \geq \text{kl}( \mathbb{P}_\mu(\hat{i}_\tau = i_*(\lambda)), \mathbb{P}_\lambda(\hat{i}_\tau = i_*(\lambda)) \) \]

\[ \leq \delta \]
\[ \geq 1 - \delta \]

with \( \text{kl}(x, y) = \text{KL}(\mathcal{B}(x), \mathcal{B}(y)) \). [Garivier et al., 2019b]
Best achievable sample complexity

\( \mathcal{R} = \bigcup_{i=1}^{l} \mathcal{R}_i \) forms a partition

\( i^*(\mu) \): unique region that contains \( \mu \).

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

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

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

with

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

\( \Sigma_A = \{w \in [0, 1]^A : \sum_{a \in [A]} w_a = 1\} \)

\( \text{Alt}(\mu) = \{\lambda : i^*(\lambda) \neq i^*(\mu)\} \)

**Proof.** change of distribution between \( \mu \) and \( \lambda : i^*(\lambda) \neq i^*(\mu)\)

\[
\text{KL} \left( \mathbb{P}_{\mu}^{X_1, \ldots, X_{\tau}}, \mathbb{P}_{\lambda}^{X_1, \ldots, X_{\tau}} \right) \geq \text{kl}(\delta, 1 - \delta)
\]

with \( \text{kl}(x, y) = \text{KL} (\mathcal{B}(x), \mathcal{B}(y)) \).  

[Garivier et al., 2019b]
Best achievable sample complexity

\[ \mathcal{R} = \bigcup_{i=1}^{l} \mathcal{R}_i \] forms a partition
\[ i_\star(\mu): \text{unique region that contains } \mu. \]

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

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

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

with

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

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

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

\[ \text{KL} \left( \mathbb{P}_{\mu}^{X_1, \ldots, X_\tau}, \mathbb{P}_{\lambda}^{X_1, \ldots, X_\tau} \right) \geq \log(1/(3\delta)) \]
Best achievable sample complexity

\[ \mathcal{R} = \bigcup_{i=1}^{l} \mathcal{R}_i \] forms a partition

\( i_* (\mu) \): unique region that contains \( \mu \).

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

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

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

with

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

\( \Sigma_A = \{ w \in [0, 1]^A : \sum_{a \in [A]} w_a = 1 \} \)

\( \text{Alt}(\mu) = \{ \lambda : i_* (\lambda) \neq i_* (\mu) \} \)

**Proof.** change of distribution between \( \mu \) and \( \lambda : i_* (\lambda) \neq i_* (\mu) \)

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

$$\mathcal{R} = \bigcup_{i=1}^{l} \mathcal{R}_i$$ forms a partition

$i_{\ast}(\mu)$: unique region that contains $\mu$.

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

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

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

with

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

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

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

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

\[ \mathcal{R} = \bigcup_{i=1}^{l} \mathcal{R}_i \text{ forms a partition} \]
\[ i_\star(\mu): \text{unique region that contains } \mu. \]

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

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

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

with

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

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

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

\[ \mathbb{E}_\mu[\tau] \times \left[ \inf_{\lambda \in \text{Alt}(\mu)} \sum_{a \in [A]} \frac{\mathbb{E}_\mu[N_a(\tau)]}{w_a} d(\mu_a, \lambda_a) \right] \geq \log(1/(3\delta)) \]
Best achievable sample complexity

\( \mathcal{R} = \bigcup_{i=1}^{l} \mathcal{R}_i \) forms a partition

\( i_\star(\mu) \): unique region that contains \( \mu \).

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

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

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

with

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

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

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

\[
\mathbb{E}_\mu [\tau] \times \left[ \sup_{w \in \Sigma_A} \inf_{\lambda \in \text{Alt}(\mu)} \sum_{a \in [A]} w_a d(\mu_a, \lambda_a) \right] \geq \log(1/(3\delta))
\]
An algorithm matching the lower bound should satisfy

$$\forall a \in [A], \frac{E_{\mu}[N_a(\tau)]}{E_{\mu}[\tau]} \simeq w^*_a(\mu)$$

for a vector of optimal proportions

$$w^*(\mu) \in \arg\max_{w \in \Sigma_A} \inf_{\lambda \in \text{Alt}(\mu)} \sum_{a \in [A]} w_a d(\mu_a, \lambda_a).$$

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

- may be non-unique
- may be hard to compute
For the Best Arm Identification (BAI) problem, we propose an efficient algorithm to compute \( w^*(\mu) \) for any \( \mu \).

The Tracking sampling rule:

\[
A_{t+1} \in \begin{cases} 
\arg\min_{a \in U_t} N_a(t) & \text{if } U_t \neq \emptyset \quad \text{(forced exploration)} \\
\arg\max_{a \in [A]} \left[ w^*_a(\hat{\mu}(t)) - \frac{N_a(t)}{t} \right] & \text{else.} \quad \text{(tracking)}
\end{cases}
\]

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

Lemma

Under the Tracking sampling rule,

\[
P_\mu \left( \lim_{t \to \infty} \frac{N_a(t)}{t} = w^*_a(\mu) \right) = 1.
\]
Optimal Best Arm Identification

The Parallel GLRT for BAI:

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

Characteristic time:

\[
(T^*(\mu))^{-1} = \sup_w \inf_{\lambda \in \text{Alt}(\mu)} \sum_{a \in [A]} w_a d(\mu_a, \lambda_a)
\]

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

\( \rightarrow \) an asymptotically optimal algorithm for fixed-confidence BAI!
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 $w^*(\hat{\mu}(t))$ 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 $\theta \sim \Pi_t$, $a_\star(\theta) = \arg \max_a \theta_a$
- with probability $\beta$, select $A_{t+1} = a_\star(\theta)$
- with probability $1 - \beta$, re-sample the posterior $\theta' \sim \Pi_t$ until $a_\star(\theta') \neq a_\star(\theta)$, select $A_{t+1} = a_\star(\theta')$

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

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

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

where

\[
(T^{*}_{\beta}(\mu))^{-1} = \sup_{w \in \Sigma_A} \inf_{\lambda \in \text{Alt}(\mu)} \sum_{a \in [A]} w_a d(\mu_a, \lambda_a)
\]

\( w^{*}_{a*}(\mu) \Rightarrow \beta \)

→ oracle tuning \( \beta = w^{*}_{a*}(\mu) \) needed for asymptotic optimality...
Comparing the Smallest Mean to a Threshold

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

$$\mathcal{R}_\prec = \{ \mu \in \mathcal{I}^A : \mu_{\text{min}} < \gamma \}$$

or to

$$\mathcal{R}_\succ = \{ \mu \in \mathcal{I}^A : \mu_{\text{min}} > \gamma \}?$$

Algorithm:

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

Goal: $\mathbb{P}_\mu(\hat{m}_\tau \neq m^*(\mu)) \leq \delta$, small sample complexity $\tau$. 
For any $\delta$-correct strategy,

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

Oracle allocation: $w^*(\mu) = \arg\max_{w \in \Sigma_A} \inf_{\lambda \in \text{Alt}(\mu)} \sum_{a=1}^{A} w_a d(\mu_a, \lambda_a)$.

Closed-form expression for the optimal allocation:

$$w^*_a(\mu) = \begin{cases} 
1_{(a=a_{\text{min}})} & \text{if } \mu \in \mathcal{R}_< \\
\frac{1}{d(\mu_a, \gamma)} & \text{if } \mu \in \mathcal{R}_>
\end{cases}$$

and the characteristic time

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

Two different ideas to converge to those sampling profiles:

- Thompson Sampling
  \[ \theta_t^* \sim \Pi_t \]
  \[ A_{t+1} = \text{arg min}_a \theta_a(t) \]

- LCB algorithm
  \[ \text{Compute } \mu_a \]
  \[ A_{t+1} = \text{arg min}_a \text{LCB}_a(t) \]

Two different ideas to converge to those sampling profiles:

- **Thompson Sampling**
  Sample $\theta(t) \sim \Pi_t$
  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 $A_{t+1} = \arg \min_a \text{LCB}_a(t)$

  (Lower Confidence Bound on $\mu_a$)
A Solution: Murphy Sampling!

Murphy Sampling

Sample $\theta(t) \sim \Pi_t (\cdot \mid \min_a \theta_a < \gamma)$
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 $\mu$, Murphy Sampling satisfies, for all $a$,

$$\frac{N_a(t)}{t} \to w^*_a(\mu).$$

**Sampling rule:**

<table>
<thead>
<tr>
<th></th>
<th>Thompson Sampling</th>
<th>Lower Confidence Bound</th>
<th>Murphy Sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>$&lt;\ )$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$&gt;\ )$</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

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

Murphy Sampling combined with a “good” stopping rule satisfies

$$\limsup_{\delta \to 0} \frac{\tau_\delta}{\log \frac{1}{\delta}} \leq T_*(\mu), \text{ a.s.}$$
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?


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

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.
More explicit expression for BAI

**Characteristic time:** (for $a_*(\mu) = 1$)

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

\[
= \sup_{w \in \Sigma_A} \min_{a \neq 1} \left[ w_1 d\left(\mu_1, \frac{w_1 \mu_1 + w_a \mu_a}{w_1 + w_a}\right) + w_a d\left(\mu_a, \frac{w_1 \mu_1 + w_a \mu_a}{w_1 + w_a}\right) \right]
\]

**Parallel GLRT:**

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

with

\[
\hat{Z}(t) = \inf_{\lambda \in \text{Alt}(\hat{\mu}(t))} \sum_{a \in [A]} w_a d(\hat{\mu}_a(t), \lambda_a)
\]

\[
= \min_{a \neq \hat{a}_*(t)} \left[ N_{\hat{a}_*}(t)(t)d(\hat{\mu}_{\hat{a}_*}(t)(t), \hat{\mu}_{\hat{a}_*}(t), a(t)) + N_{a}(t)d(\hat{\mu}_a(t), \hat{\mu}_{\hat{a}_*}, a(t)) \right],
\]

letting $\hat{\mu}_{a,b}(t) = \frac{N_a(t)\hat{\mu}_a(t) + N_b(t)\hat{\mu}_b(t)}{N_a(t) + N_b(t)}$. 
Using the right stopping rule can make a big difference in practice!

- $\mu_1 = [0.5 \ 0.45 \ 0.43 \ 0.4]$, such that
  \[ w_\star(\mu_1) = [0.417 \ 0.390 \ 0.136 \ 0.057] \]

- $\mu_2 = [0.3 \ 0.21 \ 0.2 \ 0.19 \ 0.18]$, such that
  \[ w_\star(\mu_2) = [0.336 \ 0.251 \ 0.177 \ 0.132 \ 0.104] \]

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

<table>
<thead>
<tr>
<th></th>
<th>Track-and-Stop</th>
<th>GLRT-SE*</th>
<th>KL-LUCB</th>
<th>KL-SE*</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_1$</td>
<td>4052</td>
<td>4516</td>
<td>8437</td>
<td>9590</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>1406</td>
<td>3078</td>
<td>2716</td>
<td>3334</td>
</tr>
</tbody>
</table>

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

* Successive Elimination
How to prove

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

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

\[ \forall \lambda \in \Lambda, \forall t \in \mathbb{N}, M_\lambda^a(t) \geq e^{\lambda X_a(t) - g(\lambda)} \]

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

\[ \Rightarrow \text{ Cramer-Chernoff method + Doob inequality easily yields} \]

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

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 $\tau^{Box} := \min(\tau_{<}; \tau_{>})$ where

\[
\tau_{<} = \inf\{t : \exists a : UCB_{a}(t) < \gamma\} \quad \tau_{>} = \inf\{t : \forall a, LCB_{a}(t) > \gamma\}
\]
Good stopping rules for the Smallest Minimum

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

\[
\tau_\prec = \inf \{ t : \exists a : \text{UCB}_a(t) < \gamma \} \quad \tau_\succ = \inf \{ t : \forall a, \text{LCB}_a(t) > \gamma \}
\]

**The Parallel GLRT?**

\[
\tau_{\succ}^{\text{GLRT}} = \inf \left\{ t \in \mathbb{N}^* : \min_{a \in [A]} N_a(t) d(\hat{\mu}_a(t), \gamma) \mathbb{I}(\hat{\mu}_a(t) \geq \gamma) > \beta(t, \delta) \right\}
\]
**Sufficient for asymptotic guarantees:** a simple stopping rule based on individual confidence intervals $\tau^{\Box} := \min (\tau_\prec; \tau_\succ)$ where

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

The Parallel GLRT?

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

\[ (\mu \in \mathcal{R}_<) \]
Convergence of Murphy Sampling

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

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

Sampling proportions vs oracle, \( \delta = e^{-23} \). Sampling proportions vs oracle, \( \delta = e^{-7} \).