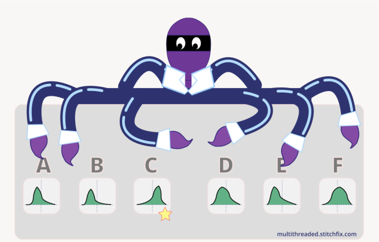
# Efficient Stopping Rules for Bandit Pure Exploration

Emilie Kaufmann

Algorithmic Statistics Workshop
Oxford, November 2025

# The Multi Armed Bandit (MAB) model

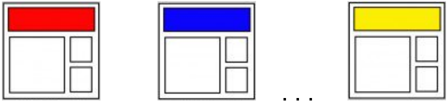
- $K$ *unknown* distributions $\nu_1, \dots, \nu_K$ called *arms*
- a time $t$, select an arm $A_t$ and collect an observation $X_t \sim \nu_{A_t}$




**Sequential strategy** / algorithm : $A_{t+1}$ can depend on:

- previous observation $A_1, X_1, \dots, A_t, X_t$
- some external randomization $U_t \sim \mathcal{U}([0,1])$
- some knowledge about the possible distributions: $\nu_a \in \mathcal{D}$

# Example: A/B/n Testing



$p_1$ $p_2$ $p_K$

$p_a$: probability that a visitor seeing version $a$ buys a product

For the $t$-th visitor:

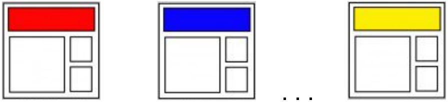
- choose a version $A_t$ to display
- observe $X_t = 1$ if a product is bought, 0 otherwise

**Objective 1: Maximizing Rewards**

- observation = reward
- maximize $\mathbb{E}[\sum_{t=1}^{T} X_t]$ for some (possibly unknown) $T$

**Regret minimization** in bandits: UCB, Thompson Sampling...

# Example: A/B/n Testing



$p_1$ $p_2$ $p_K$

$p_a$: probability that a visitor seeing version $a$ buys a product

For the $t$-th visitor:

- choose a version $A_t$ to display
- observe $X_t = 1$ if a product is bought, 0 otherwise

**Objective 2: Pure Exploration**

- identify quickly some interesting arms
- e.g. $a_\star = \arg\max_a \; p_a$ (best arm identification)

This talk: a generic receipe for **pure exploration**

# Pure Exploration

Possible bandit models: $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K) \in \mathcal{B}$

(e.g. independent sub-Gaussian arms, or Bernoulli arms)

Possible vectors of arms means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K) \in \mathcal{M}$

## Identification task

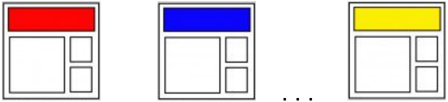
Given a correct answer function

$$\begin{aligned} i_\star : \mathcal{M} &\longrightarrow \mathcal{I} \\ \boldsymbol{\mu} &\mapsto i_\star(\boldsymbol{\mu}) \end{aligned}$$

find a correct answer with high probability.

# Examples of correct answers

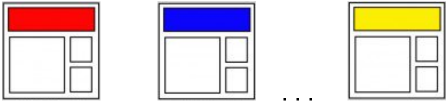
- Best Arm Identification [Even-Dar et al., 2006]



$$i_\star(\boldsymbol{\mu}) = \arg\max_{a \in [K]} \mu_a$$

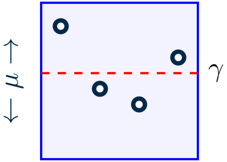
# Examples of correct answers

- Best Arm Identification [Even-Dar et al., 2006]



$$i_\star(\boldsymbol{\mu}) = \operatorname*{arg\,max}_{a \in [K]} \mu_a$$

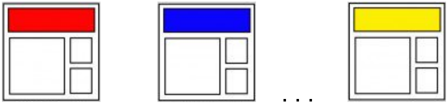
- Threshold-based questions: which means are below $\gamma$? [Locatelli et al., 2016]




$$i_\star(\boldsymbol{\mu}) = (\mathbb{1}(\mu_1 > \gamma), \dots, \mathbb{1}(\mu_K > \gamma)) \in \{0,1\}^K$$

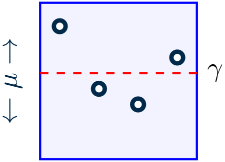
# Examples of correct answers

- Best Arm Identification [Even-Dar et al., 2006]



$$i_\star(\boldsymbol{\mu}) = \operatorname*{arg\,max}_{a \in [K]} \mu_a$$

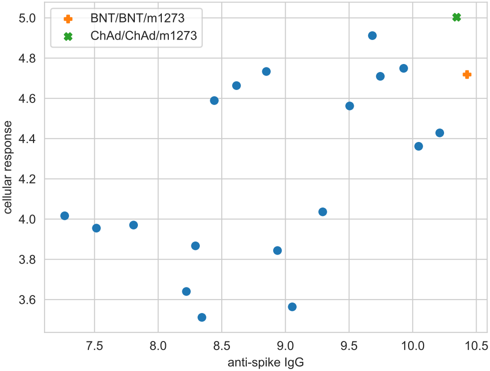
- Threshold-based questions: is there a mean below $\gamma$? [Kaufmann et al., 2018]




$$i_\star(\boldsymbol{\mu}) = \mathbb{1}(\min_i \mu_i < \gamma) \in \{0, 1\}$$

# Examples of correct answers

- Pareto Set Identification [Auer et al., 2016]




➜ arms are multi-variate distributions
➜ $i_\star(\boldsymbol{\mu})$ is the Pareto Set of the means

# Pure Exploration with Fixed Confidence

An algorithm is made of:

- a sampling rule $A_t \in [K]$ : what is the next arm to explore?

➜ get a new observation $X_t \sim \nu_{A_t}$

- a recommendation rule $\hat{\imath}_t$ : a guess for the correct answer
- a stopping rule $\tau$: when to stop the data collection?

**Definition**

An algorithm is $\delta$-correct if, for all $\boldsymbol{\mu} \in \mathcal{M}$, $\mathbb{P}_{\boldsymbol{\mu}}(\hat{\imath}_\tau \neq i_\star(\boldsymbol{\mu})) \leq \delta$.

**Goal:** a $\delta$-correct algorithm with small sample complexity $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$

# Outline

1. (Optimal) Pure Exploration: A General Receipe
2. Best Arm Identification
3. Pareto Set Identification

# Outline

# A lower bound on the sample complexity

**Setting:** independent arms, parametrized by their means

$$d(\mu, \mu') := \mathrm{KL}(\nu_\mu, \nu_{\mu'})$$

**Theorem** [Garivier and Kaufmann, 2016]

For any $\delta$-correct algorithm,

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

where

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

with

$$\Delta_K = \left\{\boldsymbol{w} \in [0,1]^K : \sum_{i=1}^{K} w_i = 1\right\}$$

$$\mathrm{Alt}(i) = \left\{\boldsymbol{\lambda} \in \mathcal{M} : i_\star(\boldsymbol{\lambda}) \neq i\right\}$$

# Optimal proportions

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

- $N_a(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = a)$: number of selections of arm $a$

The proof of the lower bound further suggests that the vector

$$\left( \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_1(\tau)]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}, \dots, \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_K(\tau)]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau]} \right)$$

should belong to

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

➜ **algorithmic strategy: let's make this happen!**

# The GLR stopping rule

Given a candidate best arm $i$, the (log) Generalized Likelihood Ratio statistic associated to

$$\mathcal{H}_0 = (\mu \in \mathrm{Alt}(i)) \quad \text{against} \quad \mathcal{H}_1 : (\mu \notin \mathrm{Alt}(i))$$

is

$$\begin{aligned} Z_i(t) &= \log \frac{\sup_{\boldsymbol{\lambda} \in \mathcal{M}} \ell(X_1, \dots, X_t; \boldsymbol{\lambda})}{\sup_{\boldsymbol{\lambda} \in \mathrm{Alt}(i)} \ell(X_1, \dots, X_t; \boldsymbol{\lambda})} \\ &= \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(i)} \log \frac{\ell(X_1, \dots, X_t; \hat{\boldsymbol{\mu}}(t))}{\ell(X_1, \dots, X_t; \boldsymbol{\lambda})} \\ &= \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(i)} \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \lambda_a) \end{aligned}$$

for exponential families (Bernoulli, Gaussian with known variance, etc.)

**Idea:** stop the first time that one of the $Z_i(t)$ is large enough

# A stopping rule aligned with the lower bound

## GLR stopping rule

Given a threshold function $\beta(t,\delta)$:

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

with the recommendation rule $\hat{\imath}_t^\star = i_\star(\hat{\boldsymbol{\mu}}(t))$

➜ reminiscent of

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

- if $\frac{N_a(t)}{t} \simeq w_a^\star(\boldsymbol{\mu})$ and $\beta(t,\delta) \simeq \log(1/\delta)$, we get

$$\tau_\delta \simeq T_\star(\mu) \log(1/\delta)$$

# Converging to the optimal proportions

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

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

## Lemma

Assume that

- for all $\boldsymbol{\mu} \in \mathcal{M}$, $|w^\star(\boldsymbol{\mu})| = 1$ (unique optimal allocation)
- $\boldsymbol{\mu} \mapsto w^\star(\boldsymbol{\mu})$ is continuous in all $\boldsymbol{\mu} \in \mathcal{M}$

Under the Tracking sampling rule,

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

# An asymptotically optimal algorithm

**Theorem** [Garivier and Kaufmann, 2016, Kaufmann and Koolen, 2021]

When the arm distributions belong to a one-dimensional exponential family, the Track-and-Stop strategy, that uses

- the Tracking sampling rule
- the GLR stopping rule with

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

- and the recommendation rule $\hat{\imath}_t = i_\star(\hat{\boldsymbol{\mu}}(t))$

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

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

# Calibration of the Stopping Rule

$$
\begin{aligned}
&\mathbb{P}_{\boldsymbol{\mu}}(\tau < \infty, \hat{\imath}^\star_\tau \neq i_\star(\boldsymbol{\mu})) \\
&\leq \mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \in \mathbb{N} : \hat{\imath}^\star_\tau \neq i_\star(\boldsymbol{\mu}), \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\imath}^\star_\tau)} \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \lambda_a) > \beta(t, \delta)\right) \\
&\leq \mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \in \mathbb{N} : \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \mu_a) > \beta(t, \delta)\right)
\end{aligned}
$$

Needed:

- a time uniform deviation inequality
- where the deviations are measured with KL-divergence
- and aggregated over arms

**Solution:** (a product of) e-processes

# Calibration of the Stopping Rule

$$
\begin{aligned}
&\mathbb{P}_{\boldsymbol{\mu}}(\tau < \infty, \hat{\imath}^\star_\tau \neq i_\star(\boldsymbol{\mu})) \\
&\leq \mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \in \mathbb{N} : \hat{\imath}^\star_\tau \neq i_\star(\boldsymbol{\mu}), \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\imath}^\star_\tau)} \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \lambda_a) > \beta(t,\delta)\right) \\
&\leq \mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \in \mathbb{N} : \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \mu_a) > \beta(t,\delta)\right)
\end{aligned}
$$

Needed:

- a time uniform deviation inequality
- where the deviations are measured with KL-divergence
- and aggregated over arms

**Solution:** (a product of) e-processes

# Calibration of the Stopping Rule

$$
\begin{aligned}
&\mathbb{P}_{\boldsymbol{\mu}}(\tau < \infty, \hat{\imath}^\star_\tau \neq i_\star(\boldsymbol{\mu})) \\
&\leq \mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \in \mathbb{N} : \hat{\imath}^\star_\tau \neq i_\star(\boldsymbol{\mu}), \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\imath}^\star_\tau)} \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \lambda_a) > \beta(t, \delta)\right) \\
&\leq \mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \in \mathbb{N} : \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \mu_a) > \beta(t, \delta)\right)
\end{aligned}
$$

Needed:

- a time uniform deviation inequality
- where the deviations are measured with KL-divergence
- and aggregated over arms

**Solution:** (a product of) e-processes

# Calibration of the Stopping Rule

$$
\begin{aligned}
&\mathbb{P}_{\boldsymbol{\mu}}(\tau < \infty, \hat{\imath}^\star_\tau \neq i_\star(\boldsymbol{\mu})) \\
&\leq \mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \in \mathbb{N} : \hat{\imath}^\star_\tau \neq i_\star(\boldsymbol{\mu}), \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\imath}^\star_\tau)} \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \lambda_a) > \beta(t, \delta)\right) \\
&\leq \mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \in \mathbb{N} : \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \mu_a) > \beta(t, \delta)\right)
\end{aligned}
$$

Needed:

- a time uniform deviation inequality
- where the deviations are measured with KL-divergence
- and aggregated over arms

**Solution:** (a product of) e-processes

# Calibration of the Stopping Rule

$$X_a(t) = N_a(t) d(\hat{\mu}_a(t), \mu_a) - 3 \log(1 + \log(N_a(t)))$$

**Step 1: $e^{\lambda X_a(t)}$ is (almost) an e-process**

$$\forall \lambda \in \Lambda : \quad M_a^\lambda(t) \geq e^{\lambda X_a(t) - g(\lambda)}$$

where $M_a^\lambda(t)$ is a test martingale and $g$ a correction function.

# Calibration of the Stopping Rule

$$X_a(t) = N_a(t)d(\hat{\mu}_a(t), \mu_a) - 3\log(1 + \log(N_a(t)))$$

**Step 1: $e^{\lambda X_a(t)}$ is (almost) an e-process**

$$\forall \lambda \in \Lambda : \quad M_a^\lambda(t) \geq e^{\lambda X_a(t) - g(\lambda)}$$

where $M_a^\lambda(t)$ is a test martingale and $g$ a correction function.

1. Link $X_a(t)$ to the martingale $W_a^\eta(t) = e^{\eta S_a(t) - \phi_{\mu_a}(\eta) N_a(t)}$

$$S_a(t) = \sum_{s=1}^{t} X_s \mathbb{1}(A_s = a) \qquad \phi_{\mu_a} = \log \mathbb{E}_{X \sim \nu_{\mu_a}}\left[e^{\lambda X}\right]$$

[Robbins, 1970]

# Calibration of the Stopping Rule

$$X_a(t) = N_a(t)d(\hat{\mu}_a(t), \mu_a) - 3\log(1 + \log(N_a(t)))$$

**Step 1: $e^{\lambda X_a(t)}$ is (almost) an e-process**

$$\forall \lambda \in \Lambda : \quad M_a^\lambda(t) \geq e^{\lambda X_a(t) - g(\lambda)}$$

where $M_a^\lambda(t)$ is a test martingale and $g$ a correction function.

1. Link $X_a(t)$ to the martingale $W_a^\eta(t) = e^{\eta S_a(t) - \phi_{\mu_a}(\eta) N_a(t)}$

$$S_a(t) = \sum_{s=1}^{t} X_s \mathbb{1}(A_s = a) \qquad \phi_{\mu_a} = \log \mathbb{E}_{X \sim \nu_{\mu_a}}\left[e^{\lambda X}\right]$$

[Robbins, 1970]

For exponential families:

If $N_a(t) \in [(1+\xi)^{i-1}, (1+\xi)^i]$, there exists some $\eta \in \{\eta_i^\pm(x)\}$:

$$\{N_a(t)d(\hat{\mu}_a(t), \mu_a) \geq x\} \subseteq \left\{W_t^\eta(t) \geq e^{\frac{x}{1+\xi}}\right\}$$

# Calibration of the Stopping Rule

$$X_a(t) = N_a(t)d(\hat{\mu}_a(t), \mu_a) - 3\log(1 + \log(N_a(t)))$$

**Step 1: $e^{\lambda X_a(t)}$ is (almost) an e-process**

$$\forall \lambda \in \Lambda : \quad M_a^\lambda(t) \geq e^{\lambda X_a(t) - g(\lambda)}$$

where $M_a^\lambda(t)$ is a test martingale and $g$ a correction function.

1. Link $X_a(t)$ to the martingale $W_a^\eta(t) = e^{\eta S_a(t) - \phi_{\mu_a}(\eta) N_a(t)}$

$$S_a(t) = \sum_{s=1}^{t} X_s \mathbb{1}(A_s = a) \qquad \phi_{\mu_a} = \log \mathbb{E}_{X \sim \nu_{\mu_a}}\left[e^{\lambda X}\right]$$

[Robbins, 1970]

For exponential families:

The martingale $Z_a^{(x)}(t) = \sum_i \frac{c}{i^2}\left[W_a^{\eta_i^-(x)}(t) + W_a^{\eta_i^+(x)}(t)\right]$ satisfies

$$\{X_a(t) - f(\xi) \geq x\} \subseteq \left\{Z_a^{(x)}(t) \geq e^{\frac{x}{1+\xi}}\right\}$$

# Calibration of the Stopping Rule

$$X_a(t) = N_a(t) d(\hat{\mu}_a(t), \mu_a) - 3\log(1 + \log(N_a(t)))$$

**Step 1: $e^{\lambda X_a(t)}$ is (almost) an e-process**

$$\forall \lambda \in \Lambda : \quad M_a^\lambda(t) \geq e^{\lambda X_a(t) - g(\lambda)}$$

where $M_a^\lambda(t)$ is a test martingale and $g$ a correction function.

1. Link $X_a(t)$ to the martingale $W_a^\eta(t) = e^{\eta S_a(t) - \phi_{\mu_a}(\eta) N_a(t)}$

$$S_a(t) = \sum_{s=1}^{t} X_s \mathbb{1}(A_s = a) \qquad \phi_{\mu_a} = \log \mathbb{E}_{X \sim \nu_{\mu_a}} \left[ e^{\lambda X} \right]$$

[Robbins, 1970]

For exponential families:

The martingale $Z_a^{(x)}(t) = \sum_i \frac{c}{i^2} \left[ W_a^{\eta_i^-(x)}(t) + W_a^{\eta_i^+(x)}(t) \right]$ satisfies

$$\left\{ e^{\lambda X_a(t) - \lambda f(\xi)} \geq z \right\} \subseteq \left\{ \widetilde{Z}_a^{(\lambda, z)}(t) \geq 1 \right\}$$

# Calibration of the Stopping Rule

$$X_a(t) = N_a(t) d(\hat{\mu}_a(t), \mu_a) - 3\log(1 + \log(N_a(t)))$$

**Step 1: $e^{\lambda X_a(t)}$ is (almost) an e-process**

$$\forall \lambda \in \Lambda : \quad M_a^\lambda(t) \geq e^{\lambda X_a(t) - g(\lambda)}$$

where $M_a^\lambda(t)$ is a test martingale and $g$ a correction function.

1. Link $X_a(t)$ to the martingale $W_a^\eta(t) = e^{\eta S_a(t) - \phi_{\mu_a}(\eta) N_a(t)}$

$$S_a(t) = \sum_{s=1}^{t} X_s \mathbb{1}(A_s = a) \qquad \phi_{\mu_a} = \log \mathbb{E}_{X \sim \nu_{\mu_a}} \left[ e^{\lambda X} \right]$$

[Robbins, 1970]

For exponential families:

The final test martingale is

$$M_a^\lambda(t) \propto 1 + \int_1^\infty \widetilde{Z}_a^{(\lambda,z)}(t) dz$$

$$X_a(t) = N_a(t) d(\hat{\mu}_a(t), \mu_a) - 3\log(1 + \log(N_a(t)))$$

## Step 2: Product martingales

$$\forall \lambda \in \Lambda : M^\lambda(t) = \prod_{a \in [K]} M_a^\lambda(t) \text{ is still a test martingale}$$

➜ Chernoff method + Ville's inequality

$$X_a(t) = N_a(t)d(\hat{\mu}_a(t), \mu_a) - 3\log(1+\log(N_a(t)))$$

## Step 2: Product martingales

$$\forall \lambda \in \Lambda : M^\lambda(t) = \prod_{a\in[K]} M_a^\lambda(t) \text{ is still a test martingale}$$

➜ Chernoff method + Ville's inequality

$$\begin{aligned}
\mathbb{P}\left(\exists t \in \mathbb{N} : \sum_{a=1}^K X_a(t) > x\right) &\leq \mathbb{P}\left(\exists t \in \mathbb{N} : e^{\sum_{a=1}^K (\lambda X_a(t) - g(\lambda))} > e^{\lambda x - Kg(\lambda)}\right) \\
&\leq \mathbb{P}\left(\exists t \in \mathbb{N} : M(t) > e^{\lambda x - Kg(\lambda)}\right) \\
&\leq e^{-\lambda x + Kg(\lambda)}
\end{aligned}$$

Then optimize over $\lambda$:

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \sum_{a=1}^K X_a(t) > K \min_{\lambda\in\Lambda} \frac{g(\lambda) + \log(1/\delta)/K}{\lambda}\right) \leq \delta.$$

# Final result

**Correctness [Kaufmann and Koolen, 2021]**

When the arm distributions belong to a one-dimensional exponential family, there exists a threshold such that

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

for which, $\mathbb{P}_{\boldsymbol{\mu}}(\tau < \infty, \hat{\imath}_\tau \neq i_\star(\boldsymbol{\mu})) \leq \delta$.

(the factor $K$ may be reduced for some particular identification tasks)

**Wait! Can we actually implement it?**

Track-and-Stop requires the computation in every round $t$ of the "minimal distance"

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

for checking the stopping rule, and

$$\arg\max_{w \in \boldsymbol{\Sigma}_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \lambda_a)$$

for the sampling rule.

# Outline

# GLR stopping rule

$$i_\star(\boldsymbol{\mu}) = a_\star(\boldsymbol{\mu}) = \arg\max_{a\in[K]} \mu_a$$

Using that $\mathrm{Alt}(\boldsymbol{\mu}) = \bigcup_{a\neq a_\star(\boldsymbol{\mu})} \{\boldsymbol{\lambda} : \lambda_a > \lambda_{a_\star}\}$, the minimal distance can be computed in closed-form:

$$\inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} w_i d(\mu_i, \lambda_i)$$

$$= \min_{a\neq a_\star} \inf_{\boldsymbol{\lambda}:\lambda_a>\lambda_{a_\star}} \sum_{i=1}^{K} w_i d(\mu_i, \lambda_i)$$

$$= \min_{a\neq a_\star} \min_{\lambda\in(\mu_a,\mu_{a_\star})} \left[w_{a_\star} d(\mu_{a_\star}, \lambda) + w_a d(\mu_a, \lambda)\right]$$

$$= \min_{a\neq a_\star} \left[w_{a_\star} d\left(\mu_{a_\star}, \frac{w_{a_\star}\mu_{a_\star} + w_a\mu_a}{w_{a_\star} + w_a}\right) + w_a d\left(\mu_a, \frac{w_{a_\star}\mu_{a_\star} + w_a\mu_a}{w_{a_\star} + w_a}\right)\right]$$

for exponential families.

# Example: Gaussian bandits

For Gaussian bandits with variance $\sigma^2$:

$$\inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} w_i \frac{(\mu_i - \lambda_i)^2}{2\sigma^2} = \min_{a \neq a_\star} \frac{(\mu_\star - \mu_a)^2}{2\sigma^2 \left( \frac{1}{w_{a_\star}} + \frac{1}{w_a} \right)}$$

hence

$$\tau_\delta = \inf \left\{ t \in \mathbb{N} : \min_{a \neq \hat{a}_t^\star} \frac{(\hat{\mu}_{\hat{a}_t^\star}(t) - \hat{\mu}_a(t))^2}{2\sigma^2 \left( \frac{1}{N_{\hat{a}_t^\star}(t)} + \frac{1}{N_a(t)} \right)} > \beta(t, \delta) \right\}$$

# Example: Gaussian bandits

For Gaussian bandits with variance $\sigma^2$:

$$\inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} w_i \frac{(\mu_i - \lambda_i)^2}{2\sigma^2} = \min_{a \neq a_\star} \frac{(\mu_\star - \mu_a)^2}{2\sigma^2 \left( \frac{1}{w_{a_\star}} + \frac{1}{w_a} \right)}$$

hence

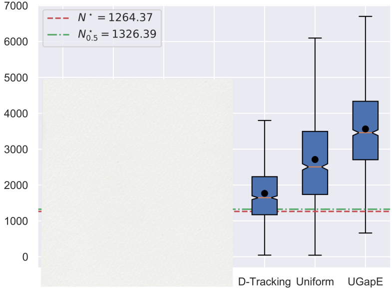
$$\tau_\delta = \inf \left\{ t \in \mathbb{N} : \min_{a \neq \hat{a}_t^\star} \frac{(\hat{\mu}_{\hat{a}_t^\star}(t) - \hat{\mu}_a(t))^2}{2\sigma^2 \left( \frac{1}{N_{\hat{a}_t^\star}(t)} + \frac{1}{N_a(t)} \right)} > \beta(t, \delta) \right\}$$

But $w_\star(\boldsymbol{\mu})$ still doesn't have a closed form

➜ we propose an efficient approximation algorithm for exponential families in [Garivier and Kaufmann, 2016]

# In practice

Empirical distribution of $\tau_\delta$ for $\delta = 0.01$ for different algorithms on $\boldsymbol{\mu} = [1, 0.8, 0.75, 0.7]$, $\sigma^2 = 1$, estimated on 1000 runs



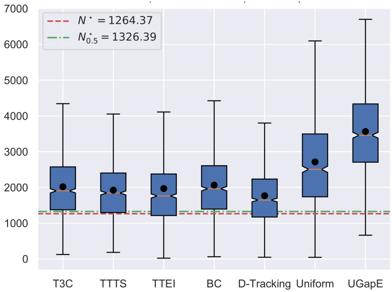

Using the right stopping rule makes a difference:

$$\text{UGapE}: \quad \forall a \neq \hat{\imath}^\star_t \quad , \; \hat{\mu}_{\hat{\imath}^\star_t}(t) - \hat{\mu}_a(t) > \sqrt{\frac{2\sigma^2\beta(t,\delta)}{N_{a,t}}} + \sqrt{\frac{2\sigma^2\beta(t,\delta)}{N_{a,t}}}$$

$$\text{GLR}: \quad \forall a \neq \hat{\imath}^\star_t \quad , \; \hat{\mu}_{\hat{\imath}^\star_t}(t) - \hat{\mu}_a(t) > \sqrt{2\sigma^2\beta(t,\delta)\left(\frac{1}{N_{a,t}} + \frac{1}{N_{a,t}}\right)}$$

<u>Limitation</u>: Computing $w^\star$ is costly

Empirical distribution of $\tau_\delta$ for $\delta = 0.01$ for different algorithms on $\boldsymbol{\mu} = [1, 0.8, 0.75, 0.7]$, $\sigma^2 = 1$, estimated on 1000 runs




Using the right stopping rule makes a difference:

$$\text{UGapE}: \quad \forall a \neq \hat{\imath}_t^\star \quad , \ \hat{\mu}_{\hat{\imath}_t^\star}(t) - \hat{\mu}_a(t) > \sqrt{\frac{2\sigma^2\beta(t,\delta)}{N_{a,t}}} + \sqrt{\frac{2\sigma^2\beta(t,\delta)}{N_{a,t}}}$$

$$\text{GLR}: \quad \forall a \neq \hat{\imath}_t^\star \quad , \ \hat{\mu}_{\hat{\imath}_t^\star}(t) - \hat{\mu}_a(t) > \sqrt{2\sigma^2\beta(t,\delta)\left(\frac{1}{N_{a,t}} + \frac{1}{N_{a,t}}\right)}$$

Efficient alternatives to Tracking exist, e.g. Top Two algorithms

# Outline

# Multi-objective bandit

**Bandit model**

- $K$ arms $\nu_1, \ldots, \nu_K$
- $\nu_k$ is a multi-variate distribution in $\mathbb{R}^D$ with mean $\boldsymbol{\mu}_k \in \mathbb{R}^D$
- Assumption: each marginal of $\nu_k$ is *sub-Gaussian*

In each round $t$, an agent selects and arm $A_t \in [K]$ and observes a response $\boldsymbol{X}_t \sim \nu_{A_t}$, independently from past observations.

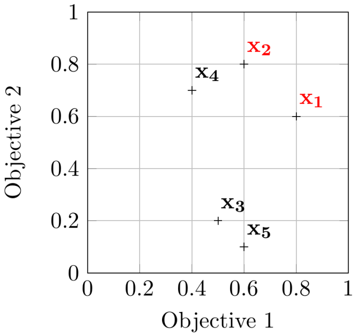
➜ What is a "good set of arms"?
a possiblity: the **Pareto Set** of the mean vectors

$$\mathcal{S}^\star(\boldsymbol{\mu}) = \{\text{"all arms that are not uniformly worse than any other arm"}\}$$

# Pareto Set

Let $\mathcal{X} \subset \mathbb{R}^D$ a set of vectors. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$.

- $\boldsymbol{x}$ is (strictly) dominated by $\boldsymbol{y}$ ($\boldsymbol{x} \prec \boldsymbol{y}$) if $\forall d \in [D],\ x^d < y^d$
- The Pareto Set is
  $\mathcal{P}(\mathcal{X}) := \{\boldsymbol{x} \in \mathcal{X} : \nexists\ \boldsymbol{y} \in \mathcal{X} \text{ such that } \boldsymbol{x} \prec \boldsymbol{y}\}$
- A vector $\boldsymbol{x} \in \mathcal{P}(\mathcal{X})$ is called Pareto optimal




1. $\boldsymbol{x_3} \prec \boldsymbol{x_1}$
2. $\boldsymbol{x_4} \prec \boldsymbol{x_2}$
3. $\boldsymbol{x_5} \prec \boldsymbol{x_1}$
4. $\boldsymbol{x_1} \not\prec \boldsymbol{x_2}$
5. $\boldsymbol{x_2} \not\prec \boldsymbol{x_1}$

$$\mathcal{P}(\mathcal{X}) = \{\boldsymbol{x_1}, \boldsymbol{x_2}\}$$

# Pareto Set Identification with Fixed Confidence

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K) \in (\mathbb{R}^D)^K$$
$$\mathcal{S}^\star(\boldsymbol{\mu}) = \{k \in [K] : \mu_k \in \mathcal{P}(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K)\}$$

Pareto Set Identification algorithm:

- a sampling rule $A_t \in [K]$ : what is the next arm to explore?

➜ get a new observation $\boldsymbol{X}_t \sim \nu_{A_t} \in \mathbb{R}^D$

- a recommendation rule $\widehat{S}_t$ : a guess for $\mathcal{S}^\star(\boldsymbol{\mu})$
- a stopping rule $\tau$: when to stop the data collection?

**Definition**

An algorithm is $\delta$-correct (on $\mathcal{M}$) if, for all $\boldsymbol{\nu} \in \mathcal{M}$,
$\mathbb{P}_{\boldsymbol{\nu}}(\widehat{S}_\tau \neq \mathcal{S}^\star(\boldsymbol{\mu})) \leq \delta$.

**Goal:** a $\delta$-correct algorithm with small sample complexity $\mathbb{E}_{\boldsymbol{\nu}}[\tau]$

# Sample Complexity Lower Bound

## Theorem

For arms that are multi-variate Gaussian (known covariance $\Sigma$), any $\delta$-correct algorithm for Pareto Set Identification satisfies, for all $\boldsymbol{\mu} \in (\mathbb{R}^D)^K$,

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

where

$$T^\star(\boldsymbol{\mu})^{-1} = \sup_{w \in \Delta_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\mathcal{S}^\star(\boldsymbol{\mu}))} \left(\sum_{k=1}^{K} w_k \mathrm{KL}(\mathcal{N}(\boldsymbol{\mu}_a, \Sigma), \mathcal{N}(\boldsymbol{\lambda}_a, \Sigma))\right).$$

with $\mathrm{Alt}(\mathcal{S}) = \{\boldsymbol{\lambda} \in (\mathbb{R}^D)^K : \mathcal{S}^\star(\boldsymbol{\lambda}) \neq \mathcal{S}\}$.

# Sample Complexity Lower Bound

## Theorem

For arms that are multi-variate Gaussian (known covariance $\Sigma$), any $\delta$-correct algorithm for Pareto Set Identification satisfies, for all $\boldsymbol{\mu} \in (\mathbb{R}^D)^K$,

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

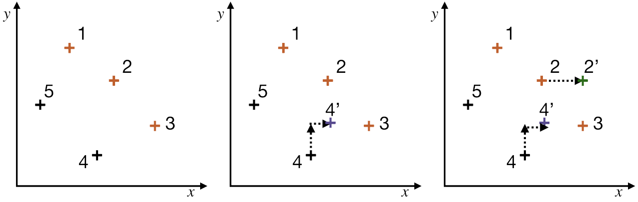
where

$$T^\star(\boldsymbol{\mu})^{-1} = \sup_{w \in \Delta_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\mathcal{S}^\star(\boldsymbol{\mu}))} \left( \sum_{k=1}^{K} w_k \frac{1}{2} \|\boldsymbol{\mu}_k - \boldsymbol{\lambda}_k\|^2_{\Sigma^{-1}} \right).$$

with $\mathrm{Alt}(\mathcal{S}) = \{\boldsymbol{\lambda} \in (\mathbb{R}^D)^K : \mathcal{S}^\star(\boldsymbol{\lambda}) \neq \mathcal{S}\}$.

# Computing the Minimal Distance

- there are many ways to alter the Pareto set




- no closed-form is known for the minimal distance

$$(1) : w \mapsto \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\mathcal{S})} \sum_k \frac{w_k}{2} \|\boldsymbol{\mu}_k - \boldsymbol{\lambda}_k\|^2_{\boldsymbol{\Sigma}^{-1}}$$

- for $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_d$, (1) can be computed by solving $O(K|S^\star(\boldsymbol{\mu})|^d)$ separably convex problems [Crepon et al., 2024]

# Track-And-Stop?

The GLR stopping rule

$$\tau = \inf \left\{ t \in \mathbb{N} : \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\widehat{S}_t^\star)} \sum_{k=1}^{K} \frac{N_k(t)}{2} \|\hat{\boldsymbol{\mu}}_k(t) - \boldsymbol{\lambda}_k\|^2_{\Sigma^{-1}} > \beta(t, \delta) \right\}$$

can be calibrated to attain correctness with

$$\beta(t,\delta) \simeq \log(1/\delta) + \log\log(1/\delta) + KD \log\log(Dt)$$

... but is computationally expansive due to the minimal distance.

The Tracking sampling rule is intractable as it further computes

$$w_\star(\boldsymbol{\mu}) = \operatorname*{arg\,max}_{w \in \Delta_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\mathcal{S}^\star(\boldsymbol{\mu}))} \sum_k \frac{w_k}{2} \|\boldsymbol{\mu}_k - \boldsymbol{\lambda}_k\|^2_{\boldsymbol{\Sigma}^{-1}}$$

➜ existing alternative approaches based on online learning e.g.[Ménard, 2019] also rely on minimal distance computation.

# The Posterior (Re)Sampling Stopping Rule

**PS Stopping rule**

For all $m \leq M(t,\delta)$, sample $\widetilde{\boldsymbol{\theta}}^m = (\widetilde{\boldsymbol{\theta}}_1^m, \ldots, \widetilde{\boldsymbol{\theta}}_K^m)$ with

$$\widetilde{\boldsymbol{\theta}}_a^m \sim \mathcal{N}\left(\hat{\boldsymbol{\mu}}_a(t), \frac{c(t,\delta)}{N_a(t)}\Sigma\right)$$

If for all $m$, $\mathcal{S}^\star(\widetilde{\boldsymbol{\theta}}^m) = \mathcal{S}^\star(\hat{\boldsymbol{\mu}}(t))$, **stop** and recommend $\mathcal{S}^\star(\hat{\boldsymbol{\mu}}(t))$

- inspired by the TS-Explore strategy for Combinatorial bandits [Wang and Zhu, 2022]
- analyzed in [Kone et al., 2025] for PSI, together with a tractable sampling rule giving asymptotic optimality

# Proving correctness

Let $\beta(t, \delta)$ be such that $\mathcal{E}_\delta$ holds w.p. at least $1 - \delta$:

$$\mathcal{E}_\delta = \bigcap_{t \geq 1} \underbrace{\left( \sum_k N_{t,k} \left\| \mu_k - \hat{\mu}_{t,k} \right\|^2_{\mathbf{\Sigma}^{-1}} < 2\beta(t, \delta) \right)}_{\mathcal{E}_\delta^t}$$

Then,

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\nu}}(\tau < \infty, \widehat{\mathcal{S}}_\tau \neq \mathcal{S}^\star) &\leq \delta/2 + \mathbb{P}_{\boldsymbol{\nu}}(\tau < \infty \text{ and } \widehat{\mathcal{S}_\tau} \neq S^\star, \mathcal{E}_{\delta/2}) \\
&\leq \delta/2 + \sum_{t \geq 1} \mathbb{P}_{\boldsymbol{\nu}}(\tau = t \text{ and } \widehat{\mathcal{S}}_t \neq \mathcal{S}^\star, \mathcal{E}^t_{\delta/2}) \\
&= \delta/2 + \sum_{t \geq 1} \mathbb{E}_{\boldsymbol{\nu}} \left[ \mathbb{1}_{\widehat{\mathcal{S}}_t \neq \mathcal{S}^\star} \mathbb{1}_{\mathcal{E}^t_{\delta/2}} \mathbb{P}_{\boldsymbol{\nu}}(\tau = t \mid \mathcal{H}_{t-1}) \right]
\end{aligned}$$

# Proving correctness

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\nu}}(\tau = t \mid \mathcal{H}_{t-1}) &\leq \mathbb{P}_{\boldsymbol{\nu}}(\forall m \leq M(t,\delta), \mathcal{S}^\star(\widetilde{\boldsymbol{\theta}}_t^m) = \widehat{\mathcal{S}}_t \mid \mathcal{H}_{t-1}) \\
&= \left(1 - \mathbb{P}_{\boldsymbol{\nu}}(S^\star(\widetilde{\boldsymbol{\theta}}_t^1) \neq \widehat{\mathcal{S}}_t \mid \mathcal{H}_{t-1})\right)^{M(t,\delta)}, \\
&= \left(1 - \Pi_t(\mathrm{Alt}(\widehat{\mathcal{S}}_t))\right)^{M(t,\delta)} \\
&\leq \exp\Big(-\Pi_t(\mathrm{Alt}(\widehat{\mathcal{S}}_t)) M(t,\delta)\Big)
\end{aligned}$$

hence the error probability is bounded by

$$\frac{\delta}{2} + \sum_{t\geq 1} \mathbb{E}_{\boldsymbol{\nu}}\left[\mathbb{1}_{\widehat{\mathcal{S}}_t \neq \mathcal{S}^\star(\boldsymbol{\mu})} \mathbb{1}_{\mathcal{E}^t_{\delta/2}} \exp\Big(-\Pi_t(\mathrm{Alt}(\widehat{\mathcal{S}}_t)) M(t,\delta)\Big)\right].$$

The tricky part of the proof is then to get a lower bound on $\Pi_t(\mathrm{Alt}(\widehat{\mathcal{S}}_t))$ (Gaussian anti-concentration)

# Proving correctness

$$
\begin{aligned}
\mathbb{P}_{\boldsymbol{\nu}}(\tau = t \mid \mathcal{H}_{t-1}) &\leq \mathbb{P}_{\boldsymbol{\nu}}(\forall m \leq M(t,\delta), \mathcal{S}^\star(\widetilde{\boldsymbol{\theta}}_t^m) = \widehat{\mathcal{S}}_t \mid \mathcal{H}_{t-1}) \\
&= (1 - \mathbb{P}_{\boldsymbol{\nu}}(S^\star(\widetilde{\boldsymbol{\theta}}_t^1) \neq \widehat{\mathcal{S}}_t \mid \mathcal{H}_{t-1}))^{M(t,\delta)} , \\
&= (1 - \Pi_t(\mathrm{Alt}(\widehat{\mathcal{S}}_t)))^{M(t,\delta)} \\
&\leq \exp\left(-\Pi_t(\mathrm{Alt}(\widehat{\mathcal{S}}_t)) M(t,\delta)\right)
\end{aligned}
$$

hence the error probability is bounded by

$$
\frac{\delta}{2} + \sum_{t \geq 1} \mathbb{E}_{\boldsymbol{\nu}}\left[\mathbb{1}_{\widehat{\mathcal{S}}_t \neq \mathcal{S}^\star(\boldsymbol{\mu})} \mathbb{1}_{\mathcal{E}^t_{\delta/2}} \exp\left(-\Pi_t(\mathrm{Alt}(\widehat{\mathcal{S}}_t)) M(t,\delta)\right)\right] .
$$

The tricky part of the proof is then to get a lower bound on $\Pi_t(\mathrm{Alt}(\widehat{\mathcal{S}}_t))$ (Gaussian anti-concentration)

**Lemma** [Kone et al., 2025]

The PS Stopping rule is $\delta$ correct for

$$
c(t,\delta) \simeq \frac{\log(\log(t)/\delta)}{\log(1/\delta)} \text{ and } \mathrm{M}(t,\delta) \simeq \frac{\log(t/\delta)}{\delta}
$$

# Conclusion

Two generic stopping rules for pure exploration tasks in bandits:

- the GLR stopping rule that is easy to calibrate (for exponential families)
- the PS stopping rule that can be easier to compute (but harder to calibrate)

In these approaches, e-processes are hidden in the proofs... but the resulting calibration are a bit conservative in practice.

➜ Tigther calibrations?

➜ When is PS "better" than GLR?

# References


- Aurélien Garivier, Emilie Kaufmann
  *Optimal Best Arm Identification with Fixed Confidence*
  (COLT 2016)
- Emilie Kaufmann, Wouter M. Koolen
  *Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals* (JMLR 2021)




- Marc Jourdan, Cyrille Koné, Emilie Kaufmann
  *Pareto Set Identification with Posterior Sampling.* (AISTATS 2025)



Auer, P., Chiang, C., Ortner, R., and Drugan, M. M. (2016).
Pareto front identification from stochastic bandit feedback.
In *AISTATS*.

Crepon, É., Garivier, A., and Koolen, W. M. (2024).
Sequential learning of the pareto front for multi-objective bandits.
In *AISTATS*.

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*.

Kaufmann, E. and Koolen, W. (2021).
Mixture martingales revisited with applications to sequential tests and confidence intervals.
*Journal of Machine Learning Research*, 22(246).

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)*.

Kone, C., Jourdan, M., and Kaufmann, E. (2025).
Pareto set identification with posterior sampling.
In *AISTATS*.

Locatelli, A., Gutzeit, M., and Carpentier, A. (2016).

An optimal algorithm for the thresholding bandit problem.
In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 1690–1698.

Ménard, P. (2019).
Gradient ascent for active exploration in bandit problems.
*arXiv 1905.08165.*

Robbins, H. (1970).
Statistical Methods Related to the law of the iterated logarithm.
*Annals of Mathematical Statistics*, 41(5):1397–1409.

Wang, S. and Zhu, J. (2022).
Thompson sampling for (combinatorial) pure exploration.
In *International Conference on Machine Learning*. PMLR.

# On the effect of correlation

We evaluate the performance of PSIPS on a 5-arm, 2-dimensional Gaussian instance with correlated objectives.

- Covariance matrix: $\boldsymbol{\Sigma}_\rho$ with unit variances and correlation $\rho \in (-1, 1)$.
- $\rho = 0$: objectives are independent.
- $\rho \to +1$ (resp. $\rho \to -1$): strongly positively (resp. negatively) correlated objectives.