Solving Pure Exploration Problems in Bandits and Beyond

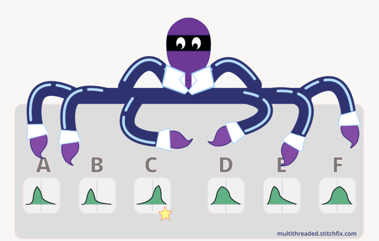
Emilie Kaufmann

CRIStAL
Centre de Recherche en Informatique, Signal et Automatique de Lille

Université de Lille
cnrs
centralelille

The Multi Armed Bandit (MAB) model

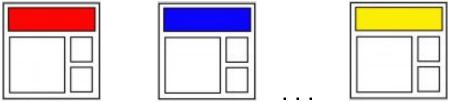
- K *unknown* distributions $\nu_1, \dots, \nu_K$ called *arms*
- a time t, select an arm A_t and collect an observation $X_t \sim \nu_{A_t}$



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 : $\nu_a \in \mathcal{D}$

Two classical bandit problems



p_1 $\quad$ p_2 $\quad$ p_K

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

For the t-th visitor :

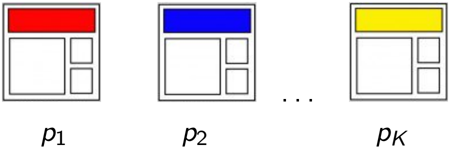
- choose a version A_t to display
- observe $X_t = 1$ if a product is bought, 0 otherwise

Objective 1 : observation = reward $\rightarrow$ maximize rewards

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

a *reinforcement learning* problem

Two classical bandit problems



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 : best arm identification

- identify quickly $a_\star = \arg\max_a \ p_a$
- find the best version (in order to keep displaying it)

a *pure exploration* problem

Other applications

- clinical trials

  $\rightarrow$ observation : success/failure (Bernoulli distribution)

- movie recommendation

  $\rightarrow$ observation : rating (multinomial)

- recommendation in agriculture

  $\rightarrow$ observation : yield (complex, non-parametric distribution)

A Detour : Maximizing Rewards

Bandit instance : $\nu = (\nu_1, \nu_2, \dots, \nu_K)$, mean $\mu_a = \mathbb{E}_{X\sim\nu_a}[X]$.

$$\mu_\star = \max_{a\in\{1,\dots,K\}} \mu_a \qquad a_\star = \operatorname*{arg\,max}_{a\in\{1,\dots,K\}} \mu_a.$$

Maximizing rewards $\leftrightarrow$ selecting $a_\star$ as much as possible

$\leftrightarrow$ minimizing the regret [Robbins, 1952]

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

A Detour : Maximizing Rewards

$N_a(t)$: number of selections of arm a in the first t rounds
$\Delta_a := \mu_\star - \mu_a$: sub-optimality gap of arm a

Regret decomposition

$$\mathcal{R}_\nu(\mathcal{A}, T) = \sum_{a=1}^{K} \Delta_a \mathbb{E}\left[N_a(T)\right].$$

A strategy with small regret should :

- select not too often arms for which $\Delta_a > 0$
- ... which requires to try all arms to estimate their Δ_a's

$\Rightarrow$ Exploration / Exploitation trade-off

The need for Exploration

Follow the Leader (or Greedy strategy)

Select each arm once, then exploit the current knowledge :

$$A_{t+1} = \underset{a \in [K]}{\arg\max} \; \hat{\mu}_a(t)$$

where

- $N_a(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = a)$ is the number of selections of arm a
- $\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} X_s \mathbb{1}(A_s = a)$ is the empirical mean of the rewards collected from arm a

The need for Exploration

Follow the Leader (or Greedy strategy)

Select each arm once, then exploit the current knowledge :

$$A_{t+1} = \underset{a \in [K]}{\arg\max} \ \hat{\mu}_a(t)$$

where

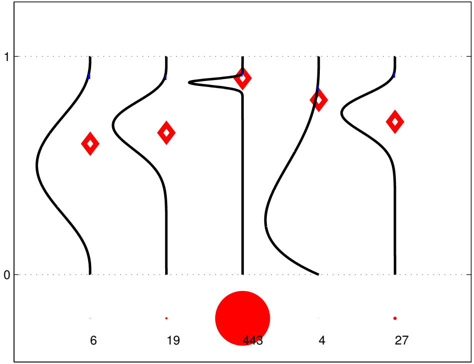
- $N_a(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = a)$ is the number of selections of arm a
- $\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} X_s \mathbb{1}(A_s = a)$ is the empirical mean of the rewards collected from arm a

Beeing greedy can fail ! $\nu_1 = \mathcal{B}(\mu_1), \nu_2 = \mathcal{B}(\mu_2)$, $\mu_1 > \mu_2$

$$\mathbb{E}[N_2(T)] \geq (1 - \mu_1)\mu_2 \times (T - 1)$$

Thompson Sampling

A Bayesian strategy : encodes uncertainty with posterior distributions



In each round, TS samples a possible bandit model from the posterior and selects the best arm in the sampled model

[Thompson, 1933, Russo et al., 2018]

Thompson Sampling

Example : Bernoulli bandit with means $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$

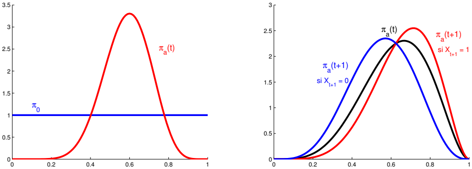
- prior distribution : $\mu_a \overset{\text{i.i.d.}}{\sim} \mathcal{U}([0,1])$

➜ posterior distribution :

$$\begin{aligned} \pi_a(t) &= \mathcal{L}\left(\mu_a | X_1, \dots, X_t\right) \\ &= \mathrm{Beta}\Big(\underbrace{S_a(t)}_{\#ones} + 1, \underbrace{N_a(t) - S_a(t)}_{\#zeros} + 1\Big) \end{aligned}$$

$N_a(t) = \sum_{s=1}^{t} \mathbb{1}_{(A_s = a)}$ number of observations from arm a

$S_a(t) = \sum_{s=1}^{t} X_s \mathbb{1}_{(A_s = a)}$ sum of the rewards from arm a



Thompson Sampling

Example : Bernoulli bandit with means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$

- prior distribution : $\mu_a \overset{\text{i.i.d.}}{\sim} \mathcal{U}([0,1])$

➜ posterior distribution :

$$\begin{aligned} \pi_a(t) &= \mathcal{L}\left(\mu_a | X_1, \ldots, X_t\right) \\ &= \mathrm{Beta}\Big(\underbrace{S_a(t)}_{\#ones}+1, \underbrace{N_a(t) - S_a(t)}_{\#zeros}+1\Big) \end{aligned}$$

Thompson Sampling

In round $t+1$:

$$\forall a \in [K], \quad \widetilde{\theta}_a(t) \sim \pi_a(t)$$

$$A_{t+1} = \underset{a \in [K]}{\arg\max}\ \widetilde{\theta}_a(t)$$

Thompson Sampling

Example : Gaussian bandit with means $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$, var σ^2

- prior distribution : $\mu_a \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \kappa^2)$

→ posterior distribution :

$$\begin{aligned}\pi_a(t) &= \mathcal{L}\left(\mu_a | X_1, \dots, X_t\right) \\ &= \mathcal{N}\left(\frac{S_a(t)}{N_a(t) + \frac{\sigma^2}{\kappa^2}}, \frac{\sigma^2}{N_a(t) + \frac{\sigma^2}{\kappa^2}}\right)\end{aligned}$$

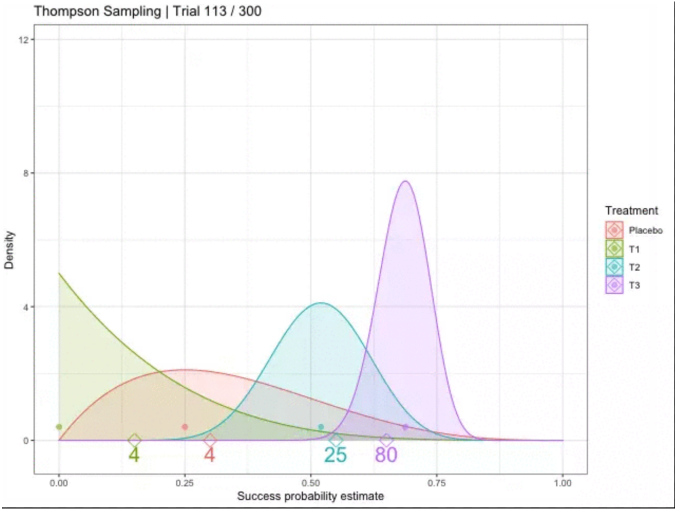
Thompson Sampling

In round $t + 1$:

$$\forall a \in [K], \quad \widetilde{\theta}_a(t) \sim \pi_a(t)$$

$$A_{t+1} = \underset{a \in [K]}{\arg\max}\ \widetilde{\theta}_a(t)$$

Thompson Sampling in action



source : Wikipedia

Thompson Sampling : Theory

Upper bound on sub-optimal selections

$$\forall a \neq a_\star, \quad \mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] \leq \frac{\log(T)}{\mathrm{KL}(\nu_a, \nu_{a_\star})} + o_\mu(\log(T)).$$

where $\mathrm{KL}(\nu_a, \nu_{a_\star})$ is the KL divergence between ν_a and $\nu_{a_\star}$

- proved for Bernoulli bandits, with a uniform prior [Kaufmann et al., 2012, Agrawal and Goyal, 2013]
- for 1-dimensional exponential families, with a conjuguate prior [Agrawal and Goyal, 2017, Korda et al., 2013]
- a nice non-parametric extension for bounded rewards [Riou and Honda, 2020]

→ Thompson Sampling is **asymptotically optimal** in these cases

Is Thompson Sampling finding the Best Arm ?

At time t, TS is selecting

$$A_t = \underset{a \in [K]}{\arg\max}\ \widetilde{\theta}_a(t-1)$$

Is it a reasonnable guess for the best arm ?

Is Thompson Sampling finding the Best Arm ?

At time t, TS is selecting

$$A_t = \underset{a \in [K]}{\arg\max}\ \widetilde{\theta}_a(t-1)$$

Is it a reasonnable guess for the best arm ?

Less explorative **recommendation rules** :

- empirical best arm : $B_t = \arg\max_{a \in [K]}\ \hat{\mu}_a(t)$
- most played arm : $B_t = \arg\max_{a \in [K]}\ N_a(t)$
- a smoother (randomized) version

$$\mathbb{P}(B_t = b | \mathcal{H}_t) = \frac{N_b(t)}{t}$$

Is Thompson Sampling finding the Best Arm ?

For Thompson Sampling + $B_t \sim \left(\frac{N_1(t)}{t}, \dots, \frac{N_K(t)}{t}\right)$

$$
\begin{aligned}
\mathbb{P}_\nu(B_t \neq a_\star) &= \sum_{b \neq a_\star} \mathbb{P}_\nu(B_t = b) = \sum_{b \neq a_\star} \mathbb{E}_\nu\left[\mathbb{P}_\nu(B_t = b | \mathcal{H}_t)\right] \\
&= \sum_{b \neq a_\star} \frac{\mathbb{E}_\nu[N_b(t)]}{t} \leq C_\nu \frac{\log(t)}{t}
\end{aligned}
$$

Is Thompson Sampling finding the Best Arm ?

For Thompson Sampling $+$ $B_t \sim \left(\frac{N_1(t)}{t}, \dots, \frac{N_K(t)}{t}\right)$

$$\begin{aligned}\mathbb{P}_\nu(B_t \neq a_\star) &= \sum_{b \neq a_\star} \mathbb{P}_\nu(B_t = b) = \sum_{b \neq a_\star} \mathbb{E}_\nu\left[\mathbb{P}_\nu(B_t = b | \mathcal{H}_t)\right] \\ &= \sum_{b \neq a_\star} \frac{\mathbb{E}_\nu[N_b(t)]}{t} \leq C_\nu \frac{\log(t)}{t}\end{aligned}$$

How good is this decay rate ?

- worse than uniform sampling $+$ empirical best arm (exponential decay)
- in order to guarantee $\mathbb{P}_\nu(B_t \neq a_\star) \leq \delta$, t has to be chosen as a function of the unknown instance ν

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


Pure Exploration

Arms : simple distributions parameterized by their means

(Bernoulli, Gaussian with known variance)

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.

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 τ : when to stop the data collection ?

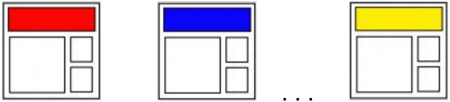
Definition

An algorithm is δ-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 δ-correct algorithm with small sample complexity $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$

Examples of identification tasks

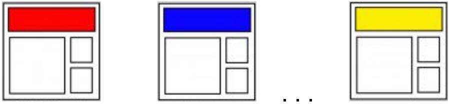
- Best Arm Identification [Even-Dar et al., 2006]



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

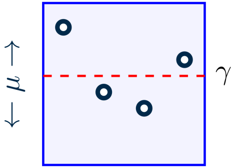
Examples of identification tasks

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



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

- Thresholding bandit : classify the arms above/ below a threshold [Locatelli et al., 2016]

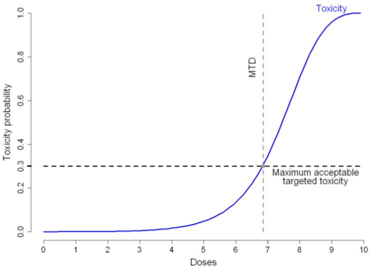


$$i_\star(\boldsymbol{\mu}) = (\mathbb{1}(\mu_1 > \gamma), \dots, \mathbb{1}(\mu_K > \gamma)) \in \{0,1\}^K$$

Examples of identification tasks

- Other threshold-based questions

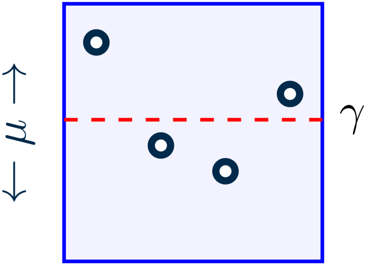
Which arm is the closest to γ ?



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

[Garivier et al., 2019a]

Is there an arm below γ ?

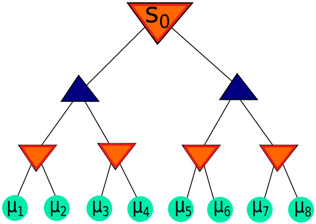


$$i_\star(\boldsymbol{\mu}) = \mathbb{1}(\min_a \mu_a < \gamma) \in \{0, 1\}$$

[Kaufmann et al., 2018]

Examples of identification tasks

- Finding the best move in a maxmin game tree



$$V_s(\boldsymbol{\mu}) = \begin{cases} \mu_s & \text{if s } \in \mathcal{L}, \\ \max_{c \in \mathcal{C}(s)} V_c & \text{for s MAX}, \\ \min_{c \in \mathcal{C}(s)} V_c & \text{for s MIN}. \end{cases}$$

$$i_\star(\boldsymbol{\mu}) = \operatorname*{argmax}_{s \in \mathcal{C}(s_0)} V_s(\boldsymbol{\mu})$$

[Teraoka et al., 2014, Kaufmann and Koolen, 2017]

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 τ : when to stop the data collection ?

Definition

An algorithm is δ-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 δ-correct algorithm with small sample complexity $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$

A lower bound on the sample complexity

Lemma ([Garivier and Kaufmann, 2016, Garivier et al., 2019b])

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

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

where KL is the Kullback-Leibler divergence and

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

A lower bound on the sample complexity

Lemma ([Garivier and Kaufmann, 2016, Garivier et al., 2019b])

$\boldsymbol{\mu} \in \mathcal{M}$ and $\boldsymbol{\lambda} \in \mathcal{M}$ two different bandit instances.
τ a stopping time and $\mathcal{E}$ an event depending on $X_1, \ldots, X_\tau$.

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

where KL is the Kullback-Leibler divergence and

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

Assumption : Arm distributions parameterized by their means

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

A lower bound on the sample complexity

Lemma ([Garivier and Kaufmann, 2016, Garivier et al., 2019b])

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

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

where KL is the Kullback-Leibler divergence and

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

Assumption : Arm distributions parameterized by their means

$$d(\mu, \mu') = \frac{(\mu - \mu')^2}{2\sigma^2} \text{ (Gaussian with variance } \sigma^2\text{)}$$

A lower bound on the sample complexity

Lemma ([Garivier and Kaufmann, 2016, Garivier et al., 2019b])

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

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

where KL is the Kullback-Leibler divergence and

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

Assumption : Arm distributions parameterized by their means

$$d(\mu, \mu') = \mathrm{kl}(\mu, \mu') \text{ (Bernoulli distributions)}$$

A lower bound on the sample complexity

Lemma ([Garivier and Kaufmann, 2016, Garivier et al., 2019b])

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

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

where KL is the Kullback-Leibler divergence and

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

Under a δ-correct algorithm,

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

A lower bound on the sample complexity

Lemma

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

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

- Let $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} \in \mathcal{M} : i_\star(\boldsymbol{\lambda}) \neq i_\star(\boldsymbol{\mu})\}$.

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

A lower bound on the sample complexity

Theorem [Garivier and Kaufmann, 2016]

For any δ-correct algorithm,

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

where

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

where

$$\begin{aligned} \Delta_K &= \left\{ \boldsymbol{w} \in [0,1]^K : \sum_{i=1}^{K} w_i = 1 \right\} \\ \mathrm{Alt}(\boldsymbol{\mu}) &= \{ \boldsymbol{\lambda} \in \mathcal{M} : i_\star(\boldsymbol{\lambda}) \neq i_\star(\boldsymbol{\mu}) \} \end{aligned}$$

Optimal proportions

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

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

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

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

The Track-and-Stop principle

First ingredient : A stopping rule aligned with the lower bound

Are we confident enough in the empirical best answer $\hat{\imath}_t = i_\star(\hat{\boldsymbol{\mu}}(t))$?

$$\hat{\boldsymbol{\mu}}(t) = (\mu_1(t), \dots, \mu_K(t))$$

➜ yes, for high values of the Generalized (log) Likelihood Ratio

$$\ln \frac{\sup_{\boldsymbol{\lambda} \in \mathcal{M}} \ell(X_1, \dots, X_t; \boldsymbol{\lambda})}{\sup_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} \ell(X_1, \dots, X_t; \boldsymbol{\lambda})} = \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\mu}(t))} \ln \frac{\ell(X_1, \dots, X_t; \hat{\boldsymbol{\mu}}(t))}{\ell(X_1, \dots, X_t; \boldsymbol{\lambda})}$$

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

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

The Track-and-Stop principle

GLR stopping rule with threshold function $\beta(t,\delta)$:

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

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

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)

The Track-and-Stop principle

Second ingredient : A mechanism to make the empirical allocation converge to $w^\star(\boldsymbol{\mu})$

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

Requirements :

- 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 $\boldsymbol{\mu} \in \mathcal{M}$
- $\boldsymbol{\mu} \mapsto w^\star(\boldsymbol{\mu})$ can be computed efficiently, for all $\boldsymbol{\mu} \in \mathcal{M}$

The Track-and-Stop principle

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

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

Lemma

Under the Tracking sampling rule,

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

An asymptotically optimal algorithm

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

The Track-and-Stop strategy, that uses

- the Tracking sampling rule
- the GLRT stopping rule with

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

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

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

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

Why ?

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

An asymptotically optimal algorithm

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

The Track-and-Stop strategy, that uses

- the Tracking sampling rule
- the GLRT stopping rule with

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

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

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

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

Why ?

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

An asymptotically optimal algorithm

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

The Track-and-Stop strategy, that uses

- the Tracking sampling rule
- the GLRT stopping rule with

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

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

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

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

Why ?

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

An asymptotically optimal algorithm

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

The Track-and-Stop strategy, that uses

- the Tracking sampling rule
- the GLRT stopping rule with

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

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

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

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

Why ?

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

Computational aspects

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

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

for the sampling rule.

➜ Both can be challenging to compute for arbitrary identification tasks, especially the second one .

Track-and-Stop for Best Arm Identification

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

Using that $\mathrm{Alt}(\boldsymbol{\mu}) = \bigcup_{a\neq a_\star(\boldsymbol{\mu})} \{\boldsymbol{\lambda} : \lambda_a > \lambda_{a_\star}\}$ yields

$$\begin{aligned}
&\inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} w_i d(\mu_i, \lambda_i) \\
&= \min_{a\neq a_\star} \inf_{\boldsymbol{\lambda}:\lambda_a>\lambda_{a_\star}} \sum_{i=1}^{K} w_i d(\mu_i, \lambda_i) \\
&= \min_{a\neq a_\star} \inf_{\boldsymbol{\lambda}:\lambda_a>\lambda_{a_\star}} \sum_{i\in\{a,a_\star\}} w_i d(\mu_i, \lambda_i) \\
&= \min_{a\neq a_\star} \underbrace{\min_{\lambda\in(\mu_a,\mu_{a_\star})} \left[w_{a_\star} d(\mu_{a_\star}, \lambda) + w_a d(\mu_a, \lambda)\right]}_{\text{"transportation cost" associated to arm } a}
\end{aligned}$$

The min in λ is further attained in $\lambda = \frac{w_{a_\star}\mu_{a_\star} + w_a\mu_a}{w_{a_\star} + w_a}$.

Track-and-Stop for Best Arm Identification

In order to compute $w^\star(\boldsymbol{\mu})$, we further need to compute

$$\underset{\boldsymbol{w}\in\Delta_K}{\arg\max}\ \min_{a\neq a_\star} \underbrace{\left[w_{a_\star} d\left(\mu_{a_\star}, \frac{w_{a_\star}\mu_{a_\star} + w_a\mu_a}{w_{a_\star} + w_a}\right) + w_a d\left(\mu_a, \frac{w_{a_\star}\mu_{a_\star} + w_a\mu_a}{w_{a_\star} + w_a}\right)\right]}_{:= T_a(\boldsymbol{w})}$$

which can be done efficiently[1] by noting that at the optimum in $\boldsymbol{w}$ all the $T_a(\boldsymbol{w})$ are equal, and optimizing for their common value.

➜ efficient evaluation of $w^\star(\boldsymbol{\mu})$

[1] By computing the root of a real-valued function whose evaluation is linear in K

Track-and-Stop for Best Arm Identification

Example : BAI in Gaussian bandits with variance 1, for which

$$d(x, y) = \frac{(x-y)^2}{2}$$

we get

$$T^\star(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{w} \in \Delta_K} \min_{a \neq a_\star} \frac{(\mu_{a_\star} - \mu_a)^2}{2\left(\frac{1}{w_{a_\star}} + \frac{1}{w_a}\right)}$$

GLR stopping rule :

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

Track-and-Stop for Best Arm Identification

Example : BAI in Gaussian bandits with variance 1, for which

$$d(x,y) = \frac{(x-y)^2}{2}$$

we get

$$T^\star(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{w}\in\Delta_K} \min_{a\neq a_\star} \frac{(\mu_{a_\star}-\mu_a)^2}{2\left(\frac{1}{w_{a_\star}}+\frac{1}{w_a}\right)}$$

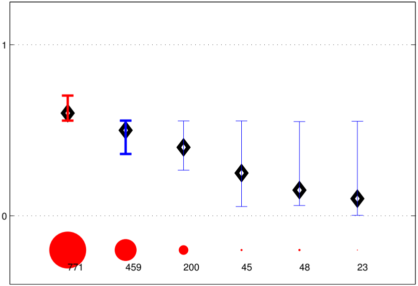
Lemma [Garivier and Kaufmann, 2016]

Recalling the gap $\Delta_a = \mu_\star - \mu_a$ for $a \neq a_\star$ and $\Delta_{a_\star} = \min_{a\neq a_\star} \Delta_a$,

$$\sum_{a=1}^{K} \frac{1}{\Delta_a^2} \leq T^\star(\mu) \leq 2\sum_{a=1}^{K} \frac{1}{\Delta_a^2}$$

Baseline : LUCB

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



- In round t, draw

$$B_t = \arg\max_b \hat{\mu}_b(t)$$
$$C_t = \arg\max_{c \neq B_t} \mathrm{UCB}_c(t)$$

- Stop at round t if

$$\mathrm{LCB}_{B_t}(t) > \mathrm{UCB}_{C_t}(t)$$

Theorem [Kalyanakrishnan et al., 2012]

For (sub)-Gaussian arms and well-chosen confidence intervals, $\mathbb{P}_{\boldsymbol{\mu}}(B_\tau \neq a_\star(\boldsymbol{\mu})) \leq \delta$ and

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] = \mathcal{O}\left(\left[\sum_{a=1}^{K} \frac{1}{\Delta_a^2}\right] \ln\left(\frac{1}{\delta}\right)\right)$$

Numerical experiments

Experiments on two Bernoulli bandit models :

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

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

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

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

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

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

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

Limitations

Track-and-Stop works really well for best arm identification but

- its computational cost is still an order of magnitude larger than existing baselines
- its performance guarantees are only asymptotic
(even if it works well for moderate values of δ)
- computing $w^\star(\boldsymbol{\mu})$ is not always doable for arbitrary identification tasks

Alternative to Track-and-Stop

[Degenne et al., 2019] leverage the interpretation of the lower bound as the value of a two-player zero-sum game

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

and propose to use two online learning algorithms to converge to it :

- The w-player gets $w^t \in \Delta_K$ from an online learning algorithm
- The λ-player best responds to it :

$$\boldsymbol{\lambda}^t = \underset{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))}{\arg\min} \sum_{a=1}^{t} w_a^t d(\hat{\mu}_a(t), \lambda_a)$$

- The online learner is fed with (an upper bound on) $g_t(w) = \sum_{a=1}^{K} w_a d(\hat{\mu}_a(t), \lambda_a^t)$

Alternative to Track-and-Stop

[Degenne et al., 2019] leverage the interpretation of the lower bound as the value of a two-player zero-sum game

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

and propose to use two online learning algorithms to converge to it :

- The w-player gets $w^t \in \Delta_K$ from an online learning algorithm
- The λ-player best responds to it :

$$\boldsymbol{\lambda}^t = \underset{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))}{\arg\min} \sum_{a=1}^{t} w_a^t d(\hat{\mu}_a(t), \lambda_a)$$

- The online learner is fed with (an upper bound on) $g_t(w) = \sum_{a=1}^{K} w_a d(\hat{\mu}_a(t), \lambda_a^t)$

Other idea : Thompson Sampling ?

References

- Xuedong Shang, Rianne De Heide, Pierre Ménard, Emilie Kaufmann, Michal Valko
  Fixed-Confidence Guarantees for Bayesian Best Arm Identification
  AISTATS 2020

- Marc Jourdan, Rémy Degenne, Dorian Baudry, Rianne De Heide, Emilie Kaufmann
  Top Two Algorithms Revisited
  NeurIPS 2022
- Marc Jourdan, Rémy Degenne, Emilie Kaufmann
  An ε-Best-Arm Identification Algorithm for Fixed-Confidence and Beyond
  NeurIPS 2023

Top Two Thompson Sampling

$\Pi_t = (\pi_1(t), \dots, \pi_K(t))$ posterior distribution on $(\mu_1, \dots, \mu_K)$

Top-Two Thompson Sampling (TTTS) [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 β, 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 $A_{t+1} = a_\star(\boldsymbol{\theta}')$

Top Two Thompson Sampling

Bayesian analysis of TTTS

[Russo, 2016] proves that, for exponential families,

$$\Pi_t\left(\{\boldsymbol{\theta} : a_\star(\boldsymbol{\theta}) \neq a_\star\}\right) \lesssim C\exp\left(-t/T^\star_\beta(\boldsymbol{\mu})\right) \quad \text{a.s.}$$

where

$$T^\star_\beta(\boldsymbol{\nu})^{-1} = \sup_{\substack{\boldsymbol{w}\in\triangle_K \\ w_{a_\star}=\beta}} \min_{a\neq a^\star} \inf_{\lambda\in(\mu_a,\mu_{a_\star})} \left[w_{a_\star} d(\mu_{a_\star},\lambda) + w_a d(\mu_a,\lambda)\right].$$

Links with our (frequentist) characteristic time $T^\star(\boldsymbol{\mu})$:

- $T^\star(\boldsymbol{\mu}) = \min_\beta T^\star_\beta(\boldsymbol{\mu})$
- $T^\star(\boldsymbol{\mu}) \leq T^\star_{1/2}(\boldsymbol{\mu}) \leq 2T^\star(\boldsymbol{\mu})$ (hence $\beta = 1/2$ is never too bad)

Sample complexity of TTTS

For Gaussian bandits, we first analyzed TTTS with the posterior

$$\pi_a(t) = \mathcal{N}\left(\hat{\mu}_a(t), \frac{\sigma^2}{N_a(t)}\right)$$

coupled with the GLR stopping rule

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

Theorem [Shang et al., 2020]

TTTS(β) is δ-correct and

$$\forall \boldsymbol{\mu}, \quad \lim_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \leq T_\beta^\star(\boldsymbol{\mu})$$

The Top Two structure

Top Two algorithm

Given a parameter $\beta \in (0, 1)$, in round t :

- define a leader $B_t \in [K]$
- define a challenger $C_t \neq B_t$
- select arm $A_t \in \{B_t, C_t\}$ at random :

$$\mathbb{P}(A_t = B_t) = \beta \quad \mathbb{P}(A_t = C_t) = 1 - \beta$$

In Top Two Thompson Sampling,

- TS leader : $B_t^{\mathrm{TS}} = a_\star(\boldsymbol{\theta})$ with $\boldsymbol{\theta} \sim \Pi_{t-1}$
- Re-Sampling (RS) challenger : $C_t^{\mathrm{RS}} = a_\star(\boldsymbol{\theta}')$ where

$$\boldsymbol{\theta}' \sim \Pi_{t-1} | \left(a_\star(\boldsymbol{\theta}') \neq B_t\right)$$

The Top Two structure

Top Two algorithm

Given a parameter $\beta \in (0, 1)$, in round t :

- define a leader $B_t \in [K]$
- define a challenger $C_t \neq B_t$
- select arm $A_t \in \{B_t, C_t\}$ at random :

$$\mathbb{P}(A_t = B_t) = \beta \quad \mathbb{P}(A_t = C_t) = 1 - \beta$$

In Top Two Thompson Sampling,

- TS leader : $B_t^{\mathrm{TS}} = a_\star(\boldsymbol{\theta})$ with $\boldsymbol{\theta} \sim \Pi_{t-1}$
- Re-Sampling (RS) challenger : $C_t^{\mathrm{RS}} = a_\star(\boldsymbol{\theta}')$ where

$$\boldsymbol{\theta}' \sim \Pi_{t-1} | \left(a_\star(\boldsymbol{\theta}') \neq B_t\right)$$

➜ re-sampling can be costly. Do we even need a posterior ?

Approximating Re-Sampling

$$\mathbb{P}\left(C_t^{\mathrm{RS}} = a | B_t = b\right) = \frac{p_{t,a}}{\sum_{i \neq b} p_{t,i}}$$

where $p_{t,a} = \Pi_t\left(\theta_a = \max_j \theta_j\right)$. For Gaussian bandits

$$p_{t,a} \simeq \Pi_t\left(\theta_a > \theta_b\right) \simeq \exp\left(-t \frac{\left(\hat{\mu}_b(t) - \hat{\mu}_a(t)\right)^2}{2\sigma^2\left(\frac{1}{N_b(t)} + \frac{1}{N_a(t)}\right)}\right)$$

when $\hat{\mu}_b(t) \geq \hat{\mu}_a(t)$.

The Transportation Cost Challenger [Shang et al., 2020]

Idea : select the mode from this distribution instead of sampling !

$$C_t^{\mathrm{TC}} = \underset{a \neq B_t}{\arg\min} \frac{\left(\hat{\mu}_{B_t}(t) - \hat{\mu}_a(t)\right)^2}{2\sigma^2\left(\frac{1}{N_{B_t}(t)} + \frac{1}{N_a(t)}\right)} \mathbb{1}\left(\hat{\mu}_{B_t}(t) \geq \hat{\mu}_a(t)\right)$$

Another (non Bayesian) interpretation

Recall that TTTS was analyzed with

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

➜ another interpretation : C_t^{TC} minimizes the Empirical Transportation Cost (TC) featured in the stopping rule

Another (non Bayesian) interpretation

Recall that TTTS was analyzed with

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

- ➜ another interpretation : C_t^{TC} minimizes the Empirical Transportation Cost (TC) featured in the stopping rule
- ➜ could we use $B_T^{\mathrm{EB}} = \hat{a}_t^\star$, i.e. Empirical Best leader ?

Asymptotically... yes!

Theorem

Combining the GLR stopping rule with a Top Two sampling rule with any pair of leader/challenger satisfying some properties yields a δ-correct algorithm satisfying for all $\boldsymbol{\nu} \in \mathcal{D}^K$ with distincts means

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

Distributions	TS	EB	RS	TC	TCI
Gaussian KV	✓	✓	✓	✓	✓
Bernoulli	✓	✓	✓	✓	✓
Exponential families	?	✓	?	✓	✓
Gaussian UV	?	✓	?	✓	✓
Bounded	✓	✓	✓	✓	✓

[Jourdan et al., 2022, Jourdan et al., 2023a]

But exploration is nice in practice

TS-TC

$$B_t \sim \operatorname*{arg\,max}_{a\in[K]} \widetilde{\theta}_a(t) \quad \widetilde{\boldsymbol{\theta}}(t) \sim \Pi_t$$

$$C_t = \operatorname*{arg\,min}_{a\neq B_t} \frac{(\hat{\mu}_{B_t}(t) - \hat{\mu}_a(t))_+^2}{2\sigma^2\left(\frac{1}{N_{B_t}(t)} + \frac{1}{N_a(t)}\right)}$$

EB-TCI

$$B_t = \operatorname*{arg\,max}_{a\in[K]} \hat{\mu}_a(t)$$

$$C_t = \operatorname*{arg\,min}_{a\neq B_t} \left[\frac{(\hat{\mu}_{B_t}(t) - \hat{\mu}_a(t))_+^2}{2\sigma^2\left(\frac{1}{N_{B_t}(t)} + \frac{1}{N_a(t)}\right)} + \log N_a(t)\right]$$

Numerical experiments

Error parameter $\delta = 0.1$. Top Two algorithms with $\beta = 1/2$.

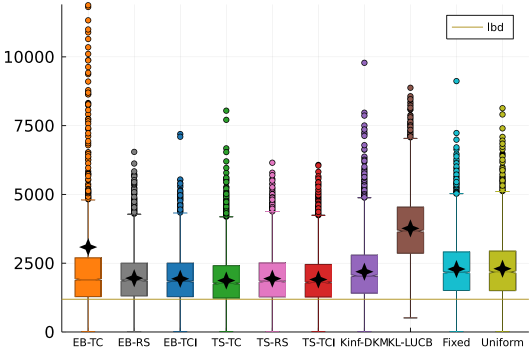
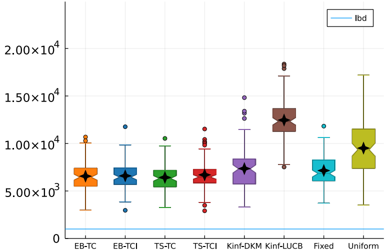


FIGURE – Empirical sample complexity averaged over 5000 random (Bernoulli) instances with $K = 8$ and $\Delta_{\min} \geq 0.01$.

Numerical experiments

arm = planting date / observation = yield

Error parameter $\delta = 0.01$. Top Two algorithms with $\beta = 1/2$.



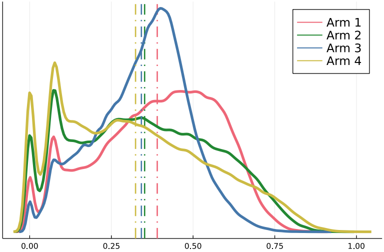


FIGURE – Empirical stopping time (a) on scaled DSSAT instances with their density and mean (b).

Top Two algorithms beyond Fixed Confidence

EB-TC$_{\varepsilon_0}$

$$B_t = \arg\max_{a \in [K]} \hat{\mu}_a(t)$$

$$C_t = \arg\min_{a \neq B_t} \left[\frac{\hat{\mu}_{B_t}(t) - \hat{\mu}_a(t) + \varepsilon_0}{\sqrt{\frac{1}{N_{B_t}(t)} + \frac{1}{N_a(t)}}} \right]$$

[Jourdan et al., 2023b]

- motivated by the lower bound for (ε_0, δ)-PAC identification
- can be used for (ε, δ)-PAC identification[1] for $\varepsilon \neq \epsilon_0$
- first guarantees in the anytime setting...

[1] $\mathbb{P}\left(\mu_{\hat{a}_\tau} > \mu_\star - \varepsilon\right) \geq 1 - \delta$

Top Two algorithms beyond Fixed Confidence

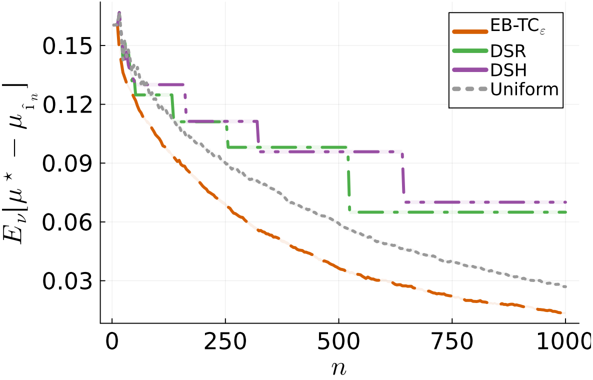


FIGURE – Simple regret as a function of time on an instance $\mu \in \{0.4, 0.6\}^{10}$ with 2 best arms

(... but the theory is just saying that the algorithm is not too much worse than uniform sampling...)

References

- Cyrille Koné, Emilie Kaufmann, Laura Richert
  Adaptive Algorithms for Relaxed Pareto Set Identification
  NeurIPS 2023
- Cyrille Koné, Marc Jourdan, Emilie Kaufmann
  Pareto Set Identification with Posterior Sampling
  AISTATS 2025



Motivation : Clinical Trials

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

For the t-th patient in a clinical trial,

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

Goal : maximize the expected number of patients healed (regret) or identifiy the best treatment $a = \arg\max_a \mu_a$ (best arm identification)

Motivation : Clinical Trials





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

For the t-th patient in a clinical trial,

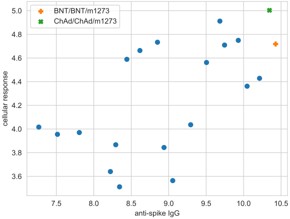
- choose a treatment A_t
- observe a response $X_t \in \{0, 1\} : \mathbb{P}(X_t = 1 | A_t = a) = \mu_a$

➜ an (idealized) model for Phase III trials, but bandits could also be useful for early stage clinical trials in which *several* indicators of safety and biological efficacy are jointly monitored

Early stage clinical trials in vaccinology

Several indicators of immunogenicity are typically measured :

- binding antibodies
- neutralising antibodies for different variants
- cellular responses (T-cells ...)



$K = 20$ combinations of Covid vaccines (COVBOOST)

Sampling an arm = giving a vaccine to a patient and measuring (15 days later) all the D indicators of interest ($X_t \in \mathbb{R}^D$)

Multi-objective bandits

Given K multi-dimensional distributions with means $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K \in \mathbb{R}^D$, what are "good arm(s)" ?

- Given a preference function $g : \mathbb{R}^D \to \mathbb{R}$, a maximizer of $g(\boldsymbol{\mu}_a)$ (such a function is in general hard to define)
- An arm maximizing one of the objectives under some (linear) constraints on the others [Katz-Samuels and Scott, 2019]
- All the arms that are not uniformly worse than the others

➜ the **Pareto set**[Auer et al., 2016]

Pareto Set

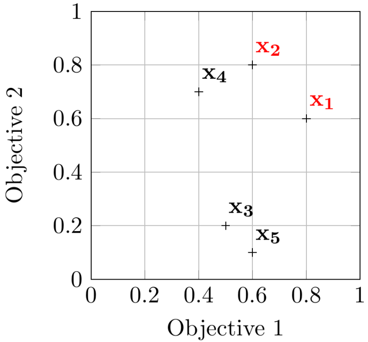
Let $\mathcal{X} \subset \mathbb{R}^D$ a set of vectors. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$.

- $\boldsymbol{x}$ 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

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

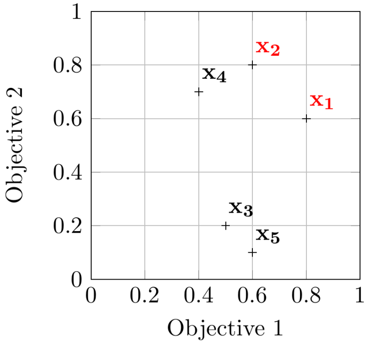


1. $\boldsymbol{x}_3 \prec \boldsymbol{x}_1$
2. $\boldsymbol{x}_4 \prec \boldsymbol{x}_2$
3. $\boldsymbol{x}_5 \prec \boldsymbol{x}_1$

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

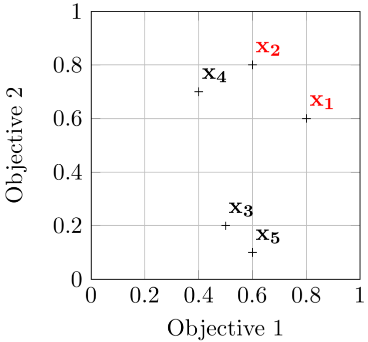


1. $\boldsymbol{x}_3 \prec \boldsymbol{x}_1$
2. $\boldsymbol{x}_4 \prec \boldsymbol{x}_2$
3. $\boldsymbol{x}_5 \prec \boldsymbol{x}_1$
4. $\boldsymbol{x}_1 \not\prec \boldsymbol{x}_2$
5. $\boldsymbol{x}_2 \not\prec \boldsymbol{x}_1$

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

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K) \in (\mathbb{R}^D)^K$$

An algorithm is made of :

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

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

- a recommendation rule $\hat{S}_t$: a guess for the Pareto Set
- a stopping rule τ : when to stop the data collection ?

Definition

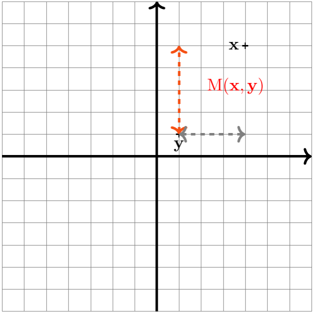
An algorithm is δ-correct if, for all $\boldsymbol{\mu} \in \mathcal{M}$, $\mathbb{P}_{\boldsymbol{\mu}}(\hat{S}_\tau \neq \mathcal{P}^\star(\boldsymbol{\mu})) \leq \delta$.

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

Adaptive Pareto Exploration

First ingredient : a non-dominance measure

$$
\begin{aligned}
\boldsymbol{x} \not\prec \boldsymbol{y} \quad & \Leftrightarrow \quad \exists\, d,\ x^d \geq y^d, \\
& \Leftrightarrow \quad \exists\, d,\ x^d - y^d \geq 0, \\
& \Leftrightarrow \quad \underbrace{\max_{d \in [D]} (x^d - y^d)}_{:=\mathrm{M}(\boldsymbol{x},\boldsymbol{y})} > 0,
\end{aligned}
$$

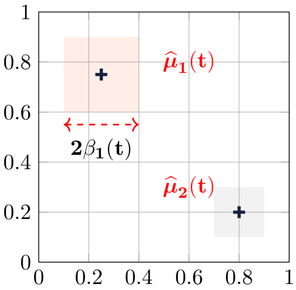


- The larger $\mathrm{M}(\boldsymbol{x}, \boldsymbol{y})$ the "further" $\boldsymbol{y}$ is from dominating $\boldsymbol{x}$

Adaptive Pareto Exploration

Second ingredient : confidence regions

- $\hat{\boldsymbol{\mu}}_k(t) \in \mathbb{R}^D$ the empirical mean vector of arm k at time t



Confidence bonus for arm k :

$$\beta_k(t) \simeq \sqrt{\frac{\log(N_k(t)/\delta)}{N_k(t)}}$$

such that, w.p. larger than $1-\delta$, all means $\boldsymbol{\mu}_k$ belong to the highlighted regions, for all t

Letting $M(i,j) = M(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$ and $M(i,j;t) = M(\hat{\boldsymbol{\mu}}_i(t), \hat{\boldsymbol{\mu}}_j(t))$

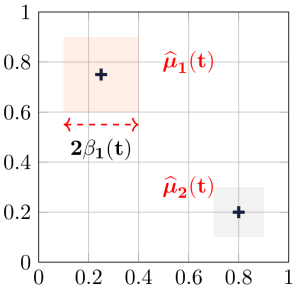
$$M(i,j) \leq M^+(i,j;t) := M(i,j;t) + \beta_i(t) + \beta_j(t)$$

with high probability.

Adaptive Pareto Exploration

Second ingredient : confidence regions

- $\hat{\boldsymbol{\mu}}_k(t) \in \mathbb{R}^D$ the empirical mean vector of arm k at time t



Confidence bonus for arm k :

$$\beta_k(t) \simeq \sqrt{\frac{\log(N_k(t)/\delta)}{N_k(t)}}$$

such that, w.p. larger than $1-\delta$, all means $\boldsymbol{\mu}_k$ belong to the highlighted regions, for all t

Letting $M(i,j) = M(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$ and $M(i,j;t) = M(\hat{\boldsymbol{\mu}}_i(t), \hat{\boldsymbol{\mu}}_j(t))$

$$M^-(i,j;t) := M(i,j;t) - \beta_i(t) - \beta_j(t) \leq M(i,j)$$

with high probability.

Adaptive Pareto Exploration

$$\mathrm{OPT}(t) := \{i \in [K] : \forall j \in [K] \backslash \{i\}, \mathsf{M}^-(i, j; t) > 0\}$$

We define

- a potentially Pareto optimal arm

$$b_t = \operatorname*{arg\,max}_{i \in [K] \backslash \mathrm{OPT}(t)} \min_{j \neq i} \mathsf{M}^+(i, j; t)$$

- the arm that is the closest to potentially dominate it

$$c_t := \operatorname*{arg\,min}_{j \neq b_t} \mathsf{M}^-(b_t, j; t)$$

Adaptive Pareto Exploration (APE)

selects the least sampled among these two candidate arms :
$A_{t+1} = \arg\min_{a \in \{b_t, c_t\}} N_a(t)$

Stopping rule

Letting $\hat{S}(t) = \mathcal{P}^\star(\hat{\boldsymbol{\mu}}_1(t), \dots, \hat{\boldsymbol{\mu}}_K(t))$, the algorithm stops and recommends $\hat{S}_t = \hat{S}(t)$ when

- all arms in $\hat{S}(t)$ are confidently non-dominated :

$$Z_1(t) := \min_{i \in \hat{S}(t)} \min_{j \neq i} M^-(i,j;t) > 0$$

- all arms in $(\hat{S}(t))^c$ are confidently dominated :

$$Z_2(t) := \min_{i \notin \hat{S}(t)} \max_{j \neq i} \left[-M^+(i,j;t)\right] > 0$$

Stopping rule for (exact) PSI

$$\tau = \inf\left\{t \in \mathbb{N} : Z_1(t) > 0, Z_2(t) > 0\right\}$$

Stopping rule

Letting $\hat{S}(t) = \mathcal{P}^\star(\hat{\boldsymbol{\mu}}_1(t), \dots, \hat{\boldsymbol{\mu}}_K(t))$, the algorithm stops and recommends $\hat{S}_t = \hat{S}(t)$ when

- all arms in $\hat{S}(t)$ are confidently non-dominated :

$$Z_1{}^\delta(t) := \min_{i \in \hat{S}(t)} \min_{j \neq i} M_\delta{}^-(i, j; t) > 0$$

- all arms in $(\hat{S}(t))^c$ are confidently dominated :

$$Z_2{}^\delta(t) := \min_{i \notin \hat{S}(t)} \max_{j \neq i} \left[-M_\delta{}^+(i, j; t)\right] > 0$$

Stopping rule for (exact) PSI

$$\tau_\delta = \inf\left\{t \in \mathbb{N} : Z_1{}^\delta(t) > 0, Z_2{}^\delta(t) > 0\right\}$$

Results : Theory

Theorem [Kone et al., 2023]

Assume the observations are bounded in $[0,1]^D$. Then, with probability larger than $1-\delta$, APE with the stopping rule τ_δ outputs $\hat{S}_\tau = \mathcal{P}^\star(\boldsymbol{\mu})$ using at most

$$\sum_{a=1}^{K} \frac{32}{\tilde{\Delta}_a^2} \log\left(\frac{2KD}{\delta} \log\left(\frac{32}{\tilde{\Delta}_a^2}\right)\right),$$

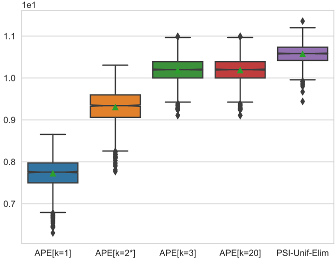
samples, where $\tilde{\Delta}_a$ is an appropriate notion of gap [Auer et al., 2016]

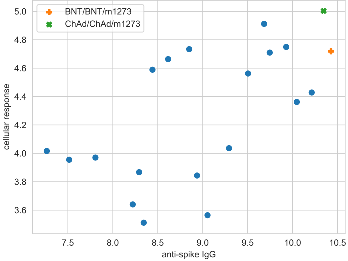
APE can further be combined with different stopping rules to tackle different relaxations of PSI, e.g. $\min(\tau, \tau^k)$ where

$$\tau^k = \inf\{t \in \mathbb{N} : |\mathrm{OPT}(t)| \geq k\}$$

to identify at most k Pareto optimal arms.

Results : Practice





(Log) Empirical sample complexity of APE (with a k-relaxation) compared to the algorithm of [Auer et al., 2016] on simulated CovBoost data [Munro et al., 2021]

- improved practical performance
- the k-relaxation (provably) reduces the sample complexity

Optimality ?

For arms that are multi-variate Gaussian, could we further try to match the lower bound ?

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

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

➜ The structure of the alternative is very complex for the PSI problem, making even the computation of "minimal distance" (needed for the stopping rule) challenging...

[Crepon et al., 2024]

Optimality ?

For arms that are multi-variate Gaussian, could we further try to match the lower bound ?

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

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

➜ The structure of the alternative is very complex for the PSI problem, making even the computation of "minimal distance" (needed for the stopping rule) challenging...

[Crepon et al., 2024]

... but we don't need to compute it !

A Fully Sampling-Based Approach

Posterior Sampling for PSI (PSIPS) [Kone et al., 2025]

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

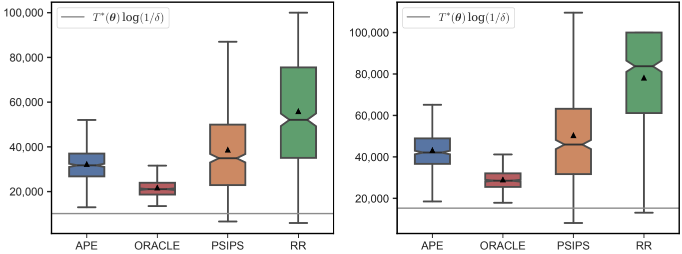
$$\widetilde{\boldsymbol{\theta}}_a^m \sim \mathcal{N}\left(\hat{\boldsymbol{\mu}}_a(t), \frac{c(t,\delta)}{N_a(t)}\Sigma\right)$$

- If for all m, $\mathcal{P}^\star(\widetilde{\boldsymbol{\theta}}^m) = \mathcal{P}^\star(\hat{\boldsymbol{\mu}}(t))$, stop and return $\hat{S}_t = \mathcal{P}^\star(\hat{\boldsymbol{\mu}}(t))$
- Else, take the first m such that $\mathcal{P}^\star(\widetilde{\boldsymbol{\theta}}^m) \neq \mathcal{P}^\star(\hat{\boldsymbol{\mu}}(t))$
  Update an online learning algorithm on Δ_K with the gain $g_t(w) = \sum_{a=1}^K w_a \frac{1}{2} \|\hat{\boldsymbol{\mu}}_a(t) - \widetilde{\boldsymbol{\theta}}_a^m\|^2_{\Sigma^{-1}}$ to get w_t
  Select arm $A_t \sim w_t$

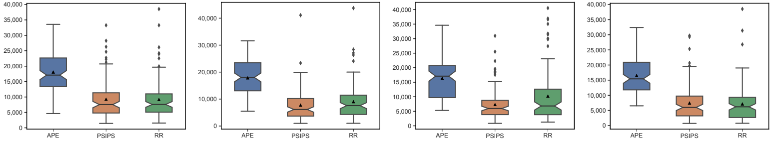
➜ For $c(t,\delta) \simeq \frac{\log(\log(t)/\delta)}{\log(1/\delta)}$ and $\mathrm{M}(t,\delta) \simeq \frac{\log(t/\delta)}{\delta}$, PSIPS satisfies $\limsup_{\delta\to 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq T_\star(\boldsymbol{\mu})$ when arms are Σ-subGaussian

Experiments

- CovBoost dataset ($d = 3$) for $\delta = 0.1$ (left) and $\delta = 0.01$ (right)



- Random Gaussian instances with $K = 10$ for $d \in \{3, 4, 5, 6\}$



Conclusion

The "follow the lower bound" approach made of

- a Tracking sampling rule
- the Generalized Likelihood Ratio (GLR) stopping rule

can reach the minimal sample complexity in a regime of small error δ, for quite general pure exploration tasks.

For some particular tasks (e.g. Best Arm Identification) :

- "Top-Two" sampling rules are easier to implement, perform well for moderate values of δ and are near-optimal for $\delta \to 0$
- we analyzed a sampling-based stopping rule as an interesting alternative to the GLR (e.g. for PSI)

Perspectives

- A better understanding of the moderate confidence regime : is there a price for asymptotic optimality ?
- Investigate the trade-off between optimality and some algorithmic constraints (privacy, fairness)
- Can bandits help for adaptive clinical trials, for real ?

Beyond bandits ?

- We made some progress on the characterization of the complexity of near-optimal policy identification in a (finite) Markov Decision Process
  [Al Marjani et al., 2023, Tuynman et al., 2024]
- But the gap between theory (tabular MDPs) and practice (deep reinforcement learning) is huge...

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

Agrawal, S. and Goyal, N. (2017).
Near-optimal regret bounds for thompson sampling.
J. ACM, 64(5) :30 :1–30 :24.

Al Marjani, A., Tirinzoni, A., and Kaufmann, E. (2023).
Active coverage for PAC reinforcement learning.
In *Proceedings of the 36th Conference On Learning Theory (COLT)*.

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*.

Degenne, R., Koolen, W. M., and Ménard, P. (2019).
Non-asymptotic pure exploration by solving games.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

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.

Jourdan, M., Degenne, R., Baudry, D., de Heide, R., and Kaufmann, E. (2022).
Top two algorithms revisited.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jourdan, M., Degenne, R., and Kaufmann, E. (2023a).
Dealing with unknown variances in best-arm identification.
In *Algorithmic Learning Theory (ALT)*.

Jourdan, M., Degenne, R., and Kaufmann, E. (2023b).
An ε-best-arm identification algorithm for fixed confidence and beyond.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Katz-Samuels, J. and Scott, C. (2019).

Top feasible arm identification.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

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)*.

Kaufmann, E. and Koolen, W. M. (2017).
Monte-Carlo tree search by best arm identification.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

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

Kone, C., Kaufmann, E., and Richert, L. (2023).
Adaptive algorithms for relaxed pareto set identification.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Korda, N., Kaufmann, E., and Munos, R. (2013).
Thompson Sampling for 1-dimensional Exponential family bandits.

In *Advances in Neural Information Processing Systems*.

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.

Munro, A.-P.-S., Janani, L., Cornelius, V., and et al. (2021).
Safety and immunogenicity of seven COVID-19 vaccines as a third dose (booster) following two doses of ChAdOx1 nCov-19 or BNT162b2 in the UK (COV-BOOST) : a blinded, multicentre, randomised, controlled, phase 2 trial.
The Lancet, 398(10318) :2258–2276.

Riou, C. and Honda, J. (2020).
Bandit algorithms based on thompson sampling for bounded reward distributions.
In *Algorithmic Learning Theory (ALT)*.

Robbins, H. (1952).
Some aspects of the sequential design of experiments.
Bulletin of the American Mathematical Society, 58(5) :527–535.

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

Russo, D., Roy, B. V., Kazerouni, A., Osband, I., and Wen, Z. (2018).
A tutorial on thompson sampling.
Foundations and Trends in Machine Learning, 11(1) :1–96.

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)*.

Teraoka, K., Hatano, K., and Takimoto, E. (2014).
Efficient sampling method for monte carlo tree search problem.
IEICE Transactions on Infomation and Systems, pages 392–398.

Thompson, W. (1933).
On the likelihood that one unknown probability exceeds another in view of the evidence of two samples.
Biometrika, 25 :285–294.

Tuynman, A., Degenne, R., and Kaufmann, E. (2024).
Finding good policies in average-reward markov decision processes without prior knowledge.
In *Advances in Neural Information Processing Systems (NeurIPS)*.