Multi-Objective Bandits Revisited

Emilie Kaufmann

based on collaborations with

Cyrille Koné, Laura Richert and Marc Jourdan

Workshop on Regret, Optimization and Games, 2025

Bandits for adaptive clinical trials?





$\mathcal{B}(p_1)$ $\mathcal{B}(p_2)$ $\mathcal{B}(p_3)$ $\mathcal{B}(p_4)$ $\mathcal{B}(p_5)$

For the t-th patient in a clinical trial,

- choose a treatment (arm) A_t
- observe its efficacy (reward/response)
  $X_t \in \{0, 1\} : \mathbb{P}(X_t = 1 | A_t = a) = p_a$

Adaptive treatment allocation / sampling rule:
A_t can be chosen based on past outcomes $A_1, X_1, \ldots, A_{t-1}, X_{t-1}$

Bandits for adaptive clinical trials?

$\mathcal{B}(p_1)$ $\mathcal{B}(p_2)$ $\mathcal{B}(p_3)$ $\mathcal{B}(p_4)$ $\mathcal{B}(p_5)$

For the t-th patient in a clinical trial,

- choose a treatment (arm) A_t
- observe its efficacy (reward/response)
  $X_t \in \{0,1\} : \mathbb{P}(X_t = 1 | A_t = a) = p_a$

Adaptive treatment allocation / sampling rule:
A_t can be chosen based on past outcomes $A_1, X_1, \dots, A_{t-1}, X_{t-1}$

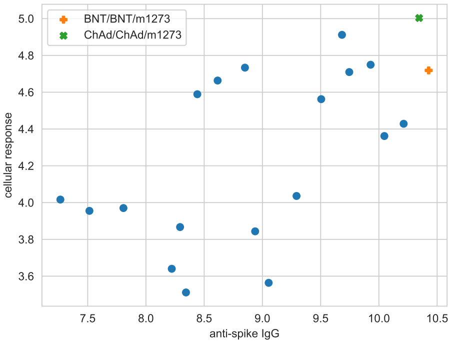
➜ an idealized model for a *Phase III* (confirmatory) trial

Specificities of early stage (*Phase I/II*) trials

Multiple responses are typically measured:

- side effects (toxicity)
- different indicators of biological efficacy (blood tests)

Vaccine design: different indicators of the immune response:



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

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

Outline

1. Pure Exploration in Multi-objective bandits
2. Best Arm Identification ($D = 1$)
3. Adaptive Pareto Exploration
4. Towards Optimal Algorithms

Outline

Multi-objective bandit

Bandit model

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

Bandit (Pure Exploration) Algorithm

- (*sampling rule*) how is A_t selected based on past observation?
- (*recommendation rule*) guess $\widehat{S}_t$ for a "good set of arms"
- (*stopping rule*) decide whether to stop collecting observations

➜ Goal: make a confident guess with few samples

What is a good set of arms?

$$\mathcal{S}^\star = \mathcal{S}^\star(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K) \subseteq [K]$$

- $k_\star = \arg\max_k g(\boldsymbol{\mu}_k)$ for some preference function $g : \mathbb{R}^D \to \mathbb{R}$, e.g. $g(\boldsymbol{\mu}_k) = \sum_{d=1}^D w_d \mu_k^d$
- Feasible Set: all arms that satisfy some linear constraints [Katz-Samuels and Scott, 2018]
- Top Feasible Arm: a feasible arm maximizing one of the objectives [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

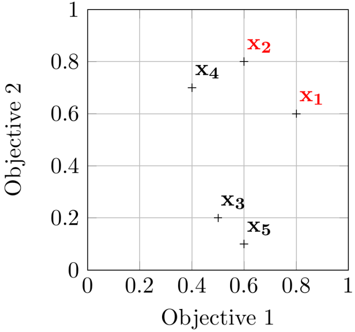
Let $\mathcal{X} \subset \mathbb{R}^D$ a set of vectors. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$.

- $\boldsymbol{x}$ 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

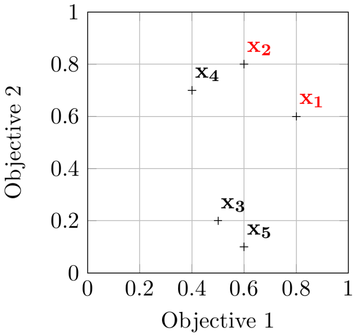


1. $\boldsymbol{x}_3 \prec \boldsymbol{x}_1$
2. $\boldsymbol{x}_4 \prec \boldsymbol{x}_2$
3. $\boldsymbol{x}_5 \prec \boldsymbol{x}_1$

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

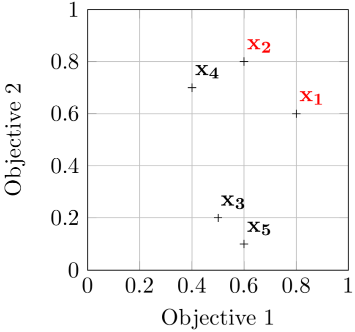


1. $\boldsymbol{x}_3 \prec \boldsymbol{x}_1$
2. $\boldsymbol{x}_4 \prec \boldsymbol{x}_2$
3. $\boldsymbol{x}_5 \prec \boldsymbol{x}_1$
4. $\boldsymbol{x}_1 \not\prec \boldsymbol{x}_2$
5. $\boldsymbol{x}_2 \not\prec \boldsymbol{x}_1$

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

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

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

Definition

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

Outline

Best Arm Identification with Fixed Confidence

$$\begin{aligned} \boldsymbol{\mu} &= (\mu_1, \ldots, \mu_K) \in \mathbb{R}^K \\ i_\star(\boldsymbol{\mu}) &= \underset{k \in [K]}{\arg\max}\ \mu_k \end{aligned}$$

Best Arm Identification algorithm:

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

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

- a recommendation rule $\hat{\imath}_t$: a guess for $i_\star(\boldsymbol{\mu})$
- a stopping rule τ: when to stop the data collection?

Definition

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

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

3 approaches to Best Arm Identification

- Uniform sampling + Eliminations
  Successive Eliminations [Even-Dar et al., 2006]
- Adaptive sampling based on Confidence Intervals
  LUCB [Kalyanakrishnan et al., 2012], UGapE [Gabillon et al., 2012] ...
- Lower Bound Inspired Algorithms
  e.g., [Garivier and Kaufmann, 2016, Degenne et al., 2019, Jourdan et al., 2022]

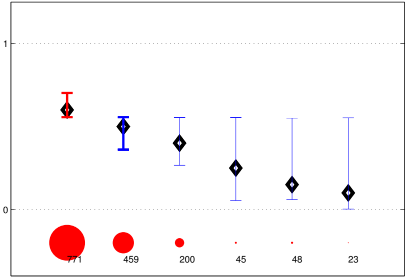
All algorithms rely on

$$N_k(t) := \sum_{s=1}^{t} \mathbb{1}(A_t = k), \quad \hat{\mu}_k(t) := \frac{1}{N_k(t)} \sum_{s=1}^{t} Y_{k,s}$$

where $(Y_{k,s})$ are the successive observations from arm k

LUCB: Lower and Upper Confidence Bounds

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



- At round t, draw

$$B_t = \underset{b \in [K]}{\arg\max}\ \hat{\mu}_b(t)$$

$$C_t = \underset{c \neq B_t}{\arg\max}\ \mathrm{UCB}_c(t)$$

- Stop at round t if

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

Theorem [Kalyanakrishnan et al., 2012]

For well-chosen confidence intervals, $\mathbb{P}_\nu(B_\tau = i_\star(\boldsymbol{\mu})) \geq 1 - \delta$ and

$$\mathbb{E}[\tau_\delta] = \mathcal{O}\left(\left[\sum_{a=1}^{K} \frac{1}{\Delta_a{}^2}\right] \ln\left(\frac{1}{\delta}\right)\right) \qquad \Delta_k = \begin{cases} \mu_\star - \mu_k, & k \neq i_\star \\ \min_{i \neq i_\star} \Delta_i, & k = i_\star \end{cases}$$

A Sample Complexity Lower Bound

Lower Bound [Garivier and Kaufmann, 2016]

For δ-correct algorithms for Gaussian bandits of variance σ^2,

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

where

$$(T_\star(\boldsymbol{\mu}))^{-1} = \sup_{w \in \Delta_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(i_\star(\mu))} \sum_{a \in [K]} w_a \frac{(\mu_a - \lambda_a)^2}{2\sigma^2}$$

with

$$\begin{aligned} \Delta_K &= \{\boldsymbol{w} \in [0,1]^K : \sum_a w_a = 1\} \\ \mathrm{Alt}(i) &= \{\boldsymbol{\lambda} \in \mathbb{R}^K : i_\star(\boldsymbol{\lambda}) \neq i\}. \end{aligned}$$

A Sample Complexity Lower Bound

The "minimal distance" has a closed form:

$$\inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(i_\star(\mu))} \sum_{a \in [K]} w_a \frac{(\mu_a - \lambda_a)^2}{2\sigma^2} = \min_{a \neq i_\star} \frac{(\mu_a - \mu_{i_\star})^2}{2\sigma^2 \left(\frac{1}{w_a} + \frac{1}{w_{i_\star}}\right)}$$

but not the characteristic time

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

Approximation of the characteristic time

$$\sum_{a=1}^{K} \frac{2\sigma^2}{\Delta_a^2} \leq T_\star(\boldsymbol{\mu}) \leq 2 \left(\sum_{a=1}^{K} \frac{2\sigma^2}{\Delta_a^2} \right)$$

➜ Can we still match this (non-explicit) lower bound?

Track-and-Stop

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

Yes, with an appropriate stopping rule

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

where $\hat{\imath}_t^\star$ is the empirical best arm at time t

Track-and-Stop

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

Yes, with an appropriate GLR stopping rule

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

where $\hat{\imath}_t^\star$ is the empirical best arm at time t

➜ Generalized Likelihood Ratio Statistic for testing

$$\mathcal{H}_0 : (i_\star(\boldsymbol{\mu}) \neq \hat{\imath}_t) \quad \text{against} \quad \mathcal{H}_1 : (i_\star(\boldsymbol{\mu}) = \hat{\imath}_t)$$

Track-and-Stop

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

Yes, with an appropriate GLR stopping rule

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

where $\hat{\imath}_t^\star$ is the empirical best arm at time t
... and a sampling rule satisfying

$$\left(\frac{N_1(t)}{t}, \dots, \frac{N_K(t)}{t}\right) \to w^\star(\boldsymbol{\mu})$$

where $w^\star(\boldsymbol{\mu})$ is the maximizer in $w \in \Delta_K$

Track-and-Stop

Tracking sampling rule: letting $U_t = \left\{a : N_a(t) < \sqrt{t}\right\}$,

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

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

The Track-and-Stop strategy, that uses

- the Tracking sampling rule
- the GLR stopping rule with $\beta(t, \delta) \simeq \log\left(\frac{K \log(t)}{\delta}\right)$
- and recommends $\hat{\imath}_t = i_\star(\hat{\boldsymbol{\mu}}(t))$

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

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

Back to Pareto Set Identification

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

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

Definition

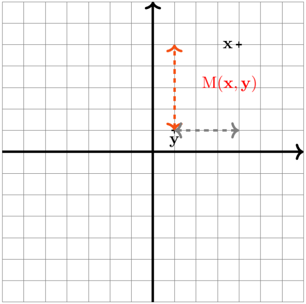
An algorithm is δ-correct if, for all $\boldsymbol{\nu}$, $\mathbb{P}_{\boldsymbol{\nu}}(\widehat{S}_\tau \neq \mathcal{S}^\star(\boldsymbol{\mu})) \leq \delta$.

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

Outline

A non-dominance measure

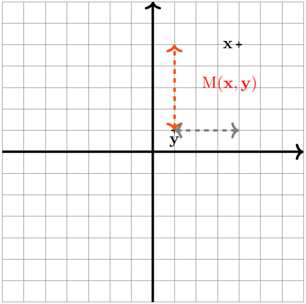
$$\begin{aligned}
\boldsymbol{x} \nprec \boldsymbol{y} \quad &\Leftrightarrow \quad \exists\, d,\ x^d \geq y^d, \\
&\Leftrightarrow \quad \exists\, d,\ x^d - y^d \geq 0, \\
&\Leftrightarrow \quad \underbrace{\max_{d \in [D]} \left(x^d - y^d\right)}_{:=\mathrm{M}(\boldsymbol{x},\boldsymbol{y})} > 0,
\end{aligned}$$



Interpretation: The larger $\mathrm{M}(\boldsymbol{x}, \boldsymbol{y})$ the “further” $\boldsymbol{y}$ is from dominating $\boldsymbol{x}$

A non-dominance measure

$$
\begin{aligned}
\boldsymbol{x} \nprec \boldsymbol{y} \quad &\Leftrightarrow \quad \exists\, d,\ x^d \geq y^d, \\
&\Leftrightarrow \quad \exists\, d,\ x^d - y^d \geq 0, \\
&\Leftrightarrow \quad \underbrace{\max_{d \in [D]} \left(x^d - y^d\right)}_{:=\mathrm{M}(\boldsymbol{x},\boldsymbol{y})} > 0,
\end{aligned}
$$



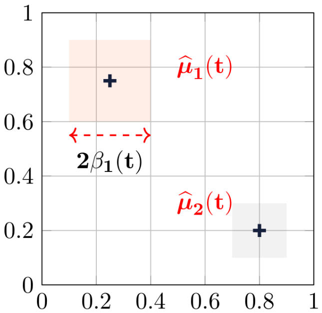
Interpretation: The larger $\mathrm{M}(\boldsymbol{x}, \boldsymbol{y})$ the "further" $\boldsymbol{y}$ is from dominating $\boldsymbol{x}$

$$M(i, j) := M(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$$

Confidence Regions on $M(i,j)$

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

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



Confidence bonus for $\boldsymbol{\mu}_k$

$$\beta_k(t) \simeq \sqrt{2\sigma^2 \log\left(\frac{K \log(N_k(t))}{\delta}\right) \frac{1}{N_k(t)}}$$

and for $\boldsymbol{\mu}_i - \boldsymbol{\mu}_j$

$$\beta_{i,j}(t) \simeq \sqrt{2\sigma^2 \log\left(\frac{K^2 \log(N_k(t))}{\delta}\right) \left(\frac{1}{N_i(t)} + \frac{1}{N_j(t)}\right)}$$

Lemma

With probability $1-\delta$, for all i, j, t,

$$\begin{aligned} M(i,j) &\geq M^-(i,j;t) := \mathrm{M}(i,j;t) - \beta_{i,j}(t) \\ M(i,j) &\leq M^+(i,j;t) := M(i,j;t) + \beta_{i,j}(t) \end{aligned}$$

Adaptive Pareto Exploration

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

Two interesting arms to explore:

- a potentially Pareto optimal arm

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

- the arm that is the closest to potentially dominate it

$$C_t := \operatorname*{arg\,min}_{j \neq B_t} \mathrm{M}^-(B_t, j; t)$$

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 $\hat{S}(t) = \mathcal{P}^\star(\hat{\boldsymbol{\mu}}_1(t), \ldots, \hat{\boldsymbol{\mu}}_K(t))$, 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

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

Sample complexity bound

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 $\hat{S}_\tau = \mathcal{S}^\star(\boldsymbol{\mu})$ and satisfies

$$\tau_\delta \leq \sum_{a=1}^{K} \frac{32}{{\Delta_a}^2} \log\left(\frac{2KD}{\delta} \log\left(\frac{32}{{\Delta_a}^2}\right)\right),$$

for an appropriate notion of "Pareto gap".

➜ same scaling as the bound of [Auer et al., 2016] for an elimination-based algorithm, with better constants and a $\log\log(1/\Delta)$ versus $\log(1/\Delta)$

APE for relaxed PSI

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.

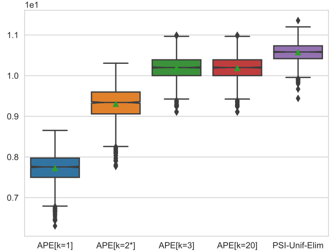
Theorem [Kone et al., 2023]

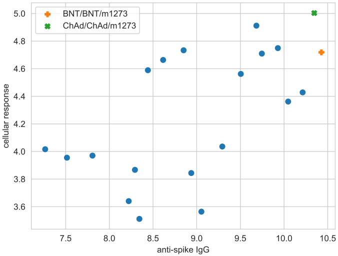
Assume the observations are bounded in $[0,1]^D$. Then, with probability larger than $1-\delta$, APE with the stopping rule $\tau_\delta \wedge \tau^k$ outputs $\hat{S}_\tau = \mathcal{P}^\star(\boldsymbol{\mu})$ and satisfies

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

for a relaxation $\widetilde{\Delta}_a = \max(\Delta_a, \omega_k)$.

Numerical results





(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

Outline

Optimality?

For arms that are multi-variate Gaussian (known covariance Σ), could we further try to match the lower bound?

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

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

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

➜ The structure of the alternative is complex for PSI, making even the computation of "minimal distance" challenging...

Optimality?

For arms that are multi-variate Gaussian (known covariance Σ), could we further try to match the lower bound?

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

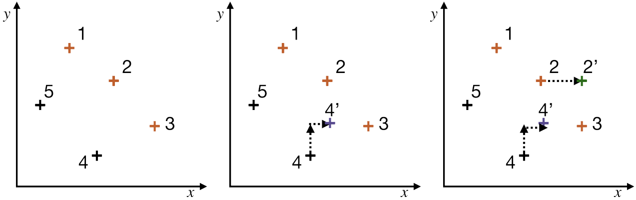
$$T^\star(\boldsymbol{\mu})^{-1} = \sup_{w \in \Delta_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\mathcal{S}^\star(\boldsymbol{\mu}))} \left(\sum_{k=1}^{K} w_k \frac{1}{2} \|\boldsymbol{\mu}_k - \boldsymbol{\lambda}_k\|^2_{\Sigma^{-1}} \right).$$

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

➜ The structure of the alternative is complex for PSI, making even the computation of "minimal distance" challenging...

Computing the Minimal Distance

- there are many ways to alter the Pareto set



- no closed-form is known for the minimal distance

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

- 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

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

is already computationally expansive due to the minimal distance.

The Tracking sampling rule is intractable as it further computes

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

➜ existing alternative approaches based on online learning [Ménard, 2019, Degenne et al., 2019] also rely on minimal distance computation.

A Fully Sampling-Based Approach

Posterior Sampling for PSI (PSIPS) *(simplified)*

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 return $\hat{S}_t = \mathcal{S}^\star(\hat{\boldsymbol{\mu}}(t))$
- Else, take the first m such that $\mathcal{S}^\star(\widetilde{\boldsymbol{\theta}}^m) \neq \mathcal{S}^\star(\hat{\boldsymbol{\mu}}(t))$
  Update an online learning algorithm on Δ_K with the gain

$$g_t(w) = \sum_{a=1}^{K} w_a \frac{1}{2} \|\hat{\boldsymbol{\mu}}_a(t) - \widetilde{\boldsymbol{\theta}}_a^m\|_{\Sigma^{-1}}^2$$

to get w_t. Select arm $A_t \sim (1-\gamma_t)w_t + \gamma_t w_{\exp}$

[Kone et al., 2025], inspired by PEPS [Li et al., 2024]

A Fully Sampling-Based Approach

Posterior Sampling for PSI (PSIPS) (*simplified*)

For all $m \leq M(t,\delta)$, sample $\widetilde{\boldsymbol{\theta}}^m = (\widetilde{\boldsymbol{\theta}}^m_1, \ldots, \widetilde{\boldsymbol{\theta}}^m_K)$ with

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

- If for all m, $\mathcal{S}^\star(\widetilde{\boldsymbol{\theta}}^m) = \mathcal{S}^\star(\hat{\boldsymbol{\mu}}(t))$, **stop** and return $\hat{S}_t = \mathcal{S}^\star(\hat{\boldsymbol{\mu}}(t))$
- Else, take the first m such that $\mathcal{S}^\star(\widetilde{\boldsymbol{\theta}}^m) \neq \mathcal{S}^\star(\hat{\boldsymbol{\mu}}(t))$
  Update an online learning algorithm on Δ_K with the gain

$$g_t(w) = \sum_{a=1}^{K} w_a \frac{1}{2} \|\hat{\boldsymbol{\mu}}_a(t) - \widetilde{\boldsymbol{\theta}}^m_a\|^2_{\Sigma^{-1}}$$

to get w_t. Select arm $A_t \sim (1-\gamma_t)w_t + \gamma_t w_{\exp}$

[Kone et al., 2025], inspired by PEPS [Li et al., 2024]

Theory: Sample Complexity

Sample complexity

Using budget M and inflation c such that

$$\limsup_{\delta \to 0} \frac{c(t,\delta) \log M(t,\delta)}{\log(1/\delta)} \leq 1,$$

PSIPS satisfies

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_{\mathrm{PS}}]}{\log(1/\delta)} \leq T_\star(\boldsymbol{\mu})$$

Rationale. the truncated prosterior density is close to

$$q_t(\boldsymbol{\lambda}) \propto \exp\left(-\sum_k N_{t,k} \left\|\boldsymbol{\mu}_k - \boldsymbol{\lambda}_k\right\|^2_{\boldsymbol{\Sigma}^{-1}}\right) \cdot \mathbb{1}_{\boldsymbol{\lambda} \in Alt(S_t)}$$

$$\propto q_{t-1}(\boldsymbol{\lambda}) \cdot \exp\left(-\left\|\boldsymbol{\mu}_{A_t} - \boldsymbol{\lambda}_{A_t}\right\|^2_{\boldsymbol{\Sigma}^{-1}}\right)$$

which mirrors the behavior of the continuous Exponential Weights algorithm under quadratic loss.

Theory: Sample Complexity

Sample complexity

Using budget M and inflation c such that

$$\limsup_{\delta \to 0} \frac{c(t,\delta) \log M(t,\delta)}{\log(1/\delta)} \leq 1,$$

PSIPS satisfies

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_{\mathrm{PS}}]}{\log(1/\delta)} \leq T_\star(\boldsymbol{\mu})$$

Rationale. the truncated prosterior density is close to

$$q_t(\boldsymbol{\lambda}) \propto \exp\left(-\sum_k N_{t,k} \left\|\boldsymbol{\mu}_k - \boldsymbol{\lambda}_k\right\|^2_{\boldsymbol{\Sigma}^{-1}}\right) \cdot \mathbb{1}_{\boldsymbol{\lambda} \in Alt(S_t)}$$

$$\propto q_{t-1}(\boldsymbol{\lambda}) \cdot \exp\left(-\left\|\boldsymbol{\mu}_{A_t} - \boldsymbol{\lambda}_{A_t}\right\|^2_{\boldsymbol{\Sigma}^{-1}}\right)$$

which mirrors the behavior of the continuous Exponential Weights algorithm under quadratic loss.

Theory: Correctness

The calibration of the PS stopping rule is not as easy as the GLR: it requires a lower bound on

$$\Pi_t(\mathrm{Alt}(S_t)^c) \quad \text{when } S_t \neq \mathcal{S}^\star$$

and thus some anti-concentration results.

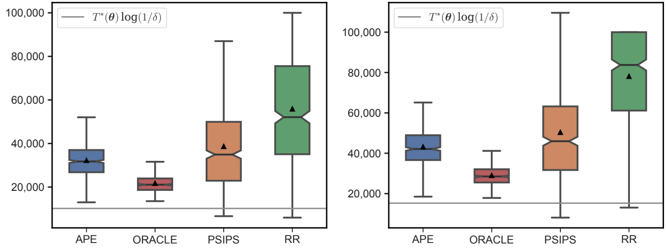
Lemma

For `PSIPS` to be δ-correct we can choose

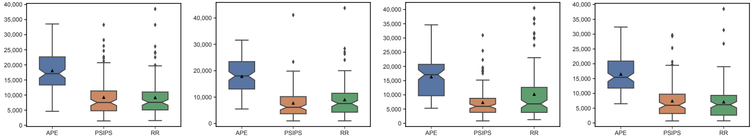
$$c(t,\delta) \simeq \frac{\log(\log(t)/\delta)}{\log(1/\delta)} \text{ and } \mathrm{M}(t,\delta) \simeq \frac{\log(t/\delta)}{\delta}$$

Practice

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



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



Conclusion

We proposed two approaches to Pareto Set Identification in the Fixed Confidence Setting:

- Adaptive Pareto Exploration: finite time bound, sub-optimal in the asymptotic regime $\delta \to 0$
- PSIPS, a (tractable !) Lower Bound Inspired algorithm, optimal in the asymptotic regime

➜ which one should we use in practise?

The sampling-based stopping rule is an interesting alternative to the GLR stopping rule for any complex pure exploration problem

Perspective: multi-objective bandit algorithms always sample *all* the marginals of the chosen arm $\to$ can we also adaptively select which marginals to observe?

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.

Gabillon, V., Ghavamzadeh, M., and Lazaric, A. (2012).
Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence.
In *Advances in Neural Information Processing Systems*.

Garivier, A. and Kaufmann, E. (2016).
Optimal best arm identification with fixed confidence.
In *Proceedings of the 29th Conference On Learning Theory*.

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)*.

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. (2018).
Feasible arm identification.
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).

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)*.

Li, Z., Jamieson, K., and Jain, L. (2024).
Optimal exploration is no harder than Thompson sampling.
In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. PMLR.

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

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.

The Pareto gaps

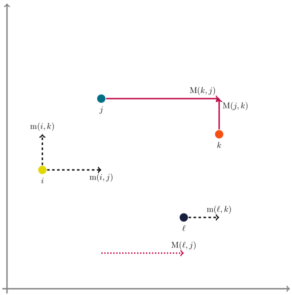
[Auer et al., 2016]

For sub-optimal arms $i \notin \mathcal{S}^\star(\boldsymbol{\mu})$,

$$\Delta_i := \max_{j \in \mathcal{S}^\star} \mathrm{m}(i,j), \quad \mathrm{m}(i,j) = -M(j,i)$$

while for optimal arms $i \in \mathcal{S}^\star$, $\Delta_i = \min(\delta_i^+, \delta_i^-)$ where

$$\begin{aligned} \delta_i^+ &:= \min_{j \in \mathcal{S}^\star \setminus \{i\}} \min(\mathrm{M}(i,j), \mathrm{M}(j,i)) \\ \delta_i^- &:= \min_{j \in [K] \setminus \mathcal{S}^\star} \{[\mathrm{M}(j,i)]_+ + \Delta_j\} \end{aligned}$$



On the effect of correlation

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

- Covariance matrix: $\mathbf{\Sigma}_\rho$ 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.