Adaptive Algorithms for Relaxed Pareto Set Identification

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper we revisit the fixed-confidence identification of the Pareto optimal set in a multi-objective multi-armed bandit model. As the sample complexity to identify the exact Pareto set can be very large, a relaxation allowing to output some additional near-optimal arms has been studied. In this work we also tackle alternative relaxations that allow instead to identify a relevant *subset* of the Pareto set. Notably, we propose a single sampling strategy, called Adaptive Pareto Exploration, that can be used in conjunction with different stopping rules to take into account different relaxations of the Pareto Set Identification problem. We analyze the sample complexity of these different combinations, quantifying in particular the reduction in sample complexity that occurs when one seeks to identify at most k Pareto optimal arms. We showcase the good practical performance of Adaptive Pareto Exploration on a real-world scenario, in which we adaptively explore several vaccination strategies against Covid-19 in order to find the optimal ones when multiple immunogenicity criteria are taken into account.

1 Introduction

In a multi-armed bandit model, an agent sequentially collects samples from several unknown distributions, called arms, in order to learn about these distributions (pure exploration), possibly under the constraint to maximize the sample collected, viewed as rewards (regret minimization). These objectives have been extensively studied for different types of univariate arms distributions [1]. In this paper, we consider the less common setting in which arms are multi-variate distributions. We are interested in the *Pareto Set Identification* (PSI) problem. In this pure exploration problem, the agent seeks to identify the arms that are *(Pareto) optimal*, i.e. such that their expected values for all objectives are not uniformly worse than those of another arm.

We formalize this as a fixed-confidence identification problem: in each round t the agent selects an arm A_t using an adaptive *sampling rule* and observes a sample $\mathbf{X}_t \in \mathbb{R}^D$ from the associated distribution. It further uses an adaptive stopping rule τ to decide when to stop sampling and output a set of arms $\hat{S}_\tau$ which is her guess for (an approximation of) the true Pareto set $\mathcal{S}^\star$. Given a risk parameter $\delta \in (0, 1)$, this guess should be correct with high probability, e.g. satisfy $\mathbb{P}(\hat{S}_\tau = \mathcal{S}^\star) \geq 1 - \delta$ for exact Pareto set identification, while requiring a small *sample complexity* τ. This generalizes the well-studied fixed-confidence Best Arm Identification (BAI) problem [2, 3, 4] to multiple objectives.

Our motivation to study multi-objective adaptive identification stems from the design of adaptive early-phase clinical trials. In phase I/II trials, the effects of a molecule in humans are explored, and several biological endpoints may be assessed at the same time as indicative markers of efficacy. In particular, in the context of vaccine development, early-phase trials usually assess multiple immunogenicity endpoints (i.e. various markers of the effects of the vaccine on the immune system, such as different facets of antibody responses or other immune parameters). In the absence of a known correlate of

Submitted to 16th European Workshop on Reinforcement Learning (EWRL 2023). Do not distribute.

protection during early clinical development, these endpoints may not have a clear *a priori* hierarchy, may not all be correlated, which makes an examination of the Pareto set of different vaccinal strategies particularly relevant. In addition, given the availability of various vaccine platforms (such as mRNA vaccines, viral-vector vaccines, protein vaccines), as exemplified by Covid-19 vaccines, there may be a need to adaptively screen the various resulting vaccine strategies to select the most promising ones. Apart from clinical trials, active Pareto Set Identification can be meaningful in many real-word contexts, and we refer the reader to the various examples given by [5], such as hardware or software design.

For many applications, the sample complexity of exact PSI can be prohibitive, either when there are many close to optimal arms or when the Pareto set is very large, and different relaxations have been considered in the literature [6, 7]. Going back to our application, in an adaptive trial that aims at pre-selecting a certain number of treatments or vaccine strategies for further investigations in clinical trials, practical constraints (the cost and feasibility of the trials) impose a constraint on the maximal number of interesting arms that can be identified. This motivates the introduction of a new setting where the agent is asked to identify *at most* k Pareto optimal arms. Interestingly the sampling rule that we propose for this setting can be used to solve (some generalizations of) other relaxations considered in the literature.

Related work The work most closely related to ours is that of Auer et al. [6], who propose a relaxation, which we refer to as ε_1-PSI: their algorithm returns a set $\hat{S}$ that contains w.h.p. all the Pareto optimal arms and possibly some sub-optimal arms, which when increased by ε_1 coordinate-wise become Pareto optimal. For arms that have sub-Gaussian marginals, they provide an instance-dependent sample complexity bound scaling with some notion of sub-optimality gap for each arm. The work of Zaluaga et al. [5, 7] studies a structured variant of fixed-confidence PSI in which the means are regular functions of arms' descriptors. They use Gaussian process modeling and obtain worse-case sample complexity bounds. In particular [7] considers the identification of an ε-cover of the Pareto set, which is a representative subset of the (ε)-Pareto set that will be related to our $(\varepsilon_1, \varepsilon_2)$-PSI criterion. The algorithms of [6] and those of [5, 7] in the unstructured setting[1] have the same flavor: they sample uniformly from a set of active arms and remove arms that have been found sub-optimal (or not representative). [6] further adds an acceptation mechanism to stop sampling some of the arms that have been found (nearly-)optimal and are guaranteed not to dominate an arm of the active set. In this paper, we propose instead a more adaptive exploration strategy, which departs from such accept/reject mechanisms and is suited for different types of relaxation, including our novel k-relaxation.

Adaptive Pareto Exploration (APE) leverages confidence intervals on the differences of arms' coordinates in order to identify a single arm to explore, in the spirit of the LUCB [3] or UGapEc [8] algorithms for Top-m identification in (one-dimensional) bandit models. These algorithms have been found out to be preferable in practice to their competitors based on uniform sampling and eliminations [9], an observation that will carry over to APE. Besides the multi-dimensional observations, we emphasize that a major challenge of the PSI problem with respect to e.g. Top m identification is that the number of arms to identify is not known in advance. Moreover, when relaxations are considered, there are multiple correct answers. In the one-dimensional settings, finding optimal algorithms in the presence of multiple correct answers is notoriously hard as discussed by the authors of [10], and their lower-bound based approach becomes impractical in our multi-dimensional setting. Finally, we remark that the k-relaxation can be viewed as an extension of to the problem of identifying any k-sized subset out of the best m arms in a standard bandit [11].

Beyond Pareto set identification, other interesting multi-objective bandit identification problems have been studied in the literature. For example [12] propose an algorithm to identify some particular arms in the Pareto set through a scalarization technique [13]. The idea is to turn the multi-objective pure-exploration problem into a single-objective one (unique optimal arm) by using a real-valued preference function which is only maximized by Pareto optimal arms (see e.g [13] for some examples of these functions). In practice, a family of those functions can be used to identify many arms of the Pareto set but it is not always possible to identify the entire Pareto set using this technique (see e.g [14] for *weighted sum* with a family of weights vectors). In a different direction, the authors of [15] introduce the feasible arm identification problem, in which the goal is to identify the set of arms

[1] PAL relies on confidence intervals that follow from Gaussian process regression, but can also be instantiated with simpler un-structured confidence intervals as those used in our work and in Auer's

whose mean vectors belong to a known polyhedron $P \subset \mathbb{R}^D$. In a follow up work [16], they propose a fixed-confidence algorithm for finding feasible arms that further maximize a given weighted sum of the objectives. In clinical trials, this could be used to find treatments maximizing efficacy (or a weighted sum of different efficacy indicators), under the constraint that the toxicity remains below a threshold. However, in the presence of multiple indicators of biological efficacy, choosing the weights may be difficult, and an examination of the Pareto set could be more suitable. Finally, some papers consider extensions of the Pareto optimality condition. The authors of [17] tackle the identification of the set of non-dominated arms of any partial order defined by an $\mathbb{R}^D$ polyhedral ordering cone (the usual Pareto dominance corresponds to using the cone defined by the positive orthant $\mathbb{R}^D_+$), and they provide worst-case sample complexity in the PAC setting. The work of [18] studies the identification of the set of non-dominated elements in a *partially ordered set* under the dueling bandit setting, in which the observations consists in pairwise comparison between arms.

Outline and contributions First, we formalize in Section 2 different relaxations of the PSI problem: ε_1-PSI, as introduced by [6], $\varepsilon_1, \varepsilon_2$-PSI, of which a particular case was studied by [7] and ε_1-PSI-k, a novel relaxation that takes as input an upper bound k on the maximum number of ε_1-optimal arms that can be returned. Then, we introduce in Section 3 Adaptive Pareto Exploration, a simple, adaptive sampling rule which can simultaneously tackle all three relaxations, when coupled with an appropriate stopping rule that we define for each of them. In Section 4, we prove high-probability upper bounds on the sample complexity of APE under different stopping rules. For ε_1-PSI, our bound slightly improves upon the state-of-the-art. Our strongest result is the bound for ε_1-PSI-k, which leads to a new notion of sub-optimality gap, quantifying the reduction in sample complexity that is obtained. Then, Section 5 presents the result of a numerical study on synthetic datasets, one of them being inspired by a Covid-19 vaccine clinical trial. It showcases the good empirical performance of APE compared to existing algorithms, and illustrates the impact of the different relaxations.

2 Problem Setting

In this section, we introduce the *Pareto Set Identification* (PSI) problem and its relaxations. Fix $K, D \in \mathbb{N}^\star$. Let $\nu_1, \dots, \nu_K$ be distributions over $\mathbb{R}^D$ with means $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K \in \mathbb{R}^D$. Let $\mathbb{A} := [K] := \{1, \dots, K\}$ denote the set of arms. Let $\nu := (\nu_1, \dots, \nu_K)$ and $\mathcal{X} := (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K)$. We use boldfaced symbols for $\mathbb{R}^D$ elements. Letting $\mathbf{X} \in \mathbb{R}^D, u \in \mathbb{R}$, for any $d \in \{1, \dots, D\}$, X^d denotes the $d-$th coordinate of $\mathbf{X}$ and $\mathbf{X} + u := (X^1 + u, \dots, X^D + u)$. In the sequel, we will assume that $\nu_1, \dots, \nu_K$ have $1-$subgaussian marginals [2].

Definition 1. *Given two arms $i, j \in \mathbb{A}$, i is weakly (Pareto) dominated by j (denoted by $\boldsymbol{\mu}_i \leq \boldsymbol{\mu}_j$) if for any $d \in \{1, \dots, D\}$, $\mu_i^d \leq \mu_j^d$. The arm i is (Pareto) dominated by j ($\boldsymbol{\mu}_i \preceq \boldsymbol{\mu}_j$ or $i \prec j$) if i is weakly dominated by j and there exists $d \in \{1, \dots, D\}$ such that $\mu_i^d < \mu_j^d$. The arm i is strictly (Pareto) dominated by j ($\boldsymbol{\mu}_i \prec \boldsymbol{\mu}_j$ or $i \prec j$) if for any $d \in \{1, \dots, D\}$, $\mu_i^d < \mu_j^d$.*

For $\boldsymbol{\varepsilon} \in \mathbb{R}^D_+$, the *$\boldsymbol{\varepsilon}$-Pareto set* $\mathcal{S}^\star_{\boldsymbol{\varepsilon}}(\mathcal{X})$ is the set of $\boldsymbol{\varepsilon}$-Pareto optimal arms, that is:

$$\mathcal{S}^\star_{\boldsymbol{\varepsilon}}(\mathcal{X}) := \{i \in \mathbb{A} \text{ s.t } \nexists j \in \mathbb{A} : \boldsymbol{\mu}_i + \boldsymbol{\varepsilon} \prec \boldsymbol{\mu}_j\}.$$

In particular, $\mathcal{S}^\star_{\mathbf{0}}(\mathcal{X})$ is called the *Pareto set* and we will simply write $\mathcal{S}^\star(\mathcal{X})$ to denote $\mathcal{S}^\star_{\mathbf{0}}(\mathcal{X})$. When it is clear from the context, we write $\mathcal{S}^\star$ (or $\mathcal{S}^\star_\varepsilon$) to denote $\mathcal{S}^\star(\mathcal{X})$ (or $\mathcal{S}^\star_\varepsilon(\mathcal{X})$). The goal of the learner is to identify the *Pareto set*. By abuse of notation we write $\mathcal{S}^\star_\varepsilon$ when $\varepsilon \in \mathbb{R}^+$ to denote $\mathcal{S}^\star_{\boldsymbol{\varepsilon}}$, $\boldsymbol{\varepsilon} := (\varepsilon, \dots, \varepsilon)$.

In each round $t = 1, 2, \dots$, the agent chooses an arm A_t and observes an independent draw $\mathbf{X}_t \sim \nu_{A_t}$ with $\mathbb{E}(\mathbf{X}_{A_t}) = \boldsymbol{\mu}_{A_t}$. We denote by $\mathbb{P}_\nu$ the law of the stochastic process $(\mathbf{X}_t)_{t \geq 1}$ and by $\mathbb{E}_\nu$, the expectation under $\mathbb{P}_\nu$. Let $\mathcal{F}_t := \sigma(A_1, \mathbf{X}_1, \dots, A_t, \mathbf{X}_t)$ the $\sigma-$algebra representing the history of the process. An algorithm for PSI consists in : i) a *sampling rule* which determines which arm to sample at time t based on history up to time $t-1$, ii) a *stopping rule* τ which is a stopping time w.r.t the filtration $(\mathcal{F}_t)_{t \geq 1}$ and iii) a *recommendation rule* which is a $\mathcal{F}_\tau$-measurable random set $\hat{S}_\tau$ representing the guess of the learner. The goal of the learner is to make a correct guess with high probability, using as few sample τ as possible. Before formalizing this, we introduce the different notion of correctness considered in this work, depending on parameters $\varepsilon_1 \geq, \varepsilon_2 \geq 0$ and $k \in [K]$. Our first criterion is the one considered by [6].

[2] A random variable X is $\sigma-$subgaussian if for any $\lambda \in \mathbb{R}$, $\mathbb{E}(\exp(\lambda(X - \mathbb{E}(X))) \leq \exp(\frac{\lambda^2 \sigma^2}{2})$.

Definition 2. *$\hat{S} \subset \mathbb{A}$ is correct for ε_1-PSI if $\mathcal{S}^\star \subset \hat{S} \subset \mathcal{S}^\star_{\varepsilon_1}$.*

To introduce our second criterion, we need the following definition.

Definition 3. *Let $\varepsilon_1, \varepsilon_2 \geq 0$. A subset $S \subset \mathbb{A}$ is an $(\varepsilon_1, \varepsilon_2)$-cover of the* Pareto set *if : $S \subset \mathcal{S}^\star_{\varepsilon_1}$ and for any $i \notin S$ either $i \notin \mathcal{S}^\star$ or $\exists j \in S$ such that $\boldsymbol{\mu}_i \prec \boldsymbol{\mu}_j + \varepsilon_2$.*

The ε-accurate set of [7] is a particular case of $(\varepsilon_1, \varepsilon_2)-$cover for which $\varepsilon_1 = \varepsilon_2 = \varepsilon$. Allowing $\varepsilon_1 \neq \varepsilon_2$ generalizes the notion of ε-correct set and can be useful, e.g., in scenarios when we want to identify the exact Pareto set (setting $\varepsilon_1 = 0$) but allow some optimal arms to be discarded if they are too close (parameterized by ε_2) to another optimal arm already returned. We note however that the *sparse cover* of [6] in not an $(\varepsilon_1, \varepsilon_2)$-cover but a "small" subset of nearly optimal arms that represents well the *Pareto set*. Identifying a sparse cover from samples requires in particular to identify $\mathcal{S}^\star_{\varepsilon_1}$ hence it can not be seen as a relaxation of ε_1-PSI.

Definition 4. *$\hat{S} \subset \mathbb{A}$ is correct for $(\varepsilon_1, \varepsilon_2)$-PSI if it is an $(\varepsilon_1, \varepsilon_2)$-cover of the Pareto set.*

Definition 5. *$\hat{S} \subset \mathbb{A}$ is correct for ε_1-PSI-k if either* i) $|\hat{S}| = k$ *and $\hat{S} \subset \mathcal{S}^\star_{\varepsilon_1}$ or* ii) $|\hat{S}| < k$ *and $\mathcal{S}^\star \subset \hat{S} \subset \mathcal{S}^\star_{\varepsilon_1}$ holds.*

Given a specified objective (ε_1-PSI, $(\varepsilon_1, \varepsilon_2)$-PSI or ε_1-PSI-k), and a target risk parameter $\delta \in (0, 1)$, the goal of the agent is to build a δ-correct algorithm, that is to guarantee that with probability larger than $1 - \delta$, her guess $\hat{S}_\tau$ is correct for the given objective, while minimizing the number of samples τ needed to make the guess, called the *sample complexity*.

We now introduce two important quantities to characterize the (Pareto) optimality or sub-optimality of the arms. For any two arms i, j, we let

$$\mathrm{m}(i,j) \quad := \quad \min_{1 \leq d \leq D} (\mu_j^d - \mu_i^d), \text{ and } \mathrm{M}(i,j) := \max_{1 \leq d \leq D} (\mu_i^d - \mu_j^d),$$

which have the following interpretation. If $i \preceq j$, $\mathrm{m}(i,j)$ is the minimal quantity $\alpha \geq 0$ that should be added component-wise to $\boldsymbol{\mu}_i$ so that $\boldsymbol{\mu}_i + \boldsymbol{\alpha} \not\prec \boldsymbol{\mu}_j$, $\boldsymbol{\alpha} := (\alpha, \ldots, \alpha)$. Moreover, $\mathrm{m}(i,j) > 0$ if and only if $i \prec j$. Then, for any arms i, j, if $i \not\prec j$, $\mathrm{M}(i,j)$ is the minimum quantity α' such $\boldsymbol{\mu}_i \leq \boldsymbol{\mu}_j + \boldsymbol{\alpha'}$, $\boldsymbol{\alpha'} := (\alpha', \ldots, \alpha')$. We remark that $\mathrm{M}(i,j) < 0$ if and only if $i \prec j$. Our algorithms, presented in the next section, rely on confidence intervals on these quantities.

3 Adaptive Pareto Exploration

We describe in this section our sampling rule, Adaptive Pareto Exploration, and present three stopping and recommendation rules to which it can be combined to solve each of the proposed relaxation. Let $T_k(t) = \sum_{s=1}^t \mathbb{1}(A_s = k)$ be the number of times arm k has been pulled by the end of round t and $\widehat{\boldsymbol{\mu}}_k(t) := T_k(t)^{-1} \sum_{s=1}^{T_k(t)} \mathbf{X}_{k,s}$ the empirical mean of this arm at time t, where $\mathbf{X}_{k,s}$ denotes the s-th observation drawn *i.i.d* from ν_k. For any arms $i, j \in \mathbb{A}$, we let

$$\mathrm{m}(i,j,t) := \min_d (\widehat{\mu}_j^d(t) - \widehat{\mu}_i^d(t)) \quad \text{and} \quad \mathrm{M}(i,j,t) := \max_d (\widehat{\mu}_i^d(t) - \mu_j^d(t)).$$

The empirical Pareto set is defined as

$$\begin{aligned} S(t) \quad &:= \quad \{i \in \mathbb{A} : \nexists j \in \mathbb{A} : \widehat{\boldsymbol{\mu}}_i(t) \prec \widehat{\boldsymbol{\mu}}_j(t)\}, \\ &= \quad \{i \in \mathbb{A} : \forall j \in \mathbb{A} \backslash \{i\}, M(i,j,t) > 0\} \ . \end{aligned}$$

3.1 Generic algorithm(s)

Adaptive Pareto Exploration relies on a *lower/upper confidence bound* approach, similar to UGapEc [8] lil'UCB [19] or LUCB [3]. The idea is to identify at any round two contentious arms and sample both or one of them. To define those, we suppose that there exists confidence intervals $[L_{i,j}^d(t,\delta), U_{i,j}^d(t,\delta)]$ on the difference in expected values for each pair of arms (i,j) and each objective $d \in D$, such that introducing

$$\mathcal{E}_t := \bigcap_{i=1}^K \bigcap_{j \neq i} \bigcap_{d=1}^D \{L_{i,j}^d(t,\delta) \leq \mu_i^d - \mu_j^d \leq U_{i,j}^d(t,\delta)\}, \text{ and } \mathcal{E} = \bigcap_{t=1}^\infty \mathcal{E}_t.$$

we have $\mathbb{P}(\mathcal{E}) \geq 1-\delta$. Concrete choices of these confidence intervals will be discussed in Section 3.2.

To ease the notation, we drop the dependency in δ in the confidence intervals and further define

$$\mathrm{M}^-(i,j,t) := \max_d L^d_{i,j}(t) \quad \text{and} \quad \mathrm{M}^+(i,j,t) := \max_d U^d_{i,j}(t) \tag{1}$$

$$\mathrm{m}^-(i,j,t) := -\mathrm{M}^+(i,j,t) \quad \text{and} \quad \mathrm{m}^+(i,j,t) := -\mathrm{M}^-(i,j,t). \tag{2}$$

Lemma 1. *For any round $t \geq 1$, if $\mathcal{E}_t$ holds, then for any $i,j \in \mathbb{A}$, $\mathrm{M}^-(i,j,t) \leq \mathrm{M}(i,j) \leq \mathrm{M}^+(i,j,t)$ and $\mathrm{m}^-(i,j,t) \leq \mathrm{m}(i,j) \leq \mathrm{m}^+(i,j,t)$.*

Noting that $\mathcal{S}^\star_{\varepsilon_1} = \{i \in \mathbb{A} : \forall j \neq i, M(i,j) + \varepsilon_1 > 0\}$, we define the following set of arms that are likely to be ε_1-Pareto optimal:

$$\mathrm{OPT}^{\varepsilon_1}(t) := \{i \in \mathbb{A} : \forall j \in \mathbb{A}\backslash\{i\}, \mathrm{M}^-(i,j,t) + \varepsilon_1 > 0\}.$$

Sampling rule In round t, Adaptive Pareto Exploration samples a_t, the least pulled arm among b_t and c_t given by

$$b_t := \operatorname*{argmax}_{i \in \mathbb{A}\backslash \mathrm{OPT}^{\varepsilon_1}(t)} \min_{j \neq i} \mathrm{M}^+(i,j,t),$$

$$c_t := \operatorname*{argmin}_{j \neq b_t} \mathrm{M}^-(b_t,j,t)$$

Indeed, b_t can be seen as the arm which is *optimistically* the most likely to be (nearly) Pareto optimal among the arms that are not yet identified as (nearly) optimal with high probability and c_t can be seen as the most likely to dominate b_t. In particular, we show in Appendix F that for $D = 1$, this sampling rule takes a simple form that is close (but not identical) to LUCB and UGapEc.

Stopping and recommendation rule(s) Depending on the objective, Adaptive Pareto Exploration can be plugged in with different stopping rules, that are summarized in Table 1 with their associated recommendations. To define those, we define for all $i \in \mathbb{A}$, $\varepsilon_1, \varepsilon_2 \geq 0$,

$$g^{\varepsilon_2}_i(t) := \max_{j \neq i} \mathrm{m}^-(i,j,t) + \varepsilon_2 \mathbb{1}\{j \in \mathrm{OPT}^{\varepsilon_1}(t)\} \quad \text{and} \quad h^{\varepsilon_1}_i(t) := \min_{j \neq i} \mathrm{M}^-(i,j,t) + \varepsilon_1.$$

and let $g_i(t) := g^0_i(t)$. Introducing

$$Z^{\varepsilon_1}_1(t) := \min_{i \in S(t)} h^{\varepsilon_1}_i(t), \text{ and } Z^{\varepsilon_1}_2(t) := \min_{i \in S(t)^\mathsf{c}} \max(g_i(t), h^{\varepsilon_1}_i(t)),$$

for ε_1-PSI, our stopping rule is $\tau_{\varepsilon_1} := \inf\{t \geq K : Z^{\varepsilon_1}_1(t) > 0 \wedge Z^{\varepsilon_1}_2(t) > 0\}$ and the associated recommendation is $\mathcal{O}(\tau_{\varepsilon_1})$ where

$$\mathcal{O}(t) := S(t) \cup \{i \in S(t)^\mathsf{c} : \nexists j \neq i : \mathrm{m}^-(i,j,t) > 0\}$$

consists of the current empirical Pareto set plus some additional arms that have not yet been formally identified as sub-optimal. Those arms should be (ε_1)-Pareto optimal.

For $(\varepsilon_1, \varepsilon_2)$-PSI we define a similar stopping rule $\tau_{\varepsilon_1,\varepsilon_2}$ where the stopping statistics are respectively replaced with

$$Z^{\varepsilon_1,\varepsilon_2}_1(t) := \min_{i \in S(t)} \max(g^{\varepsilon_2}_i(t), h^{\varepsilon_1}_i(t)) \text{ and } Z^{\varepsilon_1,\varepsilon_2}_2(t) := \min_{i \in S(t)^\mathsf{c}} \max(g^{\varepsilon_2}_i(t), h^{\varepsilon_1}_i(t))$$

with the convention $\min_\emptyset = +\infty$, and the recommendation is $\mathrm{OPT}^{\varepsilon_1}(\tau_{\varepsilon_1,\varepsilon_2})$.

To tackle the ε_1-PSI-k relaxation, we propose to couple τ_{ε_1} with an additional stopping condition checking whether $\mathrm{OPT}^{\varepsilon_1}(t)$ already contains k arms. That is, we stop at $\tau^k_{\varepsilon_1} := \min\left(\tau_{\varepsilon_1}, \tau^k\right)$ where $\tau^k := \inf\{t \geq K : |\mathrm{OPT}^{\varepsilon_1}(t)| \geq k\}$ with associated recommendation $\mathrm{OPT}^{\varepsilon_1}(\tau^k)$. Depending of the reason for stopping (τ_{ε_1} or τ^k), we follow the corresponding recommendation.

Lemma 2. *Assume $\mathcal{E}$ holds. For ε_1-PSI (resp. $(\varepsilon_1,\varepsilon_2)$-PSI , ε_1-PSI-k), Adaptive Pareto Exploration combined with the stopping rule τ_{ε_1} (resp. $\tau_{\varepsilon_1,\varepsilon_2}$, resp. $\tau^k_{\varepsilon_1}$) outputs a correct subset.*

Remark 1. *We decoupled the presentation of the sampling rule to that of the "sequential testing" aspect (stopping and recommendation). We could even go further and observe that multiple tests could actually be run in parallel, for free. If we collect samples with APE (which only depends on ε_1), whenever one of the three stopping conditions given in Table 1 triggers,* for any values of ε_2 or k, *we can decide to stop and make the corresponding recommendation or continue and wait for another "more interesting" stopping condition to be satisfied. If $\mathcal{E}$ holds, a recommendation made at any such time will be correct for the objective associated to the stopping criterion (third column in Table 1).*

	Stopping condition	Recommendation	Objective
τ_{ε_1}	$Z_1^{\varepsilon_1}(t) > 0 \ \wedge Z_2^{\varepsilon_1}(t) > 0$	$\mathcal{O}(\tau_{\varepsilon_1})$	ε_1-PSI
$\tau_{\varepsilon_1,\varepsilon_2}$	$Z_1^{\varepsilon_1,\varepsilon_2}(t) > 0 \ \wedge Z_2^{\varepsilon_1,\varepsilon_2}(t) > 0$	$\mathrm{OPT}^{\varepsilon_1}(\tau_{\varepsilon_1,\varepsilon_2})$	$(\varepsilon_1,\varepsilon_2)$-PSI
τ^k	$\lvert\mathrm{OPT}^{\varepsilon_1}(t)\rvert \geq k$	$\mathrm{OPT}^{\varepsilon_1}(\tau^k)$	ε_1-PSI-k

Table 1: Stopping conditions and associated recommendation

3.2 Our instantiation

We propose to instantiate the algorithms with confidence interval on the difference of pair of arms. For any pair $i, j \in \mathbb{A}$, we define a function $\beta_{i,j}$ such that for any $d \in [D]$, $U_{i,j}^d(t) = \widehat{\mu}_i^d(t) - \widehat{\mu}_j^d(t) + \beta_{i,j}(t)$ and $L_{i,j}^d(t) = \widehat{\mu}_i^d(t) - \widehat{\mu}_j^d(t) - \beta_{i,j}(t)$. We take from [20] the following confidence bonus for time-uniform concentration:

$$\beta_{i,j}(t) := 2\sqrt{\left(C^g\left(\frac{\log\left(\frac{K_1}{\delta}\right)}{2}\right) + \sum_{a\in\{i,j\}} \log(4+\log(T_a(t)))\right)\left(\sum_{a\in\{i,j\}} \frac{1}{T_a(t)}\right)}, \tag{3}$$

where $K_1 := K(K-1)D/2$ and $C^g \approx x + \log(x)$ is a calibration function. They result in the simple expressions $\mathrm{M}^{\pm}(i,j,t) = \mathrm{M}(i,j,t) \pm \beta_{i,j}(t)$ and $\mathrm{m}^{\pm}(i,j,t) = \mathrm{m}(i,j,t) \pm \beta_{i,j}(t)$. As an example, we state in Algorithm 1 the pseudo-code of APE combined the stopping rule suited for the k-relaxation of ε_1-PSI, which we refer to as ε_1-APE-k.

Algorithm 1: ε_1-APE-k: Adaptive Pareto Exploration for ε_1-PSI-k

Data: parameter $\varepsilon_1 \geq 0$, $k \in [K]$
initialize : sample each arm once, set $t = K$, $T_i(t) = 1$ for any $i \in \mathbb{A}$
for $t = K, \ldots$ **do**
  $S(t) = \{i \in \mathbb{A} : \forall j \in \mathbb{A}\backslash\{i\}, \mathrm{M}(i,j,t) > 0\}$;
  $\mathrm{OPT}^{\varepsilon_1}(t) = \{i \in \mathbb{A} : \forall j \in \mathbb{A}\backslash\{i\}, \mathrm{M}(i,j,t) - \beta_{i,j}(t) + \varepsilon_1 > 0\}$;
  if $\lvert\mathrm{OPT}^{\varepsilon_1}(t)\rvert \geq k$ **then**
    break and output $\mathrm{OPT}^{\varepsilon_1}(t)$
  if $Z_1^{\varepsilon_1}(t) > 0 \ \wedge Z_2^{\varepsilon_1}(t) > 0$ **then**
    break and output $\mathcal{O}(t) = S(t) \bigcup \{i \in S(t)^{\mathsf{c}} : \nexists j \neq i : \mathrm{m}(i,j,t) - \beta_{i,j}(t) > 0\}$
  $b_t := \mathrm{argmax}_{i\in\mathbb{A}\backslash\mathrm{OPT}^{\varepsilon_1}(t)} \min_{j\neq i} [\mathrm{M}(i,j,t) + \beta_{i,j}(t)]$;
  $c_t := \mathrm{argmin}_{i\neq b_t} [\mathrm{M}(b_t,j,t) - \beta_{b_t,j}(t)]$;
  sample $a_t := \mathrm{argmin}_{i\in\{b_t,c_t\}} T_i(t)$;

In Appendix G, we also study a different instantiation based on confidence bounds of the form $U_{i,j}(t) = U_i(t) - L_j(t)$ where $[L_i(t), U_i(t)]$ is a confidence interval on μ_i. This is the approach followed by LUCB for $D = 1$ and prior work on Pareto identification [6, 7]. In practice we advocate the use of the pairwise confidence intervals defined above, even if our current analysis does not allow to quantify their improvement. For the LUCB-like instantiation, we further derive in Appendix G an upper bound on the expected stopping time of APE for the different stopping rules.

4 Theoretical analysis

In this section, we state our main theorem on the sample complexity of our algorithms and give a sketch of its proof. First let us introduce some quantities that are needed to state the theorem. The sample complexity of the algorithm proposed by [6] for (ε_1)-Pareto set identification scales as a sum over the arms i of $1/(\Delta_i \vee \varepsilon_1)^2$ where Δ_i is called the sub-optimality gap of arm i and is defined as follows. For a sub-optimal arm $i \notin \mathcal{S}^\star(\mathcal{X})$,

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

which is the smallest quantity that should be added component-wise to $\boldsymbol{\mu}_i$ to make i appear Pareto optimal w.r.t $\{\boldsymbol{\mu}_i : i \in \mathbb{A}\}$. For a Pareto optimal arm $i \in \mathcal{S}^\star(\mathcal{X})$, the definition is more involved:

$$\Delta_i := \begin{cases} \min_{j \in \mathbb{A} \setminus \{i\}} \Delta_j & \text{if } \mathcal{S}^\star := \{i\} \\ \min(\delta_i^+, \delta_i^-) & \text{else,} \end{cases}$$

where

$$\delta_i^+ := \min_{j \in \mathcal{S}^\star \setminus \{i\}} \min(\mathrm{M}(i,j), \mathrm{M}(j,i)) \text{ and } \delta_i^- := \min_{j \in \mathbb{A} \setminus \mathcal{S}^\star} \{(\mathrm{M}(j,i))^+ + \Delta_j\}.$$

For $x \in \mathbb{R}, (x)+ := \max(x, 0)$. We also introduce some additional notion needed to express the contribution of the k-relaxation. Let $1 \leq k \leq K$. For any arm i, let $\omega_i = \min_{j \neq i} \mathrm{M}(i,j)$ and define

$$\omega^k := \max_{i \in \mathbb{A}}^{k} \omega_i, \qquad \mathcal{S}^{\star,k} := \operatorname*{argmax}_{i \in \mathbb{A}}^{1 \ldots k} \omega_i,$$

with the k-th max and first to k-th argmax operators. Observe that $w^k > 0$ if and only if $|\mathcal{S}^\star(\mathcal{X})| \geq k$.

Theorem 1. *Fix a risk parameter $\delta \in (0,1)$, $\varepsilon_1 \geq 0$, let $k \leq K$ and ν a bandit with $1-$subgaussian marginals. With probability at least $1-\delta$, ε_1-APE-k recommends a correct set for the ε_1-PSI-k objective and stops after at most*

$$\sum_{a \in \mathbb{A}} \frac{88}{\tilde{\Delta}_a^2} \log\left(\frac{2K(K-1)D}{\delta} \log\left(\frac{12e}{\tilde{\Delta}_a}\right)\right), \tag{4}$$

samples, where for each $a \in \mathbb{A}$, $\tilde{\Delta}_a := \max(\Delta_a, \varepsilon_1, \omega^k)$.

First, when $k = K$, observing that ε_1-APE-K provides a δ-correct algorithm for ε_1-PSI, our bound improves the result of [6] for the ε_1-PSI problem in terms of constant multiplicative factors and $\log\log\Delta^{-1}$ terms instead of $\log\Delta^{-2}$ and nearly matches the lower bound for the ε_1-PSI problem (Theorem 17 in [6]). It also shows the impact of the k-relaxation on the sample complexity. In particular, we can remark that for any arm $i \in \mathcal{S}^\star \setminus \mathcal{S}^{\star,k}$, $\max(\Delta_i, \omega_k) = \omega_k$. Intuitively, it says that we shouldn't pay more than the cost of identifying the k-th optimal arm, ordered by the ω_i's. A similar result has been obtained for the *any $k-$sized subset of the best m* problem [11]. But the authors have shown the relaxation only for the best m arms while our result shows that even the sub-optimal arms should be sampled less. In Appendix E, we prove a lower bound showing that in some scenarios, ε_1-APE-k is optimal for the k-relaxation (up to $D\log(K)$ and constant multiplicative terms). In Appendix D, we prove that Theorem 1 without the ω_k terms also holds for $(\varepsilon_1, \varepsilon_2)$-APE. This does not justifies the reduction in sample complexity when setting $\varepsilon_2 > 0$ in $(\varepsilon_1, \varepsilon_2)$-PSI observed in our experiments but it at least guarantees that the ε_2-relaxation doesn't make things worse.

Furthermore, since our algorithm allows $\varepsilon_1 = 0$, it is also an algorithm for BAI when $D = 1, \varepsilon_1 = 0$. We prove in Appendix F that in this case, the gaps $\Delta_i' s$ matches the classical gaps in BAI ([21, 22]) and we derive its sample complexity from Theorem 1 showing that it is similar in theory to LUCB, UGap, LUCB++ [3, 8, 23] but have better empirical performance.

Sketch of proof Using Proposition 24 of [20] we prove that the choice of $\beta_{i,j}$ in (3) yields $\mathbb{P}(\mathcal{E}) \geq 1-\delta$. Combining this result with Lemma 2 proves that ε_1-APE-k is correct with probability at least $1-\delta$. The idea of the remaining proof is to show that under the event $\mathcal{E}$, for our different stopping rules, if APE has not stopped at the end of round t, then a_t has not been explored enough.

Lemma 3. *Let $k \leq K$. If $\mathcal{E}_t$ holds and $t < \tau_{\varepsilon_1}^k$ then $\omega^k \leq 2\beta_{a_t,a_t}(t)$.*

Lemma 4. *If $\mathcal{E}_t$ holds and $t < \min(\tau_{\varepsilon_1}, \tau_{\varepsilon_1}^k, \tau_{\varepsilon_1,\varepsilon_2})$ then $\Delta_{a_t} \leq 2\beta_{a_t,a_t}(t)$.*

The following lemma holds for each of the stopping times $\tau_{\varepsilon_1}, \tau_{\varepsilon_1,\varepsilon_2}$ and $\tau_{\varepsilon_1}^k$.

Lemma 5. *If the algorithm has not stopped at the end of round t then $\varepsilon_1 \leq 2\beta_{a_t,a_t}(t)$.*

Combining these lemmas we prove that

$$\tau_{\varepsilon_1}^k \mathbb{1}\{\mathcal{E}\} \leq \sum_{a \in \mathbb{A}} \inf\left\{n \geq 2 : \tilde{\Delta}_a > 2\beta^n\right\}, \tag{5}$$

where β^n is the expression of $\beta_{i,j}(t)$ when $T_i(t) = T_j(t) = n$. A careful upper-bounding of the RHS of (5) completes the proof of Theorem 1.

5 Experiments

We evaluate the performance of Adaptive Pareto Exploration on a real-world scenario and on synthetic random Bernoulli instances. For a fair comparison, Algorithm 1 of [6], referred to as PSI-Unif-Elim and APE are both run with our confidence bonuses $\beta_{i,j}(t)$ on pairs of arms, which considerably improve single-arm confidence bonuses[3]. As anytime confidence bounds are known to be conservative, we use $K_1 = 1$ in (3) instead of its theoretical value coming from a union bound. Still, in all our experiments, the empirical error probability was (significantly) smaller than the target $\delta = 0.1$.

Real-world dataset COV-BOOST [24] is phase 2 trial which was conducted on 2883 participants to measure the immunogenicity of different Covid-19 vaccines as third dose (booster) in various combinations of initially received vaccines (first two doses). This resulted in a total of 20 vaccination strategies being assessed, each of them defined by the vaccines received as first, second and third dose. The authors have reported the average responses induced by each candidate strategy on cohorts of participants, measuring several immunogenicity markers. From this study, we extract and process the average response of each strategy to 3 specific immunogenicity indicators: two markers of antibody response and one of the cellular response. The outcomes are assumed to have a log-normal distribution [24]. We use the average (log) outcomes and their variances to simulate a multivariate Gaussian bandit with $K = 20, D = 3$. We give in Appendix I.2 some additional details about the processing of the data, and report the means and variances of each arm. In Appendix I.1 we further explain how APE can be simply adapted when the marginals distributions of the arms have different variances. In this experiment, we set $\varepsilon_1 = 0, \delta = 0.1$ and compare PSI-Unif-Elim to 0-APE-k (called

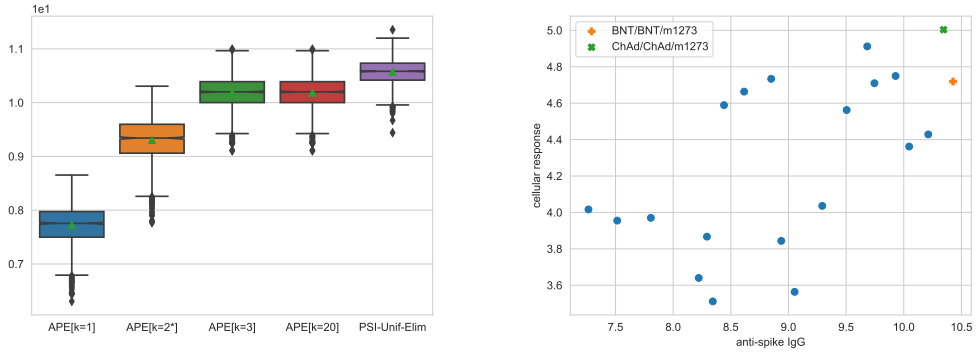


Figure 1: On the left is the log of the empirical sample complexity of PSI-Unif-Elim and APE on the real-world scenario plot (right) for 2 out of the 3 immunogenicity indicators.

APE-k in the sequel) for different values of k. The empirical distribution of the sample complexity of the algorithms, averaged over 2000 independent runs, are reported in Figure 1. The results are shown in log-scale (y-axis is the log of the sample complexity) to fit in the same figure. As $|\mathcal{S}^\star| = 2$, we first observe that, without the relaxation (i.e. for $k > 3$), APE outperforms its state-of-the-art competitor PSI-Unif-Elim . Moreover for $k = 1$ or $k = 2$, the sample complexity of APE-k is significantly reduced. For $k = 2$ when the stopping time τ^k is reached some sub-optimal arms have possibly not yet been identified as such, while for $k = 3$, even if the optimal arms have been identified, the remaining arms have to be sampled enough to ensure that they are sub-optimal before stopping. This explains the gap in sample complexity between $k = 2$ and $k = 3$. In Appendix I.3, we compare APE to an adaptation of PSI-Unif-Elim for the k-relaxation, showing that APE is always preferable.

Experiments on random instances To further illustrate the impact of the k-relaxation and to strengthen the comparison with PSI-Unif-Elim , we ran the previous algorithms on 2000 randomly generated multi-variate Bernoulli instances, with $K = 5$ arms and different values of the dimension D. We set $\delta = 0.1$ and $\varepsilon_1 = 0.005$ (to have reasonable running time). The averaged sample complexities are reported in Table 2. We observe that APE (with $k = K$) uses 20 to 25% less samples

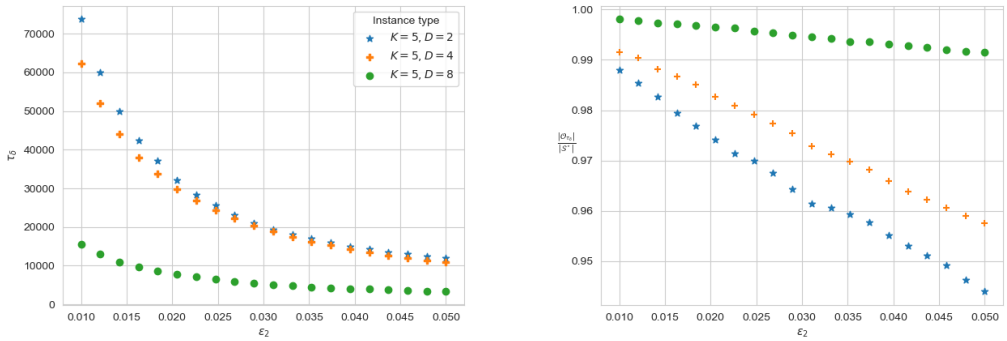
[3] In their experiments, [6] already proposed the heuristic use of confidence bonuses of this form

than PSI-Unif-Elim and tends to be more efficient as the dimension increases (and likely the size of the Pareto set, since the instances are randomly generated). We also note that identifying a k-sized subset of the Pareto set requires considerably less samples than exact PSI. In Appendix I.3 we also provide examples of instances for which APE takes up to 3 times less samples than PSI-Unif-Elim .

	ε_1-APE-1	ε_1-APE-2	ε_1-APE-3	ε_1-APE-4	ε_1-APE-5	ε_1-PSI-Unif-Elim
$D = 2$	811	39530	109020	145777	150844	190625
$D = 4$	214	6410	19908	68061	124001	157584
$D = 8$	119	204	405	1448	20336	27270

Table 2: Average sample complexity over 2000 random Bernoulli instances with $K = 5$ arms. On average the size of the Pareto set was $(2.295, 4.0625, 4.931)$ respectively for the dimensions $2, 4, 8$.

To illustrate the impact of the ε_2 relaxation, setting $\varepsilon_1 = 0$ we report the sample complexity of APE associated with the stopping time τ_{0,ε_2} for 20 equally spaced values of $\varepsilon_2 \in [0.01, 0.05]$, averaged over 2000 random Bernoulli instances. Figure 2 shows the decrease of the average sample complexity when ε_2 increases (left) and the average ratio of the size of the returned set to the size of the Pareto set (right). Note that for $\varepsilon_1 = 0$, we have $\mathcal{O}(\tau_{0,\varepsilon_2}) \subset \mathcal{S}^\star$. The average sample complexity reported decreases up to 86% for the instance with $K = 5, D = 2$ while the returned set contains more than 90% of the Pareto optimal arms. In Appendix I.3, we further illustrate the behavior of APE with the ε_2 relaxation on a fixed instance in dimension 2.



(a) Average sample complexity vs ε_2

(b) Average ratio $|\mathcal{O}(\tau_{0,\varepsilon_2})|/|\mathcal{S}^\star|$ vs ε_2

Figure 2: APE with τ_{0,ε_2} averaged over 2000 random Bernoulli instance with $K = 5$ arms.

6 Conclusion and perspective

We proposed and analyzed APE, an adaptive sampling rule in multi-variate bandits models that when coupled with different stopping rules can tackle different relaxations of the fixed-confidence Pareto Set Identification problem. Our experiments revealed the good performance of the resulting algorithms compared to the state-of-the-art PSI algorithm as well as the great reductions in sample complexity brought by the relaxations. In future work, we intend to make our algorithms more practical for possible applications to clinical trials. For this purpose, as measuring efficacy takes time, we will investigate its adaptation to a batch setting, following, e.g. the work of [25] for BAI. We will also investigate the use of APE beyond the fixed-confidence setting, to the possibly more realistic fixed-budget [21] or anytime exploration [26] settings. To the best of our knowledge, no algorithms exists in the literature for PSI in such settings. Finally, following the works of [5, 6], we defined the $\varepsilon_1, \varepsilon_2$ relaxations with scalar values, so that the same slack applies to all components. Although we could easily modify our algorithms to tackle vectorial values $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2$, so far we could only prove a dependence on $\min_d \varepsilon_1^d$ in the sample complexity. We intend to study the right quantification in the sample complexity when $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ are vectorial.

References

[1] T. Lattimore and C. CSzepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[2] E. Even-Dar, S. Mannor, and Y. Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(39):1079–1105, 2006.

[3] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 227–234. Omnipress, 2012.

[4] A. Garivier and E. Kaufmann. Optimal best arm identification with fixed confidence. In *29th Annual Conference on Learning Theory*, volume 49, pages 998–1027. PMLR, 2016.

[5] M. Zuluaga, G. Sergent, A. Krause, and M. Püschel. Active learning for multi-objective optimization. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 462–470. PMLR, 2013.

[6] P. Auer, C.-K. Chiang, R. Ortner, and M.-M. Drugan. Pareto front identification from stochastic bandit feedback. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pages 939–947. PMLR, 2016.

[7] M. Zuluaga, A. Krause, and M. Püschel. e-pal: An active learning approach to the multi-objective optimization problem. *Journal of Machine Learning Research*, 17(104):1–32, 2016.

[8] V. Gabillon, M. Ghavamzadeh, and A. Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[9] E. Kaufmann and S. Kalyanakrishnan. Information complexity in bandit subset selection. In *Conference On Learning Theory*, volume 30. JMLR: Workshop and Conference Proceedings, 2013.

[10] R. Degenne and W.M. Koolen. Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[11] A. Roy Chaudhuri and S. Kalyanakrishnan. PAC identification of many good arms in stochastic multi-armed bandits. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 991–1000. PMLR, 2019.

[12] M.-M. Drugan and A. Nowe. Designing multi-objective multi-armed bandits algorithms: A study. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2013.

[13] K. Miettinen. Nonlinear multiobjective optimization. In *International Series in Operations Research and Management Science*, 1998.

[14] M. Ehrgott. *Multicriteria Optimization*. Springer, second edition edition, 2005.

[15] J. Katz-Samuels and C. Scott. Feasible arm identification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 2535–2543. PMLR, 2018.

[16] J. Katz-Samuels and C. Scott. Top feasible arm identification. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 1593–1601. PMLR, 2019.

[17] C. Ararat and C. Tekin. Vector optimization with stochastic bandit feedback. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 2165–2190. PMLR, 2023.

[18] J. Audiffren and L. Ralaivola. Bandits dueling on partially ordered sets. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[19] K. Jamieson and R. Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6, 2014.

[20] E. Kaufmann and W.-M. Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021.

[21] J.-Y. Audibert and S. Bubeck. Best arm identification in multi-armed bandits. In *COLT - 23th Conference on Learning Theory - 2010*, page 13 p., 2010.

[22] E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

[23] M. Simchowitz, K. Jamieson, and B. Recht. The simulator: Understanding adaptive sampling in the moderate-confidence regime. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 1794–1834. PMLR, 2017.

[24] A.-P.-S. Munro, L. Janani, V. Cornelius, and et al. 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, 2021.

[25] K.-S. Jun, K. Jamieson, R. Nowak, and X. Zhu. Top arm identification in multi-armed bandits with batch arm pulls. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pages 139–148. PMLR, 2016.

[26] K-S. Jun and R. Nowak. Anytime exploration for multi-armed bandits using confidence information. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 974–982. PMLR, 2016.

[27] M. Jourdan, R. Degenne, D. Baudry, R. de Heide, and E. Kaufmann. Top two algorithms revisited. In *Advances in Neural Information Processing Systems*, volume 35, pages 26791–26803. Curran Associates, Inc., 2022.

[28] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil' ucb : An optimal exploration algorithm for multi-armed bandits. In *Proceedings of The 27th Conference on Learning Theory*, volume 35, pages 423–439. PMLR, 2014.

Contents

1 Introduction **1**

2 Problem Setting **3**

3 Adaptive Pareto Exploration **4**

3.1 Generic algorithm(s) 4

3.2 Our instantiation 6

4 Theoretical analysis **6**

5 Experiments **8**

6 Conclusion and perspective **9**

A Outline and Notation **14**

B Correctness for different stopping rules **14**

B.1 Proof of Lemma 1 14

B.2 Proof of Lemma 2 15

B.3 Calibration of the confidence intervals 15

C Sample complexity analysis **16**

C.1 Proof of Lemma 3 18

C.2 Proof of Lemma 4 18

C.3 Proof of Lemma 5 20

C.4 Auxiliary results 20

D Algorithm for finding an $(\varepsilon_1, \varepsilon_2)$-cover **21**

E Lower Bound **23**

F Best Arm Identification **25**

G LUCB1-like instantiation of APE **27**

H Technical Lemmas **31**

I Implementation and Additional Experiments **34**

I.1 Implementation 34

I.2 Data processing 35

I.3 Additional experiments 36

I.3.1 Additional experiments for ε_1-APE-k 36

I.3.2 $(\varepsilon_1, \varepsilon_2)$-APE 37

I.3.3 Comparing ε_1-APE-k to an adaptation of PSI-Unif-Elim 39

I.3.4 Comparison to some BAI algorithms 40

A Outline and Notation

In this section, we provide an outline of the supplemental material and define some additional notation. Appendix B proves the correctness of ours algorithms and some concentration lemmas. In Appendix C, we prove Theorem 1 and the lemmas used in its proof. In Appendix D we analyze the correctness and sample complexity of APE associated to the stopping time $\tau_{\varepsilon_1,\varepsilon_2}$. Appendix E describes our worst-case lower bound and in Appendix F we relate our algorithm to other algorithms for BAI. In Appendix G we derive an upper-bound on the expectation of the sample complexity of ε_1-APE-k with a LUCB1-like instantiation and in Appendix H we recall or prove some technical lemmas that are used in the main proofs. Finally, in Appendix I we give further details about the experiments together with additional experimental results.

Notation	Type	Description
$\mathbb{A}$		Set of arms
K	$\mathbb{N}^\star$	Number of arms
D	$\mathbb{N}^\star$	Dimension or number of attributes
$[n]$		$\{1,\dots,n\}$
m	$\mathbb{A}^2 \to \mathbb{R}$	$\mathrm{m}(i,j) := \min\{\mu_j^d - \mu_i^d : d \in [D]\}$
M	$\mathbb{A}^2 \to \mathbb{R}$	$\mathrm{M}(i,j) := \max\{\mu_i^d - \mu_j^d : d \in [D]\}$
$(x)^+$	$\mathbb{R} \to \mathbb{R}^+$	$\max(0,x)$
$\boldsymbol{\mu}_a$	$\mathbb{R}^D$	Mean of arm $a \in \mathbb{A}$
$\mathcal{S}_\varepsilon^\star$		$\{i \in \mathbb{A} : \nexists j \neq i \in \mathbb{A} : \boldsymbol{\mu}_i + \varepsilon \prec \boldsymbol{\mu}_j\}$

Table 3: Table of notation.

B Correctness for different stopping rules

In this section, we gather and prove results that are related to the correctness of our algorithms, either for their generic form (Lemma 1 and Lemma 2) or some specific calibration. We recall the definition of the events

$$\mathcal{E}_t = \bigcap_{i=1}^{K} \bigcap_{j\neq i} \bigcap_{d=1}^{D} \{L_{i,j}^d(t) \leq \mu_i^d - \mu_j^d \leq U_{i,j}^d(t)\} \text{ and } \mathcal{E} = \bigcap_{t=1}^{\infty} \mathcal{E}_t \ .$$

B.1 Proof of Lemma 1

Lemma 1. *For any round $t \geq 1$, if $\mathcal{E}_t$ holds, then for any $i,j \in \mathbb{A}$, $\mathrm{M}^-(i,j,t) \leq \mathrm{M}(i,j) \leq \mathrm{M}^+(i,j,t)$ and $\mathrm{m}^-(i,j,t) \leq \mathrm{m}(i,j) \leq \mathrm{m}^+(i,j,t)$.*

Proof of Lemma 1. This result simply follows from the definition of $\mathcal{E}_t$. Since

$$\mathcal{E}_t := \bigcap_{i=1}^{K} \bigcap_{j\neq i} \bigcap_{d=1}^{D} \{L_{i,j}^d(t) \leq \mu_i^d - \mu_j^d \leq U_{i,j}^d(t)\},$$

if $\mathcal{E}_t$ holds, then for any i,j

$$\mathrm{M}^-(i,j,t) := \max_d L_{i,j}^d(t) \leq \mathrm{M}(i,j) := \max_d(\mu_i^d - \mu_j^d) \leq \max_d U_{i,j}^d(t) := \mathrm{M}^+(i,j,t),$$

and the second point follows by noting that $\mathrm{m}(i,j) = -\mathrm{M}(i,j)$ and $\mathrm{m}^+(i,j,t) := -\mathrm{M}^-(i,j,t); \mathrm{m}^-(i,j,t) := -\mathrm{M}^+(i,j,t)$ for any pair of arms. □

We remark that when the algorithm uses confidence bonus of form $(\widehat{\mu}_i^d(t) - \widehat{\mu}_j^d(t)) \pm \beta_{i,j}(t)$,

$$\begin{aligned}
\mathrm{M}^+(i,j,t) &:= \max_d U_{i,j}^d(t) = \max_d(\widehat{\mu}_i^d(t) - \widehat{\mu}_j^d(t)) + \beta_{i,j}(t) = \mathrm{M}(i,j,t) + \beta_{i,j}(t), \\
\mathrm{M}^-(i,j,t) &:= \max_d L_{i,j}^d(t) = \max_d(\widehat{\mu}_i^d(t) - \widehat{\mu}_j^d(t)) - \beta_{i,j}(t) = \mathrm{M}(i,j,t) - \beta_{i,j}(t),
\end{aligned}$$

and the previous lemma implies that on $\mathcal{E}_t$,

$$|\mathrm{M}(i,j) - \mathrm{M}(i,j,t)| \leq \beta_{i,j}(t) \quad \text{and} \quad |\mathrm{m}(i,j) - \mathrm{m}(i,j,t)| \leq \beta_{i,j}(t),$$

which is extensively used in our sample complexity analyses.

B.2 Proof of Lemma 2

Lemma 2. *Assume $\mathcal{E}$ holds. For ε_1-PSI (resp. $(\varepsilon_1, \varepsilon_2)$-PSI , ε_1-PSI-k), Adaptive Pareto Exploration combined with the stopping rule τ_{ε_1} (resp. $\tau_{\varepsilon_1,\varepsilon_2}$, resp. $\tau^k_{\varepsilon_1}$) outputs a correct subset.*

We show the correctness of ε_1-PSI-k (for any k) and we derive the correctness for ε_1-PSI which is equivalent to ε_1-PSI-K. The correctness of $(\varepsilon_1, \varepsilon_2)$-PSI is shown separately in Lemma 11 (see Appendix D).

Proof of Lemma 2. Assume $\mathcal{E}$ holds. Let $t = \tau^k_{\varepsilon_1}$ and $i \in \mathrm{OPT}^{\varepsilon_1}(t)$. Since $i \in \mathrm{OPT}^{\varepsilon_1}(t)$, for any $j \neq i$,

$$\mathrm{M}(i,j) + \varepsilon_1 \overset{\mathcal{E}}{\geq} \mathrm{M}^-(i,j,t) + \varepsilon_1 > 0,$$

that is $i \in \mathcal{S}^\star_{\varepsilon_1}$. Therefore, on the event $\mathcal{E}$, $\mathrm{OPT}^{\varepsilon_1}(t) \subset \mathcal{S}^\star_{\varepsilon_1}$. Thus, if the stopping has occurred because $|\mathrm{OPT}^{\varepsilon_1}(t)| \geq k$, since in this case $\mathcal{O}(t) \subset \mathrm{OPT}^{\varepsilon_1}(t) \subset \mathcal{S}^\star_{\varepsilon_1}$, all the recommended arms will be (ε_1)-Pareto optimal. On the contrary, if $|\mathrm{OPT}^{\varepsilon_1}(t)| < k$, then from the definition of $\tau^k_{\varepsilon_1}$ it holds that

$$Z^{\varepsilon_1}_1(t) > 0 \quad \text{and} \quad Z^{\varepsilon_1}_2(t) > 0,$$

and the recommended set is then

$$\mathcal{O}(t) := S(t) \cup \{i \in S(t)^{\mathsf{c}} : \nexists j \neq i : \mathrm{m}^-(i,j,t) > 0\}.$$

For any $i \in \mathcal{O}(t)^{\mathsf{c}}$, by the definition of the recommended set and since $Z_2(t) > 0$,

$$\exists j \in \mathbb{A} \text{ such that } \mathrm{m}(i,j) \overset{\mathcal{E}}{\geq} \mathrm{m}^-(i,j,t) > 0,$$

so i is a sub-optimal arm. Therefore,

$$\mathcal{S}^\star \subset \mathcal{O}(t).$$

Moreover, for any $i \in \mathcal{O}(t) \cap S(t)$, since $Z^{\varepsilon_1}_1(t) > 0$ we have $h^{\varepsilon_1}_i(t) > 0$, that is

$$\min_{j \in \mathbb{A}\setminus\{i\}} \mathrm{M}(i,j) + \varepsilon_1 \overset{\mathcal{E}}{\geq} \min_{j \in \mathbb{A}\setminus\{i\}} \mathrm{M}^-(i,j,t) + \varepsilon_1 > 0. \tag{6}$$

If $i \in \mathcal{O}(t) \cap S(t)^{\mathsf{c}}$, by definition of $\mathcal{O}(t)$, we have $g_i(t) < 0$. However, since $Z^{\varepsilon_1}_2(t) > 0$, $\max(g_i(t), h^{\varepsilon_1}_i(t)) > 0$ so we also have $h^{\varepsilon_1}_i(t) > 0$ and (6) applies. Thus, for any $i \in \mathcal{O}(t)$,

$$\min_{j \in \mathbb{A}\setminus\{i\}} \mathrm{M}(i,j) + \varepsilon_1 > 0,$$

that is $i \in \mathcal{S}^\star_{\varepsilon_1}$, so $\mathcal{S}^\star \subset \mathcal{O}(t) \subset \mathcal{S}^\star_{\varepsilon_1}$. Finally we can conclude that ε_1-APE-k and ε_1-APE output a correct subset on $\mathcal{E}$. □

B.3 Calibration of the confidence intervals

In Section 3.2 we proposed to instantiate our algorithms with the confidence intervals

$$U^d_{i,j}(t) = \widehat{\mu}^d_i(t) - \widehat{\mu}^d_j(t) + \beta_{i,j}(t) \text{ and } L^d_{i,j}(t) = \widehat{\mu}^d_i(t) - \widehat{\mu}^d_j(t) - \beta_{i,j}(t)\ . \tag{7}$$

We prove below that $\mathcal{E}$ is indeed a high-probability event for a suitable choice of $\beta_{i,j}(t)$.

Lemma 6. *Let ν be a bandit with 1-subgaussian marginals. For the confidence intervals defined in (7), with*

$$\beta_{i,j}(t) = 2\sqrt{\left(C^g\left(\frac{\log\left(\frac{K_1}{\delta}\right)}{2}\right) + \sum_{a \in \{i,j\}} \log(4 + \log(T_a(t)))\right)\left(\sum_{a \in \{i,j\}} \frac{1}{T_a(t)}\right)}.$$

the event

$$\mathcal{E} = \bigcap_{t=1}^{\infty} \mathcal{E}_t \quad \textit{with} \quad \mathcal{E}_t = \bigcap_{i=1}^{K} \bigcap_{j \neq i} \bigcap_{d=1}^{D} \{L^d_{i,j}(t) \leq \mu^d_i - \mu^d_j \leq U^d_{i,j}(t)\}$$

is such that $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$.

Proof. By observing that for any pair of arm $\beta_{i,j} = \beta_{j,i}$, $\mathcal{E}_t$ can be rewritten as

$$\mathcal{E}_t = \bigcap_{\{i,j\}\in\Gamma} \bigcap_{d=1}^{D} \left\{ L^d_{i,j}(t,\delta) \leq \mu^d_i - \mu^d_j \leq U^d_{i,j}(t,\delta) \right\}, \tag{8}$$

$$= \bigcap_{\{i,j\}\in\Gamma} \bigcap_{d=1}^{D} \left\{ |(\widehat{\mu}^d_i(t) - \widehat{\mu}^d_j(t)) - (\mu^d_i - \mu^d_j)| \leq \beta_{i,j}(t) \right\}, \tag{9}$$

□

where $\Gamma := \binom{2}{[K]}$ is the set of pair of 2 elements of $[K]$, which satisfies $|\Gamma| = K(K-1)/2$. Therefore, using a union bound,

$$\begin{aligned}
\mathbb{P}(\mathcal{E}^{\mathsf{c}}) &= \mathbb{P}(\exists t \geq 1 : \mathcal{E}^{\mathsf{c}}_t \text{ holds }), \\
&= \mathbb{P}\left(\exists t \geq 1, d \in [D], \{i,j\} \in \Gamma : |(\widehat{\mu}^d_i(t) - \widehat{\mu}^d_j(t)) - (\mu^d_i - \mu^d_j)| > \beta_{i,j}(t)\right), \\
&\leq \sum_{\{i,j\}\in\Gamma} \sum_{d=1}^{D} \mathbb{P}\left(\exists t \geq 1 : |(\widehat{\mu}^d_i(t) - \widehat{\mu}^d_j(t)) - (\mu^d_i - \mu^d_j)| > \beta_{i,j}(t)\right), \\
&\leq \sum_{\{i,j\}\in\Gamma} \sum_{d=1}^{D} \frac{\delta}{K_1} \quad \text{(by Proposition 24 of [20] which we recall below in Lemma 7)}, \\
&= \delta,
\end{aligned}$$

since $K_1 := \frac{K(K-1)D}{2}$ and $|\Gamma| = K(K-1)/2$.

Lemma 7 (Proposition 24 of [20])**.** *Let X, Y be centered $1-$subgaussian random variables and $\delta \in (0,1)$. Let $X, X_1, X_2, \ldots$ be i.i.d random variables and $Y, Y_1, Y_2, \ldots$ be i.i.d random variables. With probability at least $1-\delta$, for all $p, q \geq 1$,*

$$\left| \frac{1}{p}\sum_{s=1}^{p} X_s - \frac{1}{q}\sum_{s=1}^{q} Y_s \right| \leq 2\sqrt{\left(C^g\left(\frac{\log(1/\delta)}{2}\right) + \log\log(e^4 p) + \log\log(e^4 q) \right)\left(\frac{1}{p} + \frac{1}{q}\right)}$$

where $C^g(x) \approx x + \log(x)$.

C Sample complexity analysis

In this section we prove Theorem 1 which is restated below.

Theorem 1. *Fix a risk parameter $\delta \in (0,1)$, $\varepsilon_1 \geq 0$, let $k \leq K$ and ν a bandit with $1-$subgaussian marginals. With probability at least $1-\delta$, ε_1-APE-k recommends a correct set for the ε_1-PSI-k objective and stops after at most*

$$\sum_{a\in\mathbb{A}} \frac{88}{\tilde{\Delta}^2_a} \log\left(\frac{2K(K-1)D}{\delta} \log\left(\frac{12e}{\tilde{\Delta}_a} \right) \right), \tag{4}$$

samples, where for each $a \in \mathbb{A}$, $\tilde{\Delta}_a := \max(\Delta_a, \varepsilon_1, \omega^k)$.

Proof of Theorem 1. The correctness follows from Lemma 2 and the fact that $\mathbb{P}(\mathcal{E}) \geq 1-\delta$ (Lemma 6). The upper-bound on the sample complexity is a direct consequence of Lemma 3, Lemma 4, Lemma 5 which are proved later in this section. Indeed, using these lemmas we have that, if ε_1-APE -k has not stopped during round t i.e $t < \tau^k_{\varepsilon_1}$ and the event $\mathcal{E}_t$ holds, then

a) $\omega^k \leq 2\beta_{a_t,a_t}(t)$ (Lemma 3),

b) $\Delta_{a_t} \leq 2\beta_{a_t,a_t}(t)$(Lemma 4),

c) $\varepsilon_1 \leq 2\beta_{a_t,a_t}(t)$ (Lemma 5)

hold simultaneously. Then, if we do not count the first K rounds due to initialization, and letting $\tilde{\Delta}_a := \max(\omega^k, \varepsilon_1, \Delta_a)$,

$$\begin{aligned}
\tau^k_{\varepsilon_1} \mathbb{1}\{\mathcal{E}\} - 1 &\leq \sum_{t=1}^{\infty} \mathbb{1}\{\mathcal{E}\}\mathbb{1}\{\tau^k_{\varepsilon_1} > t\}, && (10)\\
&\leq \sum_{t=1}^{\infty} \mathbb{1}\{\max(\omega^k, \varepsilon_1, \Delta_{a_t}) \leq 2\beta_{a_t,a_t}(t)\} && (11)\\
&= \sum_{t=1}^{\infty} \mathbb{1}\{\tilde{\Delta}_{a_t} \leq 2\beta_{a_t,a_t}(t)\} && (12)\\
&= \sum_{t=1}^{\infty}\sum_{a=1}^{K} \mathbb{1}\{\{a_t = a\} \wedge \{\tilde{\Delta}_a \leq 2\beta_{a,a}(t)\}\} && (13)\\
&= \sum_{a=1}^{K}\sum_{t=1}^{\infty} \mathbb{1}\{\{a_t = a\} \wedge \{\tilde{\Delta}_a \leq 2\beta_{a,a}(t)\}\} && (14)\\
&\leq \sum_{a=1}^{K} \inf\{n \geq 2 : \tilde{\Delta}_a > 2\beta^n\}, && (15)
\end{aligned}$$

where β^n is the expression of $\beta_{i,j}(t)$ when $T_i(t) = T_j(t) = n$, that is

$$\beta^n = 2\sqrt{\left(C^g\left(\frac{\log\left(\frac{K_1}{\delta}\right)}{2}\right) + 2\log(4 + \log(n))\right)\frac{2}{n}}\,.$$

Then, an inversion result given in Lemma 19 yields

$$\inf\{s \geq 2 : 2\beta^s < \tilde{\Delta}_a\} \leq \frac{88}{\tilde{\Delta}_a^2}\log\left(\frac{2K(K-1)D}{\delta}\log\left(\frac{12e}{\tilde{\Delta}_a}\right)\right).$$

Therefore,

$$\tau^k_{\varepsilon_1}\mathbb{1}\{\mathcal{E}\} \leq \sum_{a\in\mathbb{A}} \frac{88}{\tilde{\Delta}_a^2}\log\left(\frac{2K(K-1)D}{\delta}\log\left(\frac{12e}{\tilde{\Delta}_a}\right)\right). \tag{16}$$

□

We will now prove the lemmas involved in the proof of the main theorem. Two of them (Lemma 4, Lemma 5) rely on the following result, which is an important consequence of the definition of the APE sampling rule.

Lemma 8. *If $t < \min(\tau_{\varepsilon_1}, \tau^k_{\varepsilon_1}, \tau_{\varepsilon_1,\varepsilon_2})$ then for any $j \in \mathbb{A}$, $\mathrm{m}(b_t, j, t) \leq \beta_{b_t,j}(t)$.*

Proof. The proof is split into two steps.

Step 1 If $t < \min(\tau^k_{\varepsilon_1}, \tau_{\varepsilon_1})$ then for any $j \in \mathbb{A}$, $\mathrm{m}(b_t, j, t) \leq \beta_{b_t,j}(t)$.

First, note that $t < \min(\tau^k_{\varepsilon_1}, \tau_{\varepsilon_1})$ implies that $Z_1^{\varepsilon_1}(t) \leq 0$ or $Z_2^{\varepsilon_1}(t) \leq 0$. By definition of b_t and noting that $\mathrm{M}(i,j) = -\mathrm{m}(i,j)$, we have

$$b_t \in \operatorname*{argmin}_{i\in\mathrm{OPT}^{\varepsilon_1}(t)^{\mathsf{c}}} \max_{j\neq i} \mathrm{m}(i,j,t) - \beta_{i,j}(t). \tag{17}$$

so that if there exists j such that $\mathrm{m}(b_t, j, t) > \beta_{b_t,j}(t)$, then

$$\max_{j\neq b_t} \mathrm{m}(b_t, j, t) - \beta_{b_t,j}(t) > 0,$$

therefore,

$$\forall i \in \mathrm{OPT}^{\varepsilon_1}(t)^{\mathsf{c}}\,, \max_{j\neq i} \mathrm{m}(i,j,t) - \beta_{i,j}(t) > 0 \quad \text{i.e} \quad g_i(t) > 0. \tag{18}$$

Furthermore, for any $i \in \mathrm{OPT}^{\varepsilon_1}(t)$, $h_i^{\varepsilon_1}(t) > 0$. Putting things together, if there exists j such that $\mathrm{m}(b_t, j, t) > \beta_{b_t,j}(t)$ then, $Z_1^{\varepsilon_1}(t) > 0$ and $Z_2^{\varepsilon_1}(t) > 0$.

Step 2 If $t < \tau_{\varepsilon_1,\varepsilon_2}$ then for any $j \in \mathbb{A}$, $\mathrm{m}(b_t, j, t) \leq \beta_{b_t,j}(t)$.

Recall that by definition $t < \tau_{\varepsilon_1,\varepsilon_2}$ implies that $Z_1^{\varepsilon_1,\varepsilon_2}(t) \leq 0$ or $Z_2^{\varepsilon_1,\varepsilon_2}(t) \leq 0$. Using (17), if there exists j such that $\mathrm{m}(b_t, j, t) > \beta_{b_t,j}(t)$, then

$$\max_{j \neq b_t} \mathrm{m}(b_t, j, t) - \beta_{b_t,j}(t) > 0.$$

Combining this with

$$g_i^{\varepsilon_2}(t) := \max_{j \in \mathbb{A}\setminus\{i\}} \mathrm{m}(i, j, t) - \beta_{i,j}(t) + \varepsilon_2 \mathbb{1}\{j \in \mathrm{OPT}^{\varepsilon_1}(t)\},$$

yields

$$\forall i \in \mathrm{OPT}^{\varepsilon_1}(t)^{\mathsf{c}}\ , 0 < \max_{j \neq i} \mathrm{m}(i, j, t) - \beta_{i,j}(t) \leq g_i^{\varepsilon_2}(t). \tag{19}$$

Furthermore, since we have

$$\forall i \in \mathrm{OPT}^{\varepsilon_1}(t)\ , h_i^{\varepsilon_1}(t) > 0, \tag{20}$$

the initial assumption would yield that for any arm i, $\max(h_i^{\varepsilon_1}(t), g_i^{\varepsilon_2}(t)) > 0$, so $Z_1^{\varepsilon_1,\varepsilon_2}(t) > 0$ and $Z_2^{\varepsilon_1,\varepsilon_2}(t) > 0$. Therefore, if $t < \min(\tau_{\varepsilon_1}, \tau_{\varepsilon_1}^k, \tau_{\varepsilon_1,\varepsilon_2})$ then for any $i \in \mathbb{A}, \mathrm{m}(b_t, j, t) \leq \beta_{b_t,j}(t)$. □

C.1 Proof of Lemma 3

Lemma 3. *Let $k \leq K$. If $\mathcal{E}_t$ holds and $t < \tau_{\varepsilon_1}^k$ then $\omega^k \leq 2\beta_{a_t,a_t}(t)$.*

Proof of Lemma 3. First, note that if $k > |\mathcal{S}^\star|$, then the lemma holds trivially since $\omega_k < 0$. In the sequel, we assume $\mathcal{E}_t$ holds and $k \leq |\mathcal{S}^\star|$. If $t < \tau_{\varepsilon_1}^k$ then it holds that $|\mathrm{OPT}^{\varepsilon_1}(t)| < k$. So $\mathcal{S}^{\star,k} \cap \mathrm{OPT}^{\varepsilon_1}(t)^{\mathsf{c}} \neq \emptyset$. Let $i \in \mathcal{S}^{\star,k} \cap \mathrm{OPT}^{\varepsilon_1}(t)^{\mathsf{c}}$, we have

$$\begin{aligned}
\omega^k &\leq \omega_i = \min_{j \in \mathbb{A}\setminus\{i\}} \mathrm{M}(i, j), \\
&\overset{(a)}{\leq} \min_{j \in \mathbb{A}\setminus\{i\}} \mathrm{M}(i, j, t) + \beta_{i,j}(t), \\
&\overset{(b)}{\leq} \min_{j \in \mathbb{A}\setminus\{b_t\}} \mathrm{M}(b_t, j, t) + \beta_{b_t,j}(t), \\
&\leq \mathrm{M}(b_t, c_t, t) + \beta_{b_t,c_t}(t), \\
&\overset{(c)}{\leq} 2\beta_{b_t,c_t}(t), \\
&\leq 2\beta_{a_t,a_t}(t),
\end{aligned}$$

where (a) uses that $\mathcal{E}_t$ holds and Lemma 1, (b) uses the definition of b_t and (c) follows from the definition of c_t and the fact that $b_t \notin \mathrm{OPT}^{\varepsilon_1}(t)$,which yields $\mathrm{M}(b_t, c_t, t) \leq \beta_{b_t,c_t}(t)$. The last inequality follows since a_t is the least sampled among b_t, c_t and β is decreasing. □

C.2 Proof of Lemma 4

Lemma 4. *If $\mathcal{E}_t$ holds and $t < \min(\tau_{\varepsilon_1}, \tau_{\varepsilon_1}^k, \tau_{\varepsilon_1,\varepsilon_2})$ then $\Delta_{a_t} \leq 2\beta_{a_t,a_t}(t)$.*

Before proving the Lemma 4, we state the following lemma which is used to derive an upper bound on the gap of an optimal arm. Its proof is postponed to the end of the section.

Lemma 9. *For any Pareto optimal arm i, $\Delta_i \leq \min_{j \neq i} \mathrm{M}(i, j)$.*

Proof of Lemma 4. Assume that $\mathcal{E}_t$ holds. We consider four different cases depending on whether b_t and c_t are optimal or sub-optimal.

Case 1.1 b_t is a Pareto optimal arm. From the definition of the gap of an optimal arm and using Lemma 9 it follows $\Delta_{b_t} \leq \mathrm{M}(b_t, c_t)$ which on $\mathcal{E}_t$ and using Lemma 1 yields

$$\Delta_{b_t} + \varepsilon_1 \leq \mathrm{M}(b_t, c_t, t) + \beta_{b_t,c_t}(t) + \varepsilon_1 \tag{21}$$

then, noting that there exists $j \in \mathbb{A} \setminus \{b_t\}$ such that $\mathrm{M}(b_t, j, t) + \varepsilon_1 \leq \beta_{b_t,j}(t)$, by definition of c_t, we have

$$\mathrm{M}(b_t, c_t, t) + \varepsilon_1 \leq \beta_{b_t,c_t}(t), \tag{22}$$

therefore,

$$\Delta_{b_t} + \varepsilon_1 \quad \leq \quad 2\beta_{b_t,c_t}(t).$$

Case 1.2 b_t is a sub-optimal arm. By definition of c_t and using $\mathrm{M} = -\mathrm{m}$, we have

$$c_t \in \operatorname*{argmax}_{j \in \mathbb{A} \setminus \{b_t\}} \mathrm{m}(b_t, j, t) + \beta_{b_t,j}(t), \tag{23}$$

then, from the definition of the gap of a sub-optimal arm and since $\mathcal{E}_t$ holds, we know that there exists an arm $b_t^\star$ such that

$$\begin{aligned}
\Delta_{b_t} = \mathrm{m}(b_t, b_t^\star) \quad &\leq \quad \mathrm{m}(b_t, b_t^\star, t) + \beta_{b_t,b_t^\star}(t), \\
&\overset{(a)}{\leq} \quad \mathrm{m}(b_t, c_t, t) + \beta_{b_t,c_t}(t), \\
&\overset{(b)}{\leq} \quad 2\beta_{b_t,c_t}(t).
\end{aligned}$$

where (a) uses the definition of c_t and (b) uses Lemma 8.

Case 2.1 c_t is a Pareto optimal arm. If b_t is also an optimal arm, it follows that $\Delta_{c_t} \leq \mathrm{M}(b_t, c_t)$ which on $\mathcal{E}_t$ yields $\Delta_{c_t} \leq \mathrm{M}(b_t, c_t, t) + \beta_{b_t,c_t}(t)$, then, similarly to case 1.1, we have $\mathrm{M}(b_t, j, t) + \varepsilon_1 \leq \beta_{b_t,j}(t)$ so

$$\Delta_{c_t} + \varepsilon_1 \leq 2\beta_{b_t,c_t}(t).$$

Now, assume b_t is a sub-optimal arm. Then, by definition, $\Delta_{c_t} \leq \mathrm{M}(b_t, c_t)^+ + \Delta_{b_t}$. Using a similar reasoning to case 1.2, it holds that $\Delta_{b_t} \leq \mathrm{m}(b_t, c_t, t) + \beta_{b_t,c_t}(t)$, so

$$\begin{aligned}
\Delta_{c_t} \quad &\leq \quad \mathrm{M}(b_t, c_t)^+ + \Delta_{b_t}, \\
&\leq \quad (\mathrm{M}(b_t, c_t, t) + \beta_{b_t,c_t}(t))^+ + \mathrm{m}(b_t, c_t, t) + \beta_{b_t,c_t}(t), \\
&= \quad (-\mathrm{m}(b_t, c_t, t) + \beta_{b_t,c_t}(t))^+ + \mathrm{m}(b_t, c_t, t) + \beta_{b_t,c_t}(t), \\
&\overset{(a)}{\leq} \quad \max(2\beta_{b_t,c_t}(t), \mathrm{m}(b_t, c_t, t) + \beta_{b_t,c_t}(t)) \\
&\overset{(b)}{\leq} \quad 2\beta_{b_t,c_t}(t).
\end{aligned}$$

where (a) follows from $(x-y)^+ + (x+y) \leq \max(x+y, 2x)$ and (b) follows from $\mathrm{m}(b_t, c_t, t) \leq \beta_{b_t,c_t}(t)$ (Lemma 8).

Case 2.2 c_t is a sub-optimal arm. We know that there exists an arm $c_t^\star$ such that $\Delta_{c_t} = \mathrm{m}(c_t, c_t^\star)$. If $c_t^\star = b_t$ then, since $\mathrm{m}(j, i) \leq \mathrm{M}(i, j)$ (follows from the definition), we have

$$\begin{aligned}
\Delta_{c_t} = \mathrm{m}(c_t, c_t^\star) \quad &= \quad \mathrm{m}(c_t, b_t), \\
&\leq \quad \mathrm{M}(b_t, c_t), \\
&\overset{(a)}{\leq} \quad \mathrm{M}(b_t, c_t, t) + \beta_{b_t,c_t}(t), \\
&\overset{(b)}{\leq} \quad 2\beta_{b_t,c_t}(t),
\end{aligned}$$

where (a) follows from $\mathcal{E}_t$ and (b) has been already justified in the case 1.1. If $b_t \neq c_t^\star$, then by definition of c_t, we have

$$\mathrm{m}(b_t, c_t, t) + \beta_{b_t,c_t}(t) \geq \mathrm{m}(b_t, c_t^\star, t) + \beta_{b_t,c_t^\star}(t),$$

which implies that there exists $d \in [D]$ such that

$$\widehat{\mu}_{c_t}^d(t) - \widehat{\mu}_{b_t}^d(t) + \beta_{b_t,c_t}(t) \geq \widehat{\mu}_{c_t^\star}^d(t) - \widehat{\mu}_{b_t}^d(t) + \beta_{b_t,c_t^\star}(t) \overset{\mathcal{E}_t}{\geq} \mu_{c_t^\star}^d - \mu_{b_t}^d,$$

then recalling that $\beta_{i,j} = \beta_{j,i}$,

$$\mu^d_{c_t} - \mu^d_{b_t} + 2\beta_{b_t,c_t}(t) \overset{\mathcal{E}_t}{\geq} (\widehat{\mu}^d_{c_t}(t) - \widehat{\mu}^d_{b_t}(t) - \beta_{b_t,c_t}(t)) + 2\beta_{b_t,c_t}(t) \geq \mu^d_{c^\star_t} - \mu^d_{b_t}.$$

Put together, there exists $d \in [D]$ such that

$$\mu^d_{c^\star_t} - \mu^d_{c_t} \leq 2\beta_{b_t,c_t}(t),$$

so

$$\Delta_{c_t} = \min_d(\mu^d_{c^\star_t} - \mu^d_{c_t}) \leq 2\beta_{b_t,c_t}(t),$$

Putting the four case together, we have proved that if $t < \min(\tau^k_{\varepsilon_1}, \tau_{\varepsilon_1,\varepsilon_2})$ then both

$$\Delta_{b_t} \leq 2\beta_{b_t,c_t}(t) \quad \text{and} \quad \Delta_{c_t} \leq 2\beta_{b_t,c_t}(t) \tag{24}$$

holds. Further noting that a_t is the least sampled among among b_t, c_t and β is non-increasing, $\beta_{b_t,c_t}(t) \leq \beta_{a_t,a_t}(t)$, (24) yields

$$\Delta_{a_t} \leq 2\beta_{a_t,a_t}(t),$$

which achieves the proof. □

The following lemma holds for each of the stopping times $\tau_{\varepsilon_1}, \tau_{\varepsilon_1,\varepsilon_2}$ and $\tau^k_{\varepsilon_1}$.

C.3 Proof of Lemma 5

Lemma 5. *If the algorithm has not stopped at the end of round t then $\varepsilon_1 \leq 2\beta_{a_t,a_t}(t)$.*

Proof of Lemma 5. By Lemma 8, we have $\mathrm{m}(b_t, c_t, t) \leq \beta_{b_t,c_t}(t)$ or equivalently

$$\mathrm{M}(b_t, c_t, t) \geq -\beta_{b_t,c_t}(t)\ . \tag{25}$$

Then, knowing that $b_t \notin \mathrm{OPT}^{\varepsilon_1}(t)$, there exists an arm j such that $\varepsilon_1 + \mathrm{M}(b_t, j, t) \leq \beta_{b_t,j}(t)$. Using further the definition of c_t, it follows that $\varepsilon_1 + M(b_t, c_t, t) \leq \beta_{b_t,c_t}(t)$. Combining this with inequality (25) and noting that a_t is the least sampled among b_t, c_t yields

$$\beta_{a_t,a_t}(t) \geq \beta_{b_t,c_t}(t) \geq \varepsilon_1/2.$$

□

C.4 Auxiliary results

We state the following lemma which is used to prove Lemma 9.

Lemma 10. *For any sub-optimal arm a, there exists a Pareto optimal arm $a^\star$ such that $\boldsymbol{\mu}_a \prec \boldsymbol{\mu}_{a^\star}$ and $\Delta_a = \mathrm{m}(a, a^\star) > 0$. Moreover, For any $i \in \mathbb{A} \setminus \mathcal{S}^\star$, $j \in \mathcal{S}^\star$,*

i) $\max_{j\in\mathcal{S}^\star} \mathrm{m}(i, j) = \max_{j\in\mathbb{A}} \mathrm{m}(i, j)$,

ii) If $i \in \mathrm{argmin}_{a\in\mathbb{A}\setminus\{j\}} \mathrm{M}(j, a)$ then j is the unique arm such that $\boldsymbol{\mu}_i \prec \boldsymbol{\mu}_j$

Proof. Suppose that there are $p < n$ dominated vectors. Without loss of generality, we may assume they are $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_p$. Let $i_1 \leq p$. Suppose that no Pareto-optimal arm dominates $\boldsymbol{\mu}_{i_1}$. Since $\boldsymbol{\mu}_{i_1}$ is not optimal, by the latter assumption, there exists $i_2 \leq p$ such that $\boldsymbol{\mu}_{i_1} \prec \boldsymbol{\mu}_{i_2}$. If $\boldsymbol{\mu}_{i_2}$ is dominated by a Pareto optimal arm, this arm also dominates $\boldsymbol{\mu}_{i_1}$ (strict dominance is transitive) which contradicts the initial assumption. If not, there exits $i_3 \leq p$ such that $\boldsymbol{\mu}_{i_1} \prec \boldsymbol{\mu}_{i_2} \prec \boldsymbol{\mu}_{i_3}$. Again we can use the same reasoning as before for i_3. In any case we should stop in at most p steps, otherwise we would have $\boldsymbol{\mu}_{i_1} \prec \boldsymbol{\mu}_{i_2} \prec \cdots \prec \boldsymbol{\mu}_{i_p}$ and $\boldsymbol{\mu}_{i_p}$ should be dominated by a Pareto-optimal arm, otherwise it would be itself Pareto-optimal, which is not the case. Therefore, for any $a \in \mathbb{A} \setminus \mathcal{S}^\star$, there exists $a^\star \in \mathcal{S}^\star$ such that $a^\star \prec a$ and $\Delta_a = \mathrm{m}(a, a^\star) > 0$.

Letting i be a sub-optimal arm, since for any $a \in \mathbb{A} \setminus \mathcal{S}^\star$, there exists $a^\star \in \mathcal{S}^\star$ such that $a \prec a^\star$, it follows that

$$\forall d \in [D],\ \mu^d_a - \mu^d_i < \mu^d_{a^\star} - \mu^d_i,$$

which leads to $\mathrm{m}(i,a) \leq \mathrm{m}(i,a^\star)$, so

$$\max_{j\in\mathbb{A}} \mathrm{m}(i,j) = \max_{j\in\mathcal{S}^\star} \mathrm{m}(i,j) > 0,$$

which achieves the proof of the first point *i)*. For the second point, let $q \in \mathbb{A} \setminus \mathcal{S}^\star$ and $q' \neq q$ such that $q \prec q'$ and

$$q \in \operatorname*{argmin}_{a\in\mathbb{A}\setminus\{j\}} \mathrm{M}(j,a).$$

By direct algebra, since $q \prec q'$, we have

$$\mathrm{M}(j,q') < \mathrm{M}(j,q),$$

which is impossible if $q' \neq j$ (because q belongs to the argmin). Therefore, if

$$q \in \operatorname*{argmin}_{a\in\mathbb{A}\setminus\{j\}} \mathrm{M}(j,a)$$

is a sub-optimal arm, then j is the only arm such that $q \prec j$ (i.e $\boldsymbol{\mu}_q \prec \boldsymbol{\mu}_j$). □

We now prove Lemma 9 which follows from the previous lemma.

Proof of Lemma 9. If $\operatorname{argmin}_{j\neq i} \mathrm{M}(i,j) \subset \mathcal{S}^\star$, then the lemma follows from the definition of the gap of an optimal arm recalled in Section 4. If $\min_{j\neq i} \mathrm{M}(i,j) = \mathrm{M}(i,a)$, $a \notin \mathcal{S}^\star$, then, from Lemma 10, i is the unique arm which dominates a so $\Delta_a = \mathrm{m}(a,i)$ and using the definition of the gap of an optimal arm,

$$\begin{aligned}\Delta_i &\leq \mathrm{M}(a,i)^+ + \Delta_a,\\ &= 0 + \mathrm{m}(a,i) \leq \mathrm{M}(i,a),\end{aligned}$$

where we have used the the fact that $\mathrm{m}(p,q) \leq \mathrm{M}(q,p)$ for any pair of arms p,q (which follows from the definition). Therefore, for an optimal arm i, we always have

$$\Delta_i \leq \min_{j\neq i} \mathrm{M}(i,j).$$

□

D Algorithm for finding an $(\varepsilon_1,\varepsilon_2)$-cover

In this section, we analyse the sample complexity of APE when it is associated to the stopping time $\tau_{\varepsilon_1,\varepsilon_2}$ for identifying an $(\varepsilon_1,\varepsilon_2)-$cover of the Pareto set. The sampling rule remains unchanged an we prove that the algorithm does not require more samples to find an $(\varepsilon_1,\varepsilon_2)$-cover than to solve the ε_1-PSI problem.

Algorithm 2: $(\varepsilon_1,\varepsilon_2)$-APE

Data: parameters $\varepsilon_1 \geq 0$
initialize : sample each arm once, set $t = K$, $T_i(t) = 1$ for any $i \in \mathbb{A}$
for $t = K, K+1, \ldots$ **do**
  $b_t := \operatorname{argmax}_{i\in\mathbb{A}\setminus\mathrm{OPT}^{\varepsilon_1}(t)} \min_{j\neq i} \mathrm{M}^+(i,j,t)$;
  $c_t := \operatorname{argmin}_{j\neq b_t} \mathrm{M}^-(b_t,j,t)$;
  if $Z_1^{\varepsilon_1,\varepsilon_2}(t) > 0 \wedge Z_2^{\varepsilon_1,\varepsilon_2}(t) > 0$ **then**
    break and output $\mathrm{OPT}^{\varepsilon_1}(t)$;
  sample $a_t := \operatorname{argmin}_{i\in\{b_t,c_t\}} T_i(t)$;

We recall the stopping time $\tau_{\varepsilon_1,\varepsilon_2}$.

Stopping rule Let $\varepsilon_1, \varepsilon_2 \geq 0$ and $0 < \delta < 1$. Then, by ignoring the first K rounds of initialization,

$$\tau_{\varepsilon_1,\varepsilon_2} := \inf\{t \in \mathbb{N}^\star : \; Z_1^{\varepsilon_1,\varepsilon_2}(t) > 0 \; \wedge Z_2^{\varepsilon_1,\varepsilon_2}(t) > 0\}, \tag{26}$$

where,

$$\begin{aligned} Z_1^{\varepsilon_1,\varepsilon_2}(t) &:= \min_{i \in S(t)} \max(g_i^{\varepsilon_2}(t), h_i^{\varepsilon_1}(t)) \\ Z_2^{\varepsilon_1,\varepsilon_2}(t) &:= \min_{i \in S(t)^\mathsf{c}} \max(g_i^{\varepsilon_2}(t), h_i^{\varepsilon_1}(t)), \end{aligned}$$

and

$$\begin{aligned} g_i^{\varepsilon_2}(t) &:= \max_{j \in \mathbb{A}\setminus\{i\}} \mathrm{m}^-(i,j,t) + \varepsilon_2 \mathbb{1}\{j \in \mathrm{OPT}^{\varepsilon_1}(t)\} \\ h_i^{\varepsilon_1}(t) &:= \min_{j \in \mathbb{A}\setminus\{i\}} \mathrm{M}^-(i,j,t) + \varepsilon_1 \end{aligned}$$

Recommendation rule When it is associated to the stopping time $\tau_{\varepsilon_1,\varepsilon_2}$, APE recommends

$$\mathcal{O}(\tau_{\varepsilon_1,\varepsilon_2}) := \mathrm{OPT}^{\varepsilon_1}(\tau_{\varepsilon_1,\varepsilon_2}),$$

which can be understood as follows. When $\tau_{\varepsilon_1,\varepsilon_2}$ is reached, the arms that are not yet identified as (nearly) optimal are either ε_2-dominated by an arm in $\mathrm{OPT}^{\varepsilon_1}(\tau_{\varepsilon_1,\varepsilon_2})$ or sub-optimal, which is proven formally in Lemma 11.

Lemma 11. *Fix $\delta \in (0,1), \varepsilon_1, \varepsilon_2 \geq 0$ then $(\varepsilon_1, \varepsilon_2)$-APE recommends an $(\varepsilon_1, \varepsilon_2)$-cover of the Pareto set on the event $\mathcal{E}$.*

Proof of Lemma 11. Assume $\mathcal{E}$ holds. Let $t = \tau_{\varepsilon_1,\varepsilon_2}$ and $i \in \mathrm{OPT}^{\varepsilon_1}(t)$. Since $i \in \mathrm{OPT}^{\varepsilon_1}(t)$, for any $j \neq i$, $\mathrm{M}(i,j) + \varepsilon_1 \overset{\mathcal{E}}{\geq} \mathrm{M}^-(i,j,t) + \varepsilon_1 > 0$ that is $i \in \mathcal{S}^\star_{\varepsilon_1}$. Therefore, on the event $\mathcal{E}$, $\mathrm{OPT}^{\varepsilon_1}(t) \subset \mathcal{S}^\star_{\varepsilon_1}$. When the stopping time $\tau_\delta^{\varepsilon_1,\varepsilon_2}$ is reached, $Z_1^{\varepsilon_1,\varepsilon_2}(t) > 0$ and $Z_2^{\varepsilon_1,\varepsilon_2}(t) > 0$. Under this condition,

$$\mathrm{OPT}^{\varepsilon_1}(t) \neq \emptyset.$$

Indeed, since $Z_1^{\varepsilon_1,\varepsilon_2}(t) > 0$ and $Z_2^{\varepsilon_1,\varepsilon_2}(t) > 0$, if $\mathrm{OPT}^{\varepsilon_1}(t) = \emptyset$ then, by the stopping rule and since $\mathrm{OPT}^{\varepsilon_1}(t) = \emptyset$, for any arm i, we would have $h_i^{\varepsilon_1}(t) < 0$ and $g_i^{\varepsilon_2}(t) > 0$. That is, for any arm $i \in \mathbb{A}$,

$$\exists j \neq i \text{ such that } \mathrm{m}(i,j) \overset{\mathcal{E}}{>} \mathrm{m}^-(i,j,t) > 0,$$

so every arm would be strictly dominated, which is impossible since the Pareto set cannot be empty. Then, $\mathrm{OPT}^{\varepsilon_1}(t) \neq \emptyset$ and for any $i \in \mathcal{O}(t)^\mathsf{c} = \mathrm{OPT}^{\varepsilon_1}(t)^\mathsf{c}$, by the stopping rule it holds that $\max(g_i^{\varepsilon_2}(t), h_i^{\varepsilon_1}(t)) > 0$. Further noting that for such arm $i \in \mathrm{OPT}^{\varepsilon_1}(t)^\mathsf{c}$, $h_i^{\varepsilon_1}(t) < 0$, we thus have $g_i^{\varepsilon_2}(t) > 0$, that is

$$\mathrm{m}^-(i,j,t) + \varepsilon_2 \mathbb{1}\{j \in \mathcal{O}(t)\} > 0,$$

which on the event $\mathcal{E}$ yields

$$\mathrm{m}(i,j) + \varepsilon_2 \mathbb{1}\{j \in \mathcal{O}(t)\} > 0.$$

Therefore, for such arm i, either

i) $\exists j \in \mathbb{A}$ such that $\mathrm{m}(i,j) > 0$ that is $\boldsymbol{\mu}_i \prec \boldsymbol{\mu}_j$ or

ii) $\exists j \in \mathcal{O}(t)$ such that $\mathrm{m}(i,j) + \varepsilon_2 > 0$ that is $\boldsymbol{\mu}_i \prec \boldsymbol{\mu}_j + \boldsymbol{\varepsilon_2}$ with $\boldsymbol{\varepsilon_2} := (\varepsilon_2, \ldots, \varepsilon_2)$.

Put together, $\mathcal{O}(t) \subset \mathcal{S}^\star_{\varepsilon_1}$ and for any $i \notin \mathcal{O}(t)$, either $i \notin \mathcal{S}^\star$ (i is a sub-optimal arm) or there exists $j \in \mathcal{O}(t)$ such that $\boldsymbol{\mu}_i \prec \boldsymbol{\mu}_j + \boldsymbol{\varepsilon_2}$. Thus $\mathcal{O}(t)$ is an $(\varepsilon_1, \varepsilon_2)$-cover of the Pareto set and $(\varepsilon_1, \varepsilon_2)$-APE is correct for $(\varepsilon_1, \varepsilon_2)$-cover identification. □

Put together, the two lemmas restated below are used to prove identically to Theorem 1, the main theorem of this section.

Lemma 4. *If $\mathcal{E}_t$ holds and $t < \min(\tau_{\varepsilon_1}, \tau_{\varepsilon_1}^k, \tau_{\varepsilon_1,\varepsilon_2})$ then $\Delta_{a_t} \leq 2\beta_{a_t,a_t}(t)$.*

The following lemma holds for each of the stopping times $\tau_{\varepsilon_1}, \tau_{\varepsilon_1,\varepsilon_2}$ and $\tau_{\varepsilon_1}^k$.

Lemma 5. *If the algorithm has not stopped at the end of round t then* $\varepsilon_1 \leq 2\beta_{a_t,a_t}(t)$.

Theorem 2. *Fix* $\delta \in (0,1)$, $\varepsilon_1, \varepsilon_2 \geq 0$. *Then* $(\varepsilon_1, \varepsilon_2)$*-APE outputs an* $(\varepsilon_1, \varepsilon_2)$*-cover of the Pareto set with probability at least* $1-\delta$ *using at most*

$$\sum_{a \in \mathbb{A}} \frac{88}{(\Delta_a^\varepsilon)^2} \log\left(\frac{2K(K-1)D}{\delta} \log\left(\frac{12e}{\Delta_a^\varepsilon}\right)\right) \tag{27}$$

samples, where for all $a \in \mathbb{A}$, $\Delta_a^\varepsilon := \max(\Delta_a, \varepsilon_1)$.

This is the first problem-dependent sample complexity upper-bound for the $(\varepsilon_1, \varepsilon_2)$-cover of the Pareto set. In particular, this result holds for the ε-accurate Pareto set identification [7] which corresponds to the particular case $\varepsilon_1 = \varepsilon_2 = \varepsilon$ of the Pareto set cover. Therefore, $(\varepsilon, \varepsilon)$-APE could be compared to ε-PAL for ε-accurate Pareto set, which however relies on a different Gaussian process modelling assumption.

While this sample complexity result upper bound does not clearly show the dependence in ε_2, we note that for some problems, we have a nearly matching lower bound that does not depend on ε_2. In particular, consider the case $D = 1, \varepsilon_1 = 0, \varepsilon_2 > 0$ and assume there is a unique best arm (classical assumption) $a_\star$. For this setting, an algorithm for $(\varepsilon_1, \varepsilon_2)$-cover identification is required to output a set $\hat{S}$ such that $\hat{S} \subset \mathcal{S}^\star = \{a_\star\}$ and for any $i \neq a_\star$ either $\mu_i < \mu_{a_\star}$ or $\mu_i \leq \mu_{a_\star} + \varepsilon_2$ which trivially holds as long as $\hat{S} \subset \mathcal{S}^\star$. Therefore, this problem is equivalent to (exact) Best Arm Identification. Almost matched lower bounds for BAI are known and does not depend on ε_2 ([22, 23, 4]). This observation can be generalized to any configuration where there is a unique (Pareto) optimal arm. Letting $D \geq 1, \varepsilon_1 = 0, \varepsilon_2 > 0$ and ν a bandit with one Pareto optimal arm $a_\star$, any algorithm for $(\varepsilon_1, \varepsilon_2)$-covering is required to output a set $\hat{S} \subset \mathcal{S}^\star = \{a_\star\}$. And for any $i \neq a_\star$ either $\boldsymbol{\mu}_i \prec \boldsymbol{\mu}_{a_\star}$ or $\boldsymbol{\mu}_i \prec \boldsymbol{\mu}_{a_\star} + \varepsilon_2$ which trivially holds as long as $\hat{S} \subset \mathcal{S}^\star = \{a_\star\}$. So, on theses instances, $(0, \varepsilon_2)$-covering is equivalent 0-PSI and the nearly matched lower of [6] for 0-PSI does not depend on ε_2 (Theorem 17 therein).

In our experiments (see subsection I.3), we will see that in configurations with multiple Pareto optimal arms, the parameter ε_2 can still help to empirically reduce the sample complexity. Quantifying its precise impact on the sample complexity is left as future work.

E Lower Bound

In this section, we give a gap-dependent lower-bound for the $k-$relaxation in some configurations. We use the change of distribution lemma of [22] (lemma 1).

Theorem 3. *There exists a bandit instance* ν *with* $|\mathcal{S}^\star| = p \geq 3$ *such that for* $k \in \{p, p-1, p-2\}$ *any* δ*-correct algorithm for* 0*-PSI-*k *verifies*

$$\mathbb{E}_\nu(\tau_\delta) \geq \frac{1}{D} \log\left(\frac{1}{\delta}\right) \sum_{a=1}^{K} \frac{1}{(\Delta_a^k)^2},$$

where $\Delta_a^k := \Delta_a + \omega^k$ *and* τ_δ *is the stopping time of the algorithm.*

Proof. Let $p = K = |\mathcal{S}^\star|$. w.l.o.g assume $\mathcal{S}^\star = \{1, \ldots, p\}$ and $\mathcal{S}^{\star,k} = \{1, \ldots, k\}$. Let $\boldsymbol{\mu}_0 \in \mathbb{R}^D$ and for $(p-2) \leq i \leq p$, define

$$\mu_i^d := \begin{cases} 2^{p-i}\omega & \text{if d = 1} \\ -2^{p-i}\omega & \text{else if d = 2} \\ \mu_0^d & \text{else.} \end{cases},$$

for $1 \leq i \leq p-3$,

$$\mu_i^d := \begin{cases} (4+2i)\omega & \text{if d = 1} \\ -(4+2i)\omega & \text{else if d = 2} \\ \mu_0^d & \text{else.} \end{cases}$$

Let ν be a bandit where each arm i is a multivariate Gaussian with mean $\boldsymbol{\mu}_i$ and covariance matrix I_D i.e $\nu_i \sim \mathcal{N}(\boldsymbol{\mu}_i, I_D)$ (with I_D the identity matrix in dimension D). By direct calculation, for $1 \leq i, j \leq p-3$,

$$\mathrm{M}(i,j) = \mathrm{M}(j,i) = 2\omega|i-j|,$$

and for $p-2 \leq i, j \leq p$,

$$\mathrm{M}(i,j) = \mathrm{M}(j,i) = 2^p\omega|2^{-i} - 2^{-j}|,$$

for $i \leq p-3$ and $(p-2) \leq j \leq p$,

$$\mathrm{M}(i,j) = \mathrm{M}(j,i) = (4 + 2i - 2^{p-j})\omega \geq 2\omega.$$

Therefore, computing ω_i and δ_i^+ for any $i \in [p]$ yields

$$\begin{aligned} \delta_i^+ &:= \min_{j \in [p] \setminus \{i\}} \min(\mathrm{M}(i,j), \mathrm{M}(j,i)), \\ &= \begin{cases} \omega & \text{if } i = p, \\ 2^{p-i-1}\omega & \text{if } i \in \{p-2, p-1\} \\ 2\omega & \text{else,} \end{cases} \end{aligned}$$

additionally, for any $i \leq p$,

$$\begin{aligned} \omega_i &:= \min_{j \neq i} \mathrm{M}(i,j), \\ &= \begin{cases} \omega & \text{if } i = p, \\ 2^{p-i-1}\omega & \text{if } i \in \{p-2, p-1\} \\ 2\omega & \text{else.} \end{cases} \end{aligned}$$

Thus,

$$\omega^{(p)} = \omega^{(p-1)} = \omega \text{ and } \omega^{(p-2)} = 2\omega.$$

Let $\gamma > 0$. For any optimal arm i, since $\mathrm{M}(i, i+1) = \mathrm{M}(i+1, i) = \delta_i^+$, the vector

$$\boldsymbol{\mu}_i + \delta_i^+ + \gamma$$

Pareto dominates $\boldsymbol{\mu}_{i+1}$ or $\boldsymbol{\mu}_{i-1}$ and $\boldsymbol{\mu}_i - \delta_i^+ - \gamma \prec \boldsymbol{\mu}_{i+1}$ or $\boldsymbol{\mu}_{i-1}$. Moreover, it is easy to observe that for $k \in \{p-2, p-1, p\}$ and any $i \in [p]$,

$$\boldsymbol{\mu}_i + \delta_i^+ + \omega^{(k)} + \gamma$$

Pareto dominates 1 (if $k \in \{p, p-1\}$) or 2 (if $k = p-2$) other optimal arms. Letting $k \in \{p-2, p-1, p\}$, for any $i \in [p]$, we define the alternative bandit $\nu^{(i)}$ which is also Gaussian with the same covariance matrix I_D and means given by

$$\boldsymbol{\mu}_j^{(i)} = \begin{cases} \boldsymbol{\mu}_j & \text{if } j \neq i \\ \boldsymbol{\mu}_j - \delta_i^+ - \omega^{(k)} - \gamma & \text{if } j = i \text{ and } \mathbb{P}_\nu(j \in \hat{S}) \geq \frac{1}{2} \\ \boldsymbol{\mu}_j + \delta_i^+ + \omega^{(k)} + \gamma & \text{if } j = i \text{ and } \mathbb{P}_\nu(j \in \hat{S}) < \frac{1}{2}. \end{cases} \tag{28}$$

Therefore, since $\mathcal{A}$ is $\delta-$correct, and by what precedes,

- if $\mathbb{P}_\nu(i \in \hat{S}) \geq \frac{1}{2}$ then $\mathbb{P}_{\nu^{(i)}}(i \in \hat{S}) \leq \delta$ and
- if $\mathbb{P}_\nu(i \in \hat{S}) < \frac{1}{2}$ then $\mathbb{P}_{\nu^{(i)}}(i \in \hat{S}) \geq 1 - \delta$.

The first point follows simply from the definition of δ_i^+ and the fact that by design $\mathrm{M}(i,j) = \mathrm{M}(i,j)$ for $i, j \in [p]$, For the second point, if $k \in \{p, p-1\}$, in the bandit $\nu^{(i)}$, at least one arm of $\mathcal{S}^\star(\nu)$ is no longer optimal, then $|\mathcal{S}^\star(\nu^{(i)})| \leq p-1 \leq k$. So $\mathbb{P}_{\nu^{(i)}}(i \in \hat{S}) \geq 1 - \delta$. If $k = p-2$ since two arms of $\mathcal{S}^\star(\nu)$ are now dominated, we have $|\mathcal{S}^\star(\nu^{(i)})| \leq p-2 = k$, hence $\mathbb{P}_{\nu^{(i)}}(i \in \hat{S}) \geq 1 - \delta$ Letting KL denote the KL divergence and using lemma 1 of [22], on $\mathcal{F}_\tau-$measurable event

$$E_i = \begin{cases} \{i \in \hat{S}\} & \text{if } \mathbb{P}_\nu(i \in \hat{S}) \geq \frac{1}{2}, \\ \{i \notin \hat{S}\} & \text{if } \mathbb{P}_\nu(i \in \hat{S}) < \frac{1}{2}, \end{cases}$$

for which $\mathbb{P}_\nu(E_i) \geq \frac{1}{2}$ and $\mathbb{P}_{\nu^{(i)}}(E_i) \leq \delta$, it comes that

$$\sum_{a \in \mathbb{A}} \mathbb{E}_\nu(T_a(\tau_\delta)) \, \mathrm{KL}(\nu_a, \nu_a^{(i)}) \geq d(\mathbb{P}_\nu(E_i), \mathbb{P}_{\nu^{(i)}}(E_i)),$$

hence

$$\mathbb{E}_\nu(T_i(\tau_\delta)) \, \mathrm{KL}(\nu_i, \nu_i^{(i)}) \geq d(\mathbb{P}_\nu(E_i), \mathbb{P}_{\nu^{(i)}}(E_i)), \tag{29}$$

where $d(x, y) = x \log(x/y) + (1 - x) \log((1 - x)/(1 - y))$ is the binary relative entropy. Since $\mathbb{P}_\nu(E_i) \geq \frac{1}{2}$ and $\mathbb{P}_{\nu^{(i)}}(E_i) \leq \delta$, (29) yields (see [22]),

$$\begin{aligned}
\mathbb{E}_\nu(T_i(\tau_\delta)) &\geq \frac{1}{\mathrm{KL}(\nu_i, \nu_i^{(i)})} \frac{1}{2} \left(\log\left(\frac{1}{2\delta}\right) + \log\left(\frac{1}{2(1-\delta)}\right) \right) \\
&= \frac{1}{2\,\mathrm{KL}(\nu_i, \nu_i^{(i)})} \log\left(\frac{1}{\delta(1-\delta)}\right) \\
&\geq \frac{1}{2\,\mathrm{KL}(\nu_i, \nu_i^{(i)})} \log(1/\delta).
\end{aligned}$$

By direct algebra, we compute (independent marginals since the covariance is diagonal I_D),

$$\mathrm{KL}(\nu_i, \nu_i^{(i)}) = \frac{1}{2} \|\boldsymbol{\mu}_j - \delta_i^+ - \omega^{(k)} - \gamma - \boldsymbol{\mu}_j\|_2^2 = \frac{D}{2}(-\delta_i^+ - \omega^{(k)} - \gamma)^2.$$

Noting that on this instance all the arms are optimal, we have for any arm $i, \Delta_i = \delta_i^+$. Finally, letting $\gamma \longrightarrow 0$ proves that for any arm i,

$$\mathbb{E}_\nu(T_i(\tau_\delta)) \geq \frac{1}{D(\Delta_i^k)^2} \log(1/\delta),$$

further noting that $\mathbb{E}(\tau_\delta) = \sum_{i=1}^K \mathbb{E}(T_i(\tau_\delta))$ achieves the proof. We have chosen a diagonal matrix matrix simplicity, we believe that choosing carefully correlated objectives like in [6] could give a tighter lower bound especially regarding the dependence in the dimension D. □

F Best Arm Identification

In this section, we discuss the sample complexity and the performance of APE associated to the stopping rule τ_0^1 for BAI. Noting that when $D = 1$, the Pareto set is just the *argmax* over the means, BAI and PSI are the same for uni-dimensional bandits. For this setting we should expect algorithms for PSI to be competitive with existing algorithms for BAI. We will show that it is actually the case for APE . Let $D = 1$ and ν be a one-dimensional K-armed bandit. Letting $a_\star$ denote the unique optimal arm of the bandit ν, i.e $\mathcal{S}^\star = \{a_\star\}$, one can easily check that the gaps defined for PSI matches the common notion of gaps for BAI. Indeed, for any $a \neq a_\star$,

$$\begin{aligned}
\Delta_a &:= \max_{j \in \mathcal{S}^\star} \mathrm{m}(a, j), \\
&= \mathrm{m}(a, a_\star) \\
&= \mu_{a_\star} - \mu_a,
\end{aligned}$$

and

$$\Delta_{a_\star} = \min_{j \neq a_\star} \{\mathrm{M}(j, a^\star)^+ + \Delta_j\} = \min_{j \neq a_\star} \Delta_j,$$

which matches the definition of the gap in the one-dimensional bandit setting ([21, 22]). Therefore, the sample complexity of APE for BAI can be deduced from Theorem 1.

Theorem 4. *Let $\delta \in (0, 1), K \geq 2$ and ν a $K-$armed bandit with a unique best arm $a_\star$ and 1-subgaussian distributions. APE associated to be stopping time τ_0^1 identifies the best arm $a^\star$ with probability at least $1 - \delta$ using at most the following number of samples*

$$\sum_{a=1}^K \frac{88}{\Delta_a^2} \log\left(\frac{2K(K-1)}{\delta} \log\left(\frac{12e}{\Delta_a}\right)\right).$$

In particular, the k relaxation is not meaningful in this setting. Under the unique optimal arm assumption, the algorithm will always stop when the best arm has been identified. And we remark that from the definition of ω_i's

$$\omega_1 = \min_{j \neq a_\star} \mathrm{M}(a_\star, j) = \min_{j \neq a_\star} \Delta_j = \Delta_{a_\star} \quad \text{and} \quad \forall i \neq a_\star, \quad \omega_i < 0,$$

so for any $k \leq K$, $\max(\omega_k, \Delta_a) = \Delta_a$.

Remark 2. *Theorem 4 could be slightly improved. On the event $\mathcal{E}$ we consider that for any pair of arms the difference of their empirical mean does not deviate too much from its actual value. For BAI, since we know that there is a unique optimal arm (enforced by assumption), it is sufficient to control the difference between the best arm and any other arm, therefore we could replace the $K(K-1)/2$ term due to union bound in the confidence bonus by $K-1$ and we could show that this will reflect in the sample complexity by replacing $K(K-1)$ by $2(K-1)$. However, this cannot be done in general for PSI since we do not know in advance the number of optimal arms.*

When $D = 1$, APE reduces to sample at each round t, the least sampled among

$$b_t := \operatorname*{argmax}_i \left\{ \min_{j \neq i} U_{i,j}(t) \right\}, \tag{30}$$

$$c_t := \operatorname*{argmin}_{j \neq b_t} L_{b_t,j}(t), \tag{31}$$

where $U_{i,j}(t) := \widehat{\mu}_i(t) - \widehat{\mu}_j(t) + \beta_{i,j}(t)$ and $L_{i,j}(t) := \widehat{\mu}_i(t) - \widehat{\mu}_j(t) - \beta_{i,j}(t)$ are upper and lower bounds on the difference $\mu_i - \mu_j$. To be in the same setting as LUCB and UGapEc which uses confidence interval on single arms, we would have $\beta_{i,j}(t) := \beta_i(t) + \beta_j(t)$, where $\beta_i' s$ are confidence bonuses on single arms such that $L_i(t) := \widehat{\mu}_i(t) - \beta_i(t)$ and $U_i(t) := \widehat{\mu}_i(t) + \beta_i(t)$ are lower and upper confidence bounds on μ_i. Then (30) and (31) rewrite as

$$\begin{aligned} b_t &:= \operatorname*{argmax}_i \left\{ U_i(t) - \max_{j \neq i} L_j(t) \right\}, \\ c_t &:= \operatorname*{argmax}_{j \neq b_t} U_j(t), \end{aligned}$$

This resembles the sampling rule of UGap, which defines

$$\begin{aligned} b_t^{\text{UGap}} &:= \operatorname*{argmax}_i \left\{ L_i(t) - \max_{j \neq i} U_j(t) \right\}, \\ c_t^{\text{UGap}} &:= \operatorname*{argmax}_{j \neq b_t} U_i(t), \end{aligned}$$

and also pulls the least sampled so far. We note that a variant of our algorithm in which both b_t, c_t would be sampled (in the spirit of LUCB [3]) could also be analyzed using the same arguments employed in the proof of Theorem 1.

Note that when $\varepsilon_1 = 0$, for any $i \in S(t)^{\mathsf{c}}$, $g_i(t) > h_i^0(t)$. Indeed, by definition,

$$h_i^0(t) = \min_{j \neq i} (\mathrm{M}(i,j,t) - \beta_{i,j}(t)) = \min_{j \neq i} (-\mathrm{m}(i,j,t) - \beta_{i,j}(t))$$

and since $i \in S(t)^{\mathsf{c}}$, there exists $i^\star$ such that $\mathrm{m}(i, i^\star, t) > 0$ (i.e $\widehat{\boldsymbol{\mu}}_i(t) \prec \widehat{\boldsymbol{\mu}}_{i^\star}(t)$) and so

$$-\mathrm{m}(i, i^\star, t) - \beta_{i,i^\star}(t) < \mathrm{m}(i, i^\star, t) - \beta_{i,i^\star}(t).$$

Therefore,

$$\min_{j \neq i} (-\mathrm{m}(i,j,t) - \beta_{i,j}(t)) := h_i^0(t) < \max_{j \neq i} (\mathrm{m}(i,j,t) - \beta_{i,j}(t)) := g_i(t).$$

Thus for $\varepsilon_1 = 0$,

$$Z_2^0(t) = \min_{i \in S(t)^{\mathsf{c}}} g_i(t). \tag{32}$$

In the sequel, for this section, we remove the dependence on ε_1 to write $Z_i(t)$ instead of Z_i^0 for $i = 0$ and $i = 1$. In particular, when $D = 1, \varepsilon_1 = 0$, the stopping time τ_0 can be simplified to

$$\tau_0 = \inf\{t \in \mathbb{N}^\star : Z_1(t) > 0\},$$

which is a consequence of the following lemma.

Lemma 12. *For $D = 1, \varepsilon_1 = 0$,*

$$\inf\{t \in \mathbb{N}^\star : Z_1(t) > 0\} = \inf\{t \in \mathbb{N}^\star : Z_1(t) > 0 \wedge Z_2(t) > 0\}.$$

Proof of Lemma 12. Let $S(t) = \{\hat{a}_t\}$. Using the definition of h_i^0, g_i and (32), $Z_1(t)$ and $Z_2(t)$ simplifies to

$$Z_1(t) = \min_{i \neq \hat{a}_t}\{\widehat{\mu}_{\hat{a}_t}(t) - \widehat{\mu}_i(t) - \beta_{\hat{a}_t,i}(t)\}, \tag{33}$$

$$Z_2(t) = \min_{i \neq \hat{a}_t}\{\max_{j \neq i}[\widehat{\mu}_j(t) - \widehat{\mu}_i(t) - \beta_{i,j}(t)]\}. \tag{34}$$

We have :

$$\begin{aligned} Z_1(t) > 0 &\implies \forall i \neq \hat{a}_t, \widehat{\mu}_{\hat{a}_t}(t) - \widehat{\mu}_i(t) - \beta_{\hat{a}_t,i}(t) > 0, \\ &\implies \forall i \neq \hat{a}_t, \max_{j \neq i}[\widehat{\mu}_j(t) - \widehat{\mu}_i(t) - \beta_{j,i}(t)] > 0, \\ &\implies Z_2(t) > 0. \end{aligned}$$

Thus, $Z_1(t) > 0 \implies (Z_1(t) > 0 \wedge Z_2(t) > 0)$ and the reverse holds trivially. So

$$Z_1(t) > 0 \iff (Z_1(t) > 0 \wedge Z_2(t) > 0).$$

□

Letting $\hat{a}_t$ denote the empirical best arm after t rounds, the stopping rule of APE (with the instantiation proposed in Section 3.2 based on confidence intervals on pairs of arms) reduces to

$$\tau_0 = \inf\left\{t \in \mathbb{N}^\star : \forall i \neq \hat{a}_t, \frac{(\hat{\mu}_{\hat{a}_t}(t) - \hat{\mu}_i(t))^2}{2\left(\frac{1}{T_{\hat{a}_t}(t)} + \frac{1}{T_i(t)}\right)} \geq 2C^g\left(\frac{\log(K_1/\delta)}{2}\right) + 2\sum_{a \in \{\hat{a}_t, i\}} \log(4 + \log(T_a(t)))\right\}$$

which is very close to a Generalized Likelihood Ratio (GLR) stopping rule assuming Gaussian distributions with variance 1 for the rewards (which is known to be also correct for sub-Gaussian rewards) [4, 20]. This modified stopping rule compared to LUCB1 and UGapEc can partially explains the empirical improvement observed in Section I.3.

G LUCB1-like instantiation of APE

In this section we derive an upper bound on the expectation of the sample complexity $\tau_{\varepsilon_1}^k$ when APE is run with confidence bonuses similar to LUCB1 [3]. This is different from Theorem 1 for which the sample complexity is bounded only on the high-probability event $\mathcal{E}$ but as, for many algorithms in pure-exploration [23, 8, 6] we do not control what happens on $\mathcal{E}^c$. Therefore,our goal here is to upper-bound $\mathbb{E}(\tau)$ instead of $\mathbb{E}(\mathbb{1}\{\mathcal{E}\}\tau)$ which we did in Theorem 1. To adapt the strategy employed in [3], we use similar confidence bonuses, thus we define for any arm i,

$$\beta_i(t) = \sqrt{\frac{2}{T_i(t)}\log\left(\frac{5KDt^4}{2\delta}\right)}, \tag{35}$$

and for any pair $i, j \in \mathbb{A}$, $\beta_{i,j}(t) = \beta_i(t) + \beta_j(t)$. Recalling the definition of $\mathcal{E}$ and $\mathcal{E}_t$ introduced in Section 3.1,

$$\mathcal{E}_t := \bigcap_{i=1}^{K}\bigcap_{j \neq i}\bigcap_{d=1}^{D}\{L_{i,j}^d(t,\delta) \leq \mu_i^d - \mu_j^d \leq U_{i,j}^d(t,\delta)\}, \text{ and } \mathcal{E} = \bigcap_{t=1}^{\infty}\mathcal{E}_t,$$

the lemma hereafter shows that with the choice of β_i's in (35) and for 1-subgaussian marginals, $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$.

Lemma 13. *It holds that $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$.*

Proof. Letting

$$\tilde{\mathcal{E}}_t := \bigcap_{i=1}^{K} \bigcap_{d=1}^{D} |\widehat{\mu}_i^d(t) - \mu_i^d| \leq \beta_i(t),$$

we have $\tilde{\mathcal{E}}_t \subset \mathcal{E}$. Indeed, on $\tilde{\mathcal{E}}_t$, for any $i, j \in \mathbb{A}$ and $d \leq D$,

$$\widehat{\mu}_i^d(t) - \widehat{\mu}_j^d(t) - \beta_i(t) - \beta_j(t) \leq \mu_i^d - \mu_j^d \leq \widehat{\mu}_i^d(t) - \widehat{\mu}_j^d(t) + \beta_i(t) + \beta_j(t),$$

which combined with $\beta_{i,j}(t) = \beta_i(t) + \beta_j(t)$ yields $\tilde{\mathcal{E}}_t \subset \mathcal{E}_t$ so

$$\mathbb{P}(\mathcal{E}^{\mathsf{c}}) \leq \sum_{t=1}^{\infty} \mathbb{P}(\mathcal{E}_t^{\mathsf{c}}) \leq \sum_{t=1}^{\infty} \mathbb{P}(\tilde{\mathcal{E}}_t^{\mathsf{c}}). \tag{36}$$

Applying Hoeffding's inequality to the 1-subgaussian marginals yields

$$\begin{aligned}
\mathbb{P}(\tilde{\mathcal{E}}_t^{\mathsf{c}}) &\leq \sum_{i=1}^{K} \sum_{d=1}^{D} \mathbb{P}(|\widehat{\mu}_i^d(t) - \mu_i^d| > \beta_i(t)), \\
&\leq \sum_{i=1}^{K} \sum_{d=1}^{D} \sum_{s=1}^{t} \mathbb{P}(|\widehat{\mu}_{i,s}^d - \mu_i^d| > \beta^{t,s}) \quad \text{where } \beta^{t,s} = \sqrt{\frac{2}{s} \log\left(\frac{5KDt^4}{2\delta}\right)}, \\
&\leq \sum_{i=1}^{K} \sum_{d=1}^{D} \sum_{s=1}^{t} \frac{4\delta}{5KDt^4}, \\
&= \frac{4\delta}{5t^3}.
\end{aligned}$$

Finally,

$$\begin{aligned}
\mathbb{P}(\mathcal{E}^{\mathsf{c}}) &\leq \sum_{t=1}^{\infty} \mathbb{P}(\tilde{\mathcal{E}}_t^{\mathsf{c}}), \\
&\leq \frac{4\delta}{5} \sum_{t=1}^{\infty} \frac{1}{t^3}, \\
&\leq \delta.
\end{aligned}$$

□

We can now state the main theorem of this section.

Theorem 5. *Let $\varepsilon_1 \geq 0, k \leq K$ and ν a bandit with 1-subgaussian marginals. APE run with the β_i's of* (35) *and associated to the stopping time $\tau_{\varepsilon_1}^k$ outputs a valid set and its expected sample complexity is upper bounded as follows :*

$$\mathbb{E}_\nu(\tau_{\varepsilon_1}^k) \leq 64\sqrt{e}H \log\left(\frac{5KD}{2\delta}\right) + 256\sqrt{e}H \log(256H) + \frac{8\pi^2}{15} + 1,$$

with $H := \sum_{a=1}^{K} \max(\Delta_{a_t}, \varepsilon_1, \omega_k)^{-2}$.

Proof of Theorem 1. The correctness follows from Lemma 2 combined with Lemma 13. It remains to upper-bound $\mathbb{E}(\tau_{\varepsilon_1}^k)$. Note that this proof technique has been already used in [3, 9, 26] for LUCB-like algorithms. Let $n \geq 1$ to be specified later and

$$\mathcal{E}(n) = \bigcap_{t \in [\frac{1}{2}n, n]} \mathcal{E}_t. \tag{37}$$

Remark that

$$\mathcal{E}(n) \cap \{\tau_{\varepsilon_1}^k > n\} \text{ holds} \implies \sum_{t=1}^{n} \mathbb{1}\{\{\tau_{\varepsilon_1}^k > t\} \cap \mathcal{E}(n)\} = n. \tag{38}$$

We will show that for some choice of n, the RHS of (38) will be strictly less than n so the LHS does not hold. We proceed by upper-bounding the RHS

$$\sum_{t=1}^{n} \mathbb{1}\{\{\tau_{\varepsilon_1}^k > t\} \cap \mathcal{E}(n)\} \leq \frac{n}{2} + \sum_{t=\frac{n}{2}}^{n} \mathbb{1}\{\{\tau_{\varepsilon_1}^k > t\} \cap \mathcal{E}(n)\}, \tag{39}$$

$$\leq \frac{n}{2} + \sum_{t=\frac{n}{2}}^{n} \mathbb{1}\{\{\tau_{\varepsilon_1}^k > t\} \wedge \mathcal{E}_t\}. \tag{40}$$

From Lemma 4, Lemma 5 and Lemma 3, we have that for any $t \in [\frac{n}{2}, n]$,

$$\{\tau_{\varepsilon_1}^k > t\} \cap \mathcal{E}_t \Longrightarrow \max(\Delta_{a_t}, \varepsilon_1, \omega_k) \leq 2\beta_{a_t,a_t}(t),$$

with $\beta_{a_t,a_t}(t) = 2\beta_{a_t}(t)$. Therefore, using this result back in (39) and letting $c_\delta := (5KD/(2\delta))^{1/4}$, $\tilde{\Delta}_a := \max(\Delta_a, \varepsilon_1, \omega_k)$ yields

$$\begin{aligned}
\sum_{t=1}^{n} \mathbb{1}\{\{\tau_{\varepsilon_1}^k > t\} \cap \mathcal{E}(n)\} &\leq \frac{n}{2} + \sum_{t=\frac{n}{2}}^{n} \mathbb{1}\{\tilde{\Delta}_{a_t} \leq 4\beta_{a_t}(t)\} \\
&\leq \frac{n}{2} + \sum_{t=\frac{n}{2}}^{n} \sum_{a=1}^{K} \mathbb{1}\{(a_t = a) \wedge \tilde{\Delta}_a \leq 4\beta_a(t)\}, \\
&\leq \frac{n}{2} + \sum_{a=1}^{K} \sum_{t=\frac{1}{2}n}^{n} \mathbb{1}\{\{a_t = a\} \wedge \{T_a(t) \leq \frac{128}{\tilde{\Delta}_a^2} \log(c_\delta t)\}\} \\
&\leq \frac{n}{2} + \sum_{a=1}^{K} \sum_{t=\frac{n}{2}}^{n} \mathbb{1}\{\{a_t = a\} \wedge \{T_a(t) \leq \frac{128}{\tilde{\Delta}_a^2} \log(c_\delta n)\}\} \\
&\leq \frac{n}{2} + \sum_{a=1}^{K} \frac{128}{\tilde{\Delta}_a^2} \log(c_\delta n) \\
&\leq \frac{n}{2} + 128H \log(c_\delta n),
\end{aligned}$$

where $H := \sum_a \tilde{\Delta}_a^{-2}$. Then, choosing n such that

$$\frac{n}{2} + 128H \log(c_\delta n) < n,$$

that is

$$n > T^\star := \inf\left\{ s \in \mathbb{N}^\star : \frac{128H \log(c_\delta s)}{s} < \frac{1}{2} \right\}, \tag{41}$$

would yield

$$\sum_{t=1}^{n} \mathbb{1}\{\{\tau_{\varepsilon_1}^k > t\} \cap \mathcal{E}(n)\} < n, \tag{42}$$

so

$$\mathcal{E}(n) \cap \{\tau_{\varepsilon_1}^k > n\} = \emptyset,$$

which means

$$\{\tau_{\varepsilon_1}^k > n\} \subset \mathcal{E}(n)^{\mathsf{c}}.$$

Therefore, for any $n > T^\star$,

$$\{\tau_{\varepsilon_1}^k > n\} \subset \mathcal{E}(n)^{\mathsf{c}}. \tag{43}$$

Thus,

$$\begin{aligned}
\mathbb{E}_\nu(\tau_{\varepsilon_1}^k) &= \mathbb{E}_\nu\left(\tau_{\varepsilon_1}^k \mathbb{1}\{\tau_{\varepsilon_1}^k \leq T^\star\} + \tau_{\varepsilon_1}^k \mathbb{1}\{\tau_{\varepsilon_1}^k > T^\star\}\right) \\
&\leq T^\star + \mathbb{E}_\nu(\tau_{\varepsilon_1}^k \mathbb{1}\{\tau_{\varepsilon_1}^k > T^\star\} \\
&\leq T^\star + \sum_{n=T^\star+1}^{\infty} \mathbb{P}_\nu(\tau_{\varepsilon_1}^k > n) \\
&\leq T^\star + \sum_{n=T^\star+1}^{\infty} \mathbb{P}_\nu(\mathcal{E}(n)^{\mathsf{c}}),
\end{aligned}$$

using (37) and union bound yields,

$$\begin{aligned}\mathbb{P}(\mathcal{E}(n)^{\mathsf{c}}) &\leq \sum_{t=\frac{n}{2}}^{n} \frac{4\delta}{5t^3}, \\ &\leq \frac{4\delta}{5}\frac{(1/2)n}{(1/2)^3 n^3}, \\ &= \frac{16\delta}{5}\frac{1}{n^2}.\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{E}_\nu(\tau_{\varepsilon_1}^k) &\leq T^\star + \frac{16\delta}{5}\frac{\pi^2}{6} \\ &\leq T^\star + \frac{8\pi^2}{15}.\end{aligned}$$

Upper-bounding $T^\star$ will conclude the proof.

Lemma 14. *It holds that*

$$T^\star - 1 \leq \frac{1}{c_\delta}\exp\left(-W_{-1}\left(-\frac{1}{256 c_\delta H}\right)\right) \leq 256\sqrt{e}H\log\left(256 c_\delta H\right).$$

Finally,

$$\begin{aligned}\mathbb{E}_\nu(\tau_\delta) &\leq 256\sqrt{e}H\log\left(256(5KD/(2\delta))^{1/4}H\right) + \frac{8\pi^2}{15} + 1, && (44)\\ &\leq 64\sqrt{e}H\log\left(\frac{5KD}{2\delta}\right) + 256\sqrt{e}H\log\left(256H\right) + \frac{8\pi^2}{15} + 1, && (45)\end{aligned}$$

which achieves the proof.

Remark 3. *The same technique could be applied to upper-bound* $\mathbb{E}_\nu(\tau_{\varepsilon_1,\varepsilon_2})$.

Now we prove Lemma 14.

Proof of Lemma 14. We have

$$128H\frac{\log(c_\delta s)}{s} < \frac{1}{2} \iff \frac{\log(c_\delta s)}{s} < \frac{1}{256H} \tag{46}$$

then, using Lemma 15 yields

$$(46) \implies \begin{cases} s > 0 & \text{if } \frac{1}{256H} < c_\delta/e \\ 0 < s \leq \frac{1}{c_\delta} \text{ or } s \geq N^\star & \text{else,}\end{cases}$$

with

$$N^\star = \frac{1}{c_\delta}\exp\left(-W_{-1}\left(-\frac{1}{256 c_\delta H}\right)\right).$$

Therefore,

$$T^\star = \inf\left\{s \in \mathbb{N}^\star : 128H\frac{\log(c_\delta s)}{s} < \frac{1}{2}\right\} \leq \begin{cases} 1 & \text{if } \frac{1}{256H} < c_\delta/e \\ \lceil N^\star \rceil & \text{else.}\end{cases}$$

Using Corollary 1, to upper bound $N^\star$ yields

$$T^\star - 1 \leq 256\sqrt{e}H\log_+\left(256 c_\delta H\right), \tag{47}$$

where $\log_+(x) = \max(0, \log(x))$. □

□

H Technical Lemmas

Lemma 15. *Let $a, b > 0$. If $b < a/e$ then*

$$\frac{\log(ax)}{x} < b \Longrightarrow 0 < x \leq \frac{1}{a} \text{ or } x \geq \frac{1}{a}\exp\left(-W_{-1}\left(-\frac{b}{a}\right)\right).$$

Moreover, if $b \geq a/e$, then for any $x > 0$, $\log(ax)/x \leq b$.

Proof. We have

$$\begin{aligned}
\frac{\log(ax)}{x} < b \quad &\Longrightarrow \quad -\frac{1}{ax}\log\left(\frac{1}{ax}\right) < \frac{b}{a} \\
&\Longrightarrow \quad y\log(y) > -\frac{b}{a}, \quad y = \frac{1}{ax} \\
&\Longrightarrow \quad \frac{1}{ax} \geq 1 \text{ or } -\frac{b}{a} < y\log(y) < 0
\end{aligned}$$

since $-b/a > -1/e$ and the negative branch W_{-1} of the Lambert function is decreasing on $[-1/e, 0]$,

$$\begin{aligned}
\frac{\log(ax)}{x} < b \quad &\Longrightarrow \quad 0 < x \leq \frac{1}{a} \text{ or } W_{-1}(y\log(y)) \leq W_{-1}(-b/a) \\
&\Longrightarrow \quad 0 < x \leq \frac{1}{a} \text{ or } \log(y) \leq W_{-1}(-b/a) \\
&\Longrightarrow \quad 0 < x \leq \frac{1}{a} \text{ or } ax \geq \exp(-W_{-1}(-b/a)) \\
&\Longrightarrow \quad 0 < x \leq \frac{1}{a} \text{ or } x \geq \frac{1}{a}\exp(-W_{-1}(-b/a)).
\end{aligned}$$

Proving the second part of the lemma just follows from $\log(x) \leq x/e$. □

The following lemma is taken from [27]

Lemma 16 ([27])**.** *For any $x \in [0, -e^{-1}]$,*

$$-\log(-x)+\log(-\log(-x)) \leq -W_{-1}(x) \leq -\log(-x)+\log(-\log(-x))+\min\left\{\frac{1}{2}, \frac{1}{\sqrt{-x\log(-x)}}\right\}$$

Corollary 1. *Let $0 < a < 1/e$. It holds that*

$$\exp(-W_{-1}(-a)) \leq \frac{e^{1/2}}{a}\log\left(\frac{1}{a}\right).$$

We recall the following lemma which is taken from [20].

Proof. Using Lemma 16 yields,

$$-W_{-1}(-a) \leq -\log(a) + \log(-\log(a)) + \frac{1}{2},$$

and taking $\exp$ on both sides gives the result. □

Lemma 17. *Let $\Delta^2 > 0$. Then, for $t \geq 2$,*

$$t \geq \frac{1}{\Delta^2}\log\left(2\log\left(\frac{3e^2}{2\Delta^2}\right)\right) \Longrightarrow \frac{\log\log\left(e^4 t\right)}{t} < \Delta^2.$$

Proof. We note that if $\Delta^2 \geq \frac{e}{3}$, then the result follows trivially since it can be easily checked that for $t \geq 2$,

$$\log\log(e^4 t) \leq \frac{e}{3}t.$$

Therefore, in the sequel, we assume $\Delta^2 < e/3$. Let

$$t_\Delta := \frac{1}{\Delta^2} \log\left(2\log\left(\frac{3e^2}{2\Delta^2}\right)\right),$$

and

$$g(t) = t - \frac{1}{\Delta^2}\log(\log(e^4 t)).$$

Then,

$$g'(t) = 1 - \frac{1}{\Delta^2 t \log(e^4 t)},$$

and $g'(t) \geq 0$ for t such that $\Delta^2 t \log(e^4 t) \geq 1$. Using the Lambert function W_0, which is increasing on $[0, \infty)$,

$$\begin{aligned}
\Delta^2 t \log(e^4 t) \geq 1 \quad &\iff \quad e^4 t \log(e^4 t) \geq \frac{e^4}{\Delta^2} && (48)\\
&\iff \quad \log(e^4 t) \geq W_0\left(\frac{e^4}{\Delta^2}\right) && (49)\\
&\iff \quad t \geq t^0 := \frac{1}{e^4}\exp\left(W_0\left(\frac{e^4}{\Delta^2}\right)\right) && (50)
\end{aligned}$$

and by definition of W_0, we have

$$W_0(x)\exp(W_0(x)) = x,$$

so

$$\exp\left(W_0\left(\frac{e^4}{\Delta^2}\right)\right) = \frac{e^4}{\Delta^2}\frac{1}{W_0(e^4\Delta^{-2})},$$

and therefore,

$$t^0 = \frac{1}{\Delta^2}\frac{1}{W_0(e^4\Delta^{-2})}.$$

We will show that $t_\Delta > t^0$. Indeed, since W_0 is increasing,

$$\begin{aligned}
\frac{1}{\Delta^2} > 3/e \quad &\implies \quad W_0\left(\frac{e^4}{\Delta^2}\right) \geq W_0(3e^3) = 3\\
&\implies \quad \frac{1}{\Delta^2}\frac{1}{W_0(e^4\Delta^{-2})} \leq \frac{1}{3}\frac{1}{\Delta^2},
\end{aligned}$$

that is

$$t^0 \leq \frac{1}{3}\frac{1}{\Delta^2}. \tag{51}$$

On the other side,

$$\begin{aligned}
\frac{1}{\Delta^2} > 3/e \quad &\implies \quad \log\left(2\log\left(\frac{3e^2}{2\Delta^2}\right)\right) > \log(2\log(9e/2))\\
&\implies \quad t_\Delta \geq \frac{1}{\Delta^2}\log(2\log(9e/2)) > \frac{1}{3}\frac{1}{\Delta^2}
\end{aligned}$$

Therefore,

$$t^0 \leq \frac{1}{3}\frac{1}{\Delta^2} < \frac{\log(2\log(9e/2))}{\Delta^2} \leq t_\Delta.$$

Thus, we have shown that $t^0 \leq t_\Delta$ and for any $t \geq t_\Delta$, $g'(t) \geq 0$ so

$$\forall t \geq t_\Delta,\ g(t) \geq g(t_\Delta). \tag{52}$$

Showing that $g(t_\Delta) > 0$, will conclude the proof. Letting $a = 3e^2/2$, we have

$$\begin{aligned}
g(t_\Delta) > 0 \quad &\iff \quad \frac{1}{\Delta^2}\log(2\log(a/\Delta^2)) - \frac{1}{\Delta^2}\log(\log(e^4 t_\Delta)) > 0 && (53)\\
&\iff \quad \log(2\log(a/\Delta^2)) - \log(\log(e^4 t_\Delta)) > 0 && (54)\\
&\iff \quad 2\log(a/\Delta^2) - \log(e^4 t_\Delta) > 0 && (55)\\
&\iff \quad \log(a/\Delta^2) - \log(e^4 t_\Delta \Delta^2/a) > 0 && (56)\\
&\iff \quad \frac{a}{\Delta^2} - \frac{e^4}{a}\Delta^2 t_\Delta > 0 && (57)\\
&\iff \quad \frac{a}{\Delta^2} - \frac{e^4}{a}\log(2\log(a/\Delta^2)) > 0, && (58)
\end{aligned}$$

then, observing that for $x \geq 12$,

$$\log(2\log(x)) \leq \frac{x}{e^2},$$

and since

$$a/\Delta^2 > (3e^2/2) \times (3/e) > 12,$$

we have

$$\log(2\log(a/\Delta^2)) \leq \frac{1}{e^2}\frac{a}{\Delta^2} \tag{59}$$

so, using (59) yields that the LHS of (58) is larger than

$$\frac{3}{2}e^2\frac{1}{\Delta^2} - e^2\frac{1}{\Delta^2},$$

which is always positive. Therefore,

$$\forall t \geq t_\Delta,\ g(t_\Delta) > 0,$$

that is

$$\forall t \geq t_\Delta,\ \frac{\log\log(e^4 t)}{t} < \Delta^2. \tag{60}$$

□

Lemma 18. *Let $\delta \in (0,1), \Delta > 0$ and $c > 0$. Let $f : t \mapsto \sqrt{\frac{g(\delta)+c\log\log(e^4 t)}{t}}$ where g is a non-negative function. Then, for any $\alpha \in (0,1)$ and $t \geq 2$,*

$$t \geq \frac{1}{\Delta^2}\left(\frac{1}{\alpha}g(\delta) + \frac{c}{1-\alpha}\log_+\left(2\log\left(\frac{c}{(1-\alpha)\Delta^2}\right)\right)\right) \implies f(t) < \Delta. \tag{61}$$

Proof. Letting $t \geq 2$, we have

$$t \geq t_1 := \frac{1}{\alpha}\frac{1}{\Delta^2}g(\delta) \implies \frac{g(\delta)}{t} \leq \alpha\Delta^2. \tag{62}$$

Furthermore, using Lemma 17 yields

$$t \geq t_2 := \frac{c}{(1-\alpha)\Delta^2}\log_+\left(2\log\left(\frac{3e^2}{2(1-\alpha)\Delta^2}\right)\right) \implies \frac{\log\log(e^4 t)}{t} \leq (1-\alpha)\Delta^2/c. \tag{63}$$

Combining (62) and (63) yields for $t \geq 2$,

$$t \geq \max(t_1, t_2) \implies f(t)^2 < \Delta^2,$$

so

$$t \geq t_1 + t_2 \geq \max(t_1, t_2) \implies f(t) < \Delta. \tag{64}$$

□

Lemma 19. *Let $\Delta > 0$ and $\delta \in (0,1)$. Let*

$$f(t) := 4\sqrt{\frac{2C^g(\log(1/\delta)/2) + 4\log\log(e^4 t)}{t}}.$$

Then

$$\inf\{n \geq 2 : f(n) < \Delta\} \leq \frac{88}{\Delta^2}\log\left(\frac{4}{\delta}\log\left(\frac{12e}{\Delta}\right)\right).$$

Proof. We have

$$f(t) = \sqrt{\frac{32C^g(\log(1/\delta)/2) + 64\log\log(e^4 t)}{t}}.$$

Therefore, letting

$$g(\delta) := 32C^g(\log(1/\delta)/2) \quad \text{and} \quad c = 64$$

and further using Lemma 18 yields for any $\alpha \in (0,1)$ and $t \geq 2$,

$$t \geq t_\alpha \Longrightarrow f(t) < \Delta,$$

where

$$t_\alpha := \frac{1}{\Delta^2}\left(\frac{32}{\alpha}C^g(\log(1/\delta)/2) + \frac{64}{1-\alpha}\log\left(2\log(96e^2(1-\alpha)^{-1}\Delta^{-2})\right)\right).$$

Since $C^g(x) \approx x + \log(x)$ [20], and $\log(x) \leq x/e$ we have

$$\Delta^2 t_\alpha \leq \left(16 + \frac{16}{e}\right)\frac{1}{\alpha}\log(1/\delta) + \frac{64}{1-\alpha}\log(2\log(96e^2(1-\alpha)^{-1}\Delta^{-2})).$$

Taking $\alpha = \alpha^\star$ such that

$$\left(16 + \frac{16}{e}\right)\frac{1}{\alpha^\star} = \frac{64}{1-\alpha^\star},$$

that is setting

$$\alpha^\star = \frac{1+e}{1+5e},$$

yields

$$\Delta^2 t_{\alpha^\star} \leq \frac{64}{1-\alpha^\star}\log\left(\frac{2}{\delta}\log\left(\frac{96e^2}{(1-\alpha^\star)\Delta^2}\right)\right).$$

By numerical evaluation,

$$\frac{64}{1-\alpha^\star} \approx 86 < 88 \quad \text{and} \quad \frac{96}{1-\alpha^\star} \approx 130 < 12^2,$$

so

$$\Delta^2 t_{\alpha^\star} < 88\log\left(\frac{4}{\delta}\log\left(\frac{12e}{\Delta}\right)\right). \tag{65}$$

Therefore, putting these results together, for $t \geq 2$

$$t \geq t_\star := \frac{88}{\Delta^2}\log\left(\frac{4}{\delta}\log\left(\frac{12e}{\Delta}\right)\right) \Longrightarrow f(t) < \Delta$$

which yields

$$\inf\{n \geq 2 : f(n) < \Delta\} \leq \max(2, t_\star). \tag{66}$$

□

I Implementation and Additional Experiments

In this section, we give additional details about the experiments and additional experimental results.

I.1 Implementation

Setup We have implemented the algorithms mainly in `C++17` interfaced with `python` through the `cython` package. The experiments are run on an ARM64 8GB RAM/8 core/256GB disk storage computer. For the function C^g we have used the approximation $C^g(x) \approx x + \log(x)$ which is usually admitted [20]. For the experiments on real-world scenario we generate a certain number of seeds (usually 2000) and we use a different seed for each run on the same bandit. This procedure is identical for every experiment where we report the average sample complexity on the same bandit. To assess the robustness of our algorithm, the experiments on the synthetic dataset consisted in randomly uniformly sampling some bandit means for each configuration. For each sampled bandit, the algorithms compared are run once on the same instance and we note their empirical sample complexity. Finally, we report the average sample complexity across all the bandits of the same configuration.

Adaptation to bandits with marginals of different scaling We have presented the algorithm and the results specialized to the case where all the marginals are 1-subgaussian. Indeed our results can be simply extended where the marginals are instead all σ-subgaussian. Furthermore, there is a simple way to adapt the algorithm to the case where the marginals have different *known* subgaussianity parameter (i.e different scaling) but they are the same for every arm. The idea is to rescale each observation with the subgaussianity parameters. Let $\boldsymbol{\sigma} := \sigma_1, \dots \sigma_D$, $\sigma_i > 0$. Assuming that the marginal distributions of each arm are respectively $\sigma_1, \dots \sigma_D$-subgaussian, each observation $\mathbf{X}_{A_t}, s$ from arm A_t will be rescaled component-wise to $X^d_{A_t,s}/\sigma_d$ before being given to the algorithm. It is easy to see that this rescaling does not change the Pareto set since all the means are divided by the same values coordinate-wise.

Furthermore, by defining

$$\mathrm{M}^{\boldsymbol{\sigma}}(i,j) \quad := \quad \max_d \left(\frac{\mu_i^d - \mu_j^d}{\sigma_d} \right),$$

and $\mathrm{m}^{\boldsymbol{\sigma}}(i,j) =: -\mathrm{M}^{\boldsymbol{\sigma}}(i,j)$, all the results proved for 1-subgaussian distributions still holds using $\mathrm{m}^{\boldsymbol{\sigma}}$ and $\mathrm{M}^{\boldsymbol{\sigma}}$ in the definition of the gaps (Section 4).

I.2 Data processing

Dataset The dataset is extracted from [24] and some processing steps are applied to compute the covariance matrix of the distribution. First, as observed in [24], the 3 immunogenicity indicators extracted are weakly correlated, therefore, we assume the covariance matrix to be diagonal. To compute the variance of the marginals, we use the log-normal assumption as assumed for the data reported in [24]. Using this log-normal assumption the authors have provided for each arm and each indicator: the geometrical mean, the sample size and a 95% confidence interval on the geometrical mean based on the central limit theorem.

For each of the $K = 20$ arms (combination of three doses), we use these information to compute the sample variance of each immunogenicity indicator. Moreover, we compute the arithmetic average of the $\log$ outcomes which is obtained by taking the $\log$ (base e) of the geometrical empirical mean:

$$\begin{aligned} \bar{x} &= \log(\bar{x}_{\text{geometrical}}), \\ &= \log\left(\left(\prod_{i=1}^n x_i \right)^{1/n} \right), \\ &= n^{-1} \sum_{i=1}^n \log(x_i), \end{aligned}$$

where $x_1, \dots, x_n$ are the observations which are assumed to be log-normal. $\bar{x}$ represents by assumption the empirical mean of a Gaussian distribution, which we use as a proxy for its true, unknown mean. From this, we built a bandit model where each arm is a 3-dimensional Gaussian distribution with independent coordinates, whose means are given by the corresponding mean estimates (reported in Table 4) and in which the variance of each indicator is the pooled variance over the different arms (given in Table 5). Sampling an arm in this bandit simulates the measurement of the ($\log$ of the) 3 immunogenicity criteria in consideration on a new patient.

The 20 arms are classified into two groups. Each three/four-letters acronym denote a vaccine candidate. Prime BNT/BNT corresponds to giving BNT as first and second dose and similarly for Prime ChAd/ChAd. For example ChAd in the group Prime BNT/BNT means to give BNT as first and second dose and ChAd as third dose (booster).

Table 4: Table of the empirical arithmetic mean of the log-transformed immune response for three immunogenicity indicators. Each acronym corresponds to a vaccine. There are two groups of arms corresponding to the first 2 doses: one with prime BNT/BNT (BNT as first and second dose) and the second with prime ChAd/ChAd (ChAd as first and second dose). Each row in the table gives the values of the 3 immune responses for an arm (i.e. a combination of three doses).

Dose 1/Dose 2	Dose 3 (booster)	Immune response		
		Anti-spike IgG	NT_{50}	cellular response
Prime BNT/BNT	ChAd	9.50	6.86	4.56
	NVX	9.29	6.64	4.04
	NVX Half	9.05	6.41	3.56
	BNT	10.21	7.49	4.43
	BNT Half	10.05	7.20	4.36
	VLA	8.34	5.67	3.51
	VLA Half	8.22	5.46	3.64
	Ad26	9.75	7.27	4.71
	m1273	10.43	7.61	4.72
	CVn	8.94	6.19	3.84
Prime ChAd/ChAd	ChAd	7.81	5.26	3.97
	NVX	8.85	6.59	4.73
	NVX Half	8.44	6.15	4.59
	BNT	9.93	7.39	4.75
	BNT Half	8.71	7.20	4.91
	VLA	7.51	5.31	3.96
	VLA Half	7.27	4.99	4.02
	Ad26	8.62	6.33	4.66
	m1273	10.35	7.77	5.00
	CVn	8.29	5.92	3.87

Table 5: Pooled variance of each group.

	Immune response		
	Anti-spike IgG	NT_{50}	cellular response
Pooled sample variance	0.70	0.83	1.54

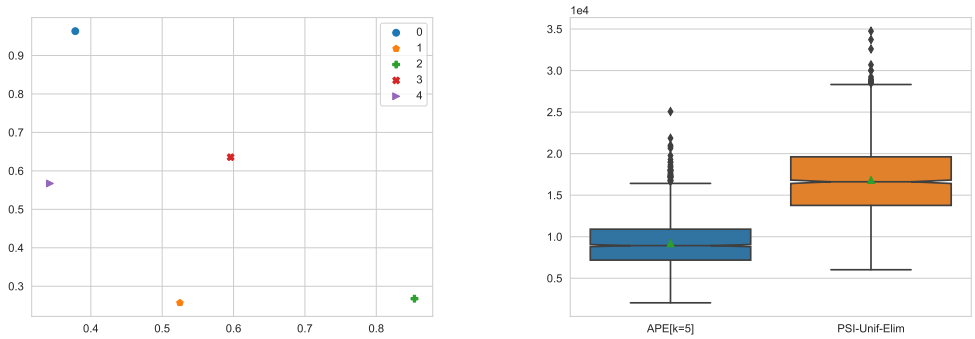
I.3 Additional experiments

I.3.1 Additional experiments for ε_1-APE-k

In this section we show that for some instances our algorithm can require up to 3 times less samples compared to PSI-Unif-Elim . This is due to the strategy of PSI-Unif-Elim which continue sampling arms identified as optimal until there are shown not to dominate any arm in the active set. For example on Figure 3a, the optimal arm 2 is "easy" to identify as such. However, since it slightly dominates the sub-optimal arm 1, PSI-Unif-Elim should continue sampling arm 2 until arm 1 is removed from the active set (likely this will happen when the algorithm "sees" that arm 1 is dominated by arm 3). We would expect our adaptive sampling rule to avoid this behaviour.

Figure 3b shows that APE takes nearly half the average sample complexity of PSI-Unif-Elim on this instance. In particular, Table 6 shows the average number of pulls taken by PSI-Unif-Elim divided by the average number of pulls taken by 0-APE-K for each arm. We can observe that the major difference in sample complexity is due to arm 2 being pulled nearly 6 times more by PSI-Unif-Elim w.r.t APE .

By increasing the number of arms and the dimension we can generate instances similar to Figure 3a where the gap between our algorithm and PSI-Unif-Elim is even larger. We chose a specific instance where $K = 12$, $D = 10$ and there are 11 optimal arms. On this instance (Figure 4), we can see that our algorithm uses 3 times less samples than PSI-Unif-Elim .



(a) Instance

(b) Sample complexity

Figure 3: For this instance (left) $\mathcal{S}^\star = \{0, 2, 3\}$ and on the right is the average sample complexity over the trials on the same instance.

Arm	0	1	2	3	4
Average ratio of pulls	3.08	1.36	5.68	1.28	1.40

Table 6: Average number of pulls taken by PSI-Unif-Elim divided by the average number of pulls taken by 0-APE-K for each arm.

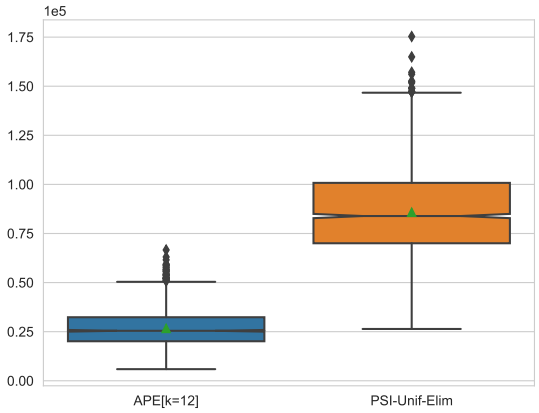
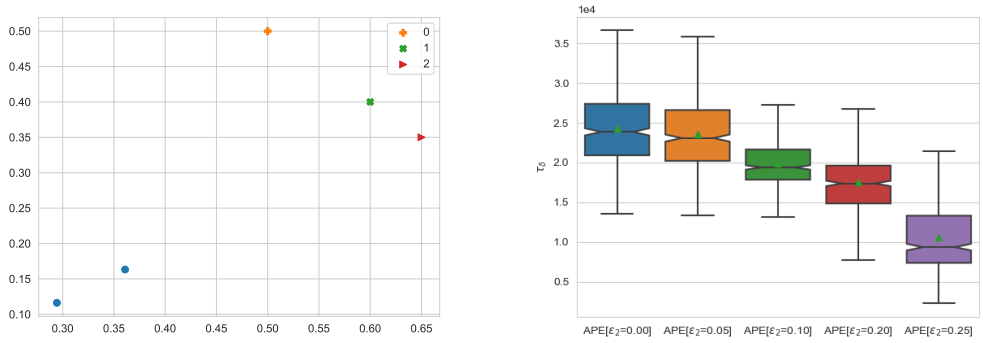


Figure 4: Average sample complexity on a specific instance with $K = 12$, $D = 10$ and $|\mathcal{S}^\star| = 11$

Finally, combining these additional experiments with the results of Table 2 we observe that on average ε_1-APE-k performs nearly 20% better than APE but there are some instances where the gap can be even larger. Of course, this also means that there should exist instances in which the improvement is smaller than 20% to compensate for instances like Figure 3a. But we note that instances like Figure 3a are very unlikely to be generated randomly so they should only be a few in the 2000 instances used in Figure 3a.

I.3.2 $(\varepsilon_1, \varepsilon_2)$-APE

We investigate the empirical behavior of $(\varepsilon_1, \varepsilon_2)$-APE for identifying an $(\varepsilon_1, \varepsilon_2)$-cover. We set $\varepsilon_1 = 0, \delta = 0.01$ and we test different values of $\varepsilon_2 \in \{0, 0.05, 0.1, 0.2, 0.25\}$. We average the results over 2000 independent trials with different seeds on the same instance (Figure 5a). We use multi-variate Bernoulli with independent marginals. The instance of Figure 5a is a toy example where $(\varepsilon_1, \varepsilon_2)$-covering can be meaningful and reduce the sample complexity. The 3 Pareto optimal vectors are chosen by hand and the last 2 vectors are randomly uniformly generated.

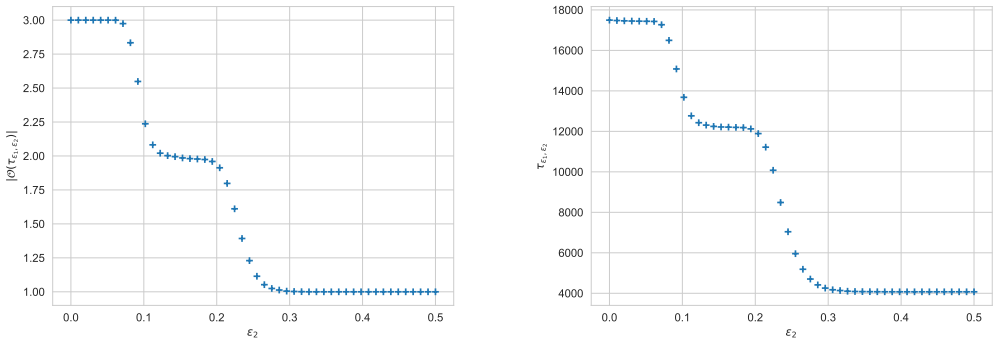


(a) Instance

(b) Sample complexity

Figure 5: For this instance (left) $\mathcal{S}^\star = \{0, 1, 2\}$. The difference in on x and y axis is 0.1 between arm 0 and 1 and 0.05 between arm 1 and 2. The rightmost figure is the sample complexity averaged over 2000 trials.

We can observe on Figure 5b that the sample complexity decreases as ε_2 increases. This is further confirmed in Figure 6 which shows the empirical sample complexity and the average size of the recommended cover versus ε_2 for 50 equally-spaced values of ε_2 between 0 and $1/2$. The drops observed in Figure 6b correspond to the values of ε_2 for which APE removes an optimal arm from the cover to save some samples (Figure 6a). A major decrease in the sample complexity corresponds more or less to an arm being removed from the recommended set. We observe in Figure 7 the histogram of occurrence of each arm in the recommended set for 3 values of ε_2 corresponding more or less to the middle of each plateau. We can see that for $\varepsilon_2 = 0.15$, arm 0 is always recommended, but the others are recommended on half of the runs. For $\varepsilon_2 = 0.4$, the algorithm nearly always recommend arm 0, which as the largest ω_i term (i.e the easiest to identify as optimal).

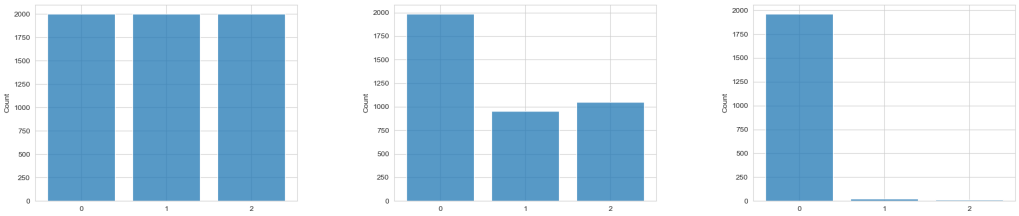


(a) Average $|\mathcal{O}(\tau_{\varepsilon_1,\varepsilon_2})|$ vs ε_2

(b) Sample complexity vs ε_2

Figure 6: On the left is the average length of $\mathcal{O}(\tau_{\varepsilon_1,\varepsilon_2})$ (over the different runs) versus ε_2 and on the right is the empirical sample complexity (averaged over the runs) versus ε_2. The empirical probability of error was equal to zero.

The plateau in the sample complexity for large values of ε_2 (> 0.3) is explained by the fact the algorithm needs to identify at least one optimal arm (which is reflected in the size of the returned set Figure 6a). Indeed, for $\varepsilon_1 = 0$ fixed, an algorithm for $(\varepsilon_1, \varepsilon_2)$-covering still need to assert that the arms in the recommended set are truly optimal which will require some samples even when ε_2 is very large. Thus, for $\varepsilon_2 > 0.3$ the algorithm need to identify at least one optimal arm and we can see on Figure 6a that for these values of ε_2, the recommended set contains only one optimal arm. Actually we can observe empirically that the "limit" sample complexity observed in Figure 6b is close to the average sample complexity of 0-APE-1 on the same instance (4073 samples).



(a) Average $\varepsilon_2 = 0.05$ (b) $\varepsilon_2 = 0.15$ (c) $\varepsilon_2 = 0.4$

Figure 7: Histogram of the number of times each arm is in the recommended set over the 2000 runs for 3 values of ε_2.

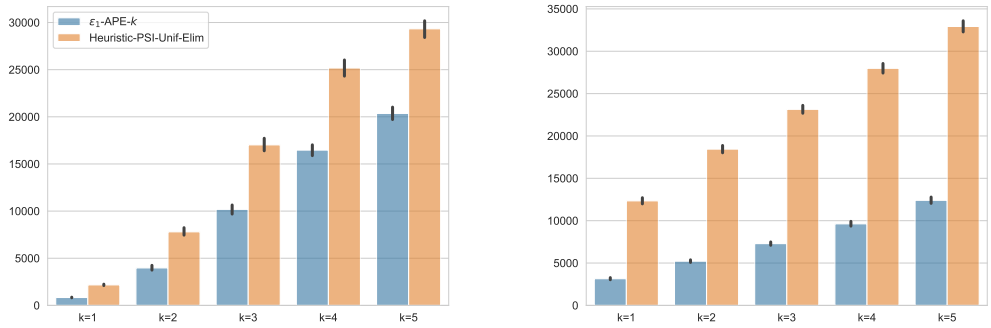
I.3.3 Comparing ε_1-APE-k to an adaptation of PSI-Unif-Elim

In this section, we compare ε_1-APE-k to an adaptation of PSI-Unif-Elim which stops earlier if at least k optimal arms have been identified. The pseudo-code of the algorithm is given in algorithm 3. As shown in [6], arms in $P_1(t)$ are already identified as optimal but when the goal is to identify the Pareto set, some of them (namely $P_1(t) \backslash P_2(t)$) need to be sampled again until all the arms they potentially dominate are removed from $A(t)$ and only then those arms will belong to $P_2(t)$ and will be removed from the active set. However, for the k-relaxation, identifying k optimal arms is enough so the algorithm can stop as soon as $|P(t-1) \cup P_1(t)| \geq k$. It this never occurs, then the algorithm will follow the initial stopping condition, that is to stop when $A(t) = \emptyset$.

Algorithm 3: Adaptation of PSI-Unif-Elim for the ε_1-APE-k objective

Data: parameter $\varepsilon_1 \geq 0, k \leq K$
initialize : $A(0) = \mathbb{A}$, $t \leftarrow 1$, $P(0) = P_1(0) = P_2(0) = \emptyset$
for $t = 1, 2, \ldots$ **do**
  sample each arm in $A(t-1)$ once ;
  $A(t) \leftarrow \{i \in A(t-1) : \forall\, j \in A(t-1), \mathrm{m}(i,j,t) \leq \beta_{i,j}(t)\}$;
  $P_1(t) \leftarrow \{i \in A_1(t) : \forall\, j \in A_1(t) \setminus \{i\}, \tilde{\mathrm{M}}_{\varepsilon_1}(i,j,t) \geq \beta_{i,j}(t)\}$;
  $P_2(t) \leftarrow \{j \in P_1(t) : \nexists\, i \in A_1(t) \setminus P_1(t) \text{ s.t } \tilde{\mathrm{M}}_{\varepsilon_1}(i,j,t) \leq \beta_{i,j}(t)\}$;
  $A(t) \leftarrow A_1(t) \setminus P_2(t)$ and $P(t) \leftarrow P(t-1) \cup P_2(t)$;
  if $A(t) = \emptyset$ **then**
    break and output $P(t)$;
  if $|P(t-1) \cup P_1(t)| \geq k$ **then**
    break and output $P(t-1) \cup P_1(t)$;

We set $\delta = 0.1$ and we compare both algorithms on 2 type of randomly uniformly generated Bernoulli instances. For the first type we set $K = 10, D = 2$ and for the second one, we set $K = 50, D = 2$. For the instances with $K = 10$ we set $\varepsilon_1 = 0.05$ and run both algorithms on 2000 random Bernoulli instances with. For the second type of instances $K = 50$, we set $\varepsilon_1 = 0.1$ and we benchmark the algorithms on 500 random instances. The average size of the Pareto set was 2.90 (for $K = 10, D = 2$) and 4.51 ($K = 50, D = 2$). Figure 8 shows the average sample complexity of the algorithms for different values of $k \in \{1, \ldots, 5\}$. We can observe that the difference between ε_1-APE-k and APE for the 5 values reported is more important for $K = 50$ than for $K = 10$. Put together, theses experiments show that our algorithm is still preferable for the k-relaxation.



(a) $K = 10, D = 2, \varepsilon_1 = 0.05$ (b) $K = 50, D = 2, \varepsilon_1 = 0.1$

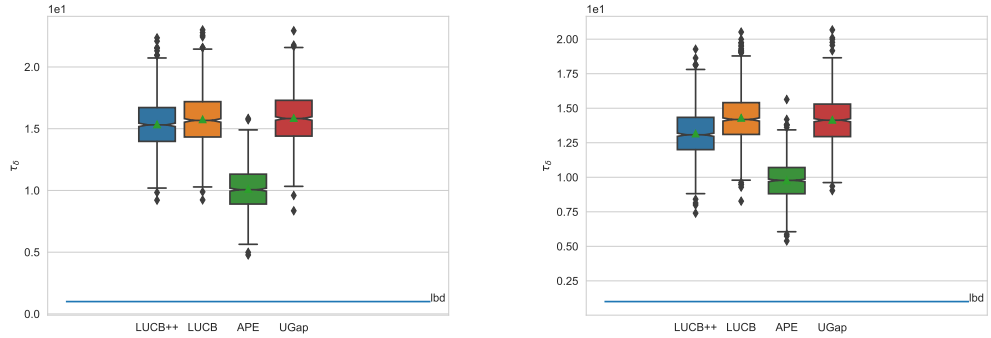
Figure 8: Average sample complexity on 2000 random instances (left) and 500 random instances (right).

I.3.4 Comparison to some BAI algorithms

We evaluate the performance of APE for Best Arm Identification ($D = 1$) on two randomly generated instances: one with $K = 5$ and means (rounded) $\mathcal{X}_1 := (0.25, 0.16, 0.87, 0.22, 0.98)$, and the second one with $K = 10$ and means (rounded) $\mathcal{X}_2 := (0.43, 0.33, 0.56, 0.85, 0.20, 0.93, 0.70, 0.82, 0.56, 0.78)$ We use the instantiation of APE with confidence bonuses on pair of arms but without the possible improvement invoked in Remark 2. UGap and LUCB are implemented with the tightest known (to our knowledge) confidence bonus taken from [16] (in spirit of the finite-time law of the iterated logarithm[28]). LUCB++ is used with the improved scheme given in [23]. We set $\delta = 0.01$ but the empirical error was way smaller. The results are averaged over 1000 independent trials.

On the y-axis is the sample complexity in units of (an approximation of) the lower bound of BAI for Gaussian rewards with $\sigma = 1/2$[4] ([22, 23]) :

$$H \log\left(\frac{1}{2.4\delta}\right) \text{ with } H = \sum_{i=1}^{K} \frac{1}{2\Delta_i^2}.$$



(a) $H = 94.84$ (b) $H = 223.20$

Figure 9: Empirical sample complexity expressed in units of the lower bound $H \log(1/2.4\delta)$ (blue line) on a random instance with $K = 5$ (left) and $K = 10$ (right), $\delta = 0.01$.

[4] as Bernoulli distributions are $1/2$-subgaussian