Preference-based Pure Exploration

Apurv Shukla[*]
Department of ECE, Texas A&M University
College Station, TX 77840
apurv.shukla@umich.edu

Debabrota Basu
Équipe School, Univ. Lille, Inria, CNRS
Centrale Lille, UMR-9189 - CRIStAL, France
debabrota.basu@inria.fr

Abstract

We study the preference-based pure exploration problem for bandits with vector-valued rewards. The rewards are ordered using a (given) preference cone $\mathcal{C}$ and our the goal is to identify the set of Pareto optimal arms. First, to quantify the impact of preferences, we derive a novel lower bound on the sample complexity for identifying the most preferred policy with confidence level $1-\delta$. Our lower bound elicits the role played by the geometry of the preference cone and punctuates the difference in hardness compared to existing best-arm identification variants of the problem. We further explicate this geometry when rewards follow Gaussian distributions. We then provide a convex relaxation of the lower bound. and leverage it to design Preference-based Track and Stop (PreTS) algorithm that identifies the most preferred policy. Finally, we show that sample complexity of PreTS is asymptotically tight by deriving a new concentration inequality for vector-valued rewards.

1 Introduction

Following COVID-19, the importance of reliable clinical trials and corresponding data acquisition to design effective drugs has gained wider recognition. However, conducting large-scale clinical trials is cost and time intensive as it requires working with large number of patients and following-up their medical conditions over time. In the past two decades, this has led to doubling in the cost to bring a drug to the market, i.e., to \$2.6 billion with a 12-year drug development horizon and 90% failure rate during the clinical trial (Mullard, 2014; Sun et al., 2022). However, due to the rise of systematic data acquisition about biological systems, pharmaceutical firms are interested in harvesting the collected data for drug discovery (Gaulton et al., 2012; Reker and Schneider, 2015). Thus, machine learning-based methods are increasingly studied and deployed as a promising avenue for identifying potentially successful drugs with less patient involvement, increasing the "hit rate", and speeding up the development process (Jayatunga et al., 2022; Smer-Barreto et al., 2023; Sadybekov and Katritch, 2023; Hasselgren and Oprea, 2024). But deciding whether a drug is successful depends on multiple and often conflicting objectives regarding safety, efficacy, and pharmacokinetic constraints (Lizotte and Laber, 2016). For example, COV-BOOST (Munro et al., 2021) demonstrates a phase II vaccine clinical trial conducted on 2883 participants to measure the immunogenicity indicators (e.g. cellular response, anti-spike IgG and NT_{50}) of different Covid-19 vaccines as a booster (third dose). Experts decide how different indicators are preferred over one another, and above different thresholds (Jayatunga et al., 2022). This motivates us to study a sequential decision-making problem, where *we aim to conduct minimum number of experiments to acquire informative data, and to reliably validate a hypothesis with multiple objectives by imposing preferences over them.*

[*]This work was done when the author was at Texas A& M University. The author is currently at the University of Michigan, Ann Arbor.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Problems of such nature can be modeled as a multi-armed bandit (in brief, bandits), which is an established framework for sequential decision-making under uncertainty (Lattimore and Szepesvári, 2020). In bandits, a learner has access to an instance of K decisions (or arms). Each arm $k \in \{1, \ldots, K\}$ corresponds to a probability distribution P_k of feedback (rewards) with unknown means μ_k. At each step $t \in \mathbb{N}$, the learner interacts with the instance by taking a decision k_t (analogously pulling an arm), and observes a noisy reward R_t from the corresponding distribution of rewards P_{k_t}. The goal of the learner is to identify the arm with the highest expected reward over a certain confidence level through minimum number of interactions with the instance. This is popularly known as a *fixed-confidence Best Arm Identification (BAI)* in bandit literature (Jamieson and Nowak, 2014; Garivier and Kaufmann, 2016; Soare et al., 2014), which is a special case of *pure exploration problems* (Even-Dar et al., 2006; Bubeck et al., 2009; Auer et al., 2016).

The bandit literature spanning over a century mostly focuses on a scalar reward, i.e., a single objective. In our problem, each reward R_t is a real-valued vector of $L \in \mathbb{N}$ objectives, and thus, the unknown mean vector of each arm $\mu_k \in \mathbb{R}^L$. Since the objectives can be often conflicting, there might not exist a single best arm. Rather, there exists a *Pareto Optimal Set of arms* (Drugan and Nowe, 2013; Auer et al., 2016). Given a set of preferences over the objectives, the Pareto Optimal Set consists of arms whose mean vectors dominate the mean vectors of any other arm outside the set. Keeping generality, we assume that preferences are defined by a cone of vectors $\mathcal{C} \subseteq \mathbb{R}^L$. Every $\mathcal{C}$ induces a set of partial or incomplete orders over the L objectives (Jahn et al., 2009; Löhne, 2011). Given the preference cone $\mathcal{C}$, we aim to *exactly* identify the *complete Pareto Optimal Set* with a confidence level $(1 - \delta) \in [0, 1)$ using as few interactions as possible. We refer to this problem *Preference-based Pure Exploration (**PrePEx**)*.

Recently, Auer et al. (2016); Kone et al. (2023a,b); crepon et al. (2024) consider a special case of PrePEx, where the preference is known. To the best of our knowledge, Ararat and Tekin (2023) and Korkmaz et al. (2023) are the only studies of PrePEx from frequentist and Bayesian angles, respectively. Here, we consider a frequentist approach as in (Ararat and Tekin, 2023). However, their goal is to identify points that are in the Pareto Optimal Set or very close to it. In contrast, we focus on exactly identifying the Pareto Optimal Set. Additionally, (Ararat and Tekin, 2023) propose gap-based elimination algorithm to solve the problem generalising the algorithm of (Even-Dar et al., 2006). But in BAI, there is another paradigm of designing efficient algorithms that solves and tracks the exact lower bound on the expected time to identify the best arm $(1 - \delta)$ correctly (Garivier and Kaufmann, 2016; Degenne and Koolen, 2019). We explore this paradigm for PrePEx and ask two questions:

What is the exact lower bound of PrePEx *for identifying the Pareto Optimal Set, and how to design a computationally tractable algorithm matching this bound?*

We address them affirmatively in our contributions:

1. Lower Bound for PrePEx. In Theorem 3.1, we study hardness of PrePEx problems by deriving the novel lower bound on the expected sample complexity of any algorithm to yield the exact Pareto Optimal Set with confidence $(1 - \delta)$. The challenge here is to extend the classical BAI lower bound (Garivier and Kaufmann, 2016) to a set of confusing instances given $\mathcal{C}$. We observe that unlike BAI, distinguishability of two arms in PrePEx depends on their projections on the cone polar to $\mathcal{C}$. We also show that our lower bound generalises the lower bound for pure exploration under known constraints (Carlsson et al., 2024). Additionally, we provide an exact characterization the lower bound further for Gaussian reward distributions in Theorem 3.2. It shows that the hardness depends on the bilinear projection of the mean matrix of arms onto the boundary of a normal cone of policies and the preferences. This is novel w.r.t. the existing gap-dependent lower bounds that hold either for a narrow range of μ_a's (Ararat and Tekin, 2023), or fixed preference (Kone et al., 2023a).

2. Algorithm Design. First, we observe that the optimisation problem in our lower bound involves minimisation over a non-convex set. We provide a convex relaxation of the problem based on ideas from disjunctive programming (Theorem 4.1 and 4.2). We then leverage this lower bound to propose a novel Track-and-Stop (Garivier and Kaufmann, 2016) style algorithm, called PreTS (Preference-based Track-and-Stop). In Theorem 4.3, we devise a new stopping rule that can handle the preference-aligned suboptimality gaps between the arms.

3. Sample Complexity Analysis. Finally, we provide an upper bound on sample complexity of PreTS. This requires us to define a distance metric between two pareto sets of arms, and proving a concentration bound with respect to this metric (Theorem 5.1). In Theorem 5.2, we prove that sample complexity of PreTS matches the convexified lower bound up to constants.

1.1 Related Works

In the past decade, works on multi-armed bandits also focuses on pure-exploration in addition to regret minimization. Regret minimization and pure-exploration differ in the sense when arms in pure-exploration are immediately discarded upon being deemed as sub-optimal, whereas, in the regret minimization setting, sub-optimal arms may still be played since they provide additional information about other arms. Pure-exploration problems have been considered in two settings: fixed-budget and fixed-confidence. The fixed-budget setting aims at bounding the probability of underestimating the best arm given a budget of samples. Audibert and Bubeck (2010) propose the first algorithm for the fixed budget setting. Here, the budget is divided into $K-1$ rounds and at the end of every round, the arms with the lowest empirical mean are discarded. On the other hand, best-arm identification is a version of the pure-exploration problem with scalar rewards (Even-Dar et al., 2006). In this setting, we are given a $\delta \in (0,1)$ and the goal is to identify the best-arm with probability at least $1-\delta$. Several strategies such as those based on elimination, adaptivity, racing, upper-confidence bounds have been proposed to minimize the number of expected pulls of an arm in the fixed confidence setting by (Kalyanakrishnan et al., 2012; Gabillon et al., 2012; Jamieson et al., 2014; Garivier and Kaufmann, 2016; Jedra and Proutiere, 2020). Arm rewards can be modeled as a vector with Gaussian Process (Zuluaga et al., 2016), linear rewards (Drugan and Nowe, 2013; Lu et al., 2019), and non-parametric rewards (Turgay et al., 2018), which can include contextual bandit formulations (Tekin and Turgay, 2017; Shukla, 2022). In recent past, the pure exploration techniques have been successfully applied in hyperparameter tuning (Li et al., 2018) and black-box optimization problems (Contal et al., 2013; Wang et al., 2021, 2022) demonstrating considerable performance gains.

In a marked deviation, given an instance of the bandit problem, the goal of this paper is to identify the entire Pareto front. A key observation in this regard is that there might be arms, which are sub-optimal for almost every objective but still lie on the Pareto front. Further, since sampling an arm returns a vector of rewards determining an arm-strategy that reduces the uncertainty in the estimate of every reward function is challenging. An immediate consequence of these differences is the fact that the complexity of identifying the Pareto front is different from that of best arm identification. Auer et al. (2016) consider the Pareto front identification problem in the multi-armed bandit model and establish sample complexity bounds for the problem in terms of relevant problem parameters in the fixed-confidence setting. The multi-armed bandit problem is further studied under cone-based preferences by Ararat and Tekin (2023). The main contribution of (Ararat and Tekin, 2023) are bounds on the sample complexity of the problem in terms of gap-based notions that depend on the cone. Karagözlü et al. (2024) builds upon this work to introduce adaptive elimination based algorithms for learning the Pareto front under incomplete preferences. When the reward vectors are Gaussian processes Korkmaz et al. (2023) propose an elimination based algorithm based for identifying the Pareto front. The goal in these works is to identify the set of arms that are ϵ close to the Pareto front as the sample complexity to identify the exact Pareto set can be very large. Kone et al. (2023a) consider the problem of identifying a relevant subset of the Pareto set using a single sampling strategy Adaptive Pareto Exploration, along with different stopping rules to consider variations of the Pareto Set Identification problem. crepon et al. (2024) consider the exact Pareto front identification problem in the multi-armed bandit setting but with fixed and known preference cone. They propose a lower bound and a computationally efficient gradient-based algorithm to implement a track-and-stop based strategy. To the best of our knowledge, ours is the first work to consider *the exact Pareto front identification problem from a pure-exploration perspective*. Therefore, our proposed framework can be used for identifying the Pareto front given a preference cone for several variants of the bandit problem including the standard multi-armed bandit problem, linear bandits, etc.

2 Preference-based Pure Exploration Problem

In this section, we formalise the fixed-confidence setting of preference-based pure exploration and introduce the notations.

Notations. For $n \in \mathbb{N}$, let $[n]$ denote the set $\{1,2,\dots,n\}$. We use $\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_\infty$ to denote the ℓ_1-norm, ℓ_2-norm and ℓ_∞-norm, respectively. For a vector z, $z^{(\ell)}$ denotes its ℓ^{th} component. Let e_ℓ denote the vector with 1 in the ℓ^{th} position and zero otherwise. Δ_K denotes the simplex on $[K]$. $d_{\mathrm{KL}}(P,Q)$ measures the KL-divergence between distributions P and Q. $\mathrm{vect}(A)$ is the vectorized version of matrix A. $\mathbf{1}$ is the vector of all 1's. Further details of notations are deferred to Appendix A.

Formulation. In PrePEx, a learner has access to a bandit instance with K arms. Each arm $k \in [K]$ corresponds to a reward distribution ν_k over $\mathbb{R}^L$ with unknown mean $\mu_k \in \mathbb{R}^L$. Here, L denotes the number of objectives corresponding to each arm. Thus, a bandit instance can be specified with the vector of mean rewards $\{\mu_k\}_{k=1}^K$. For brevity, we represent them with a matrix $M \in \mathbb{R}^{L \times K}$ such that its k^{th} column is μ_k. At each time $t \in \mathbb{N}$, the learner pulls an arm $k_t \in [K]$ and observes the corresponding reward vector R_t sampled from ν_{k_t}. In pure exploration, the learner typically focuses on finding the best arm, i.e. the arm with highest mean (Garivier and Kaufmann, 2016). In pure exploration, a more general setting of BAI, the learner aims to find a policy $\pi \in \Delta_K$ that dictates the arm-proportion to choose in order to maximize the expected reward obtained from the instance.

Following the vector optimization literature (Jahn et al., 2009; Ararat and Tekin, 2023), we assume that the learner has additionally access to an ordering cone $\mathcal{C}$.

Definition 2.1 (Ordering Cone). *A set $\mathcal{C} \subseteq \mathbb{R}^L$ is called a cone if $v \in \mathcal{C}$ implies that $\alpha v \in \mathcal{C}$ for all $\alpha \geq 0$. A solid cone has a non-empty interior, i.e., $\operatorname{int}(\mathcal{C}) \neq \emptyset$. A pointed cone contains the origin. A closed convex cone that is both pointed and solid is called an ordering cone.*

An ordering cone can be both polyhedral and non-polyhedral. Following the literature (Ararat and Tekin, 2023; Karagözlü et al., 2024), we consider access to a polyhedral ordering cone.

Definition 2.2 (Polyhederal Cone). *A cone $\mathcal{C}$ is a polyhedral cone if $\mathcal{C} \triangleq \{x \in \mathbb{R}^L \mid Ax \geq 0\}$, where $A \in \mathbb{R}^{K \times L}$ with rows $a_i^\top$. A is called the half-space representation of $\mathcal{C}$.*

Each polyhedral ordering cone induces a set of partial order on the reward vectors in $\mathbb{R}^L$. To ignore the redundancies and to focus on the bandit problem, we further assume that A is full row-rank and $\|A_i\|_2 = 1$ (Ararat and Tekin, 2023). Hereafter, we call them *preference cones*, and the vectors in the cone as the *preferences*. We refer to (Jahn et al., 2009; Löhne, 2011) for further details on cones.

Example 2.1 (Preference cones). *The positive orthant $\mathbb{R}_+^L$ is a polyhedral cone. This is the one used in pareto-set identification literature (Auer et al., 2016; Kone et al., 2023b; crepon et al., 2024). The cones with all non-negative entries are called solvency cones and used in finance (Kabanov, 2009). Another simple example is $\mathcal{C}_{\pi/3} \triangleq \{(r\cos\theta, r\sin\theta) \in \mathbb{R}^2 \mid r \geq 0 \wedge \theta \in [0, \pi/3]\}$, i.e. all the 2-dimensional vectors that makes an angle less than $\pi/3$ with the x-axis.*

Definition 2.3 (Partial Order). *For every $\mu, \mu' \in \mathbb{R}^L$, $\mu \preceq_{\mathcal{C}} \mu'$ if $\mu \in \mu' + \mathcal{C}$ and $\mu \prec_{\mathcal{C}} \mu$ if $\mu \in \mu' + int(\mathcal{C})$. Alternatively, $\mu \preceq_{\mathcal{C}} \mu'$ is equivalent to $z^\top(\mu -$*

The partial order induced by $\mathcal{C}$ induces further order over the set of arms $[K]$.

Definition 2.4 (Order over arms). *Consider two arms $i, j \in [K]$. (i) Arm i weakly dominated by arm j iff $\mu_j \preceq_{\mathcal{C}} \mu_i$. (ii) Arm i dominates arm j iff $\mu_i \prec_{\mathcal{C}\setminus\{0\}} \mu_j$. (iii) Arm i strongly dominates arm j iff $\mu_i \prec_{\mathcal{C}} \mu_j$.*

Definition 2.5 (Pareto Optimal Set). *An arm $i \in [K]$ is Pareto Optimal if it is not dominated by any other arm w.r.t. the cone $\mathcal{C}$. The Pareto Optimal Set $\mathcal{P}^*$ is defined as the set of all Pareto Optimal arms. Let $\mathcal{Z}$ be the set of all Pareto Frontiers on $[0,1]^K$.*

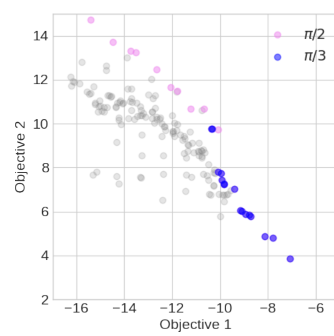

Figure 1: Effect of cone selection on size of Pareto optimal set

Given a preference cone, a learner aims to exactly identify the Pareto Optimal Set from a finite set of arms $[K]$ whose mean rewards belong to the Pareto Optimal Set w.r.t. $\mathcal{C}$. Alternatively, this vector optimization problem can be represented in the policy space as finding a policy $\pi \in \Delta_K$ supported on the Pareto optimal set of arms. This is given by the following vector optimization problem:

$$V(M) = \max_{\pi \in \Delta_K} M\pi \text{ over } \mathcal{C}. \tag{1}$$

In this context, we denote the set of Pareto optimal policies as $\Pi^*(M) \triangleq \arg\max_{\pi \in \Delta_K} M\pi$ over $\mathcal{C}$. We assume that $\Pi^*(M)$ is non-empty.

Example 2.2 (Pareto Optimal Sets for different cones). *Figure 1 illustrates the Pareto Optimal Sets among 2-dimensional mean vectors of 200 randomly selected arms under preference cones $\mathcal{C}_{\pi/2}$ and $\mathcal{C}_{\pi/3}$. We observe that the Pareto Optimal Sets for them (in pink and blue respectively), are* *completely different for same set of arms. Thus, we have to adapt to the available preferences to solve the aforementioned problem. As noted later, the geometry of this cone plays a crucial role in determining the Pareto front.*

In PrePEx, we consider the problem of Equation (1), when the mean matrix M is unknown a priori but bounded, i.e., the entries of $M,\ M_{ij} \in [M_{\min}, M_{\max}],\ M \in \mathcal{M}$. Having identified the policy, will lead us to identify the true Pareto front $\mathcal{P}^*$. In the noisy feedback setting, the reward at time t is $R_t = \mu_{k_t} + \eta_t$, where $\eta_t \in \mathbb{R}^L$ is the noise vector. We assume that the noise vectors η_t are independent of μ_{k_t} and also across time. Further, they are sub-Gaussian with parameter σ and adapted to the filtration $\mathcal{F}_t$, which is a standard assumption in the literature. A policy $\pi \in \Pi \subset \Delta^K$ is a randomized mapping from the history $\mathcal{H}_t$ to the probability simplex over the set of arms $[K]$. In Preference-based Pure EXploration (PrePEX) problem, the goal of the learner is to identify a Pareto optimal policy in Π^* (Equation (1)) given an instance M and a preference cone $\mathcal{C}$ while observing only noisy rewards from the arms, and also using as few observations as possible.

Definition 2.6 ($(1-\delta)$-correct PrePEX). *An algorithm for Preference-based Pure Exploration (PrePEx) is said to be $(1-\delta)$ correct if with probability $1-\delta$, it recommends a Pareto optimal policy $\pi \in \Pi^*$.*

For example, a pareto optimal policy for $\mathcal{C}_{\pi/2}$ would be a distribution in Δ_K with support on the arms corresponding to the pink reward vectors (Figure 1). For $\mathcal{C}_{\pi/3}$, it would be one with support on blue points.

3 Lower Bound on Sample Complexity

We begin by deriving a KL-divergence based lower bound for PrePEx using techniques from (Garivier and Kaufmann, 2016). Our lower bound is based on establishing a change-of-measure argument in the spirit of (Graves and Lai, 1997; Kaufmann et al., 2016). The lower bounds are derived first by defining a set of alternating instances Λ for a given bandit instance and then by trying to compute an optimal allocation policy $w \in \Delta_K$ that maximises the sum of minimum KL-divergence between any instance in Λ and the bandit instance under interaction. *The key insight of our work is to formulate the identification of Pareto Set problem in the policy space rather than in the arm space as done in antecedent literature.* This formulation helps us to derive the KL-based lower bound, which is more general than the existing suboptimality gap-based lower bounds (Auer et al., 2016; Ararat and Tekin, 2023; crepon et al., 2024).

The Alternating Instances with respect to Pareto Fronts. The learner needs to distinguish between all instance $\tilde{M} \in \mathcal{M} \setminus \{M\}$ for which the Pareto front associated with $\tilde{M}$ is different from the one associated with M. At first, given an optimal policy of M, say π^*, it would appear that the set of confusing instances is $\Lambda_{\pi^*}(M)^{\text{naive}} \triangleq \left\{\tilde{M} \in \mathcal{M} : \tilde{M}\pi^* \preceq_{\mathcal{C}} \max_{\pi\in\Pi} \tilde{M}\pi\right\}$. However, this is fallacious since the instances whose rewards dominate M can also confuse a policy π. Given a $\pi^\star$, the correct alternating set is the set of instances in $\mathcal{M}$ whose Pareto optimal set is not dominated by $\pi^\star$ corresponding to M.

$$\begin{aligned}\Lambda_{\pi^*}(M) &\triangleq \left\{\tilde{M} \in \mathcal{M} \setminus \{M\} : \max_{\pi\in\Pi} \tilde{M}\pi \not\preceq_{\mathcal{C}} \tilde{M}\pi^*\right\}\\ &= \left\{\tilde{M} \in \mathcal{M} \setminus \{M\} : \exists z \in \mathcal{C} \text{ s.t. } \max_{\pi\in\Pi} z^\top \tilde{M}\pi > z^\top \tilde{M}\pi^*\right\}.\end{aligned}$$

With this new alternate set defined, we now establish lower bounds on the performance of any PrePEX algorithm.

Theorem 3.1 (Lower Bound). *Given a bandit model $M \in \mathcal{M}$, a preference cone $\mathcal{C}$, and a confidence level $\delta \in [0,1)$, the expected stopping time of any $(1-\delta)$-correct PrePEx algorithm, to identify the Pareto Optimal Set is*

$$\mathbb{E}[\tau] \geq \mathcal{T}_{M,\mathcal{C}} \log\left(\frac{1}{2.4\delta}\right), \tag{2}$$

where, the expectation is taken over the stochasticity of both the algorithm and the bandit instance. Here, $\mathcal{T}_{M,\mathcal{C}}$ is called the characteristic time of the PrePEx instance $(\mathcal{M},\mathcal{C})$ and is expressed as

$$(\mathcal{T}_{M,\mathcal{C}})^{-1} \triangleq \sup_{w\in\Delta^K} \inf_{\substack{\pi\in\Pi\setminus\{\pi^*\}\\ \pi^*\in\Pi^*(M)}} \inf_{\tilde{M}\in\partial\Lambda_{\pi^*}(M)} \inf_{z\in\mathcal{C}} \sum_{k=1}^{K} w_k d_{KL}\left(z^\top M_k, z^\top \tilde{M}_k\right), \tag{3}$$

such that $\partial\Lambda_{\pi^*}(M) \triangleq \cup_{\Pi\setminus\{\pi^*\}} \left\{\tilde{M} \in \mathcal{M} : \exists\, z \in \mathcal{C},\ \langle vect\left(z(\pi-\pi^*)^\top\right), vect\big(\tilde{M}\big)\rangle = 0\right\}.$

Proof Intuition. First, we observe that an instance $\tilde{M}$ is in alternating set if there exists a $\pi \in \Pi\setminus\{\pi^*\}$ and $z \in \mathcal{C}$, such that $z^\top \tilde{M}(\pi-\pi^*) > 0$. If π and π^* were pure strategies, it would have been exactly $\inf_{z\in\mathcal{C}\setminus\{0\}} z^\top(\tilde{M}_a - \tilde{M}_{a^*}) > 0$. Let us denote the z achieving the $\inf$ as $z_{\inf}$, i.e., the preference for which $\tilde{M}_a$ and $\tilde{M}_{a^*}$ are least distinguishable. Thus, we observe that $z_{\inf}^\top \tilde{M}_a$ exactly functions as the mean of the arm a in an instance $\tilde{M}$, whereas $z_{\inf}^\top(\tilde{M}_{a^*} - M_a)$ acts as the suboptimality gap. Now, we extend this idea in the classical lower bound scheme to get a nested optimization problem with $\inf$ over $z \in \mathcal{C}$ and $\tilde{M}$ in the alternating set, and a sup over allocations $w \in \Delta_K$. We further show that the $\inf$ for $\tilde{M}$ appears at the boundary of the alternating set defined as $\partial\Lambda(M)$.

Discussions. (i) Novelty: In the best of our knowledge, this is the first lower bound for PrePEx with fixed confidence with an explicit KL-based dependence. All the existing lower bounds are gap dependent, and valid for a narrow range on mean vectors or known preference cone, i.e. the right orthant. Our proof does not need such assumptions. The gap-dependent bounds are special case of ours (cf. Theorem 3.2 for the case of Gaussian rewards).

(ii) Geometric Insights. Theorem 3.1 provides multiple geometric insights into the affect of the ordering cone $\mathcal{C}$ on the characteristic time. *First*, the alternating set $\Lambda_{\pi^*}(M)$ is piece-wise polyhedral and non-convex. This is a concern that we address later in Section 4.1. *Second*, there is an additional minimization over the vectors lying in the cone $\mathcal{C}$. We interpret the minimization over vectors in the cone as a *instance- and preference-dependent scalarization of the distance between the given instance* M *and the corresponding most-confusing instance in* $\Lambda_{\pi^*}(M)$. *Third*, in the proof, we show that the reward gap using the best policy π^* and a given policy π for the most confusing instance belongs to the polar cone $\mathcal{C}^\circ$ of the preference cone $\mathcal{C}$. The most confusing lies on the boundary of this polar cone and its projection the policy gaps ($\pi^* - \pi$). Further insights can be obtained by imagining the polar cone to be orthogonal to the cone $\mathcal{C}$. Then, the vector of reward-gaps for the most confusing instance for every objective is orthogonal to the generating rays of $\mathcal{C}$. These novel geometric insights are complementary to the existing algebraic and statistical insights available in the lower bound literature (Kone et al., 2023a; Ararat and Tekin, 2023).

3.1 Characteraization of Lower Bounds for Gaussians

To understand our lower bound better and to compare it with the literature, we present a reduction for Gaussian bandits. In Gaussian Bandits, we assume that the reward vectors of arm $a \in [K]$ are generated from a multivariate Gaussian distribution $\mathcal{N}(\mu_a, \Sigma)$, where the covariance is a diagonal matrix: $\Sigma \triangleq \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_L^2)$.

Theorem 3.2 (Lower Bound for Gaussian Bandits)**.** *1. Given any* $\pi^\star \in \Pi^*(M)$ *and* $N(\pi^*)$ *being the set of neighbouring policies of* π^**, the most confusing instance of* M *belongs to the set*

$$\left\{\tilde{M} \in \mathcal{M}\setminus\{M\} : \tilde{M} = M + \frac{1}{2}(\alpha(z)\mathbf{1}_L - \mathrm{vect}(\Sigma))(\pi^* - \pi)^\top \forall \pi \in N(\pi^*) \wedge z \in \mathcal{C}\setminus\{0\}\right\},$$

where $\alpha(z) \triangleq \frac{\sum_{i,\ell,k}\lambda_i v_i^{(\ell)}\Delta_k\mu_k^{(\ell)}}{\sum_{i,\ell,k}\frac{\lambda_i v_i^{(\ell)}\Delta_k^2}{w_k\pi_k^2}} = \frac{z^\top M(\pi-\pi^\star)}{(\sum_l z^l)\sum_k \frac{1}{w_k}\left(1-\frac{\pi_k^\star}{\pi_k}\right)^2}.$

2. The inverse of characteristic time, i.e. $(\mathcal{T}_{M,\mathcal{C}}^{\mathrm{Gauss}})^{-1}$*, for an instance* $(M, \mathcal{C})$ *is*

$$\max_{w\in\Delta^K}\ \inf_{\substack{\pi\in N(\pi^*)\\ \pi^*\in\Pi^*(M)}}\ \min_{z\in\mathcal{C}\setminus\{0\}} \left(1-\alpha(z)-\frac{1}{4\alpha(z)}\right) z^\top M(\pi^\star - \pi)\,.$$

Consequences. *First*, we observe an interesting phenomenon that a bilinear projection of mean matrix M on the preferences and policy gaps operates as an extension of suboptimality gap in classical BAI. This is a reminiscent of the lower bound for pure exploration under known linear constraints as in Carlsson et al. (2024) who show that the hardness of the problem depends only on the projection of the mean vector on the policy gap. In addition to similar projection structure, preferences introduce a novel bilinearity here. *Second*, we show how the lower bound inflates with the covariance matrix for each objective. This shows the richness of our KL-divergence based lower bound as opposed to gap-based bounds which have diffiuclty accomodating variance related terms directly.

Connection to existing results. Our result generalizes several existing lower bounds for BAI.

1. *BAI lower bound.* Our lower bound is able to recover that of Kaufmann et al. (2016) for the standard BAI problem with fixed confidence. In the case of the standard BAI problem, the ordering cone is given by $\mathcal{C} \triangleq \mathbb{R}_+$ and therefore the minimization over $\mathcal{C}$ in (3) becomes redundant. The definition of the alternating set is then given by the set of instances which have a different optimal arm than μ which is exactly the set considered in (Kaufmann et al., 2016).

2. *Pure exploration under known constraints.* Our lower bound is able to recover the lower bound of Carlsson et al. (2023) for the BAI problem with fixed confidence and linear constraints. This is the case with $L = 1$ and the ordering cone being $\mathcal{C} \triangleq \mathbb{R}_+$ making the minimization over $z \in \mathcal{C}$ in (3) redundant. $\Lambda_{\pi^*}(M)$ becomes $\Lambda_{\pi^*}(M) = \{\tilde{\mu} : \max_{\pi \in \Pi} \tilde{\mu}^\top \pi \geq \mu^\top \pi^*\}$.

4 Algorithm Design: PreTS

In this section, we propose an algorithm that tracks the lower bound. However, this is not straightforward since the alternating set is non-convex. We first propose a convex relaxation for this set and then, design a Track and Stop style algorithm, called PreTS.

4.1 Convex Relaxation of the Lower Bound

One of the major differences regarding the structure of lower bounds compared to a standard BAI problem is that $\Lambda_{\pi^*}(M)$ is a piece-wise polyhederon, i.e., a union of hyperplanes. Each hyperplance corresponds to a policy $\pi \in \Pi \setminus \{\pi^*\}$. In order to make the optimization problem tractable and obtain a convex program, we relax $\Lambda_{\pi^*}(M)$ using its convex closure, denoted by $\text{ch}(\Lambda_{\pi^*}(M))$. We note that the construction of such a convex relaxation for the purpose of track-and-stop (when the lower bound problem is non-convex) has been done in the MDP setting Al Marjani and Proutiere (2021). We define $\text{ch}(\Lambda_{\pi^*}(M))$ in Theorem 4.1 by formulating it as a disjunctive program, which we can reformulate further as a linear program (Balas, 1985).

Theorem 4.1. *Let* $\mathcal{F} \triangleq \cup_{\Pi \setminus \pi^*} \left\{\tilde{M} \in \mathcal{M} : \exists\, z \in \mathcal{C}, \langle \mathit{vect}\left(z^\top(\pi - \pi^*)\right), \mathit{vect}(\tilde{M})\rangle = 0\right\}$. *Fix* $z \in \mathcal{C}$ *such that* $z = \sum_i \alpha_i v_i$. *Then, we have* $\mathit{ch}(\mathcal{F}) = \mathcal{I}$, *where* $\mathcal{I}$ *is defined as*

$$\mathcal{I} \triangleq \{\tilde{M} \in \mathcal{M} : \gamma^\top \mathit{vect}(\tilde{M}) \geq \gamma_0, \gamma = \sum_i u_i \alpha_i \mathit{vect}(v_i^\top(\pi - \pi^*)), \gamma_0 \leq u_i \sum_i \alpha_i v_i^\top \pi^*\}. \quad (4)$$

Using the convex hull (Eq. (4)), we quantify the optimal value for a given allocation w as

$$\overline{\mathcal{V}}_{\mathcal{C}}(w, M) \triangleq \min_{\tilde{M} \in \text{ch}(\Lambda_{\pi^*}(M))} \inf_{z \in \mathcal{C}} \sum_{k=1}^{K} w_k d_{\text{KL}}\left(z^\top M, z^\top \tilde{M}\right).$$

The corresponding optimal allocation is

$$\overline{w}^*(M) = \arg\max_{w \in \Delta^K} \inf_{\substack{\pi \in \Pi \setminus \pi^* \\ \pi^* \in \Pi^*(M)}} \min_{\tilde{M} \in \text{ch}(\Lambda_{\pi^*}(M))} \inf_{z \in \mathcal{C}} \sum_{k=1}^{K} w_k d_{\text{KL}}\left(z^\top M, z^\top \tilde{M}\right). \quad (5)$$

Hereafter, we consider Equation (5) as the optimization problem to be tracked. To compute $\overline{\mathcal{V}}_{\mathcal{C}}(w, M)$, we need access to the true instance M which is not available to us. Our Track-and-Stop strategy is based on repeatedly sampling an arm to construct an estimate of M, i.e. M_t, and exploiting continuity properties of $\overline{\mathcal{V}}_{\mathcal{C}}(w, M)$ to show that $\overline{\mathcal{V}}_{\mathcal{C}}(w, M_t) \to \overline{\mathcal{V}}_{\mathcal{C}}(w, M)$ and the cumulative number of arm plays $N_{t,k} \to w_k,\ w_k \in \overline{w}^*(M)$. These properties ensure that it makes sense to design a Track and Stop style algorithm for this problem.

Theorem 4.2 (Analytical Properties)**.** *For all* $M \in \mathbb{R}^{L \times K}$ *and all preference cones* $\mathcal{C}$, *we get 1. The mapping* $(w, M) \to \overline{\mathcal{V}}_{\mathcal{C}}(w, M)$ *is continuous. 2. The characteristic time mapping* $M \to \overline{\mathcal{T}_{M,\mathcal{C}}}$ *is continuous. 3. The set valued function* $M \to \overline{w}^*(M)$ *is upper-hemicontinuous. 4. The set* $\overline{w}^*(M)$ *is convex.*

Discussion: Cost of Convexification. For Gaussian bandits, as we can get the analytical form of the most confusing instance $\tilde{M}$ (Theorem 3.2), we do not pay any extra cost of convexification. In the non-Gaussian settings, where we cannot find such analytical forms for the most confusing instances, the minimum value of the inner minimisation problem under convex hull (Equation (5)) can go lower

Algorithm 1 Preference-based Track-and-Stop (PreTS)

1: **Input:** Confidence parameter δ,
2: **if** $Z(t) \geq \beta(t, \delta)$ **then**
3: Compute $w_t \leftarrow \arg\max_{w \in \Delta^K} \overline{\mathcal{V}}_{\mathcal{C}}(w, \hat{M}_t)$
4: Play $k_t \leftarrow \arg\min_{k \in [K]} \left| N_{k,t} - \sum_{s=1}^{t} w_s \right|$
5: Observe reward r_t
6: Construct estimator $\hat{M}_t$ using Equation (6)
7: **end if**
8: Construct a Pareto Front $\hat{\mathcal{P}}_\tau$ from empirical means $\hat{M}_\tau$
9: Return: $\hat{\mathcal{P}}_\tau$

than the minimum value found in the original non-convex set of instances (Equation (3)). Thus, the characteristic time attained by solving the convex relaxation might be higher than that of the original lower bound. Hence, an algorithm solving the convex relaxation has a higher stopping time. But convexification is essential for computational feasibility of a lower bound-tracking algorithm for PrePEx. This computational-statistical trade-off will be interesting to study in the future.

4.2 Algorithm: Preference-based Track-and-Stop (PreTS)

We now construct a general recipe to design a PrePEx algorithm when we do not have access to the true instance M. The fundamental element of any such recipe is constructing an estimate of M. For a given set of observed rewards $\{R_t\}_{t=1}^{T}$, we obtain a column-wise least-squares estimator of M by solving the convex optimization problem in Equation (6).

$$\hat{M}_t^{(k)} = \arg\min_{M^{(k)} \in \mathbb{R}^L} \| \sum_{s=1}^{t} R_s^{(k)} - M^{(k)} \|_2^2 + \lambda_t \|M^{(k)}\|_2. \tag{6}$$

Now, we elaborate the three key components of our PrePEx algorithm Preference-based Track and Stop (PreTS, Algorithm 1).

1. **Sampling Rule:** For the sampling rule, we consider a Track-and-Stop strategy (Garivier and Kaufmann, 2016). It tracks the optimal proportion of arm sampling by plugging in the empirical estimates of means and empirical count $N_{k,t}$ in the convexified lower bound. This leads to an allocation policy with an improved information acquisition.

2. **Stopping Rule:** Our ultimate stopping goal is to identify arms that are on the Pareto front. Based on this, we define the confidence set as:

$$c(t, \delta) \triangleq \left\{ \tilde{M} \in \mathcal{M} : \min_{z \in \mathcal{C}} \sum_{k} N_{k,t} d_{\mathrm{KL}}(z^\top \hat{M}_t^{(k)}, z^\top \tilde{M}^{(k)}) \leq \log \frac{c_1 t^3}{\delta} \right\}, \tag{7}$$

where c_1 is specified in the appendix. Our first claim is to show that the true instance belongs to the confidence ellipsoid with high-probability.

Lemma 4.1 (Confidence Ball)**.** *There exists a constant $c_1 > 0$ such that for any $t \in \mathbb{N}$ and $c(t, \delta) \triangleq \log \frac{c_1 t^3}{\delta}$, we have $\mathbb{P}(M \notin c(t, \delta)) \leq \delta$.*

Thus, we can now formalise the corresponding Chernoff stopping rule as

$$\min_{\tilde{M} \in \mathrm{ch}\left(\Lambda_{\pi^*}(\hat{M}_t)\right)} \min_{z \in \mathcal{C}} \sum_{k} N_{k,t} d_{\mathrm{KL}} \left(z^\top \hat{M}_t^{(k)}, z^\top \tilde{M}^{(k)} \right) \geq c(t, \delta) \tag{8}$$

Given the estimates $\hat{M}_t$ constructed using Equation (6), the problem in Equation (8) can be solved efficiently. Next, we show that upon stopping with Equation (8), PreTS returns the true Pareto Front $\mathcal{P}^*$ with probability $1 - \delta$. Let $\hat{\mathcal{P}}_t$ denote the estimated Pareto Front at time t, which is constructed using estimates $\hat{M}_t$. Then at stopping time τ, we have

$$\begin{aligned} \mathbb{P}\left(\mathcal{P}^* \neq \hat{\mathcal{P}}_t\right) &\leq \mathbb{P}\left(\exists\, t \in \mathbb{N} : \sum_{k,\ell} N_{k,t} d_{\mathrm{KL}} \left(z^\top \hat{M}_t^{(k)}, z^\top M^{(k)} \right) \geq c(t.\delta) \right) \\ &\leq \sum_{t=1}^{\infty} \mathbb{P}\left(\sum_{k,\ell} N_{k,t} d_{\mathrm{KL}} \left(z^\top \hat{M}_t^{(k)}, z^\top M^{(k)} \right) \geq c(t, \delta) \right) \leq \delta \end{aligned}$$

where, the last inequality is true due to Theorem 4.3, a concentration result on the KL-divergence with preference projected mean rewards.

Theorem 4.3. *For all $\gamma \geq (KL+1)$ and $n \in \mathbb{N}$, we have that:*

$$\mathbb{P}\left[\sum_k N_{k,t} d_{KL}\left(z^\top \hat{M}_t^{(k)}, z^\top M^{(k)}\right) \geq \gamma\right] \leq \exp\left(-\gamma\right)\left(\frac{\lceil \gamma \log(\gamma) \rceil}{KL}\right)^{KL} \exp\left(KL+1\right)$$

3. **Recommendation Rule:** At the end of stopping time τ, the algorithm returns an estimate of the Pareto Front $\hat{\mathcal{P}}_\tau$.

5 Upper Bound on Sample Complexity

Now, we prove upper bound on the expected sample complexity of PreTS. This requires us to the Track-an-Stop proof technique. But the challenge is to show concentration of the pareto fronts under a suitable metric.

Concentrating to the Pareto Front. To show that upon stopping the algorithm returns the true Pareto Frontier, we need to establish a valid metric to show such convergence. Usually, the distance between sets is measured using the Hausdorff metric (Costantini and Vitolo, 1995), i.e. $d_H(\hat{\mathcal{P}}_\tau, \mathcal{P}) \triangleq \max\left\{\sup_{k \in \hat{\mathcal{P}}_\tau} \inf_{k' \in \mathcal{P}} \|\mu_k - \mu_{k'}\|_\infty,\ \sup_{k \in \mathcal{P}} \inf_{k' \in \hat{\mathcal{P}}_\tau} \|\mu_k - \mu_{k'}\|_\infty\right\}$. But the Hausdorff distance only defines a pseudo-distance between sets and $\mathcal{Z}$ may not be closed under this metric. To circumvent this issue, we build upon the notion of a gap-based metric considered in the antecedent literature (Auer et al., 2016) to measure the distance between the mean reward of an arm and a given Pareto Front. We extend it to a distance metric between elements in the space of Pareto Fronts $\mathcal{Z}$.

Definition 5.1 (Distance from Pareto Front)**.** *The distance of the mean of arm k from the Pareto Front $\mathcal{P}^*$ is $d(k, \mathcal{P}^*) \triangleq \inf_{\varepsilon \geq 0} \varepsilon$, such that $\mu_k + \varepsilon \mathbf{1} \not\preceq_{\mathcal{C}} \mu_{k'},\ k' \in \mathcal{P}^*$. Equivalently,*

$$d(k, \mathcal{P}^*) = \inf_{k' \in \mathcal{P}^*} \max\left\{0, \sup_{z \in \mathcal{C} \cap \mathbb{B}(1)} z^\top \left(\mu_{k'} - \mu_k\right)\right\}, \tag{9}$$

Definition 5.2 (Distance between Pareto Fronts)**.** *We define the metric between Pareto Fronts $d(\cdot,\cdot) : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ as $d_{\mathrm{P}}\left(\hat{P}, \mathcal{P}^*\right) \triangleq \max\left\{\sup_{k \in \hat{P}} d(k, \mathcal{P}^*), \sup_{k \in \mathcal{P}^*} d(k, \hat{P})\right\}$.*

In the appendix, we establish that (i) $d(\cdot,\cdot)$ is a valid metric on $\mathcal{Z}$, and (ii) $\mathcal{Z}$ is compact and complete under $d(\cdot,\cdot)$. Now, we leverage this metric to show that the Pareto Front defined by the arm-wise constructed estimator $\hat{M}_t$ concentrates towards the true Pareto Front.

Theorem 5.1 (Concentration of mean estimates)**.** *For any pair $(i,j) \in [K] \times [K]$ and $z \in \mathcal{C}$, we have*

$$\left\| z^\top \left(\mu_i - \mu_j\right) - z^\top \left(\hat{\mu}_{i,t} - \hat{\mu}_{j,t}\right) \right\| \leq \beta_{ij}(t)\,.$$

$\beta_{ij}^2(t) \triangleq 4\left(h\left(\frac{\log(\frac{K_1}{\delta})}{2}\right) + \sum_{a \in \{i,j\}} \log\left(4 + \log(N_a(t))\right)\right)\left(\sum_{a \in \{i,j\}} \frac{1}{N_a(t)}\right)\left(\frac{1}{N_{i,t}} \sum_{\ell \in [L]} z^{(\ell)}\right)$ *and $K_1 \triangleq \frac{K(K-1)L}{2},\ h(\cdot) \approx x + \log(1+x)$.*

Proof Sketch. This is a consequence of jointly applying a vectorial concentration result for multiple-objectives of each arm (Kaufmann and Koolen, 2021), and pairwise time-uniform concentration bounds (Kone et al., 2023a). A key observation here is that the confidence radii depends on the magnitude of the preference vector z and scales with different objectives accordingly.

Sample Complexity of PreTS. Using this new concentration result for the Pareto Front and the stopping rule in Equation 8, we derive an upper bound on the expected stopping time of PreTS.

Theorem 5.2 (Upper Bound on Sample Complexity)**.** *For any $\alpha > 0$ and $c(t,\delta)$ defined in (8), we have that the stopping time satisfies*

$$\lim_{\delta \to 0} \frac{\mathbb{E}[\tau]}{\log\left(\frac{1}{\delta}\right)} \leq \overline{\mathcal{T}_{M,\mathcal{C}}}\ \forall\ M \in \mathbb{R}^{L \times K}$$

The basic outline of the proof follows a general strategy to prove Track-and-Stop result. However, the new arguments lie in establishing that the Pareto fronts converge under a suitable metric sufficiently fast. Our proof implies that PreTS matches the convex relaxation of the lower bound asymptotically at the corresponding risk level δ. Strictly, speaking this is not *asymptotically optimal* since, we do not track the exact lower-bound.

6 Conclusion and Future Works

We study the fixed-confidence version of preference-based pure exploration problem under linear stochastic bandit feedback, where each arm corresponds to a reward vector ordered according to a preference cone. We derive a novel lower bound for this problem. We leverage the lower bound further to derive a track-and-stop based algorithm for PrePEx problem. As future work, it would be interesting to verify our results on a real-world datasets.

Additionally, it would be interesting and challenging to study how other asymptotically optimal pure exploration strategies, e.g. gamified explorers (Degenne and Koolen, 2019), top-two algorithms (), can be adapted to this setting. In general, improving the computational efficiency and studying the optimality gap with respect to the non-convex lower-bound problem would be of fundamental interest.

References

Al Marjani, A. and Proutiere, A. (2021). Adaptive sampling for best policy identification in markov decision processes. In *International Conference on Machine Learning*, pages 7459–7468. PMLR.

Ararat, C. and Tekin, C. (2023). Vector optimization with stochastic bandit feedback. In *International Conference on Artificial Intelligence and Statistics*, pages 2165–2190. PMLR.

Audibert, J.-Y. and Bubeck, S. (2010). Best arm identification in multi-armed bandits.

Auer, P., Chiang, C.-K., Ortner, R., and Drugan, M. (2016). Pareto front identification from stochastic bandit feedback. In *Artificial intelligence and statistics*, pages 939–947. PMLR.

Balas, E. (1985). Disjunctive programming and a hierarchy of relaxations for discrete optimization problems. *SIAM Journal on Algebraic Discrete Methods*, 6(3):466–486.

Berge, C. (1877). *Topological spaces: Including a treatment of multi-valued functions, vector spaces and convexity*. Oliver & Boyd.

Bubeck, S., Munos, R., and Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory: 20th International Conference, ALT 2009, Porto, Portugal, October 3-5, 2009. Proceedings 20*, pages 23–37. Springer.

Carlsson, E., Basu, D., Johansson, F., and Dubhashi, D. (2024). Pure exploration in bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 334–342. PMLR.

Carlsson, E., Basu, D., Johansson, F. D., and Dubhashi, D. (2023). Pure exploration in bandits with linear constraints. *arXiv preprint arXiv:2306.12774*.

Contal, E., Buffoni, D., Robicquet, A., and Vayatis, N. (2013). Parallel gaussian process optimization with upper confidence bound and pure exploration. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 225–240. Springer.

Costantini, C. and Vitolo, P. (1995). On the infimum of the hausdorff metric topologies. *Proceedings of the London Mathematical Society*, 3(2):441–480.

crepon, e., Garivier, A., and M Koolen, W. (2024). Sequential learning of the Pareto front for multi-objective bandits. In Dasgupta, S., Mandt, S., and Li, Y., editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3583–3591. PMLR.

Degenne, R. and Koolen, W. M. (2019). Pure exploration with multiple correct answers. In *Neural Information Processing Systems*.

Donsker, M. D. and Varadhan, S. S. (1975). Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on pure and applied mathematics*, 28(1):1–47.

Drugan, M. M. and Nowe, A. (2013). Designing multi-objective multi-armed bandits algorithms: A study. In *The 2013 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.

Even-Dar, E., Mannor, S., and Mansour, Y. (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(Jun):1079–1105.

Gabillon, V., Ghavamzadeh, M., and Lazaric, A. (2012). Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, pages 3212–3220.

Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., et al. (2012). Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107.

Graves, T. L. and Lai, T. L. (1997). Asymptotically efficient adaptive choice of control laws incontrolled markov chains. *SIAM journal on control and optimization*, 35(3):715–743.

Hasselgren, C. and Oprea, T. I. (2024). Artificial intelligence for drug discovery: Are we there yet? *Annual Review of Pharmacology and Toxicology*, 64:527–550.

Jahn, J. et al. (2009). *Vector optimization*. Springer.

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. (2014). lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439.

Jamieson, K. and Nowak, R. (2014). 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. IEEE.

Jayatunga, M. K., Xie, W., Ruder, L., Schulze, U., and Meier, C. (2022). Ai in small-molecule drug discovery: a coming wave. *Nat. Rev. Drug Discov*, 21(3):175–176.

Jedra, Y. and Proutiere, A. (2020). Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017.

Kabanov, Y. (2009). *Markets with Transaction Costs Mathematical Theory*. Springer.

Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. (2012). Pac subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662.

Karagözlü, E. M., Yıldırım, Y. C., Ararat, C., and Tekin, C. (2024). Learning the pareto set under incomplete preferences: Pure exploration in vector bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3070–3078. PMLR.

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

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

Kone, C., Kaufmann, E., and Richert, L. (2023a). Adaptive algorithms for relaxed pareto set identification. *Advances in Neural Information Processing Systems*, 36:35190–35201.

Kone, C., Kaufmann, E., and Richert, L. (2023b). Bandit pareto set identification: the fixed budget setting. *arXiv preprint arXiv:2311.03992*.

Korkmaz, İ. O., Ararat, C., and Tekin, C. (2023). Bayesian vector optimization with gaussian processes.

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. (2018). Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52.

Lizotte, D. J. and Laber, E. B. (2016). Multi-objective markov decision processes for data-driven decision support. *Journal of Machine Learning Research*, 17(210):1–28.

Löhne, A. (2011). *Vector optimization with infimum and supremum*. Springer Science & Business Media.

Lu, S., Wang, G., Hu, Y., and Zhang, L. (2019). Multi-objective generalized linear bandits. *arXiv preprint arXiv:1905.12879*.

Mullard, A. (2014). New drugs cost us $2.6 billion to develop. *Nature reviews drug discovery*, 13(12).

Munro, A. P., Janani, L., Cornelius, V., Aley, P. K., Babbage, G., Baxter, D., Bula, M., Cathie, K., Chatterjee, K., Dodd, K., et al. (2021). Safety and immunogenicity of seven covid-19 vaccines as a third dose (booster) following two doses of chadox1 ncov-19 or bnt162b2 in the uk (cov-boost): a blinded, multicentre, randomised, controlled, phase 2 trial. *The Lancet*, 398(10318):2258–2276.

Peskun, P. H. (1973). Optimum monte-carlo sampling using markov chains. *Biometrika*, 60(3):607–612.

Reker, D. and Schneider, G. (2015). Active-learning strategies in computer-assisted drug discovery. *Drug discovery today*, 20(4):458–465.

Sadybekov, A. V. and Katritch, V. (2023). Computational approaches streamlining drug discovery. *Nature*, 616(7958):673–685.

Shukla, A. (2022). *Optimization and Learning in Dynamic Environments: Models and Algorithms*. Columbia University.

Smer-Barreto, V., Quintanilla, A., Elliott, R. J., Dawson, J. C., Sun, J., Campa, V. M., Lorente-Macías, Á., Unciti-Broceta, A., Carragher, N. O., Acosta, J. C., et al. (2023). Discovery of senolytics using machine learning. *Nature communications*, 14(1):3445.

Soare, M., Lazaric, A., and Munos, R. (2014). Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, pages 828–836.

Sun, D., Gao, W., Hu, H., and Zhou, S. (2022). Why 90*Acta Pharmaceutica Sinica B*, 12(7):3049–3062.

Tekin, C. and Turgay, E. (2017). Multi-objective contextual bandits with a dominant objective. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

Turgay, E., Oner, D., and Tekin, C. (2018). Multi-objective contextual bandit problem with similarity information. In *International Conference on Artificial Intelligence and Statistics*, pages 1673–1681. PMLR.

Wang, J., Basu, D., and Trummer, I. (2022). Procrastinated tree search: Black-box optimization with delayed, noisy, and multi-fidelity feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10381–10390.

Wang, J., Trummer, I., and Basu, D. (2021). Demonstrating udo: A unified approach for optimizing transaction code, physical design, and system parameters via reinforcement learning. In *Proceedings of the 2021 International Conference on Management of Data*, pages 2794–2797.

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

Appendix

Table of Contents

A Notations

Notation	Description
$\mathcal{C}, \prec_{\mathcal{C}\setminus\{0\}}$	Given convex cone and induced partial order
K, L	number of arms and objectives
$\mathcal{P}^*, \hat{\mathcal{P}}_t$	Ground truth Pareto set and estimated Pareto set
$\mathcal{Z}$	Space of all Pareto Frontiers on $[0,1]^K$
$M \in \mathbb{R}^{K\times L}$	matrix with mean reward of K arms
$r_{k_t}, \mu_{k_t}, \eta_t$	observed reward, mean reward and noise
$c(t,\delta)$	Confidence ball at time t with confidence δ
$d_{\mathrm{H}}(X,Y)$	Hausdroff distance between sets X and Y
$d_{\mathrm{P}}\left(P,\hat{P}\right)$	Distance metric between Pareto Fronts P and $\hat{P}$
w	Allocation vector
Π	Family of policies
$\hat{\mu}_{k,t}^{(\ell)}, \mu_k^{(\ell)}$	Estimated and true of mean rewards
ch (S)	Convex hull of set S
$\Lambda_{\pi^*}(M)$	Set of alternating instances associated with M

Table 1: Table of Notations

B Proofs of Lower Bounds

B.1 Generic Lower Bound: Proof of Theorem 3.1

Proof. The proof follows the basic structure of constructing a lower bound as in Kaufmann et al. (2016). Recall that their inverse of characteristic time is given by

$$\mathcal{T} = \sup_{w\in\Pi} \inf_{\tilde{M}\in\Lambda_{\pi^*}(M)} \sum_k w_k d_{\text{KL}}\left(M_k, \tilde{M}_k\right) \tag{10}$$

Adopting the bound as is leads to incorrect definition as well as The main challenge for our setting is describing the alternating set $\Lambda_{\pi^*}(M)$ and scalarization of the given instance $M \in \mathcal{M}$. Given a matrix of arm-objective mean-rewards M, ordering cone $\mathcal{C}$ and a family of policies Π, the Pareto front is the set of optimal values (ordered wrt $\mathcal{C}$) of

$$\max_{\pi\in\Pi} M^\top \pi\,. \tag{11}$$

Let $\pi^* \in \arg\max_{\pi\in\Pi} M^\top\pi$. Throughout we assume that the optimal solution to (11) is unique.

- **Step 1: Constructing set of alternative instances**
  The set of confusing instances given Π and $\mathcal{C}$ is the set of all matrices $\tilde{M}$ which have a different Pareto front than M when using the policy π^*. Therefore, the optimal values of (11) with instance M are not-dominated by those of instance $\tilde{M}$. Hence, the set of alternative instances, $\Lambda_{\pi^*}(M)$ is given by:

  $$\Lambda_{\pi^*}(M) := \left\{\tilde{M}\in\mathcal{M} : \max_{\pi\in\Pi} \tilde{M}^\top\pi \npreceq_{\mathcal{C}} \tilde{M}^\top\pi^*\right\}$$

  Let $\pi' \in \arg\max_{\pi\in\Pi} \tilde{M}^\top\pi$ over $\mathcal{C}$ which implies $\tilde{M}^\top\pi' \npreceq_{\mathcal{C}} \tilde{M}^\top\pi^*$ or equivalently:

  $$\exists\, z\in\mathcal{C},\ \pi\in\Pi\setminus\{\pi^*\} \text{ s.t. } z^\top\tilde{M}\pi > z^\top\tilde{M}\pi^*$$

  Therefore, the alternative set can be written as:

  $$\begin{aligned}\Lambda_{\pi^*}(M) &\triangleq \cup_{\pi\in\Pi\setminus\{\pi^*\}}\left\{\tilde{M}\in\mathcal{M} :\ \exists\, z\in\mathcal{C}, z^\top\tilde{M}\pi > z^\top\tilde{M}\pi^*\right\}\\ &= \cup_{\pi\in\Pi\setminus\{\pi^*\}}\left\{\tilde{M}\in\mathcal{M} : \exists\, z\in\mathcal{C} : z^\top\tilde{M}\cdot(\pi-\pi^*) > 0\right\}\end{aligned}$$

  where, $\cdot$ represents a bilinear product and, its complement is given by:

  $$\begin{aligned}\overline{\Lambda_{\pi^*}(M)} &\triangleq \cap_{\pi\in\Pi\setminus\{\pi^*\}}\left\{\tilde{M}\in\mathcal{M} : \forall\, z\in\mathcal{C} : z^\top\tilde{M}\cdot(\pi-\pi^*) \le 0\right\}\\ &= \cap_{\pi\in\Pi\setminus\{\pi^*\}}\left\{\tilde{M}\in\mathcal{M} : \tilde{M}\cdot(\pi-\pi^*) \in \mathcal{C}^\circ\right\}\end{aligned}$$

  where, $\text{ri}(\mathcal{C}^\circ)$ denotes the relative interior of the polar cone to $\mathcal{C}$. Since $\mathcal{C}$ is a polyhederal cone, it is closed and convex and therefore, its polar cone is non-empty, closed and convex. Therefore, $\overline{\Lambda_{\pi^*}(M)}$ is non-empty. We now show that given π^*, and π, the hardest instances $\tilde{M}\in\mathcal{M}$ are such that:

  $$\tilde{M}\cdot(\pi-\pi^*) \in \text{bd}(C^\circ)$$

  and the alternating set can be further characterized as:

  $$\overline{\Lambda_{\pi^*}(M)} := \cap_{\pi\in\Pi\setminus\pi^*}\left\{\tilde{M}\in\mathcal{M} : \forall\, z\in C,\ z^\top\tilde{M}\cdot(\pi-\pi^*) = 0\right\} \tag{12}$$

- **Step 2: Hardest instance lies on the boundary**
  Fix $\pi' \in \Pi\setminus\{\pi^*\}$ and let $M' \in \left\{\tilde{M}\in\mathcal{M} : \tilde{M}\cdot(\pi'-\pi^*) \in \text{ri}(\mathcal{C}^\circ)\right\}$. Then, by convexity of $\left\{\tilde{M}\in\mathcal{M} : \exists\, z\in\mathcal{C} : z^\top\tilde{M}\cdot(\pi-\pi^*) > 0\right\}$ there exists $M'' \in \left\{\tilde{M}\in\mathcal{M} : \tilde{M}\cdot(\pi'-\pi^*) \in \text{bd}(\mathcal{C}^\circ)\right\}$ such that $\left|z^\top M_k - z^\top M'_k\right| \ge \left|z^\top M_k - z^\top M''_k\right|, \forall\, z\in\mathcal{C}$. Since $d_{\text{KL}}\left(z^\top M, \cdot\right)$ is decreasing, we have: $d_{\text{KL}}\left(z^\top M_k, z^\top M'_k\right) \ge d_{\text{KL}}\left(z^\top M_k, z^\top M''_k\right)$.

- **Step 3: Concluding arguments**

  Using the above arguments we see that the minimum argument of $\sum_k w_k d_{\text{KL}}\left(z^\top M', z^\top M''\right)$ is such that $M'' \in \left\{\tilde{M} \in \mathcal{M} : \tilde{M} \cdot (\pi - \pi^*) \in \text{bd}(\mathcal{C}^\circ)\right\}$. Hence, we can characterise $\Lambda_{\pi^*}(M)$ as:

  $$\Lambda_{\pi^*}(M) \triangleq \cup_{\Pi \setminus \pi^*} \left\{\tilde{M} \in \mathcal{M} : \exists\, z \in \mathcal{C},\ z^\top \tilde{M} \cdot (\pi - \pi^*) = 0\right\}$$

  Linearising the bilinear terms leads us to:

  $$\Lambda_{\pi^*}(M) \triangleq \cup_{\Pi \setminus \pi^*} \left\{\tilde{M} \in \mathcal{M} :\ \exists\, z \in \mathcal{C},\ \langle \text{vect}\left(z^\top (\pi - \pi^*)\right), \text{vect}(\tilde{M})\rangle = 0\right\}$$

  The optimization problem now becomes:

  $$\sup_{w \in \Delta^K} \inf_{\pi \in \Pi \setminus \pi^*} \inf_{z \in \mathcal{C}} \sum_{k=1}^{K} w_k d_{\text{KL}}\left(z^\top M_k, z^\top \tilde{M}_k\right)$$

  Finally, we observe that if we have multiple candidates $\pi^* \in \Pi^*$, the inner minimisation problem add a new layer to yield

  $$\sup_{w \in \Delta^K} \inf_{\substack{\pi \in \Pi \setminus \pi^* \\ \pi^* \in \Pi^*(M)}} \inf_{z \in \mathcal{C}} \sum_{k=1}^{K} w_k d_{\text{KL}}\left(z^\top M_k, z^\top \tilde{M}_k\right)$$

  This concludes the proof.

□

B.2 Lower Bound for Gaussians: Proof of Theorem 3.2

Proof.
Step 1. Simplifying the KL-divergence for a Gaussian bandit instance with identical variance across all objectives yields

$$\mathcal{V}_{\mathcal{C}}(w, M) = \min_{\tilde{M} \in \Lambda_{\pi^*}(M)} \min_{z \in C} \sum_{k=1}^{K} w_k \sum_{\ell=1}^{L} z^{(\ell)} \frac{\left(\mu_k^{(\ell)} - \tilde{M}_k^{(\ell)}\right)^2}{\sigma^2}. \tag{13}$$

Recalling that due to the projection lemma, the $\tilde{M}$ achieving the minimum satisfies

$$z^\top \sum_{k=1}^{K} \tilde{M}_k^\top \left(\pi_k^* - \pi_k\right) = 0,\ \forall z \in \mathcal{C} \tag{14}$$

Additionally, for $z \in C$, we have

$$z = \sum_{i=1}^{N} \lambda_i v_i,\ \text{ for } \lambda_i \geq 0 \forall i \in [N]\,. \tag{15}$$

Now, we formulate the Lagrangian of (13) with dual variables β for (14) and γ for (15) as

$$\begin{aligned} &\mathcal{L}\left(w, M, \tilde{M}, \gamma, \beta\right) \\ &= \sum_{k=1}^{K} \pi_k^2 w_k \sum_{\ell=1}^{L} z^{(\ell)} \left(\frac{\mu_k^{(\ell)} - \tilde{M}_k^{(\ell)}}{\sigma}\right)^2 + \beta \sum_{i=1}^{N} \lambda_i v_i^\top \tilde{M}\left(\pi^* - \pi\right) + \gamma^\top \left(z - \sum_{i=1}^{N} \lambda_i v_i\right) \\ &= \sum_{k=1}^{K} \pi_k^2 w_k \sum_{\ell=1}^{L} \sum_{i=1}^{N} \lambda_i v_i^{(\ell)} \left(\frac{\mu_k^{(\ell)} - \tilde{M}_k^{(\ell)}}{\sigma}\right)^2 + \beta \sum_{i=1}^{N} \lambda_i v_i^\top \tilde{M}\left(\pi^* - \pi\right) + \gamma^\top \left(z - \sum_{i=1}^{N} \lambda_i v_i\right) \end{aligned} \tag{16}$$

Step 2. Taking derivative with respect to γ, we have

$$\frac{\partial \mathcal{L}}{\partial \gamma} = 0 \quad \Longrightarrow \quad z = \sum_{i=1}^{N} \lambda_i v_i$$

Thus, we get the Lagrangian as

$$\begin{aligned}\mathcal{L}(w, M, \tilde{M}, \beta) &= \sum_{k=1}^{K} \pi_k^2 w_k \sum_{\ell=1}^{L} \left(\sum_{i=1}^{N} \lambda_i v_i^{(\ell)} \right) \left(\frac{\mu_k^{(\ell)} - \tilde{M}_k^{(\ell)}}{\sigma} \right)^2 + \beta \sum_{i=1}^{N} \lambda_i v_i^\top \tilde{M} \left(\pi^* - \pi\right) \\ &= \sum_{k=1}^{K} \pi_k^2 w_k \sum_{\ell=1}^{L} \left(\sum_{i=1}^{N} \lambda_i v_i^{(\ell)} \right) \left(\frac{\mu_k^{(\ell)} - \tilde{M}_k^{(\ell)}}{\sigma} \right)^2 + \beta \sum_{i=1}^{N} \lambda_i \sum_{\ell=1}^{L} v_i^{(\ell)} \sum_{k=1}^{K} \tilde{M}_k^{(\ell)} \left(\pi^* - \pi\right)_k\end{aligned}$$

Step 3. Taking the derivative w.r.t. $\tilde{M}$, we have

$$\frac{\partial \mathcal{L}}{\partial \tilde{M}_k^{(\ell)}} = \pi_k^2 w_k \left(\sum_{i=1}^{N} \lambda_i v_i^{(\ell)} \right) \left(\frac{-2}{\sigma^2} \right) \left(\mu_k^{(\ell)} - \tilde{M}_k^{(\ell)} \right) + \beta \sum_{i=1}^{N} \lambda_i v_i^{(\ell)} \left(\pi^* - \pi\right)_k \ ,$$

and setting it to zero, we get

$$\begin{aligned}\pi_k^2 w_k \left(\sum_{i=1}^{N} \lambda_i v_i^{(\ell)} \right) \left(\frac{2}{\sigma^2} \right) \left(\mu_k^{(\ell)} - \tilde{M}_k^{(\ell)} \right) &= \beta \sum_{i=1}^{N} \lambda_i v_i^{(\ell)} \left(\pi^* - \pi\right)_k \\ \Longrightarrow \tilde{M}_k^{(\ell)} = \mu_k^{(\ell)} - \frac{\sigma^2 \beta \left(\sum_{i=1}^{N} \lambda_i v_i^{(\ell)} \left(\pi - \pi^*\right)_k \right)}{2\pi_k^2 w_k \left(\sum_{i=1}^{N} \lambda_i v_i^{(\ell)} \right)} &= \mu_k^{(\ell)} - \frac{\sigma^2 \beta \left(\pi^* - \pi\right)_k}{2\pi_k^2 w_k} \ . \end{aligned} \tag{17}$$

The last equality holds for any $\sum_{i=1}^{N} \lambda_i v_i^{(\ell)} \neq 0$.

Therefore, the Lagrangian now becomes

$$\begin{aligned}&\mathcal{L}(w, M, \beta) \\ &= \sum_{k=1}^{K} \pi_k^2 w_k \sum_{\ell=1}^{L} \sum_{i=1}^{N} \lambda_i v_i^{(\ell)} \left(\frac{\mu_k^{(\ell)} - \tilde{M}_k^{(\ell)}}{\sigma} \right)^2 + \beta \sum_{i=1}^{N} \lambda_i \sum_{\ell=1}^{L} v_i^{(\ell)} \sum_{k=1}^{K} \tilde{M}_k^{(\ell)} \left(\pi^* - \pi\right)_k \\ &= \sum_{k=1}^{K} \pi_k^2 w_k \sum_{\ell=1}^{L} \sum_{i=1}^{N} \lambda_i v_i^{(\ell)} \frac{\left(\pi^* - \pi\right)_k^2}{\pi_k^4 w_k^2} \frac{\beta^2 \sigma^2}{4} + \beta \sum_{i=1}^{N} \sum_{\ell=1}^{L} \lambda_i v_i^{(\ell)} \sum_{k=1}^{K} \left(\pi^* - \pi\right)_k \left(\mu_k^{(\ell)} - \frac{\left(\pi^* - \pi\right)_k}{\pi_k^2 w_k} \left(\frac{\sigma^2}{2} \right) \right) \\ &= \left(\frac{\beta^2 \sigma^2}{4} - \frac{\beta \sigma^2}{2} \right) \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{\ell=1}^{L} \frac{\left(\pi^* - \pi\right)_k^2}{w_k \pi_k^2} \lambda_i v_i^{(\ell)} + \beta \sum_{i=1}^{N} \sum_{\ell=1}^{L} \sum_{k=1}^{K} \lambda_i v_i^{(\ell)} \left(\pi^* - \pi\right)_k \mu_k^{(\ell)} \\ &= \beta \left[\frac{\sigma^2}{2} \left(\frac{\beta}{2} - 1 \right) \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{\ell=1}^{L} \frac{\left(\pi^* - \pi\right)_k^2}{\pi_k^2 w_k} \lambda_i v_i^{(\ell)} + \sum_{i=1}^{N} \sum_{\ell=1}^{L} \sum_{k=1}^{K} \lambda_i v_i^{(\ell)} \left(\pi^* - \pi\right)_k \mu_k^{(\ell)} \right]\end{aligned}$$

Step 4. By taking the derivative with respect to β, we have

$$\frac{\partial \mathcal{L}}{\partial \beta} = \frac{\sigma^2}{2} \left(\beta - 1\right) \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{\ell=1}^{L} \frac{\left(\pi^* - \pi\right)_k^2}{w_k \pi_k^2} \lambda_i v_i^{(\ell)} + \sum_{i=1}^{N} \sum_{\ell=1}^{L} \sum_{k=1}^{K} \lambda_i v_i^{(\ell)} \left(\pi^* - \pi\right)_k \mu_k^{(\ell)} \ ,$$

and setting it to zero, leads to:

$$\beta = 1 - \frac{2}{\sigma^2} \frac{\sum_{i=1}^{N} \sum_{\ell=1}^{L} \sum_{k=1}^{K} \lambda_i v_i^{(\ell)} \left(\pi^* - \pi\right)_k \mu_k^{(\ell)}}{\sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{\ell=1}^{L} \frac{\lambda_i v_i^{(\ell)} \left(\pi^* - \pi\right)_k^2}{w_k \pi_k^2}} \tag{18}$$

From (17), excluding for the origin lying within the cone, we get:

$$\begin{aligned}\tilde{M}_k^{(\ell)} &= \mu_k^{(\ell)} - \frac{\beta \sigma^2}{2} \frac{\left(\pi^* - \pi\right)_k}{\pi_k^2 w_k} \\ &= \mu_k^{(\ell)} - \frac{\sigma^2}{2} \left(1 - \frac{2}{\sigma^2} \frac{\sum_{i=1}^{N} \sum_{\ell=1}^{L} \sum_{k=1}^{K} \lambda_i v_i^{(\ell)} \left(\pi^* - \pi\right)_k \mu_k^{(\ell)}}{\sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{\ell=1}^{L} \frac{\lambda_i v_i^{(\ell)} \left(\pi^* - \pi\right)_k^2}{w_k \pi_k^2}} \right) \left(\frac{\left(\pi^* - \pi\right)_k}{\pi_k^2 w_k} \right)\end{aligned}$$

Let $\Delta_k = (\pi^* - \pi)_k$.

Finally, the Lagrangian from (16) leads to

$$\begin{aligned}
&\mathcal{L}(w, M)\\
&= \left(1 - \frac{2}{\sigma^2}\frac{\sum_{i,\ell,k}\lambda_i v_i^{(\ell)}\Delta_k\mu_k^{(\ell)}}{\sum_{i,\ell,k}\frac{\lambda_i v_i^{(\ell)}\Delta_k^2}{w_k\pi_k^2}}\right)\left[-\left(\frac{\sigma^2}{4} + \frac{\sum_{i,\ell,k}\lambda_i v_i^{(\ell)}\Delta_k\mu_k^{(\ell)}}{2\sum_{i,\ell,k}\frac{\lambda_i v_i^{(\ell)}\Delta_k^2}{w_k\pi_k^2}}\right)\sum_{i,\ell,k}\frac{\Delta_k^2}{\pi_k^2 w_k}\lambda_i v_i^{(\ell)} + \sum_{i,\ell,k}\lambda_i v_i^{(\ell)}\Delta_k\mu_k^{(\ell)}\right]\\
&= \left(1 - \frac{2}{\sigma^2}\frac{\sum_{i,\ell,k}\lambda_i v_i^{(\ell)}\Delta_k\mu_k^{(\ell)}}{\sum_{i,\ell,k}\frac{\lambda_i v_i^{(\ell)}\Delta_k^2}{w_k\pi_k^2}}\right)\left[\frac{1}{2}\sum_{i,\ell,k}\lambda_i v_i^{(\ell)}\Delta_k\mu_k^{(\ell)} - \frac{\sigma^2}{4}\sum_{i,\ell,k}\frac{\Delta_k^2}{\pi_k^2 w_k}\lambda_i v_i^{(\ell)}\right]\\
&= \left(1 - \frac{2}{\sigma^2}\frac{\sum_{i,\ell,k}\lambda_i v_i^{(\ell)}\Delta_k\mu_k^{(\ell)}}{\sum_{i,\ell,k}\frac{\lambda_i v_i^{(\ell)}\Delta_k^2}{w_k\pi_k^2}}\right)\frac{1}{2}\sum_{i,\ell,k}\lambda_i v_i^{(\ell)}\Delta_k\mu_k^{(\ell)} - \frac{\sigma^2}{4}\sum_{i,\ell,k}\frac{\Delta_k^2}{\pi_k^2 w_k}\lambda_i v_i^{(\ell)} + \frac{1}{2}\sum_{i,\ell,k}\lambda_i v_i^{(\ell)}\Delta_k\mu_k^{(\ell)}\\
&= \left(1 - \frac{1}{\sigma^2}\frac{\sum_{i,\ell,k}\lambda_i v_i^{(\ell)}\Delta_k\mu_k^{(\ell)}}{\sum_{i,\ell,k}\frac{\lambda_i v_i^{(\ell)}\Delta_k^2}{w_k\pi_k^2}}\right)\sum_{i,\ell,k}\lambda_i v_i^{(\ell)}\Delta_k\mu_k^{(\ell)} - \frac{\sigma^2}{4}\sum_{i,\ell,k}\frac{\Delta_k^2}{\pi_k^2 w_k}\lambda_i v_i^{(\ell)}\\
&= \left(1 - \alpha(z) - \frac{1}{4\alpha(z)}\right)z^\top M(\pi^\star - \pi)
\end{aligned}$$

Here, $\alpha(z) \triangleq \frac{1}{\sigma^2}\frac{\sum_{i,\ell,k}\lambda_i v_i^{(\ell)}\Delta_k\mu_k^{(\ell)}}{\sum_{i,\ell,k}\frac{\lambda_i v_i^{(\ell)}\Delta_k^2}{w_k\pi_k^2}} = \frac{z^\top M(\pi^\star - \pi)}{\sigma^2(\sum_l z^l)\sum_k \frac{1}{w_k}\left(\frac{\pi_k^\star}{\pi_k} - 1\right)^2}$.

□

B.3 Proof of Theorem 4.1

This proof follows directly from Theorem 3.1 in Balas (1985). Recall that the set $\mathcal{F}$ is given by:

$$\mathcal{F} \triangleq \cup_{\Pi\setminus\{\pi^*\}}\left\{\tilde{M} \in \mathcal{M} : \langle \text{vect}(z^\top(\pi - \pi^*)), \text{vect}(\tilde{M})\rangle = 0\right\}$$

Any $z \in \mathcal{C}$, we have $z = \sum_i \alpha_i v_i$. Rewriting, every hyperplane in $\mathcal{F}$ as $P_\pi = \left\{\tilde{M} \in \mathcal{M} \middle| \langle \text{vect}\left(\sum_i \alpha_i v_i^\top \pi\right), \text{vect}(\tilde{M})\rangle = \sum_i \alpha_i v_i^\top \pi^*\right\}$. Then, by Theorem 3.1 in Balas (1985), $\mathcal{C}(\mathcal{F})$ is given by:

$$\mathcal{C}(\mathcal{F}) = \begin{cases}\gamma^\top \text{vect}(\tilde{M}) \geq \gamma_0,\ \gamma = \sum_i u_i\alpha_i \text{vect}(v_i^\top(\pi - \pi^*))\\ \gamma_0 \leq u_i \sum_i \alpha_i v_i^\top \pi^*, \tilde{M} \in \mathcal{M}\end{cases}$$

B.4 Proof of Theorem 4.2

Recall that:

$$\bar{\mathcal{V}}_{\mathcal{C}}(w, M) \triangleq \min_{\tilde{M}\in \text{ch}(\Lambda_{\pi^*}(M))}\inf_{z\in\mathcal{C}}\sum_{k=1}^K w_k d_{\text{KL}}\left(z^\top M, z^\top \tilde{M}\right)$$

$$\bar{w}^*(M) \triangleq \arg\max_{w\in\Delta^K}\inf_{\substack{\pi\in\Pi\setminus\pi^*\\ \pi^*\in\Pi^*(M)}}\min_{\tilde{M}\in\text{ch}(\Lambda_{\pi^*}(M))}\min_{z\in\mathcal{C}}\sum_{k=1}^K w_k d_{\text{KL}}\left(z^\top M, z^\top\tilde{M}\right)$$

- For (1) and (2) observe that $(z, M) \to z^\top M$ and $(z, M, \tilde{M}) \to d_{\text{KL}}\left(z^\top M, z^\top \tilde{M}\right)$ are continuous maps for all $(z, M) \in \mathcal{C} \times \mathcal{M}$ and $(z, M, \tilde{M}) \in \mathcal{C} \times \mathcal{M} \times \text{ch}\left(\Lambda_{\pi^*}(M)\right)$. Further, $\sum_k w_k d_{\text{KL}}\left(z^\top M, z^\top \tilde{M}\right)$ is continuous in all its elements. Fix a sequence $(w_t, , z_t, M_t) \in \Pi \times \mathcal{C} \times \mathcal{M}$ such that $(w_t, z_t, M_t) \to (w, z, M)$. For any ϵ, $\exists\, t' \geq 1$ such that $\|(w_t, z_t, M_t) -$

$(w, z, M)\| \leq \epsilon \ \forall \ t \geq t'$. Further, $\mathrm{ch}\left(\Lambda_{\pi^*}\left(M_t\right)\right) \rightarrow \mathrm{ch}\left(\Lambda_{\pi^*}\left(M\right)\right)$. Therefore, for every ϵ', $\exists\, t'' \geq 1$ such that $\forall\, t \geq t''$ we

$$\Big| \sum_k w_{k,t} d_{\mathrm{KL}} \left(z_t^\top M_t, z_t^\top \tilde{M}_t \right) - \sum_k w_k d_{\mathrm{KL}} \left(z^\top M, z_t^\top \tilde{M}_t \right) \Big| \leq \epsilon' \ \forall \ \tilde{M}_t \in \mathbb{R}^{K \times L}$$

Taking $t \geq \max\{t', t''\}$, we have:

$$\begin{aligned}
&\Big| \inf_{\tilde{M} \in \Lambda_{\pi^*}(M)} \inf_{z \in \mathcal{C}} \sum_k w_{k,t} d_{\mathrm{KL}} \left(z_t^\top M_t, z_t^\top \tilde{M}_t \right) - \inf_{\tilde{M} \in \Lambda_{\pi^*}(M)} \inf_{z \in \mathcal{C}} \sum_k w_k d_{\mathrm{KL}} \left(z^\top M, z_t^\top \tilde{M}_t \right) \Big| \\
&\leq \Big| \inf_{\tilde{M} \in \Lambda_{\pi^*}(M)} \inf_{z \in \mathcal{C}} \sum_k w_{k,t} d_{\mathrm{KL}} \left(z_t^\top M_t, z_t^\top \tilde{M}_t \right) - \sum_k w_k d_{\mathrm{KL}} \left(z^\top M, z_t^\top \tilde{M}_t \right) \Big| \\
&\leq \epsilon'
\end{aligned}$$

- For (3), we define $f(w, M) = \inf_{\tilde{M} \in \Lambda_{\pi^*}(M)} \inf_{z \in \mathcal{C}} \sum_k w_k d_{\mathrm{KL}} \left(z^\top M, z^\top \tilde{M} \right)$ and $C(w) = \Pi$. Then, from Berge's Theorem (Theorem G.1 in Appendix), we get $w^*(M)$ is upper-hemicontinuous.
- For (4), the convexity of $w^*(M)$ follows since the optimal solution

$$\max_{w \in \Pi} \inf_{\tilde{M} \in \Lambda_{\pi^*}(M)} \inf_{z \in \mathcal{C}} \sum_k w_k d_{\mathrm{KL}} \left(z^\top M, z^\top \tilde{M} \right)$$

is concave for any given π and π^*.

C Proof of the Stopping Time

C.1 Proof of Lemma 4.1

The proof follows by showing that $\inf_{z\in\mathcal{C}} \sum_k N_{k,t} d_{\mathrm{KL}}\left(z^\top \hat{M}_t, z^\top M\right)$ is an appropriate stochastic process and using results from Kaufmann and Koolen (2021). To this end, it can be verified that for all $k \in [K]$:

1. There exists a martingale such that $M^{\rho}_{k,t}$ such that
   (a) $M^{\rho}_{k,\ell}(t)$ is non-negative and $M^{\rho}_{k,\ell}(0) = 1$
   (b) $M^{\rho}_{k,\ell}(t)$ is such that:
   $$\forall\, t \in \mathbb{N} : M^{\rho}_{k}(t) \geq \exp\left(\rho d_{\mathrm{KL}}\left(z^\top \hat{M}_t, z^\top M\right) - g(\rho)\right)$$
   (c) For any subset $\mathcal{S} \subseteq [K]$ and ρ we have: $\Pi_{k\in\mathcal{S}} M^{\rho}_{k}(t)$ is a martingale.
2. From Lemma 4 in Kaufmann and Koolen (2021) and Theorem 4.3 on KL-concentration, we conclude the proof.

C.2 Proof of Theorem 4.3

Let $\rho > KL + 1$ and η. Define $D = \lceil \frac{\log(n)}{\log(1+\eta)} \rceil$ and set $\mathcal{D} = \{1, 2, \ldots, D\}^K$. Let:

$$A_t = \left\{ \sum_{k\in[K]} \sum_{\ell\in[L]} N_{k,t} \cdot d_{\mathrm{KL}}\left(M^{(\ell)}_k, \hat{M}^{(\ell)}_{k,t}\right) \geq \rho \right\}$$
$$B_d = \cap_{k=1}^{K} \left\{(1+\xi)^{d-1} \leq N_{k,t} \leq (1+\xi)^d\right\}$$

We have $A = \cup_{d\in\mathcal{D}} A \cap B_d$, hence $\mathbb{P}(A) \leq \sum_{d\in\mathcal{D}} \mathbb{P}(A \cap B_d)$. Using Lemma F.1 with $\eta = \frac{1}{\delta-1}$ and $\bar{t}_k = (1+\eta)^{d_k-1}$. Since $\rho \geq K+1$, for $\eta = \frac{1}{\rho-1}$ and $\rho \geq (1+\eta)K$, for all $d \in \mathcal{D}$ we have:

$$\mathbb{P}\left(A \cap B_d\right) \leq \left(\frac{\rho e}{KL}\right)^{KL} \exp\left(\frac{-\rho}{(1+\eta)}\right)$$

By a union bound on $\mathcal{D}$, we have:

$$\mathbb{P}\left(A\right) \leq \left(\frac{D\rho e}{K}\right)^{K} \exp\left(\frac{-\rho}{1+\eta}\right)$$

Noting that $\eta = \frac{1}{\rho-1}$, we get:

$$\mathbb{P}\left(A\right) \leq \exp\left(-\delta\right) \left(\frac{\delta\lceil\delta \log(n)\rceil}{KL}\right)^{KL}$$

This concludes the proof.

D Proofs for Sample Complexity Upper Bound

D.1 Pairwise Concentration Bounds

Lemma D.1 (Pairwise concentration)**.** *Consider the event:*

$$\mathcal{E}_t \triangleq \cap_{k\in[K]} \cap_{i\neq k} \cap_{\ell\in[L]} \left\{ L_{i,j}^{(\ell)}(t) \leq \mu_i^{(\ell)} - \mu_j^{(\ell)} \leq U_{i,j}^{(\ell)}(t) \right\}, \tag{19}$$

where $L_{ij}(t) = z^\top(\mu_i - \mu_j) - \beta_{ij}(t)$ and $U_{ij}(t) = z^\top(\mu_i - \mu_j) + \beta_{ij}(t)$, and $\beta_{ij}(t)$ is defined in Theorem 5.1. Then, we get $\mathbb{P}\left(\cap_{t=1}^{\infty}\mathcal{E}\right) \geq 1-\delta$.

Proof. We have the following:

$$\begin{aligned}
\mathcal{E}_t &= \cap_{(i,j)\in\mathcal{B}} \cap_{\ell=1}^{L} \left\{ L_{i,j}^{(\ell)} \leq \mu_i^{(\ell)} - \mu_j^{(\ell)} \leq U_{i,j}^{(\ell)} \right\} \\
&= \cap_{(i,j)\in\mathcal{B}} \cap_{\ell=1}^{L} \left\{ \left| \left(\hat{\mu}_{i,t}^{(\ell)} - \hat{\mu}_{j,t}^{(\ell)}\right) - (\mu_i - \mu_j) \right| \leq \beta_{ij}(t) \right\}
\end{aligned}$$

where, $\mathcal{B}$ is the set of arm pairs. By a union bound we have the following:

$$\begin{aligned}
\mathbb{P}\left(\overline{\mathcal{E}}_t\right) &= \mathbb{P}\left(\exists\, t \geq 1 : \overline{\mathcal{E}}_t \text{ holds}\right) \\
&= \mathbb{P}\left(\exists\, t \geq 1 : (i,j) \in \mathcal{B},\ \ell \in [L], (i,j) \in \mathcal{B} : \left| \left(\hat{\mu}_{i,t}^{(\ell)} - \hat{\mu}_{j,t}^{(\ell)}\right) - (\mu_i - \mu_j) \right| \geq \beta_{ij}(t)\right) \\
&\leq \sum_{(i,j)\in\mathcal{B}} \sum_{\ell\in[L]} \frac{\delta}{K(K-1)} \\
&= \delta
\end{aligned}$$

□

We now show that the Pareto fronts under the metric d_p.

Lemma D.2. *$(\mathcal{Z}, d_p)$ is a complete metric space.*

Proof. From Definition 5.2, for two Pareto fronts $\mathcal{P}_1, \mathcal{P}_2 \in \mathcal{Z}$, we have that:

$$d_p(\mathcal{P}_1, \mathcal{P}_2) \triangleq \max\left\{ \sup_{k\in\mathcal{P}_1} d(k, \mathcal{P}_1), \sup_{k\in\mathcal{P}_2} d(k, \mathcal{P}_1) \right\}$$

where,

$$d(k, \mathcal{P}) = \inf_{k'\in\mathcal{P}} \max\left\{ 0, \sup_{z\in\mathcal{C}\cap\mathbb{B}(1)} z^\top\left(\mu_{k'} - \mu_k\right) \right\}$$

1. We first show that $d_p(\mathcal{P}_1, \mathcal{P}_2)$ is a metric. Let $\mathcal{P}_1, \mathcal{P}_2 \in \mathcal{Z}$. To show that d_p is a metric, we show that:
   (a) Symmetry: $d_p(\mathcal{P}_1, \mathcal{P}_2)$ is symmetric by definition
   (b) Triangle Inequality: We show that $d_p(\mathcal{P}_1, \mathcal{P}_3) \leq d_p(\mathcal{P}_1, \mathcal{P}_2) + d_p(\mathcal{P}_2, \mathcal{P}_3)$.

   $$d_p\left(\mathcal{P}_1, \mathcal{P}_3\right) = \max\left\{ \max_{k\in\mathcal{P}_1} \min_{k'\in\mathcal{P}_3} \mu_{k'}(X_3) - \mu_k(X_1), \max_{k\in\mathcal{P}_3} \min_{k'\in\mathcal{P}_1} \mu_{k'}(X_1) - \mu_k(X_3) \right\}$$

   We have that:

   $$\begin{aligned}
   & \max_{k\in\mathcal{P}_1} \min_{k'\in\mathcal{P}_3} \mu_{k'}(X_3) - \mu_k(X_1) \\
   \leq\ & \max_{k\in\mathcal{P}_1} \min_{k'\in\mathcal{P}_3} \mu_{k'}(X_3) + \min_{k''\in\mathcal{P}_2} \mu_{k''}(X_2) - \max_{k''\in\mathcal{P}_2} \mu_{k''}(X_2) - \mu_k(X_1) \\
   \leq\ & \max_{k''\in\mathcal{P}(X_2)} \min_{k'\in\mathcal{P}_3} \mu_{k'}(X_3) - \mu_{k''}(X_2) + \max_{k\in\mathcal{P}_1} \min_{k''\in\mathcal{P}(X_2)} \mu_{k''}(X_2) - \mu_k(X_1)
   \end{aligned}$$

   Using a similar argument:

   $$\max_{k\in\mathcal{P}_3} \min_{k'\in\mathcal{P}_1} \mu_{k'}(X_1) - \mu_k(X_2) \leq \max_{k''\in\mathcal{P}(X_2)} \min_{k'\in\mathcal{P}_1} \mu_{k'}(X_1) - \mu_{k''}(X_2) + \max_{k\in\mathcal{P}_3} \min_{k''\in\mathcal{P}(X_2)} \mu_{k''}(X_2) - \mu_k(X_3)$$

Noting that for any positive numbers $a, b, c, d,\ \max\{a+b, c+d\} = \max\{a+c, b+d\}$, we have:

$$\begin{aligned}
d_p(\mathcal{P}_1, \mathcal{P}_3) \leq \quad & \max\Big\{ \max_{k'' \in \mathcal{P}(X_2)} \min_{k' \in \mathcal{P}_3} \mu_{k'}(X_3) - \mu_{k''}(X_2) + \max_{k \in \mathcal{P}_1} \min_{k'' \in \mathcal{P}(X_2)} \mu_{k''}(X_2) - \mu_k(X_1), \\
& \max_{k'' \in \mathcal{P}(X_2)} \min_{k' \in \mathcal{P}_1} \mu_{k'}(X_1) - \mu_{k''}(X_2) + \max_{k \in \mathcal{P}_3} \min_{k'' \in \mathcal{P}(X_2)} \mu_{k''}(X_2) - \mu_k(X_3)\Big\} \\
= \quad & \max\Big\{ \max_{k'' \in \mathcal{P}(X_2)} \min_{k' \in \mathcal{P}_1} \mu_{k'}(X_1) - \mu_{k''}(X_2) + \max_{k' \in \mathcal{P}_1} \min_{k'' \in \mathcal{P}(X_2)} \mu_{k''}(X_2) - \mu_{k'}(X_1), \\
& \max_{k'' \in \mathcal{P}(X_2)} \min_{k' \in \mathcal{P}_3} \mu_{k'}(X_3) - \mu_{k''}(X_2) + \max_{k \in \mathcal{P}_3} \min_{k'' \in \mathcal{P}(X_2)} \mu_{k''}(X_2) - \mu_k(X_3)\Big\} \\
= \quad & d_p(\mathcal{P}_1, \mathcal{P}_2) + d_p(\mathcal{P}_2, \mathcal{P}_3)
\end{aligned}$$

(c) We now show that $d_p(\mathcal{P}_1, \mathcal{P}_2) = 0 \iff \mathcal{P}_1 = \mathcal{P}_2$. The implication $\mathcal{P}_1 = \mathcal{P}_2 \implies d_p(\mathcal{P}_1, \mathcal{P}_2) = 0$ is immediate. For the other side, note that by Definition 5.2, we have:

$$\begin{aligned}
& d_p(\mathcal{P}_1, \mathcal{P}_2) = 0 \\
\implies \quad & \sup_{k \in \mathcal{P}_1} \Delta(k, \mathcal{P}_2) = 0 \text{ and } \sup_{k \in \mathcal{P}_2} \Delta(k, \mathcal{P}_1) = 0
\end{aligned}$$

Further, $\sup_{k \in \mathcal{P}_1} \Delta(k, \mathcal{P}_2) = 0$ implies:

$$\forall\, k \in \mathcal{P}_1, k \not\preceq_{\mathcal{C}} k',\ k' \in \mathcal{P}_2 \iff \forall\, k \in \mathcal{P}_1,\ k \in \mathcal{P}_2$$

A similar agrument using $\sup_{k \in \mathcal{P}_1} \Delta(k, \mathcal{P}_1) = 0$ implies that $\forall\, k \in \mathcal{P}_2,\ k \not\preceq_{\mathcal{C}} k', k' \in \mathcal{P}_1$.

2. We now show that $\mathcal{Z}$ is compact under the metric d_p. Consider a sequence of Pareto fronts $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n \in \mathcal{Z}$ and $\mathcal{P}$ be the candidate for limiting Pareto front.
   - Boundedness of $\mathcal{P}$ is immediate.
   - $\mathcal{P}$ is convex since $\mathcal{P}_n, \mathcal{P}_{n+1}$ are convex and $\lambda \mathcal{P}_n + (1-\lambda)\mathcal{P}_{n+1}$ is also convex for all $\lambda \in [0, 1]$.
   - $\mathcal{P}_n \to \mathcal{P},$ therefore, $\forall\, \epsilon > 0, \exists\, N(\epsilon)$ s.t. $\forall\, n > N(\epsilon)$ and $d_p(P_n, P) < \epsilon$. Let μ_k be a limit point of $\mathcal{P}$, i.e., $\exists$ a sequence $\mu_{k,n} \in \mathcal{P}$ such that $\mu_{k,n} \to \mu_k$. Since $d_p(\mathcal{P}_n, \mathcal{P}) \to 0$ for each $\mu_{k,n} \in \mathcal{P}$ there exists $\mu_{k,n,m} \in \mathcal{P}_n$ s.t. $\mu_{k,n,m} \to \mu_{k,n}$. Using a diagonalization argument, we can obtain a subsequence $\mu_{k,n,m} \to \mu_k$. Since $\mathcal{P}_n$ is compact, μ_k must lie in $\mathcal{P}$ and therefore, $\mathcal{P}$ is closed.

□

D.2 Proof of Theorem 5.2

Lemma D.3. *There exists constants C_1, C_2 such that:*

$$\mathbb{P}\left(\bar{\mathcal{G}}_T\right) \leq \exp\left(-CT^{1/8}\right)$$

Proof. We then have that:

$$\begin{aligned}
\mathbb{P}\left(d_p\left(\hat{\mathcal{P}}_t, \mathcal{P}^*\right) \geq \epsilon\right) &= \mathbb{P}\left(\max\left\{d\left(\hat{\mathcal{P}}_t, \mathcal{P}^*\right), d\left(\mathcal{P}^*, \hat{\mathcal{P}}_t\right)\right\} \geq \epsilon\right) \\
&\leq \mathbb{P}\left(d\left(\hat{\mathcal{P}}_t, \mathcal{P}^*\right) \geq \epsilon\right) + \mathbb{P}\left(d\left(\mathcal{P}^*, \hat{\mathcal{P}}_t\right) \geq \epsilon\right)
\end{aligned}$$

Focusing on the first term we have:

$$\begin{aligned}
d\left(\hat{\mathcal{P}}_t, \mathcal{P}^*\right) &= \inf_{k \in \hat{\mathcal{P}}_t} \sup_{k' \in \mathcal{P}^*} \max\left\{0, \max_{z \in \mathcal{C} \cap \mathbb{B}(1)} z^{(\ell)}\left(\mu_{k'}^{(\ell)} - \hat{\mu}_{k,t}^{(\ell)}\right)\right\} \geq \sum_{(k,k')} \beta_{kk'}(t) \\
&\overset{(a)}{=} \min_{k \in \hat{\mathcal{P}}_t} \max_{k' \in \mathcal{P}^*} \max\left\{0, \max_{z \in \mathcal{C} \cap \mathbb{B}(1)} z^{(\ell)}\left(\mu_{k'}^{(\ell)} - \hat{\mu}_{k',t}^{(\ell)} + \hat{\mu}_{k',t}^{(\ell)} - \hat{\mu}_{k,t}^{(\ell)}\right)\right\} \\
&\overset{(b)}{\leq} \max_{k' \in \mathcal{P}^*} \max\left\{0, \max_{z \in \mathcal{C} \cap \mathbb{B}(1)} z^{(\ell)}\left(\mu_{k'}^{(\ell)} - \hat{\mu}_{k',t}^{(\ell)}\right)\right\} \\
&\quad + \max_{k' \in \mathcal{P}^*} \max\left\{0, \max_{z \in \mathcal{C} \cap \mathbb{B}(1)} z^{(\ell)}\left(\hat{\mu}_{k',t}^{(\ell)} - \hat{\mu}_{k,t}^{(\ell)}\right)\right\}
\end{aligned}$$

From Lemma D.1 with probability $1-\delta$ we have:

$$\left| z^\top \left(\mu_k^{(\ell)} - \mu_{k'}^{(\ell)}\right) - z^\top \left(\hat{\mu}_{k,t}^{(\ell)} - \hat{\mu}_{k',t}^{(\ell)}\right) \right| \leq \beta_{kk'}(t)$$

Therefore, with probability $1-\delta$, we have:

$$d_p\left(\hat{\mathcal{P}}_t, \mathcal{P}^*\right) \leq \sum_{k\in[K]} \beta_k + \sum_{(k,k')} \beta_{kk'}$$

An identical argument shows that with probability $1-\delta$,

$$d_p\left(\mathcal{P}^*, \hat{\mathcal{P}}_t\right) \leq \sum_{k\in[K]} \beta_k + \sum_{(k,k')} \beta_{kk'}$$

Now, we observe that $\beta_k(t) = O(\sqrt{\frac{\log t}{t}}$. This allows us to hereafter follow the similar arguments as in Lemma 19 of (Garivier and Kaufmann, 2016), and prove that

$$\mathbb{P}\left(\bar{\mathcal{G}}_T\right) \leq \exp\left(-CT^{1/8}\right).$$

□

Theorem D.1 (Restating Theorem 5.2)**.** *For any $\alpha > 0$ and $c(t,\delta)$ defined in* (8)*, we have that the stopping time satisfies :*

$$\lim_{\delta\to 0} \frac{\mathbb{E}[\tau]}{\log\left(\frac{1}{\delta}\right)} \leq \alpha \bar{\mathcal{T}}_{\mathcal{F}}(M)\ \forall\ M \in \mathbb{R}^{K\times L}$$

Proof. **Step 1: Good Event** Let $T\in\mathbb{N}$, and $h(T) = \sqrt{T}$, define the good event:

$$\mathcal{G}_T = \cap_{t=h(T)}^T \left\{ d_p(\hat{\mathcal{P}}_t, \mathcal{P}^*) \leq f(\epsilon)\right\} \tag{20}$$

where, $f(\epsilon)$ is such that:

$$d_p(\hat{\mathcal{P}}_t, \mathcal{P}^*) \leq f(\epsilon) \implies \sup_{w'\in w^*(\hat{\mathcal{P}}_\tau)} \sup_{w\in w^*(\mathcal{P}^*)} \|w'-w\| \leq \epsilon$$

Step 2: Concentration of Good Event In Lemma D.3, we show that:

$$\mathbb{P}\left(\bar{\mathcal{G}}_T\right) \leq \exp\left(-cT^{\frac{1}{8}}\right)$$

Step 3: Tracking Lemma From (Garivier and Kaufmann, 2016), we have that:

$$\max_k \left| N_{k,t} - \sum_t w_{k,t}\right| \leq K\left(1+\sqrt{t}\right)$$

Step 4: Complexity of the good event

Assume $t \geq T_\epsilon$, and let:

$$C_\epsilon(M) \triangleq \inf_{w',M'} \overline{\mathcal{V}}_{\mathcal{C}}(w,M),\ \forall (w,M) \text{ s.t. } \|w'-w\| \leq 3\epsilon,\ d_{\mathrm{P}}\left(\hat{P}_t, \mathcal{P}^*\right) \leq f(\epsilon)$$

Then, we have that:

$$\overline{\mathcal{V}}_{\mathcal{C}}\left(N_t, \hat{M}_t\right) \geq tC_\epsilon(M)$$

Step 5: Bounding the stopping time for good and bad events

Let τ_δ be the stopping time, then:

$$\min\{\tau_\delta, T\} \leq \sqrt{T} + \sum_{t=T_\epsilon}^T \mathbb{1}_{\tau_\delta \geq t}$$

From the stopping rule (Equation (8)), we get that:

$$\begin{aligned} T_\epsilon + \sum_{t=T_\epsilon}^{T} \mathbb{1}_{\overline{\mathcal{V}}_{\mathcal{C}}\left(N_t, \hat{M}_t\right) \leq c(t,\delta)} &\leq \sqrt{T} + \sum_{t=T_\epsilon}^{T} \mathbb{1}_{tC_\epsilon(M) \leq c(t,\delta)} \\ &\leq \sqrt{T} + \frac{c(t,\delta)}{C_\epsilon(M)} \end{aligned}$$

Define $T_\delta = \inf\left\{T \in \mathbb{N} : \sqrt{T} + \frac{c(t,\delta)}{C_\epsilon(M)} \leq T\right\}$. Hence, we have:

$$\mathbb{E}\left[\tau_\delta\right] \leq T_\epsilon + T_\epsilon + \sum_{T=1}^{\infty} BT \exp\left(-CT^{-1/8}\right) \leq T_\epsilon + T_\delta + T'$$

Let $C(\eta) = \inf\{T : T - \sqrt{T} \geq \frac{T}{(1+\eta)}\}$. Then:

$$T_\delta \leq C(\eta) + \inf\left\{T \in \mathbb{N} : \frac{TC_\epsilon(M)}{(1+\eta)} \geq c(t,\delta)\right\}$$

Step 6: Obtaining the asymptotic bounds Taking limits:

$$\lim_{\delta \to 0} \inf \frac{\mathbb{E}\left[\tau_\delta\right]}{\log\left(\frac{1}{\delta}\right)} \leq \alpha \mathcal{T}(M) \;\forall\; \alpha \geq 1$$

□

E Reduction to Best-arm Identification

We briefly discuss how the metric $d_{\mathrm{P}}(\cdot,\cdot)$ extends existing notions of gap in best-arm and Pareto-front identification literature. Specifically, we proceed with the three following observations.

1. Observe that $d_{\mathrm{P}}(\mathcal{P}^*, \mathcal{P}^*) = 0$.
2. Further, iff $\mathcal{C}$ represents the component-wise ordering as in Pareto-front identification (Auer et al., 2016; Kone et al., 2023a), then $z^{(\ell)} = 1,\ \forall\, \ell \in [L]$.

$$
\begin{aligned}
&d_{\mathrm{P}}\left([K]\setminus\mathcal{P}^*, \mathcal{P}^*)\right) \\
&= \max\left\{\sup_{k\in[K]\setminus\mathcal{P}^*} d\left(k,\mathcal{P}^*\right),\ \sup_{k\in\mathcal{P}^*} d\left([K]\setminus\mathcal{P}^*, k\right)\right\} \\
&= \max\left\{\sup_{k\in[K]\setminus\mathcal{P}^*}\ \inf_{k'\in\mathcal{P}^*} \max\left\{0, \min_{\ell\in[L]}\left(\mu_k^{(\ell)} - \mu_{k'}^{(\ell)}\right)\right\},\right. \\
&\qquad\qquad \left.\sup_{k\in\mathcal{P}^*}\ \inf_{k'\in[K]\setminus\mathcal{P}^*} \max\left\{0, \min_{\ell\in[L]}\left(\mu_k^{(\ell)} - \mu_{k'}^{(\ell)}\right)\right\}\right\} \\
&= \max\left\{\sup_{k\in[K]\setminus\mathcal{P}^*}\ \inf_{k'\in\mathcal{P}^*} \max\left\{0, m(k',k)\right\},\right. \\
&\qquad\qquad \left.\sup_{k\in\mathcal{P}^*}\ \inf_{k'\in[K]\setminus\mathcal{P}^*} \max\left\{0, M(k,k')\right\}\right\} \\
&= \sup_{k\in\mathcal{P}^*}\ \inf_{k'\in[K]\setminus\mathcal{P}^*} \max\left\{0, M(k,k')\right\}\ .
\end{aligned}
$$

3. Finally, when there is only a single objective, i.e., $|L| = 1$ and assuming an unique optimal arm (fairly common assumption in BAI literature), we have:

$$
\begin{aligned}
d_{\mathrm{P}}\left([K]\setminus\mathcal{P}^*, \mathcal{P}^*)\right) &= \max\left\{\sup_{k\in[K]\setminus\mathcal{P}^*} d\left(k,\mathcal{P}^*\right),\ \sup_{k\in\mathcal{P}^*} d\left([K]\setminus\mathcal{P}^*, k\right)\right\} \\
&= \max\left\{\sup_{k\in[K]\setminus k^*} \max\left\{0, \left(\mu_k - \mu_{k^*}\right)\right\},\right. \\
&\qquad\qquad \left.\sup_{k'\in[K]\setminus k^*} \max\left\{0, \left(\mu_{k^*} - \mu_{k'}\right)\right\}\right\} \\
&= \min_{k'\neq k^*} \Delta_{k'}\ ,
\end{aligned}
$$

which is exactly the gap for one-dimensional bandit.

F Technical Lemmas

Lemma F.1. *For any $k = 1, 2, \ldots, K$, let $1 \le t_k \le t$. Let $\eta > 0$ and define the event:*

$$C \triangleq \cap_{k\in[K]} C_k \triangleq \cap_{k\in[K]} \{t_k \le N_{k,t} \le (1+\eta)t_k\}$$

and let $\mathbb{1}_{C_k}$ denote that the event holds. For $\rho \ge (1+\eta)K$, we have:

$$\mathbb{P}\left[\mathbb{1}_{C_k} \sum_{k\in[K]} \sum_{\ell\in[L]} N_{k,t} d_{KL}\left(M_k^{(\ell)}, \hat{M}_{k,t}^{(\ell)}\right) \ge \rho\right] \le \left(\frac{\rho e}{KL}\right)^{KL} \exp\left(\frac{-\rho}{1+\eta}\right).$$

Proof. Fix $\zeta \in \mathbb{R}_+^{K\times L}$ and $t \ge 0$. Define $m_{k,t}^{(\ell)}$ such that:

$$m_{k,t}^{(\ell)} \quad = \quad \begin{cases} m, \text{ if } \exists\, 0 \le m \le M_k^{(\ell)}, \text{ s.t. } tD_{\mathrm{KL}}(m, M_k^{(\ell)}) = \zeta_k^{(\ell)} \\ 0, \text{ otherwise} \end{cases}$$

By monotonicity of $tD_{\mathrm{KL}},\ t \to m_{k,t}^{(\ell)}$ is increasing. With $t = N_{k,t}$, we have that

$$N_{k,t} d_{\mathrm{KL}}\left(M_k^{(\ell)}, m_{k,N_{k,t}}^{(\ell)}\right) = \zeta_k^{(\ell)} \le N_{k,t} d_{\mathrm{KL}}\left(M_k^{(\ell)}, \hat{M}_{k,t}^{(\ell)}\right), \implies \hat{M}_{k,t}^{(\ell)} \overset{(a)}{\le} m_{k,N_{k,t}}^{(\ell)} \overset{(b)}{\le} m_{k,(1+\eta)t_k}^{(\ell)},$$

where (a) follows from monotonicity of $D_{\mathrm{KL}}(\cdot,\cdot)$ and (b) follows from monotonicity of $m_{k,t}^{(\ell)}$. With $t_k d_{\mathrm{KL}}\left(M_k^{(\ell)}, \hat{M}_{k,t_k(1+\eta)}^{(\ell)}\right) = \frac{\zeta_k^{(\ell)}}{1+\eta}$ and non-negativity of D_{KL} we have:

$$\begin{aligned}
& \mathbb{P}\left(\cap_{k\in[K]} \cap_{\ell\in[L]} \left\{\mathbb{1}_{C_k} \cdot N_{k,t} d_{\mathrm{KL}}\left(M_k^{(\ell)}, \hat{M}_{k,t}^{(\ell)}\right) \ge \zeta_k^{(\ell)}\right\}\right) \\
\le\ & \mathbb{P}\left(\cap_{k\in[K]} \cap_{\ell\in[L]} \left\{\hat{M}_{k,t}^{(\ell)} \le m_{k,N_{k,t}}^{(\ell)}, C_k\right\}\right) \\
\le\ & \mathbb{P}\left(\cap_{k\in[K]} \cap_{\ell\in[L]} \left\{\hat{M}_{k,t}^{(\ell)} \le m_{k,(1+\eta)t_k}^{(\ell)}, C_k\right\}\right) \\
\le\ & \Pi_{k\in[K]} \Pi_{\ell\in[L]} \mathbb{P}\left(\left\{\hat{M}_{k,t}^{(\ell)} \le m_{k,(1+\eta)t_k}^{(\ell)}, C_k\right\}\right) \\
\overset{(c)}{\le}\ & \Pi_{k\in[K]} \Pi_{\ell\in[L]} \exp\left(-(1+\eta)t_k d_{\mathrm{KL}}\left(M_k^{(\ell)}, m_k^{(\ell)}(t)\right)\right) \\
=\ & \exp\left(-\sum_{k\in[K]} \sum_{\ell\in[L]} (1+\eta)t_k d_{\mathrm{KL}}\left(M_k^{(\ell)}, m_k^{(\ell)}(t)\right)\right)
\end{aligned}$$

where, the (c) follows from Lemma F.2. Using Lemma F.3, with $Z_k =$ and $a = \frac{1}{(1+)}$, we have that:

$$\mathbb{P}\left[\mathbb{1}_{C_k} \sum_{k\in[K]} \sum_{\ell\in[L]} N_{k,t} d_{\mathrm{KL}}\left(M_k^{(\ell)}, \hat{M}_{k,t}^{(\ell)}\right) \ge \rho\right] \le \left(\frac{\rho e}{KL}\right)^{KL} \exp\left(\frac{-\rho}{1+\eta}\right)$$

This concludes the proof. □

Lemma F.2. *For any $k = 1, 2, \ldots, K$ let $1 \le t_k \le t$. Then for all $0 \le C_k^{(\ell)} \le M_k^{(\ell)}$ we have*

$$\mathbb{P}\left(\cap_{k\in[K]} \cap_{\ell\in[L]} \left\{\hat{M}_{k,t}^{(\ell)} \le C^{(\ell)},\ t_k \le N_{k,t}\right\}\right) \le \exp\left(-\sum_{k\in[K]} \sum_{\ell\in[L]} t_k d_{KL}\left(M_k^{(\ell)}, \hat{M}_{k,t}^{(\ell)}\right)\right).$$

Proof. For all $k \in [K]$ and $\ell \in [L]$ define the moment generating function as:

$$\phi_k^{(\ell)}(\lambda) = \ln \mathbb{E}_{M_k^{(\ell)}}\left[\exp(\lambda X)\right],$$

where X is sampled from a single-parameter exponential family with mean $M_k^{(\ell)}$. From Donsker-Vardhan Variational Formula, (Theorem G.2 in Appendix G), we have that:

$$d_{\text{KL}}\left(M_k^{(\ell)}, \hat{M}_{k,t}^{(\ell)}\right) = \sup_{\lambda \leq 0} \mathbb{E}_{\hat{M}_{k,t}^{(\ell)}}\left[\lambda X\right] - \ln \mathbb{E}_{M_k^{(\ell)}}\left[\exp(\lambda X)\right]$$

Define the events $\mathcal{E}_1 := \cap_{k\in[K]} \{t_k \leq N_{k,t}\}$ and $\mathcal{E}_2 := \cap_{k\in[K]} \cap_{\ell\in[L]} \left\{\hat{M}_{k,t}^{(\ell)} \leq C_k^{(\ell)}\right\}$ and $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$.

Define the martingale $G_t = \exp\left(\sum_{k\in[K]}\sum_{\ell\in[L]} \lambda_k^{(\ell)} X_{k,t} - N_{k,t}\phi_k^{(\ell)}(\lambda_k^{(\ell)})\right)$. For all $t' \leq t$, $G_{t'} = G_{t'-1}\exp\left(\sum_{k\in[K]}\sum_{\ell\in[L]} \lambda_k^{(\ell)} X_{k,t'} - N_{k,t'}\phi_k^{(\ell)}(\lambda_k^{(\ell)})\right)$ We deduce $\mathbb{E}\left[G_{t'}|\mathcal{F}_{t'-1}\right] = G_{t'-1}$ and $\mathbb{E}\left[G_t\right] = 1$.

We then define

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}\right) &= \mathbb{P}\left(\cap_{k\in[K]}\cap_{\ell\in[L]}\left\{X_{k,t} \leq N_{k,t}C_k^{(\ell)}\right\} \wedge \mathcal{E}_1\right) \\
&\leq \mathbb{P}\left(\mathbb{1}_{\mathcal{E}_1} \times \mathbb{1}\left\{\sum_{k\in[K]}\sum_{\ell\in[L]} \lambda_k^{(\ell)} X_{k,t} \geq \sum_{k\in[K]}\sum_{\ell\in[L]} \lambda_k^{(\ell)} N_{k,t} C_k^{(\ell)}\right\}\right) \\
&\leq \mathbb{P}\left(\mathbb{1}_{\mathcal{E}_1} \times \mathbb{1}\left\{\exp\left(\sum_{k\in[K]}\sum_{\ell\in[L]} \lambda_k^{(\ell)} X_{k,t}\right) \geq \exp\left(\sum_{k\in[K]}\sum_{\ell\in[L]} \lambda_k^{(\ell)} N_{k,t} C_k^{(\ell)}\right)\right\}\right) \\
&\leq \mathbb{P}\left(\mathbb{1}_{\mathcal{E}_1} \times \mathbb{1}\left\{G_t \geq \exp\left(\sum_{k\in[K]}\sum_{\ell\in[L]} \lambda_k^{(\ell)} N_{k,t} C_k^{(\ell)} - \lambda_k^{(\ell)}\phi_k^{\ell}\right)\right\}\right) \\
&\leq \mathbb{P}\left(\mathbb{1}_{\mathcal{E}_1} \times \mathbb{1}\left\{G_t \geq \exp\left(\sum_{k\in[K]}\sum_{\ell\in[L]} N_{k,t}(\lambda_k^{(\ell)} C_k^{(\ell)} - \phi_k^{\ell})\right)\right\}\right) \\
&\leq \mathbb{P}\left(\mathbb{1}_{\mathcal{E}_1} \times \mathbb{1}\left\{G_t \geq \exp\left(\sum_{k\in[K]}\sum_{\ell\in[L]} N_{k,t} d_{\text{KL}}\left(M_k^{(\ell)}, (m_{k,t}^{(\ell)}\right)\right)\right\}\right)
\end{aligned}$$

Using Markov's inequality and $\mathbb{E}\left[G_t\right] = 1$, we have:

$$\begin{aligned}
\mathbb{P}\left[\mathcal{E}\right] &= \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_1} G_t\right]\exp\left(-\sum_{k\in[K]}\sum_{\ell\in[L]} N_{k,t} d_{\text{KL}}\left(M_k^{\ell}, m_k^{(\ell)}\right)\right) \\
&\leq \exp\left(-\sum_{k\in[K]}\sum_{\ell\in[L]} N_{k,t} d_{\text{KL}}\left(M_k^{\ell}, m_k^{(\ell)}\right)\right)
\end{aligned}$$

□

Lemma F.3. *Let $a > 0$ and $K \geq 2$ and $Z \in \mathbb{R}^{K\times L}$ such that for all $\xi \in \mathbb{R}_+^{K\times L}$ we have:*

$$\mathbb{P}\left(Z \geq \zeta\right) \geq \exp\left(-a\sum_{k\in[K]}\sum_{\ell\in[L]} \zeta_k^{(\ell)}\right)$$

Then, for all $\rho \geq \frac{K}{a} \in \mathbb{R}_+$, we have:

$$\mathbb{P}\left(\sum_{k\in[K]}\sum_{\ell\in[L]} Z_{k,\ell} \geq \rho\right) \geq \left(\frac{ae\rho}{KL}\right)^{KL}\exp(-a\rho)$$

G Useful Existing Results

Theorem G.1 (Berge's Maximum Theorem (Berge, 1877)). *Let $\mathcal{U}$ and $\mathcal{V}$ be topological spaces, $f : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$ and $C : \mathcal{U} \to \mathcal{V}$ be non-empty compact set for all $u \in \mathcal{U}$. Then, if C is continuous at u, $f^*(u) = \max_{v \in C(u)} f(u, v)$ is continuous and $C^*(u) = \{v \in C(u) : f^*(u) = f(u, v)\}$ is upper-hemicontinuous.*

Theorem G.2 (Donsker-Vardhan Variational Formula (Donsker and Varadhan, 1975)). *For mutual information $KL(P\|Q)$, we have that:*

$$d_{\mathrm{KL}}(P\|Q) = \sup_f \mathbb{E}_P[f] - \ln \mathbb{E}_Q[\exp(f)]$$

Lemma G.1 (Peskun Ordering (Peskun, 1973)). *For any two random variables X, Y on $\mathbb{R}^{KL}$ the following are equivalent:*

1. $X \leq_s Y$
2. *For all* $x \in \mathbb{R}^{KL}$, $\mathbb{P}[X \geq x] \leq \mathbb{P}[Y \geq x]$
3. *For all non-negative functions* $f_1, f_2, \ldots, f_k$, *we have that:* $\Pi_{i=1}^K f_i \leq \Pi_{i=1}^K f_i$

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and precede the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification:

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with humansubjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.