Learning to Explore With Lagrangians For Bandits Under Unknown Constraints

Udvas Das **Debabrota Basu**
Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 – CRIStAL F-59000 Lille, France

Abstract

Pure exploration in bandits formalises multiple real-world problems, such as tuning hyper-parameters or conducting user studies to test a set of items, where different safety, resource, and fairness constraints on the decision space naturally appear. We study these problems as pure exploration in multi-armed bandits with unknown linear constraints, where the aim is to identify an *r-optimal and feasible policy* as fast as possible with a given level of confidence. First, we propose a Lagrangian relaxation of the sample complexity lower bound for pure exploration under constraints. Second, we leverage properties of convex optimisation in the Lagrangian lower bound to propose two computationally efficient extensions of Track-and-Stop and Gamified Explorer, namely LATS and LAGEX. Then, we propose a constraint-adaptive stopping rule, and while tracking the lower bound, use optimistic estimate of the feasible set at each step. We show that LAGEX achieves asymptotically optimal sample complexity upper bound, while LATS shows asymptotic optimality up to *novel* constraint-dependent constants. Finally, we conduct numerical experiments with different reward distributions and constraints that validate efficient performance of LATS and LAGEX.

1 INTRODUCTION

Multi-Armed Bandit (MAB) serves as an archetypal framework for sequential decision-making under uncertainty that allows us to study the corresponding information-utility trade-offs (Lattimore and Szepesvári, 2020). In MAB, at each step, an agent interacts with an environment consisting of K decisions (also knows as *arms*) corresponding to K noisy feedback distributions (or *reward* distributions). At each step, the agent takes a decision, and obtains a reward from its unknown reward distribution. The goal of the agent is to compute a *policy*, i.e. a distribution over the decisions, maximising a utility metric (e.g. accumulated rewards (Auer et al., 2002), probability of identifying the best arm (Kaufmann et al., 2016) etc.).

In this paper, we focus on the *pure exploration* problem of MABs, where the agent interacts by realising a sequence of policies (or experiments) with the goal of *answering a query* about the environment *as correctly as possible*. A well-studied pure exploration problem is Best-Arm Identification (BAI), where the agent aims to identify the arm with highest expected reward (Bubeck et al., 2009; Even-Dar et al., 2002a; Jamieson and Nowak, 2014; Kaufmann et al., 2016). Kaufmann et al. (2016) derives an information-theoretic lower bound quantifying the minimum number of agent-environment interactions needed to identify the best arm with a given level of confidence. The lower bound depends on optimising the weighted sum of KL-divergences between the reward distributions of arms and their most confusing counterparts (Equation (3)). In this paper, we leverage the lower bound for algorithm design that plugs-in the empirical estimates of the reward distributions in lower bound and solves the optimisation problem on-the-go (Carlsson et al., 2024; Degenne et al., 2019b; Kaufmann et al., 2016).

BAI has been applied in hyper-parameter tuning (Li et al., 2017), communication networks (Lindståhl et al., 2022), influenza mitigation (Libin et al., 2019), finding the optimal dose of a drug (Aziz et al., 2021a) etc. However, real-world scenarios often impose constraints on the arm pulls (Carlsson et al., 2024). *For example,* Baudry et al. (2024) considers a recommendation problem with the aim to guarantee a fixed (known) minimum expected revenue per recommended

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

content while identifying the best content from bandit feedback. Additionally, if we have multiple objectives in a decision making problem, a popular approach to optimize them is finding the optimal policy for one objective while constraining the others (Fonseca and Fleming, 1998).

Pure Exploration under Constraints. The aforementioned problems motivated the study of pure exploration under a set of known and unknown constraints (Carlsson et al., 2024; Katz-Samuels and Scott, 2018; Li et al., 2023; Wang et al., 2021b; Wu et al., 2023). Specifically, we aim to find the optimal policy that maximises the expected reward obtained from the set of arms and satisfies the true constraints, with confidence $1-\delta$. This is known as the *fixed-confidence setting* of pure exploration (Carlsson et al., 2024; Wang et al., 2021b). On the other hand, there is also the fixed-budget setting, which is of independent interest (Faizal and Nair, 2022; Katz-Samuels and Scott, 2018; Li et al., 2023). Existing literature has studied either the general linear constraints when they are known (Camilleri et al., 2022a; Carlsson et al., 2024), or very specific type of unknown constraints, e.g. safety (Wang et al., 2021b), knapsack (Li et al., 2023), fairness (Wu et al., 2023), preferences (Lindner et al., 2022) etc. Here, we study the *pure exploration problem in the fixed-confidence setting subject to unknown linear constraints on the policy*, which generalises all these settings (Section 2). A detailed discussion on related works is deferred to Appendix B.1.

Recently, Carlsson et al. (2024) derives a tight lower bound and designs asymptotically optimal algorithms for this problem when the constraints are known, and show that a bandit instance might become harder or easier depending on the geometry of the constraints. They pose that *studying similar phenomenon for unknown constraints is an open problem* as constraints are also estimated.

The challenge is that the lower bound dictating the hardness of the constrained pure exploration problem is sensitive to the active constraints, and for unknown constraints, we only have access to estimated constraints with non-zero error. This affects the exact detection of the optimal feasible policy.

Example. Let us consider a bandit environment with 4 arms with Gaussian rewards – means $\boldsymbol{\mu} = [1.0, 0.8, 0.6]$ and variance 1. The constraints are $A\boldsymbol{\pi} \leq \mathbf{0}$ where $A = \begin{bmatrix} 0.0, -0.3, -0.2 \\ -0.2, 0.1, 0.1 \end{bmatrix}$. Here, the optimal policy is $\boldsymbol{\pi}^\star = [1.0, 0, 0]$, and the first constraint is active at the optimal policy. Figure 1 shows a perturbation of the active constraint by 0.005 shifts the optimal policy, and the sample complexity (Carlsson et al., 2024)

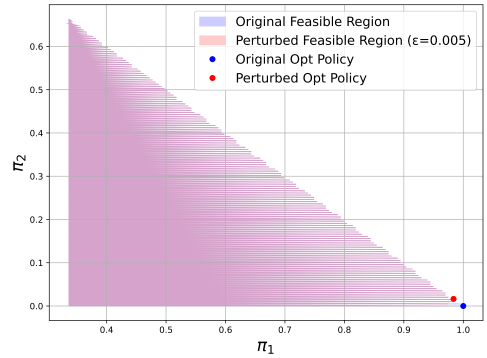

Figure 1: Effect of rank-one update of active constraint blows with $\mathcal{O}(4 \times 10^4)$.

For rigour, we relax the problem of finding the optimal feasible policy to finding an r-optimal feasible policy. For a given $r > 0$, an r-optimal policy has mean reward not more than r away from that of the optimal policy (Jourdan and Degenne, 2022; Jourdan et al., 2023; Mason et al., 2020). This leads us to two questions:

1. *How does the hardness of finding r-optimal feasible policy change under unknown constraints?*
2. *How can we design a generic algorithmic scheme to track both the constraints and optimal policy with sample- and computational-efficiency?*

Our Contributions positively address the questions.

1. *Lagrangian relaxation of the lower bound.* Minimum number of samples required to conduct r-optimal pure exploration with fixed confidence is expressed by a lower bound– an optimisation problem under known constraints (Eq. (3)). To efficiently handle unknown constraints, we propose a novel Lagrangian relaxation of this optimisation (Section 3). At every step, we construct optimistic feasible policy set, and plug it in the relaxation. Lagrangian multipliers balance the identifiability of an r-optimal policy and the feasibility under estimated constraints. We leverage results from convex analysis to show that the relaxed lower bound with optimistic feasible set preserves all the desired properties of the lower bound under known constraints, and thus, allows designing lower bound tracking algorithm.

2. *Generic algorithm design.* First, we propose a new stopping rule accommodating the estimated constraints.This ensures concentration of the estimates of mean rewards and constraints to their true values before final recommendation. We further show this stopping rule recommends a policy that is both feasible and r-optimal with confidence at least $1-\delta$. Then, we extend the Track-and-Stop (Garivier and Kaufmann, 2016) and gamified explorer (Degenne et al., 2019b) approaches with the Lagrangian lower bound to design LATS (LAgrangian Track and Stop) and LAGEX

Algorithm 1 Pure Exploration in Bandits with Unknown Linear Constraints

Input: Tolerance $r > 0$, Confidence level $\delta \in (0, 1)$
for $t = 1, \ldots$ **do**
  Decision/sampling: Play an arm $a_t \in [1, K]$
  Reward-cost Feedback: Observe reward $r_t \sim P_{a_t}$ and cost $\mathbf{c}_t \sim Aa_t + \eta_t$
  if more than $1-\delta$ confident about the estimated answer being correct **then**
    Stop and stopping time $\tau_\delta \leftarrow t$
  end if
end for
Recommendation: $\hat{\boldsymbol{\pi}} = \arg\max_{\boldsymbol{\pi} \in \Pi^r_{\hat{\mathcal{F}}_{\tau_\delta}}} \hat{\boldsymbol{\mu}}^\top_{\tau_\delta} \boldsymbol{\pi}$

(LAgrangian Gamified EXplorer), respectively in Section 4.

3. *Upper bounds on sample complexities.* We derive upper bounds on the sample complexities of LATS and LAGEX (Section 4). This requires proving a novel concentration inequality for the constraint estimates. As a consequence, LATS achieves an upper bound, which is $(1 + \mathfrak{s})$ times the asymptotic upper bound of TS under known constraints, while LAGEX exhibits optimality. $\mathfrak{s}$ is the shadow price (ratio between maximum and minimum index value of the slack vector) that quantifies its stability under perturbation. Finally, we conduct experiments across synthetic and real data. We observe that LAGEX requires the least samples among competing algorithms and exactly tracks the change in hardness due to constraints across environments (Section 5).

2 EXPLORATION: UNKNOWN CONSTRAINTS

Notations: $\mathbf{x}$, X, and $\mathcal{X}$ denote a vector, a matrix, and a set, respectively. Detailed notations are in Table 1. We augment the simplex constraints in A, and normalise each row of A, i.e. $\|A_i\|_2 = 1$ for all $i \in [d]$.

Problem Formulation. We work with a MAB instance consisting of $K \in \mathbb{N}$ arms. Each arm $a \in [K]$ has a reward distribution P_a with unknown mean $\boldsymbol{\mu}_a \in \mathbb{R}$. At each step $t \in \mathbb{N}$, the agent chooses an action $A_t \in [K]$, and observes a stochastic reward $R_t \sim P_{A_t}$. A feasible policy $\boldsymbol{\pi} \in \Delta_K$ satisfies $A\boldsymbol{\pi} \leq \mathbf{0}$ with respect to d linear constraints $A \in \mathbb{R}^{d \times K}$.

If A is known, the agent has access to the non-empty and compact set of feasible policies $\mathcal{F} \triangleq \{\boldsymbol{\pi} \in \Delta_K \mid A\boldsymbol{\pi} \leq \mathbf{0}\}$. The agent aims to identify an r-optimal optimal feasible policy, i.e. any feasible policy which belongs to $\Pi^r_{\mathcal{F}} \triangleq \{\boldsymbol{\pi} \in \mathcal{F} \mid \boldsymbol{\mu}^\top \boldsymbol{\pi} + r \geq \boldsymbol{\mu}^\top \boldsymbol{\pi}^\star_{\mathcal{F}}\}$, given

$$\boldsymbol{\pi}^\star_{\mathcal{F}} \triangleq \arg\max_{\boldsymbol{\pi} \in \mathcal{F}} \boldsymbol{\mu}^\top \boldsymbol{\pi}\,. \tag{1}$$

Definition 1 **($(1-\delta)$-correct and $(1-\delta)$-feasible r-optimal pure exploration).** *For $\delta \in [0, 1)$, an r-optimal pure exploration algorithm is called $(1-\delta)$-correct and $(1-\delta)$-feasible if the policy $\hat{\boldsymbol{\pi}}$ recommended by it satisfies $\Pr[\hat{\boldsymbol{\pi}} \notin \Pi^r_{\mathcal{F}}] \leq \delta$ and $\Pr[A\hat{\boldsymbol{\pi}} \geq \mathbf{0}] \leq \delta$.*

In our setting, we do not have access to the true set of constraints. Hence, using the observations, we construct $\hat{A}$ as an estimate of A. Then, the agent builds an estimated feasible set $\hat{\mathcal{F}} \triangleq \{\boldsymbol{\pi} \in \Delta_K \mid \hat{A}\boldsymbol{\pi} \leq \mathbf{0}\}$ to identify the optimal feasible policy as $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} \triangleq \arg\max_{\boldsymbol{\pi} \in \hat{\mathcal{F}}} \boldsymbol{\mu}^\top \boldsymbol{\pi}$. In addition, the estimated r-optimal policy set is $\Pi^r_{\hat{\mathcal{F}}} \triangleq \{\boldsymbol{\pi} \in \hat{\mathcal{F}} \mid \boldsymbol{\mu}^\top \boldsymbol{\pi} + r \geq \boldsymbol{\mu}^\top \boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}\}$. We know that obtaining accurate estimates of these quantities would require us to collect feedback of satisfying constraints over time. This poses an additional cost of using observations to estimate $\boldsymbol{\mu}$.

Goal. In order to recommend a $(1-\delta)$-correct and $(1-\delta)$-feasible policy that is r-optimal with respect ot the true optimal policy $\boldsymbol{\pi}^\star_{\mathcal{F}}$, we aim to minimise the expected number of interactions $\mathbb{E}[\tau_\delta] \in \mathbb{N}$.

2.1 Extension of Prior Bandit Problems

Now, we clarify our motivation by showing that different existing problems are special cases of our setting.

a. Thresholding Bandits. Thresholding bandits (Aziz et al., 2021a) are motivated from the safe dose finding problem, where one wants to identify the most effective dose of a drug below a known safety level. This has also motivated the safe arm identification problem (Wang et al., 2021b). Our setting generalises it further to detect the optimal combination of doses of available drugs yielding highest efficacy while staying below the safety threshold. Formally, we identify $\boldsymbol{\pi}^\star = \arg\max_{\boldsymbol{\pi} \in \Delta_K} \boldsymbol{\mu}^\top \boldsymbol{\pi}$, such that $I\boldsymbol{\pi} \leq I\boldsymbol{\theta}$ for thresholds $\boldsymbol{\theta}_a > 0$.

b. BAI with Fairness Constraints across Subpopulations (BAICS). Wu et al. (2023) aims to select an arm that must be fair across M sub-populations. Here, the arm belongs to a set $C \triangleq \{k \in [K] | \boldsymbol{\mu}_{k,m} \geq 0, m \in [M]\}$ where the observation for arm k and population m comes from $\mathcal{N}(\boldsymbol{\mu}_{k,m}, 1)$. It ensures that the chosen arm does not perform *too bad* for any sub-population. Like standard BAI, finding only the optimal arm might not be enough because it might not perform equally good for all of the sub-populations. This is similar to having M groups of patients and K drugs to administer, where we are looking for a mixture of drugs such that $\boldsymbol{\pi}^\star = \arg\max_{\boldsymbol{\pi} \in \Delta_K} \boldsymbol{\mu}^\top \boldsymbol{\pi}$, such that $\mathbb{1}^\top_{\boldsymbol{\mu}_m \geq \mathbf{0}} \boldsymbol{\pi} = 1, \forall m \in [M]$.

We refer to Section B.2 for further discussion on generalisation to the existing bandit settings.

3 LAGRANGIAN RELAXATION

Now, we derive Lagrangian relaxation of lower bound and its properties under a structural assumption.

Assumption 1 (Structures of means, policy, and constraints). *(a) The mean vector $\boldsymbol{\mu}$ is in a bounded subset $\mathcal{D}$ of $\mathbb{R}^K$. (b) There exists a unique optimal feasibly policy (Eq. (1)). (c) The true constraint A yields a non-zero slack vector Γ: $\max_{\boldsymbol{\pi}\in\Delta_K}(-A\boldsymbol{\pi}) = \Gamma$.*

We impose the unique optimal and feasible policy assumption following Carlsson et al. (2024). The assumption on slack is analogous to existence of a safe-arm (Pacchiano et al., 2020), or Slater's condition for the constraint optimisation problem (Liu et al., 2021).

3.1 Information Acquisition: Estimate Constraints

The agent acquires information at every step $t \in \mathbb{N}$ by sampling an action $\mathbf{a}_t \sim \boldsymbol{\omega}_t$. As the arms are independent, we represent the a-th arm as the a-th basis in $\mathbb{R}^K$, denoted by $\mathbf{a} \in \mathbb{R}^K$. $\boldsymbol{\omega}_t \in \Delta_K$ is called *the allocation policy*. As shown in Algorithm 1, pulling the arm a_t yields a noisy reward $r_t \in \mathbb{R}$ and cost vector $\mathbf{c}_t \in \mathbb{R}^d$. The cost vector $\mathbf{c}_t \triangleq A\mathbf{a}_t + \boldsymbol{\eta}_t$, where $\boldsymbol{\eta}_t$ is an independent and identically distributed noise vector whose each component uis generated from a 1-subGaussian distribution with mean zero.

Thus, using the observations obtained till t, we estimate the mean vector as $\hat{\boldsymbol{\mu}}_t \triangleq \Sigma_t^{-1}\left(\sum_{s=1}^{t-1} r_s \mathbf{a}_s\right)$. Here, $\Sigma_t \triangleq I + \sum_{s=1}^{\top} \mathbf{a}_s\mathbf{a}_s^\top$, is the Gram matrix or the design matrix. Similarly, the estimate of the i-th row of the constraint matrix is $\hat{A}_t^i \triangleq \Sigma_t^{-1}\left(\sum_{s=1}^{t-1} A^{i,\mathbf{a}_s}\mathbf{a}_s\right)$. But naïvely using $\hat{A}_t$ to define the feasible policy set does not ensure that for any t, the estimated feasible set $\hat{\mathcal{F}}$ is a superset of $\mathcal{F}$. Hence, we define a confidence ellipsoid around $\hat{A}_t$ that includes A with probability at least $1-\delta$, and construct a optimistic estimate of A. Formally, the confidence ellipsoid is

$$\mathcal{C}_t \triangleq \left\{A' \in \mathbb{R}^{d\times K} \mid \|A'^i - \hat{A}_t^i\|_{\Sigma_t} \le f(t,\delta)\forall i \in [d]\right\}, \quad (2)$$

where $f(\delta,t) \triangleq 1 + \sqrt{\frac{1}{2}\log\frac{K}{\delta} + \frac{1}{4}\log\det\Sigma_t}$ is a monotonically non-decreasing function of t.

Lemma 1 (Optimistic feasible sets). *At any time $t \in \mathbb{N}$, we construct the optimistic feasible policy set such that with probability $1-\delta$, $\hat{\mathcal{F}}_t \triangleq \{\boldsymbol{\pi} \in \Delta_K : \min_{A'\in\mathcal{C}_t} A'\boldsymbol{\pi} \le \mathbf{0}\}$, satisfies $\mathcal{F} \subseteq \hat{\mathcal{F}}_t$, where $\tilde{A}_t \triangleq \arg\min_{A'\in\mathcal{C}_t} A'\boldsymbol{\pi}$.*

Figure 2 plots this result using the numerical values obtained from our algorithms. Note that as we acquire more samples, estimated feasible policy set $\hat{\mathcal{F}}_t \to \mathcal{F}$.

Remark 1. *To ensure that the true optimal and feasible policy $\boldsymbol{\pi}^\star_{\mathcal{F}} \in \hat{\mathcal{F}}_t$ is inside the estimated feasible policy set, and $\hat{\mathcal{F}}_t \neq \emptyset$ for any t, we use the optimistic estimate of the feasible set $\hat{\mathcal{F}}_t$ ensuring $\mathcal{F} \subseteq \hat{\mathcal{F}}_t$ with high probability. Our construction of the optimistic estimates of feasible set resonate with the optimistic-pessimistic algorithms for regret-minimisation under constraints (Chen et al., 2022b; Liu et al., 2021; Pacchiano et al., 2024, 2020). Note that, similar confidence bounds for estimating constraints are used to test feasibility of linear programs (Gangrade et al., 2024).*

3.2 Relaxation with Estimated Constraints

Search for the optimal policy is essentially a linear programming problem when we know the mean vector $\boldsymbol{\mu}$ and a constraint matrix A. The challenge in bandit is to identify them from sequential feedback, i.e. to differentiate $\boldsymbol{\mu}$ from the other confusing instances in the same family of distributions. These are called the *alternative instances.* The strategy is to gather enough statistical evidence to rule out all such confusing instances, specifically the one that has minimum KL-divergence from $\boldsymbol{\mu}$ as observed under the allocation policy $\boldsymbol{\omega}$ (Garivier and Kaufmann, 2016). This intuition has led to the lower bound of Carlsson et al. (2024)– the expected stopping time of any $(1-\delta)$-correct and always-feasible algorithm satisfies

$$\mathbb{E}[\tau_\delta] \ge T_{\mathcal{F},0}(\boldsymbol{\mu}) \ln\frac{1}{2.4\delta}\,, \quad (3)$$

if A is known and there exists a unique optimal policy $\boldsymbol{\pi}^\star_{\mathcal{F}} = \arg\max_{\boldsymbol{\pi}\in\mathcal{F}} \boldsymbol{\mu}^\top\boldsymbol{\pi}$. $T_{\mathcal{F},\cdot}(\boldsymbol{\mu})$ is called the *characteristics time*. Its reciprocal is a max-min optimisation problem over the set of alternative instances $\Lambda_{\mathcal{F}}(\boldsymbol{\mu}) \triangleq \left\{\boldsymbol{\lambda} \in \mathcal{D} \mid \max_{\boldsymbol{\pi}\in\mathcal{F}} \boldsymbol{\lambda}^\top\boldsymbol{\pi} > \boldsymbol{\lambda}^\top\boldsymbol{\pi}^\star_{\mathcal{F}}\right\}$, where , i.e.

$$\begin{aligned} T^{-1}_{\mathcal{F},0}(\boldsymbol{\mu}) &\triangleq \sup_{\boldsymbol{\omega}\in\Delta_K} \inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{a=1}^{K} \boldsymbol{\omega}_a d\left(\boldsymbol{\mu}_a, \boldsymbol{\lambda}_a\right) \\ &\triangleq \sup_{\boldsymbol{\omega}\in\Delta_K} \inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \boldsymbol{\omega}^\top d\left(\boldsymbol{\mu}, \boldsymbol{\lambda}\right)\,. \quad (4) \end{aligned}$$

$\Lambda_{\mathcal{F}}(\boldsymbol{\mu})$, referred as the Alt-set, is the set of all bandit instances whose mean vectors are in a bounded subset $\mathcal{D} \in \mathbb{R}^K$ but the optimal policy is different than that of $\boldsymbol{\mu} \in \mathcal{D}$. Now, we inspect the change in this lower bound at any step $t > 0$, when we only have access to a optimistic estimate $\hat{\mathcal{F}}_t$ and the confidence ellipsoid $\mathcal{C}_t$ but do not know $\mathcal{F}$. For brevity, we exclude t from the subscripts for where it is clear from the context.

We realise non-uniqueness of r-optimal policies follows from the definition of $\Pi^r_{\mathcal{F}}$, thus there exists multiple "correct answers". Thus, in the spirit of (Degenne and Koolen, 2019), we define Lagrangian relaxation of the lower bound, i.e. $T^{-1}_{\mathcal{F},r}(\boldsymbol{\mu})$

$$\triangleq \sup_{\boldsymbol{\omega}\in\Delta_K} \inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \boldsymbol{\omega}^\top d\left(\boldsymbol{\mu},\boldsymbol{\lambda}\right) \leq \inf_{\boldsymbol{l}\in\mathbb{R}^d_+} \min_{A'\in\mathcal{C}} \sup_{\boldsymbol{\omega}\in\Delta_K} \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}} \inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})} \boldsymbol{\omega}^\top d\left(\boldsymbol{\mu},\boldsymbol{\lambda}\right) - \boldsymbol{l}^\top A'\boldsymbol{\omega}\,, \quad (5)$$

where the Alt-set given $\hat{\mathcal{F}}$ is defined as

$$\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) \triangleq \big\{\boldsymbol{\lambda}\in\mathcal{D} \mid \max_{\boldsymbol{\pi}\in\hat{\mathcal{F}}} \boldsymbol{\lambda}^\top\boldsymbol{\pi} - r > \boldsymbol{\lambda}^\top\boldsymbol{\pi}\big\}, \quad (6)$$

where $\boldsymbol{\pi}\in\arg\max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}} \boldsymbol{\mu}^\top\boldsymbol{\pi}$. Since $\mathcal{F}\subseteq\hat{\mathcal{F}}$, we observe that $\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})\subseteq\Lambda_{\mathcal{F}}(\boldsymbol{\mu})$ (Figure 3).

We denote this Lagrangian relaxation (Equation (5)) of the characteristic time with $\hat{\mathcal{F}}$ as $T^{-1}_{\hat{\mathcal{F}},r}(\boldsymbol{\mu})$. For non-negative Lagrange multipliers $\boldsymbol{l}\in\mathbb{R}^d_+$, the first inequality is true due to the existence of a slack for true constraints A. The inequality holds due to the optimistic choice of the estimated constraint. Equation (5) shows that the reciprocal of the Lagrangian relaxation, $T_{\hat{\mathcal{F}},r}(\boldsymbol{\mu})$, serves as a upper bound on the characteristic time $T_{\mathcal{F}}(\boldsymbol{\mu})$ for known constraints (Carlsson et al., 2024). This leads to a natural question:

> Does the dual of the optimization problem for $T^{-1}_{\hat{\mathcal{F}},r}(\boldsymbol{\mu})$ yield the same solution as the primal?

We formalise the strong duality result and self-bounding property of the Lagrangian multiplier of the relaxation in Equation (5) in Theorem 1 below.

Theorem 1 (Strong Duality and Range of Lagrange Multipliers)**.** *The optimisation problem in Equation (5) satisfies*

$$\inf_{\boldsymbol{l}\in\mathbb{R}^d_+} \min_{A'\in\mathcal{C}} \sup_{\boldsymbol{\omega}\in\Delta_K} \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}} \inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})} \boldsymbol{\omega}^\top d\left(\boldsymbol{\mu},\boldsymbol{\lambda}\right) - \boldsymbol{l}^\top A'\boldsymbol{\omega} = \sup_{\boldsymbol{\omega}\in\Delta_K} \min_{\boldsymbol{l}\in\mathcal{L}} \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}} \inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})} \boldsymbol{\omega}^\top d\left(\boldsymbol{\mu},\boldsymbol{\lambda}\right) - \boldsymbol{l}^\top\tilde{A}\boldsymbol{\omega}\,. \quad (7)$$

Here, $\mathcal{L}\triangleq\{\boldsymbol{l}\in\mathbb{R}^d_+ \mid 0\leq\|\boldsymbol{l}\|_1\leq\frac{1}{\gamma}\mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\mathcal{F})\}$, where $\gamma\triangleq\min_{i\in[1,d]}\{-\tilde{A}^i\hat{\boldsymbol{\pi}}^\star\}$, i.e. the minimum slack w.r.t. the estimated optimal feasible policy.

Detailed proof is in Appendix C. Hereafter, we use the RHS of Eq. (7) as $T^{-1}_{\hat{\mathcal{F}},r}(\boldsymbol{\mu})$. Theorem 1 provides a hypercube to search for the Lagrangian multipliers, which is a linear programming problem.

Remark 2 (Connections with Lagrangian-based Methods in Bandits.)**.** *Regret minimisation literature leverages Lagrangian-based optimistic-pessimistic methods (Slivkins et al., 2023; Tirinzoni et al., 2020) to obtain both the sub-linear regret and constraint violation guarantees (Bernasconi et al., 2024; Liu et al., 2021). Our proposed algorithm LAGEX (Algorithm 3) is a prime example where the "self-boundedness" of the dual variables results in tighter constraint violation guarantees (Figure 8 and 9).*

I. The Inner Optimisation Problem. Now, we peel the layers of the optimisation problem in Eq. (7) and focus on obtaining

$$\mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}}) \triangleq \min_{\boldsymbol{l}\in\mathcal{L}} \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}} \inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})} \boldsymbol{\omega}^\top d\left(\boldsymbol{\mu},\boldsymbol{\lambda}\right) - \boldsymbol{l}^\top\tilde{A}\boldsymbol{\omega}\,.$$

For known constraints and $r=0$, Carlsson et al. (2024) has leveraged results from convex analysis (Boyd and Vandenberghe, 2004) to show that the most confusing instance for $\boldsymbol{\mu}$ lie in the boundary of the normal cone $\Lambda_{\mathcal{F}}(\boldsymbol{\mu})^C$ spanned by the active constraints $A_{\boldsymbol{\pi}^\star_{\mathcal{F}}}$ for $\boldsymbol{\pi}^\star_{\mathcal{F}}$. $A_{\boldsymbol{\pi}^\star_{\mathcal{F}}}$ is a sub-matrix of A consisting at least K linearly independent rows. *This is called the projection lemma.* Specifically, $\mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\mathcal{F}\mid r=0) = \min_{\boldsymbol{\pi}'\in\nu_{\mathcal{F}}(\boldsymbol{\pi})} \min_{\boldsymbol{\lambda}:\boldsymbol{\lambda}^\top(\boldsymbol{\pi}-\boldsymbol{\pi}')=0} \boldsymbol{\omega}^\top d\left(\boldsymbol{\mu},\boldsymbol{\lambda}\right)$.

In our setting, we are sequentially estimating both the mean vectors and the constraints, and thus, the normal cone. Now, we derive the projection lemma for our optimistic feasible set and $r\neq 0$.

Proposition 1 (Projection Lemma for Unknown Constraints)**.** *For any $\boldsymbol{\omega}\in\hat{\mathcal{F}}$ and $\boldsymbol{\mu}\in\mathcal{D}$, the following projection lemma holds for the Lagrangian relaxation in Equation (5), $\mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}}) =$*

$$\min_{\boldsymbol{l}\in\mathcal{L}} \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi})} \min_{\boldsymbol{\lambda}:\boldsymbol{\lambda}^\top(\boldsymbol{\pi}-\boldsymbol{\pi}')=r} \boldsymbol{\omega}^\top d\left(\boldsymbol{\mu},\boldsymbol{\lambda}\right) - \boldsymbol{l}^\top\tilde{A}\boldsymbol{\omega}\,. \quad (8)$$

Proposition 1 reduces the inner minimisation problem to a less intensive discrete optimisation, where we only have to search over the neighbouring vertices of the optimal policy in $\hat{\mathcal{F}}$ for a solution. Now, a natural question arises around this formulation:

> Can we track $\mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}})$ in the projection lemma over time as we sequentially estimate the constraints?

To answer this, we first prove convergence of estimated feasible set and alternating instance in order to prove convergence of $\mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}}_t)$ in the theorem below. Detailed proof is in Appendix D.

Theorem 2. *For a sequence $\{\hat{\mathcal{F}}_t\}_{t\in\mathbb{N}}$ and $\{\hat{\boldsymbol{\lambda}}_t\}_{t\in\mathbb{N}}$, we show that (a) $\lim_{t\to\infty}\hat{\mathcal{F}}_t\to\mathcal{F}$, (b) $\boldsymbol{\lambda}^\star$ is unique, and (c) $\lim_{t\to\infty}\hat{\boldsymbol{\lambda}}_t\to\boldsymbol{\lambda}^\star$. Then, for any $\boldsymbol{\omega}\in\mathcal{F}$ and $\boldsymbol{\mu}$, $\lim_{t\to\infty}\mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}}_t)\to\mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\mathcal{F})$, where $\boldsymbol{\lambda}^*\in\arg\min_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})}\boldsymbol{\omega}^\top d\left(\boldsymbol{\mu},\boldsymbol{\lambda}\right)$.*

II. The Outer Optimisation Problem. As we guarantee the convergence of $\mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}}_t)$ as $\hat{\mathcal{F}}_t\to\mathcal{F}$,

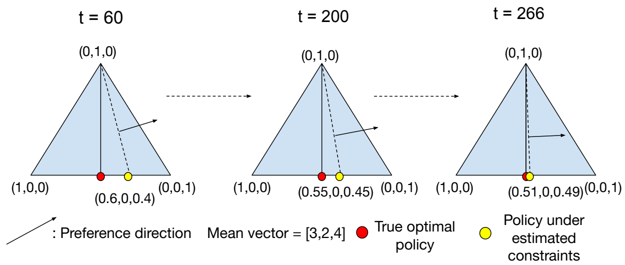

Figure 2: Convergence of the optimistic feasible set and optimal policy.

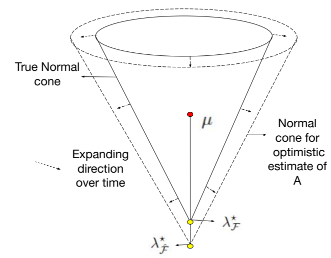

Figure 3: Normal cones over time.

we are left with the outer optimisation problem in Equation (7). Since it is a linear problem in $\boldsymbol{\omega}$, we can use a linear programming method leading to a vertex of $\hat{\mathcal{F}}_t$. But to be sure of an existence of a solution at each $t \in \mathbb{N}$, we need well-behavedness properties of the optimal allocation $\boldsymbol{\omega}^*(\boldsymbol{\mu})$. First, we observe that our estimates of the mean vector converge to $\boldsymbol{\mu}$ as $t \to \infty$. Hence, we also get $\lim_{t\to\infty} \mathcal{D}(\boldsymbol{\omega}, \hat{\boldsymbol{\mu}}_t, \hat{\mathcal{F}}_t) \to \mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \mathcal{F})$. Now, we ensure well-behavedness and existence of an optimal allocation for all $t > 0$.

Theorem 3 (Existence of unique optimal allocation)**.** *For all $\boldsymbol{\mu} \in \mathcal{D}$, $\boldsymbol{\omega}^\star(\boldsymbol{\mu})$ satisfies: 1. Both the sets $\hat{\mathcal{F}}$ and $\boldsymbol{\omega}^\star(\boldsymbol{\mu})$ are closed and convex. 2. For all $\boldsymbol{\mu} \in \mathcal{D}$ and $\boldsymbol{\omega} \in \hat{\mathcal{F}}$, $\lim_{t\to\infty} \mathcal{D}(\boldsymbol{\omega}, \hat{\boldsymbol{\mu}}_t, \hat{\mathcal{F}}_t)$ is continuous. 3. Reciprocal of the characteristic time $\lim_{t\to\infty} T^{-1}_{\hat{\mathcal{F}}_t, r}(\boldsymbol{\mu})$ is continuous for all $\boldsymbol{\mu} \in \mathcal{D}$. 4. For all $\boldsymbol{\mu} \in \mathcal{D}$, $\boldsymbol{\mu} \to \boldsymbol{\omega}^\star(\boldsymbol{\mu})$ is upper hemi-continuous. Thus, the optimization problem $\max_{\boldsymbol{\pi}\in\hat{\mathcal{F}}} \boldsymbol{\mu}^\top \boldsymbol{\pi}$ has a unique solution.*

Characterising Lower Bound for Gaussians. Since we can derive explicit form of the optimisation problem for Gaussian reward distributions, we characterise it further to relate our lower bound with the lower bound for known constraints (Appendix F).

Theorem 4. *Let $\{P_a\}_{a\in[K]}$ be Gaussian distributions with equal variance $\sigma^2 > 0$, $T^{-1}_{\hat{\mathcal{F}},r}$ is*

$$\max_{\boldsymbol{\omega}\in\Delta_K} \min_{\boldsymbol{l}\in\mathcal{L}} \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi})} \left\{ \frac{(r - \boldsymbol{\mu}^\top(\boldsymbol{\pi}-\boldsymbol{\pi}'))^2}{2\sigma^2 \|\boldsymbol{\pi}-\boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} - \boldsymbol{l}^\top \tilde{A}\boldsymbol{\omega} \right\},$$

where $\mathrm{Diag}(1/\boldsymbol{\omega}_a)$ is a K-dimensional diagonal matrix with a-th diagonal entry $1/\boldsymbol{\omega}_a$ and $\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi})$ is the set of neighbouring policies of $\boldsymbol{\pi}$ in $\hat{\mathcal{F}}$.

We get the lower bound under known constraints (Carlsson et al., 2024), if $r = 0$, i.e we search for the 'true' optimal policy $\boldsymbol{\pi}^\star_{\mathcal{F}}$ rather than an r-optimal feasible policy.

Corollary 1. *Let $d^2_{\boldsymbol{\pi}} \triangleq \frac{\|\boldsymbol{\pi}^\star_{\mathcal{F}}-\boldsymbol{\pi}\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^\top}}{\|\boldsymbol{\pi}^\star_{\mathcal{F}}-\boldsymbol{\pi}\|^2_2}$ be the norm of the projection of $\boldsymbol{\mu}$ on the policy gap $(\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi})$.*

Part i. *Then, we get*

$$\frac{2\sigma^2 K}{C_{\text{known}}}(1 + \mathfrak{s}_{\tilde{A}}) \leq T_{\hat{\mathcal{F}},0}(\boldsymbol{\mu}) \leq \frac{2\sigma^2 K}{C_{\text{known}}},$$

where $C_{\text{known}} = \min_{\boldsymbol{\pi}''\in\nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} d^2_{\boldsymbol{\pi}''}$.
Part ii.

$$T_{\hat{\mathcal{F}},0}(\boldsymbol{\mu}) \geq \frac{H}{\kappa^2}(1 + \mathfrak{s}_{\tilde{A}}),$$

where $\mathfrak{s}_{\tilde{A}} \triangleq \frac{\max_{i\in[1,d]} \tilde{\mathbf{A}}\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}}{\min_{i\in[1,d]} \tilde{\mathbf{A}}\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}}$ is the estimated shadow price and H is inversely proportional to the sum of squares of gaps and κ_{known} is condition number of a sub-matrix of A with K linearly independent active constraints for $\boldsymbol{\pi}^\star$.

Remark 3 (Connection to Existing Lower Bounds.)**.** ***(a) Pure exploration under known constraints.*** *The upper and lower bounds on characteristic time coincides with the existing lower bound under known constraints, i.e. when $\epsilon = 0$, i.e. $\hat{\mathcal{F}} = \mathcal{F}$.* ***(b) BAI without constraints.*** *In BAI, we consider only deterministic policies (or pure strategies) of playing a single arm. Then, we get $d_{\boldsymbol{\pi}_a} = \frac{\boldsymbol{\mu}^\top(\boldsymbol{\pi}^\star_{\mathcal{F}}-\boldsymbol{\pi})_a}{|(\boldsymbol{\pi}^\star_{\mathcal{F}}-\boldsymbol{\pi})_a|} = \boldsymbol{\mu}^\star - \boldsymbol{\mu}_a$, i.e. the sub-optimality gap for arm a. Here, $\boldsymbol{\mu}^\star$ is the mean of the best arm. In our setting, if there are no constraints involved then $\tilde{A} = A$ and so $\epsilon = 0$. Then the lower bound expression in Part ii. of Corollary 1 is inversely proportional to the sum-squared sub-optimal gaps which resonates with the complexity measure in standard BAI with only simplex constraints (Kaufmann et al. (2016)). Though the lower bound stated in Corollary 1 successfully encompasses the effect of unknown linear constraint by introducing novel constraint dependent factors that scales the lower bound.*

4 ALGORITHM DESIGN

Now, we propose two algorithms to conduct pure exploration with Lagrangian relaxation of lower bound, and derive upper bounds on their sample complexities.

Assumption 2 (Distributional assumptions on rewards and constraints)**.** *We require two distributional assumptions on rewards and constraints. (i) Reward distributions $\{P_a\}_{a=1}^K$ are sub-Gaussian one parameter exponential family with mean vector $\boldsymbol{\mu} \in \mathcal{D}$. (ii)*

Each constraint follows a sub-Gaussian K-parameter exponential family parameterised by A^i for $i \in [d]$.

These distributional assumptions are standard in bandits under constraints (Carlsson et al., 2024; Degenne and Koolen, 2019; Pacchiano et al., 2024, 2020).

Algorithm Design. Any algorithm in pure exploration setting consists of three main components. The first one is a sequential hypothesis testing that decides whether we need to keep sampling or not, popularly known as the "stopping criterion". To come up with such a criterion we need to make sure that we have gathered sufficient information about all the parameters in estimation, specifically $\boldsymbol{\mu}$ and A in our case. So it seems natural to enforce criterion on both mean and constraint concentration.

Lemma 2. *If the recommended policy is $(1-\delta)$-correct then it is $(1-\delta)$-feasible.*

Thus, while implementing Algorithm 2 and 3, we just need to **check the first condition** to stop sampling.

Component 1: Stopping Rule. During exploration, once we gather enough statistical information about the parameters in the system, the test statistic crosses the stopping threshold with the chosen confidence δ, and we stop to recommend the optimal policy.

Theorem 5. *The Chernoff stopping rule to ensure $(1-\delta)$-correctness and $(1-\delta)$-feasibility is*

$$\max_{\boldsymbol{\pi} \in \Pi^r_{\hat{\mathcal{F}}_t}} \inf_{\boldsymbol{\lambda} \in \Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t, \boldsymbol{\pi})} \sum_{a=1}^{K} N_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\lambda}_a) > \beta(t,\delta)\,,$$

where $\beta(t,\delta) \triangleq 3S_0 \log(1 + \log N_{a,t}) + S_0 \mathcal{T}\left(\frac{(K \wedge d) + \log\frac{1}{\delta}}{S_0}\right)$ *and* $0 \leq S_0 \leq K$.

Component 2: Recommendation rule. Once the stopping rule is fired, the agent recommends a policy based on the current estimate of $\hat{\boldsymbol{\mu}}_t$ according the rule $\hat{\boldsymbol{\pi}} = \arg\max_{\boldsymbol{\pi} \in \Pi^r_{\hat{\mathcal{F}}_{\tau_\delta}}} \hat{\boldsymbol{\mu}}_t^\top \boldsymbol{\pi}$.

Component 3: Sampling Strategy. We present two novel sampling algorithms: LATS and LAGEX.

LATS. The algorithm LATS (Algorithm 2) uses the Track and Stop strategy adapted to the unknown constraint setting.

The algorithm warms up the parameter estimates by playing each arm once. Until the stopping rule is fired, it computes the optimal allocation (Line 4) under the estimated feasible space by solving the Lagrangian relaxed optimization problem in Proposition 1. By plugging the optimal allocation we perform a bounded optimisation on the Lagrangian multiplier using Theorem 1. The algorithm further uses C-tracking (Garivier and Kaufmann, 2016) to track actions taken per step. Finally, we observe the instantaneous reward and cost feedback to update the parameter estimates (Line 9).

Algorithm 2 LATS - **LA**grangian **T**rack and **S**top

1: **Input :** Tolerance $r > 0$, Confidence level $\delta > 0$
2: **Initialization :** $\hat{A}_0 = \mathbf{0}_{d \times K}$, $\hat{\boldsymbol{\mu}}_0 = \mathbf{0}_K$, $\Sigma_0 = v\mathbf{I}_K$, l_0
3: Play each arm once to set $\boldsymbol{\mu}_1$ and $\hat{A}_1$.
4: **while** $\beta(t-1,\delta) > \mathcal{D}(\boldsymbol{\omega}^\star_{t-1}, \hat{\boldsymbol{\mu}}_{t-1}, \hat{\mathcal{F}}_{t-1}, \boldsymbol{l}^\star_{t-1})$ **do**
5: $\boldsymbol{\pi}_t = \arg\max_{\boldsymbol{\pi} \in \Pi^r_{\hat{\mathcal{F}}_t}} \hat{\boldsymbol{\mu}}_{t-1}^\top \boldsymbol{\pi}$
6: **Optimal allocation:** $\boldsymbol{\omega}^\star_t \in \arg\max_{\boldsymbol{\omega} \in \hat{\mathcal{F}}_t} \mathcal{D}(\boldsymbol{\omega}_{t-1}, \hat{\boldsymbol{\mu}}_{t-1}, \boldsymbol{\pi}_t, \hat{\mathcal{F}}_{t-1}, \boldsymbol{l}^\star_{t-1})$
7: **Optimize Lagrangian Multiplier :** $\boldsymbol{l}^\star_t \in \arg\min_{\boldsymbol{l} \in \mathcal{L}_t} \mathcal{D}(\boldsymbol{\omega}^\star_t, \hat{\boldsymbol{\mu}}_{t-1}, \boldsymbol{\pi}_{t-1}, \hat{\mathcal{F}}_{t-1}, \boldsymbol{l})$
8: **C-Tracking:** Play $a_t \in \arg\min_{a \in [1,K]} N_{a,t-1} - \sum_{s=1}^t \boldsymbol{\omega}^\star_{a,s}$
9: **Feedback :** Observe reward r_t and cost $\mathbf{c}_t$, and update Σ_t, $\hat{\boldsymbol{\mu}}_t$ and $\hat{A}_t$
10: **end while**
11: **Recommended policy:** $\hat{\boldsymbol{\pi}} = \arg\max_{\boldsymbol{\pi} \in \Pi^r_{\hat{\mathcal{F}}_t}} \hat{\boldsymbol{\mu}}_t^\top \boldsymbol{\pi}$

Theorem 6. *Let $\mathfrak{s}$ be the shadow price $\mathfrak{s} \triangleq \frac{\Gamma_{\max}}{\Gamma_{\min}} = \frac{\max_{i\in[1,d]}(-A\boldsymbol{\pi}^\star_{\mathcal{F}})}{\min_{i\in[1,d]}(-A\boldsymbol{\pi}^\star_{\mathcal{F}})}$ of the slack Γ. Under Assumption 1 and 2, the expected stopping time of LATS satisfies*

$$\lim_{\delta\to 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq \alpha T_{\mathcal{F},r}(\boldsymbol{\mu})(1+\mathfrak{s})$$

for any $\alpha > 1$.

Implications. Theorem 6 suggests that LATS (Algorithm 2) is asymptotically optimal up to problem dependent constants. This upper bound captures the effect of unknown linear constraints in shadow price, novel in the constrained bandit literature. If any one of the constraints in the bandit environment becomes more and more sensitive, minimum slack decreases, consequently $\mathfrak{s}$ increases and the identification problem becomes harder to solve. Thus, $\mathfrak{s}$ behaves as a *sensitivity or stability parameter* that explains the hardness of the pure exploration problem in hand through the structure of the true constraints. **LAGEX.** Algorithms based on track and stop mechanism tend to fail in case of larger problems where efficient optimization becomes a challenge due to the use of a max-min oracle per step. To improve on this, we leverage the two-player zero sum game approach introduced in Degenne et al. (2019b). Algorithm 3 also starts by playing each arm once to warm up parameter estimates. Then it uses a **allocation player** (We have

Algorithm 3 LAGEX- **LA**grangian **G**amified **EX**plorer

1: **Input :** Tolerance $r > 0$, Confidence level $\delta > 0$
2: Play each arm once to set $\boldsymbol{\mu}_1$ and $\hat{A}_1$.
3: **while** $\beta(t-1, \delta) > \mathcal{D}(\boldsymbol{\omega}^\star_{t-1}, \hat{\boldsymbol{\mu}}_{t-1}, \hat{\mathcal{F}}_{t-1}, \boldsymbol{l}^\star_{t-1})$ **do**
4: $\quad \boldsymbol{\pi}_t = \arg\max_{\boldsymbol{\pi} \in \Pi^r_{\hat{\mathcal{F}}_t}} \hat{\boldsymbol{\mu}}_{t-1}^\top \boldsymbol{\pi}$
5: $\quad$ **Optimal allocation** $\boldsymbol{\omega}^\star_t \rightsquigarrow$ Using AdaGrad via Theorem 4
6: $\quad$ **Optimize Lagrangian Multiplier:** $\boldsymbol{l}^\star_t \in \arg\min_{\boldsymbol{l} \in \mathcal{L}_t} \mathcal{D}(\boldsymbol{\omega}_{t-1}, \hat{\boldsymbol{\mu}}_{t-1}, \boldsymbol{\pi}_t, \hat{\mathcal{F}}_{t-1}, \boldsymbol{l})$
7: $\quad$ **C-Tracking:** Play $a_t \in \arg\min_{a \in [1,K]} N_{a,t-1} - \sum_{s=1}^t \boldsymbol{\omega}^\star_{a,s}$
8: $\quad$ **Feedback :** Observe reward r_t and cost $\mathbf{c}_t$, and update Σ_t, $\hat{\boldsymbol{\mu}}_t$ and $\hat{A}_t$
9: $\quad$ **Compute confusing instance :** $\boldsymbol{\lambda}_t \rightsquigarrow$ Via Proposition 1 plugging in $\boldsymbol{\omega}^\star_t$, $\boldsymbol{l}^\star_t$
10: $\quad$ **Confidence intervals:** for all $a \in [K]$
$\quad [\alpha_{t,a}, \beta_{t,a}] : \{\zeta : N_{a,t} d(\hat{\boldsymbol{\mu}}_{t,a}, \zeta) \leq g(t)\}$
$\quad U^a_t = \max\left\{\frac{g(t)}{N_{a,t}}, d(\alpha_{t,a}, \boldsymbol{\lambda}_{t,a}), d(\beta_{t,a}, \boldsymbol{\lambda}_{t,a})\right\}$
11: $\quad$ **Update loss for regret minimizer:** Update with $L_t = \langle \boldsymbol{\omega}^\star_t, U_t \rangle - \boldsymbol{l}^{\star\top}_t \tilde{A}_t \boldsymbol{\omega}^\star_t$
12: **end while**
13: **Recommended policy:** $\hat{\boldsymbol{\pi}} = \arg\max_{\boldsymbol{\pi} \in \Pi^r_{\hat{\mathcal{F}}_t}} \hat{\boldsymbol{\mu}}_t^\top \boldsymbol{\pi}$

used AdaGrad) to optimize the allocation $\boldsymbol{\omega}_t$ in Line 5 against the most confusing instance w.r.t current estimate of $\boldsymbol{\mu}$ optimised by a **instance player** which minimizes $\sum_{a=1}^K \boldsymbol{\omega}_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\lambda}_a) - \boldsymbol{l}^{\star\top}_t \tilde{\mathbf{A}}_t \boldsymbol{\omega}_t$ with respect to $\boldsymbol{\lambda} \in \Lambda_{\hat{\mathcal{F}}_t}(\boldsymbol{\mu}, \boldsymbol{\pi})$. Since our search space in closed and convex, the allocation player enjoys sub-linear regret of $\mathcal{O}(\sqrt{t \log t})$, whereas the instance player computes the best confusing instance using Proposition 1. Then in Line 11, Adagrad loss function is updated with a loss by introducing optimism as U_t defined in Line 10. Rest of the mechanism goes as usual

Theorem 7 (Asymptotic Optimality of LAGEX)**.** *Under Assumption 1 and 2, the expected sample complexity of LAGEX satisfies*

$$\lim_{\delta \to 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq T_{\mathcal{F},r}(\boldsymbol{\mu}).$$

Detailed proofs of this section are in Appendix G.

5 EXPERIMENTAL ANALYSIS

Now, we empirically test performance of proposed algorithms and the baselines. We refer to Appendix I for additional experiments and other details necessary for reproducibility. Code is available at this Link.[1]

Synthetic Data: Setup 1 We evaluate with two environments having means $[1.5, 1.0, \mu_3, 0.4, 0.3, 0.2, 0.1]$.

Observation 1: Universality. We vary μ_3 from 0.5 to 2.5. For each environment, we plot the corresponding unconstrained BAI lower bounds (in red) and lower bounds under constraints (in blue) in Figure 4. We observe that the constraint problem gets easier with increasing μ_3. In contrast, the BAI problem changes non-monotonically. BAI problem gets harder when μ_3 is around 1.5 as the suboptimality gap gets very small. But the constraint problem stays easier than BAI. In Figure 4, we also plot the median sample complexity of LAGEX across these environments over 500 runs. LAGEX grows parallel to the lower bound under constraints and can track it across environments.

Observation 2: Efficiency. We run all algorithms in two environment: (i) **hard** with $\mu_3 = 0.5$ and (ii) **easy** with $\mu_3 = 1.3$. We call the first environment hard as it is harder than solving BAI and similarly, the second environment easy. In Figure 5 and 6 we observe that (i) among the algorithms with unknown constraints LAGEX incur the least sample complexity, and (ii) we pay a minimal cost than the known constraint Lagrangian algorithms in **hard env** whereas the price of estimating constraints is prominent in **easy env**.

Real Data: IMDB-50K Dataset We evaluate LATS, LAGEX, and the baselines on IMDB 50K dataset (Maas et al., 2011). For ease of comparison, we use the same bandit environment as in (Carlsson et al., 2024). We use 12 movies. We search for the optimal policy which allocates weight at most 0.3 to action movies and at least 0.3 to family and drama movies.

Observation: LAGEX has better sample complexity. From Figure 7, we observe that LAGEX performs better than other algorithms in the unknown constraints setting. LATS also performs well on the IMDB environment but notably we cannot distinguish its performance from that of the Uniform explorer.

6 DISCUSSION & FUTURE WORK

We study pure exploration under unknown linear constraints. We encode the coupled effect of estimating both mean vector and constraints via a Lagrangian relaxation of the lower bound for known constraints. We further design an optimistic estimate of the feasible set to ensure identification of the r-optimal feasible policy- leading to two algorithms, LATS and LAGEX. We prove their sample complexity upper bounds and conduct numerical experiments to observe that LAGEX

[1] **Baselines:** “CTnS-WLag” and “CGE-WLag” is the CTnS and CGE algorithm of Carlsson et al. (2024) under unknown constraints without Lagrangian relaxation. “PTnS” is Projected Track and Stop algorithm.

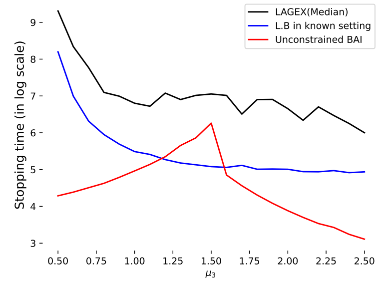

Figure 4: Lower bounds with and without constraints, and LAGEX for $\mu_3 \in [0.5, 2.5]$ in Setup 1.

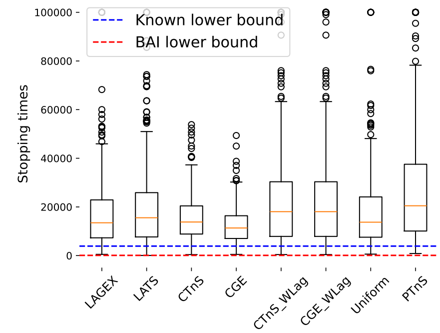

Figure 5: Sample complexity (median±std.) of algorithms for **hard env.** in Setup 1.

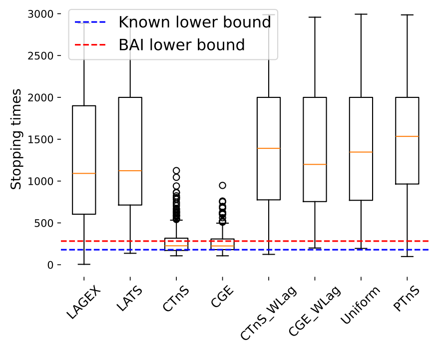

Figure 6: Sample complexity (median±std.) of algorithms for **easy env.** in Setup 1.

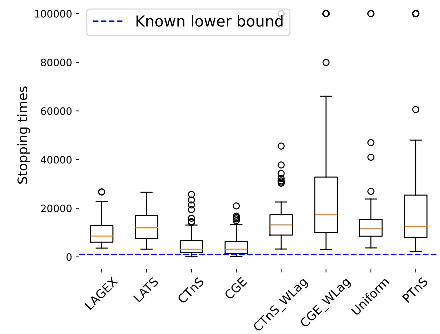

Figure 7: Sample complexity (median±std.) of algorithms for **IMDB-50K**.

is the most efficient among baselines.

From the lower bound perspective, one might be interested to derive the lower bounds when both constraints and rewards are considered as feedback. Algorithmically, it would be intriguing to extend our Lagrangian technique to non-linear constraints.

Acknowledgments

We acknowledge the ANR JCJC project REPUBLIC (ANR-22-CE23-0003-01), the PEPR project FOUNDRY (ANR23-PEIA-0003), and the Inria-ISI Kolkata associate team SeRAI for supporting the project. We also thank Achraf Azize for interesting discussions.

References

Abbasi-yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Agrawal, S. and Devanur, N. (2016). Linear contextual bandits with knapsacks. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Agrawal, S. and Devanur, N. R. (2014). Bandits with concave rewards and convex knapsacks. EC '14, page 989–1006, New York, NY, USA. Association for Computing Machinery.

Agrawal, S., Devanur, N. R., and Li, L. (2016). An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In Feldman, V., Rakhlin, A., and Shamir, O., editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 4–18, Columbia University, New York, New York, USA. PMLR.

Amani, S., Alizadeh, M., and Thrampoulidis, C. (2019). Linear stochastic bandits under safety constraints.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

Aziz, M., Kaufmann, E., and Riviere, M.-K. (2021a). On multi-armed bandit designs for dose-finding clinical trials. *The Journal of Machine Learning Research*, 22(1):686–723.

Aziz, M., Kaufmann, E., and Riviere, M.-K. (2021b).

On multi-armed bandit designs for dose-finding trials. *Journal of Machine Learning Research*, 22(14):1–38.

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. (2018). Bandits with knapsacks. *J. ACM*, 65(3).

Baudry, D., Merlis, N., Molina, M. B., Richard, H., and Perchet, V. (2024). Multi-armed bandits with guaranteed revenue per arm. In *International Conference on Artificial Intelligence and Statistics*, pages 379–387. PMLR.

Bechhofer, R. E. and Blumenthal, S. (1962). A sequential multiple-decision procedure for selecting the best one of several normal populations with a common unknown variance, ii: Monte carlo sampling results and new computing formulae. *Biometrics*, 18(1):52–67.

Berge, C. (1963). *Topological Spaces: Including a Treatment of Multi-valued Functions, Vector Spaces and Convexity*. Macmillan.

Bernasconi, M., Castiglioni, M., and Celli, A. (2024). No-regret is not enough! bandits with general constraints through adaptive regret minimization.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

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.

Bubeck, S., Munos, R., and Stoltz, G. (2010). Pure exploration for multi-armed bandit problems.

Camilleri, R., Wagenmaker, A., Morgenstern, J. H., Jain, L., and Jamieson, K. G. (2022a). Active learning with safety constraints. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 33201–33214. Curran Associates, Inc.

Camilleri, R., Wagenmaker, A., Morgenstern, J. H., Jain, L., and Jamieson, K. G. (2022b). Active learning with safety constraints. *Advances in Neural Information Processing Systems*, 35:33201–33214.

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.

Chen, S., Lin, T., King, I., Lyu, M. R., and Chen, W. (2014). Combinatorial pure exploration of multi-armed bandits. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Chen, T., Gangrade, A., and Saligrama, V. (2022a). Doubly-optimistic play for safe linear bandits. *arXiv preprint arXiv:2209.13694*.

Chen, T., Gangrade, A., and Saligrama, V. (2022b). Strategies for safe multi-armed bandits with logarithmic regret and risk. In *International Conference on Machine Learning*, pages 3123–3148. PMLR.

Cheshire, J., Menard, P., and Carpentier, A. (2021). The influence of shape constraints on the thresholding bandit problem.

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

Degenne, R., Koolen, W. M., and Ménard, P. (2019a). Non-asymptotic pure exploration by solving games. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Degenne, R., Koolen, W. M., and Ménard, P. (2019b). Non-asymptotic pure exploration by solving games.

Even-Dar, E., Mannor, S., and Mansour, Y. (2002a). Pac bounds for multi-armed bandit and markov decision processes. In *Annual Conference Computational Learning Theory*.

Even-Dar, E., Mannor, S., and Mansour, Y. (2002b). Pac bounds for multi-armed bandit and markov decision processes. In *Computational Learning Theory: 15th Annual Conference on Computational Learning Theory, COLT 2002 Sydney, Australia, July 8–10, 2002 Proceedings 15*, pages 255–270. Springer.

Faizal, F. Z. and Nair, J. (2022). Constrained pure exploration multi-armed bandits with a fixed budget. *arXiv preprint arXiv:2211.14768*.

Fiez, T., Jain, L., Jamieson, K. G., and Ratliff, L. (2019). Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems*, 32.

Fonseca, C. M. and Fleming, P. J. (1998). Multiobjective optimization and multiple constraint handling with evolutionary algorithms. i. a unified formulation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 28(1):26–37.

Gangrade, A., Gopalan, A., Saligrama, V., and Scott, C. (2024). Testing the feasibility of linear programs with bandit feedback.

Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In Feldman, V., Rakhlin, A., and Shamir, O., editors, *29th*

Annual Conference on Learning Theory, volume 49 of *Proceedings of Machine Learning Research*, pages 998–1027, Columbia University, New York, New York, USA. PMLR.

Garivier, A. and Kaufmann, E. (2021). Non-asymptotic sequential tests for overlapping hypotheses and application to near optimal arm identification in bandit models.

Garivier, A., Ménard, P., Rossi, L., and Menard, P. (2018). Thresholding bandit for dose-ranging: The impact of monotonicity.

Hutchinson, S., Turan, B., and Alizadeh, M. (2024). Directional optimism for safe linear bandits.

Immorlica, N., Sankararaman, K., Schapire, R., and Slivkins, A. (2022). Adversarial bandits with knapsacks. *J. ACM*, 69(6).

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.

Jourdan, M. and Degenne, R. (2022). Choosing answers in epsilon-best-answer identification for linear bandits. In *International Conference on Machine Learning*, pages 10384–10430. PMLR.

Jourdan, M., Degenne, R., and Kaufmann, E. (2023). An ε-best-arm identification algorithm for fixed-confidence and beyond. *Advances in Neural Information Processing Systems*, 36:16578–16649.

Jourdan, M., Mutný, M., Kirschner, J., and Krause, A. (2021). Efficient pure exploration for combinatorial bandits with semi-bandit feedback. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132.

Katz-Samuels, J. and Scott, C. (2018). Feasible arm identification. In *International Conference on Machine Learning*, pages 2535–2543. PMLR.

Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the complexity of best arm identification in multi-armed bandit models.

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.

Kazerouni, A., Ghavamzadeh, M., Abbasi-Yadkori, Y., and Roy, B. V. (2017). Conservative contextual linear bandits.

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

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. (2017). Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1):6765–6816.

Li, S., Zhang, L., Yu, Y., and Li, X. (2023). Optimal arms identification with knapsacks. In *International Conference on Machine Learning*, pages 20529–20555. PMLR.

Li, X., Sun, C., and Ye, Y. (2021). The symmetry between arms and knapsacks: A primal-dual approach for bandits with knapsacks. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6483–6492. PMLR.

Libin, P. J., Verstraeten, T., Roijers, D. M., Grujic, J., Theys, K., Lemey, P., and Nowé, A. (2019). Bayesian best-arm identification for selecting influenza mitigation strategies. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part III 18*, pages 456–471. Springer.

Lindner, D., Tschiatschek, S., Hofmann, K., and Krause, A. (2022). Interactively learning preference constraints in linear bandits. In *International Conference on Machine Learning*, pages 13505–13527. PMLR.

Lindståhl, S., Proutiere, A., and Johnsson, A. (2022). Measurement-based admission control in sliced networks: A best arm identification approach. In *GLOBECOM 2022-2022 IEEE Global Communications Conference*, pages 1484–1490. IEEE.

Liu, X., Li, B., Shi, P., and Ying, L. (2021). An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints.

Ma, W. (2014). *Improvements and Generalizations of Stochastic Knapsack and Multi-Armed Bandit Approximation Algorithms: Extended Abstract*, pages 1154–1163.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In Lin, D., Matsumoto, Y., and Mihalcea, R., editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Magureanu, S., Combes, R., and Proutiere, A. (2014). Lipschitz bandits: Regret lower bounds and optimal algorithms.

Mason, B., Jain, L., Tripathy, A., and Nowak, R. (2020). Finding all ε-good arms in stochastic bandits. *Advances in Neural Information Processing Systems*, 33:20707–20718.

Moradipari, A., Amani, S., Alizadeh, M., and Thrampoulidis, C. (2020). Safe linear thompson sampling with side information.

Pacchiano, A., Ghavamzadeh, M., and Bartlett, P. (2024). Contextual bandits with stage-wise constraints. *arXiv preprint arXiv:2401.08016*.

Pacchiano, A., Ghavamzadeh, M., Bartlett, P., and Jiang, H. (2020). Stochastic bandits with linear constraints.

Paulson, E. (1964). A sequential procedure for selecting the population with the largest mean from k normal populations. *The Annals of Mathematical Statistics*, 35(1):174–180.

Sankararaman, K. A. and Slivkins, A. (2018). Combinatorial semi-bandits with knapsacks. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1760–1770. PMLR.

Shang, X., Colin, I., Barlier, M., and Cherkaoui, H. (2023). Price of safety in linear best arm identification.

Singh, A. and Joachims, T. (2019). Policy learning for fairness in ranking. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Slivkins, A., Sankararaman, K. A., and Foster, D. J. (2023). Contextual bandits with packing and covering constraints: A modular lagrangian approach via regression. In Neu, G. and Rosasco, L., editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 4633–4656. PMLR.

Slutsky, E. (1925). Über stochastische Asymptoten und Grenzwerte. Metron 5, Nr. 3, 3-89 (1925).

Tirinzoni, A., Pirotta, M., Restelli, M., and Lazaric, A. (2020). An asymptotically optimal primal-dual incremental algorithm for contextual linear bandits. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1417–1427. Curran Associates, Inc.

Tran-Thanh, L., Chapman, A., Rogers, A., and Jennings, N. R. (2012). Knapsack based optimal policies for budget-limited multi-armed bandits. In *Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-12) (22/07/12 - 22/07/12)*, pages 1134–1140.

Wang, L., Bai, Y., Sun, W., and Joachims, T. (2021a). Fairness of exposure in stochastic bandits. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10686–10696. PMLR.

Wang, Z., Wagenmaker, A., and Jamieson, K. (2021b). Best arm identification with safety constraints.

Wu, Y., Zheng, Z., and Zhu, T. (2023). Best arm identification with fairness constraints on subpopulations. In *2023 Winter Simulation Conference (WSC)*, pages 540–551. IEEE.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] (Added as a link in Section 5)
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [Not Applicable]
(b) The license information of the assets, if applicable. [Not Applicable]
(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
(d) Information about consent from data providers/curators. [Not Applicable]
(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]
(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Notations

Notation	Definition
Δ_K	K-simplex
K	Number of Arms
A	True constraint set
d	Number of constraints
$\mathcal{F}$	True feasible set w.r.t A, $\mathcal{F} = \{A \in \mathbb{R}^{d\times K} : A\boldsymbol{\pi} \leq 0\}$
$\tilde{A}_t$	Optimistic estimate of constraint set at time t, $\tilde{A}_t = \hat{A} - f(t,\delta)\|\|\boldsymbol{\omega}_t\|\|_{\Sigma_t^{-1}}$
$\hat{\mathcal{F}}_t$	Estimated feasible set w.r.t pessimistic estimate $\tilde{A}$ at time t, $\mathcal{F} = \{\tilde{A}_t \in \mathbb{R}^{d\times K} : \tilde{A}_t\boldsymbol{\pi} \leq 0\}$
$\mathcal{A}$	The action set of K possible choices
$\boldsymbol{\omega}_t$	Allocation chosen at time t
a_t	Action at time t among K possible actions
N	Number of Constraints
Γ	Slack of the optimisation problem, i.e., $\Gamma \triangleq \min_{i\in[d]}(-A\boldsymbol{\pi}^\star_{\mathcal{F}})$
σ^2	Variance of the reward distribution (Gaussian) of arms
$r_t, \mathbf{c}_t$	Reward and cost observed at time t
δ	Chosen confidence level
$\boldsymbol{l}_t$	The Lagrangian multiplier at time t
Σ_t	The covariance matrix (Gram matrix) at round t
$\Lambda_{\mathcal{F}}(\boldsymbol{\mu})$	Set of alternative (confusing) instances for bandit instance $\boldsymbol{\mu}$
$\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})$	Estimated set of alternative (confusing) instances for bandit instance $\boldsymbol{\mu}$
$\nu(\boldsymbol{\pi}^\star)$	Neighboring set of optimal r-good feasible policy $\boldsymbol{\pi}^\star$
$\nu(\hat{\boldsymbol{\pi}}^\star)$	Neighboring set of estimated optimal r-good feasible policy $\hat{\boldsymbol{\pi}}^\star$
τ_δ	Stopping time of $(1-\delta)$-correct algorithm
$\boldsymbol{\pi}^\star_{\mathcal{F}}$	True optimal policy w.r.t actual constraint set A
$\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$	Optimal Policy for the estimated feasible set
$\boldsymbol{\pi}$	Optimal r-good feasible policy with respect to $\boldsymbol{\pi}^\star_{\mathcal{F}}$
$\hat{\boldsymbol{\pi}}^\star$	Optimal r-good feasible policy with respect to $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$
$\mathfrak{s}$	Shadow price $\triangleq \frac{\Gamma_{\max}}{\Gamma_{\min}}$

Table 1: Summary of Notations

B Additional discussion on problem setting

B.1 Extended Related Work

Historical pioneering works. Literature on bandits has come a long way since the problem of optimal sequential sampling started with the works of Bechhofer and Blumenthal (1962) and Paulson (1964) with the assumption of the populations being normally distributed. To talk about pure exploration setting, Even-Dar et al. (2002b), Bubeck et al. (2010) should be mentioned as the first ones who worked in this specific setting for stochastic bandits.

Existing work on adapting known constraints. In Multi-armed bandit literature, people often introduce constraints as a notion of safety where they impose known constraints on the chosen arm or on the exploration process. Wang et al. (2021b) considers pure strategy (only one co-ordinate as chosen action) and imposes a safety threshold on the linear cost feedback of the chosen arm. On the other hand, the setting considered in Carlsson et al. (2024) is closer as it tracks an optimal policy w.r.t to a known set of known constraints. On the other hand, Liu et al. (2021) (Improvement over Pacchiano et al. (2020) in MAB setting) generalized the known constraint regret minimization setting by assuming existence of a set of general constraints. Our work captures the hardness of not knowing the constraint set while tracking the lower bound and also in sample complexity upper bounds of Algorithm 2 and 3. Our work also introduce *shadow price* as a novel term in pure exploration literature which characterises the extra cost that arises due to tracking the unknown constraints.

Learning unknown constraints. Lindner et al. (2022) considers constrained linear best-arm identification arm are vectors with known rewards and a single unknown constraint (representing preferences) on the actions. Works on adapting to unknown constraints is discussed in the related work section of the main paper.

Transductive Linear Bandit. In this setting formalised by Fiez et al. (2019), Camilleri et al. (2022b) studies this setting with unknown linear constraints where we have to find the best safe arm in a finite set $\mathcal{Z}$ different than actual arm set $\mathcal{A}$. Our setting generalises the setting in the sense that the finite feasible set $\mathcal{Z}$ is not static, rather we track $\mathcal{Z}_t$ per time step $t \in \mathbb{N}$ and explore within that set to find the optimal allocation. At the end of exploration after hitting the stopping criterion at τ_δ the agent recommends the optimal policy inside the set $\mathcal{Z}_{\tau_\delta}$.

Regret Minimization with Unknown constraints. In bandit literature, constraints are often introduced in the setting to study regret minimization. Moradipari et al. (2020) studies regret minimization using Linear Thompson Sampling (LTS) imposing known safety constraints on the chosen action. Amani et al. (2019) studies contextual bandits under unknown and unobserved linear constraint, whereas Kazerouni et al. (2017) Pacchiano et al. (2020) studies UCB based algorithms for regret minimization for linear bandits which assumes existence of a safe action space in case of unknown anytime linear constraint. In line with these, recent works Hutchinson et al. (2024) Pacchiano et al. (2024) Shang et al. (2023) improved on regret guarantees and the first one relaxed the assumption of existence of a pessimistic safe space. Chen et al. (2022a) introduces doubly-optimistic setting to study safe linear bandit(SLB). Liu et al. (2021) generalised the setting of Pacchiano et al. (2020) not only relaxing the condition of having a safe action but also considered a set of general constraints and also captured the notion of both anytime and end-of-time constraints which we also see in Carlsson et al. (2024). Liu et al. (2021) also shows the trade-off between maximising reward or minimizing regret and constraint violation using Lyapunov drift. In this work we do not focus on regret guaranties but finding the optimal policy with sample complexity as least as possible while tracking and satisfying a set of unknown linear constraints.

BAI with Fairness Constraint. Considering fairness constraint in our setting can be an interesting application to our setting. Recently Wu et al. (2023) studied Best Arm Identification with fairness Constraints on Subpopulations (BAICS), where they have discussed the trade-off in the standard BAI complexity if there are finite number of subpopulations are given and the best chosen arm must perform well (not too bad) on all those subpopulations. Another important line of work Wang et al. (2021a) Singh and Joachims (2019) explores regret analysis of BAI with positive merit-based exposure of fairness constraints where the chosen policy has to satisfy some fairness constraint across all its indices. Our setting comes as a direct application to these settings. Further discussion in Section B.2.

BAI with Knapsack constraint. While the existing literature on bandit with knapsack Badanidiyuru et al. (2018) Agrawal and Devanur (2016) Immorlica et al. (2022), Agrawal and Devanur (2014) Agrawal et al. (2016) Sankararaman and Slivkins (2018) Ma (2014) focused on mainly regret minimization, our setting aligns more as a special case of the Optimal Arm identification with Knapsack setting in Li et al. (2023, 2021); Tran-Thanh et al. (2012). Though we aim to find the best policy rather than a specific arm in the constraint space. Our setting should be considered as a special case of these settings. Further discussion in Section B.2.

Algorithms on Pure exploration. Algorithm 2 is an extension of the Track and Stop(TnS) strategy from (Garivier and Kaufmann, 2016), while the motivation for Algorithm 3 comes from the Gamified Explorer strategy from (Degenne et al., 2019b) where the lower bound is treated as a zero-sum game between the allocation and the instance player. We refer to (Garivier and Kaufmann, 2021),(Garivier and Kaufmann, 2016),(Kaufmann et al., 2016),(Degenne and Koolen, 2019), Boyd and Vandenberghe (2004),Jourdan et al. (2021) etc for important concentration inequalities, tracking lemmas.

Dose-finding and Thresholding Bandits. Another special case of our setting is Dose-finding or Thresholding bandits in structured MAB literature Chen et al. (2014) generalized the problem, then a line of work Aziz et al. (2021b) Garivier et al. (2018) Cheshire et al. (2021) aims to find the maximum safe dose for a specific drug in early stages of clinical trials. In some sense our setting generalizes this setting. If we have to administer more than drugs to a patient, our setting generalises to track the best possible proportion in which the drugs should be administered with maximum efficacy. Further discussions in Section B.2.

B.2 Motivations: Reductions to and Generalisations of The Existing BAI Settings

Before delving into the details of the lower bounds and algorithms, we first clarify our motivation by showing how different setups studied in literature and their variations are special case of our setting.

Thresholding Bandits. Our setting encompasses the thresholding bandit problem (Aziz et al., 2021a). Thresholding bandit is motivated from the safe dose finding problem in clinical trials, where one wants to identify the highest dose of a drug that is below a known safety level. This has also motivated the studies on safe arm identification (Wang et al., 2021b). Our setting generalises it further to detect the dose of the drug with highest efficacy while it is still below the safety level. We can formulate it as identifying $\boldsymbol{\pi}^\star = \arg\max \boldsymbol{\mu}^\top \boldsymbol{\pi}$, such that $I\boldsymbol{\pi} \leq I\boldsymbol{\theta}$. Rather, generalising the classical thresholding bandits, our formulation can further model the safe doses for the optimal cocktail of drugs, and θ can have different values across drugs, i.e we can consider different thresholds for different drugs.

Optimal policy under Knapsack. Bandits under knapsack constraints have been studied both in best-arm identification (Li et al., 2023, 2021; Tran-Thanh et al., 2012) and regret minimisation (Agrawal and Devanur, 2016, 2014; Agrawal et al., 2016; Badanidiyuru et al., 2018; Immorlica et al., 2022; Ma, 2014; Sankararaman and Slivkins, 2018) literature. BAI under knapsacks is motivated by the fact that detecting an optimal arm might have additional resource constraints in addition to the number of required samples. This has led to study of BAI with knapsacks only under fixed-budget settings (Li et al., 2023). But as in regret-minimisation literature (Ma, 2014; Sankararaman and Slivkins, 2018), one might want to recommend a policy that maximises utility while satisfying knapsack constraints. For example, we want to manage caches where the recommended memory allocation should satisfy a certain resource budget. Thus, the recommended policy has to satisfy $\boldsymbol{\pi}^\star_\tau = \arg\max_{\boldsymbol{\pi}\in C_A} \hat{\boldsymbol{\mu}}_{\tau_\delta}^\top \boldsymbol{\pi}$, where $C_A \triangleq \{A\boldsymbol{\pi}_{\tau_\delta} \leq c\}$. Naturally, this is a special case of our problem setting.

Feasible arm selection. We look at the pure exploration problem of feasible arm selection studied by Katz-Samuels and Scott (2018). Here, we think of a problem of workers having a multi-dimension vector representation where each index denotes the accuracy of that worker being able to identify a specific class label in a classification task in hand. The problem turns to be a feasible arm selection from a simple BAI problem when we impose a feasibility constraint that for example, the chosen worker should show more than 90% accuracy across all labels. We can generalise this setting in the sense that we are now not looking for a specific worker, rather we want to make a team of workers that has the highest utility. The recommended policy at time $t \in \mathbb{N}$, $\max_{\boldsymbol{\pi}\in\Delta_{K-1}} \boldsymbol{\mu}_t^\top \boldsymbol{\pi}$ such that $f^\top \boldsymbol{\pi} \geq \tau$ where τ is the desired threshold level. The generalisation of the setting pitch in as thresholds of τ can have different values corresponding to different workers.

BAI with fairness across sub-populations. The Best Arm Identification with fairness Constraints on Sub-population (BAICS) studied in Wu et al. (2023) aims on selecting an arm that must be fair across all sub-populations rather than the whole population in standard BAI setting. Let, there are $\boldsymbol{l}$ sub-populations and μ_a are the means corresponding to the a-th arm. Finding only the optimal arm $K_{\text{BAI}} = \arg\max_{k\in[K]} \boldsymbol{\mu}_k$ may not be enough because it may not perform equally good for all the l sub-populations. Then the arm should belong to a set $C := \{k \in [K] | \boldsymbol{\mu}_{k,m} \geq 0, m \in [l]\}$ where the observation for arm k and population m comes from $\mathcal{N}(\mu_{k,m}, 1)$ It ensures that the chosen arm does not perform *too bad* for any sub-population. Let us think of a problem where there are l sub-groups of patients and we have K number of drugs to administer with reward means $\mu_k, k \in [K]$. We are looking for a combination of drugs rather than a single drug to administer as $\boldsymbol{\pi}^\star = \arg\max_{\pi\in\Delta_K} \boldsymbol{\mu}^\top \pi$ such that $\mathbb{1}^\top_{\boldsymbol{\mu}_m \geq 0}\boldsymbol{\pi} = 1, \forall m \in [l]$. Thus, BAICS is a special case of ours.

Fairness of exposure in bandits. Wang et al. (2021a) introduced positive merit based exposure of fairness constraints (Singh and Joachims, 2019) in stochastic bandits standing against the winner-takes-all allocation strategy that are historically studied. The chosen allocation in this setting should satisfy the fairness constraint $\frac{\boldsymbol{\pi}^\star_a}{f(\mu^\star_a)} = \frac{\boldsymbol{\pi}^\star_{a'}}{f(\mu^\star_{a'})}, \forall a' \in [K]$ where $f(.)$ transform reward of an arm to a positive merit. Though Wang et al. (2021a) studied this setting in regret analysis, this setting in BAI setting is a direct application of our setting as we are looking for an optimal policy $\boldsymbol{\pi}^\star = \arg\max \boldsymbol{\mu}^\top \boldsymbol{\pi}$ such that $\boldsymbol{\pi}^\star$ satisfies $A^\top_{f(\boldsymbol{\mu})}\boldsymbol{\pi} = 0$ where $A_{f(\boldsymbol{\mu})}$ is of order $\frac{K(K-1)}{2} \times K$ and $A_{f(\boldsymbol{\mu})}$ is expressed as,

$$(A_{f(\boldsymbol{\mu})})_{ij} = \begin{cases} \frac{1}{f(\mu_a)} & \text{if } aK - \frac{1}{2}(a-1)(a-2) \leq j \leq aK - \frac{1}{2}a(a-1) \text{ and } i = a\,, \\ \frac{-1}{f(\mu_j)} & \text{if } aK - \frac{1}{2}(a-1)(a-2) \leq j \leq aK - \frac{1}{2}a(a-1) \text{ and } a < i \leq K\,, \\ 0 & \text{otherwise.} \end{cases}$$

For example, when $K = 3$, $A_{f(\boldsymbol{\mu})} = [[\frac{1}{\mu_1}, -\frac{1}{\mu_2}, 0], [0, \frac{1}{\mu_2}, -\frac{1}{\mu_3}], [\frac{1}{\mu_1}, 0, -\frac{1}{\mu_3}]]$.

C Strong Duality and the Lagrangian Multiplier: Proof of Theorem 1

Theorem 1. *For a bounded sequence of $\{l_t\}_{t\in\mathbb{N}}$, strong-duality holds for the optimisation problem stated in Equation (5) i.e.*

$$\inf_{\boldsymbol{l}\in\mathbb{R}^d_+}\min_{A'\in\mathcal{C}}\sup_{\boldsymbol{\omega}\in\Delta_K}\max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}}\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})}\boldsymbol{\omega}^\top d(\boldsymbol{\mu},\boldsymbol{\lambda})-\boldsymbol{l}^\top A'\boldsymbol{\omega}$$
$$=\sup_{\boldsymbol{\omega}\in\Delta_K}\min_{\boldsymbol{l}\in\mathcal{L}}\max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}}\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})}\boldsymbol{\omega}^\top d(\boldsymbol{\mu},\boldsymbol{\lambda})-\boldsymbol{l}^\top\tilde{A}\boldsymbol{\omega}\,.$$

Here, $\mathcal{L}\triangleq\{\boldsymbol{l}\in\mathbb{R}^d\mid 0\le\|\boldsymbol{l}\|_1\le\frac{1}{\gamma}\mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\mathcal{F})\}$, where $\gamma\triangleq\min_{i\in[1,d]}\{-\tilde{A}^i\hat{\boldsymbol{\pi}}^\star\}$, i.e. the minimum slack for optimistic constraints w.r.t. the estimated optimal feasible policy.

Proof. This proof involves three steps. In the first step we prove convexity and other properties of the sets involved in the main optimisation problem 7. In the next step, we show that Slater's sufficient conditions hold for $\boldsymbol{\pi}$ as a consequence of these properties. Once we prove the unique optimality of $\boldsymbol{\pi}$ we state bounds on the L1-norm of the Lagrangian multiplier. We conclude by establishing strong duality and proving the statement of the theorem.

Step 1: Properties of perturbed feasible set and alt-set. Let us first check the properties of $\hat{\mathcal{F}}$, $\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})$ and $\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi})$. For that, let us remind the definitions of these sets. The estimated feasible set is defined as $\hat{\mathcal{F}}\triangleq\left\{\boldsymbol{\pi}\in\Delta_K:\tilde{A}\boldsymbol{\pi}\le 0\right\}$. The set of alternative (confusing) instances for the optimal policy $\hat{\boldsymbol{\pi}}^\star=\arg\max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}}\boldsymbol{\mu}^\top\boldsymbol{\pi}$ is $\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})\triangleq\left\{\boldsymbol{\lambda}\in\mathbb{D}:r+\max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}}\boldsymbol{\lambda}^\top\boldsymbol{\pi}>\boldsymbol{\lambda}^\top\hat{\boldsymbol{\pi}}^\star\right\}$. For $\boldsymbol{\pi}'$ being a neighbour of $\hat{\boldsymbol{\pi}}^\star$ or in other words, an extreme point in $\hat{\mathcal{F}}$, we decompose the alternative set as the union of half-spaces as,

$$\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})=\bigcup_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\hat{\boldsymbol{\pi}}^\star)}\left\{\boldsymbol{\lambda}:\boldsymbol{\lambda}^\top(\hat{\boldsymbol{\pi}}^\star-\boldsymbol{\pi}')<r\right\}$$

We should note, $\boldsymbol{\pi}'$ shares at least $(K-1)$ active constraints with $\hat{\boldsymbol{\pi}}^\star$. It is clear that $\hat{\mathcal{F}}$ is bounded and convex in $\boldsymbol{\pi}$. Since, convex combination of any two extreme point $\boldsymbol{\pi}'_1,\boldsymbol{\pi}'_2$ in the neighborhood of the optimal policy $\nu_{\hat{\mathcal{F}}}(\hat{\boldsymbol{\pi}}^\star)$ also shares $K-1$ active constraints with $\hat{\boldsymbol{\pi}}^\star$, so $\nu_{\hat{\mathcal{F}}}(\hat{\boldsymbol{\pi}}^\star)$ is convex in $\boldsymbol{\pi}'$.

Let, $\boldsymbol{\pi}'_1$ and $\boldsymbol{\pi}'_2$ are two policies in the neighborhood of $\hat{\boldsymbol{\pi}}^\star$, such that for any alternative instance $\boldsymbol{\lambda}$, $\boldsymbol{\lambda}^\top(\boldsymbol{\pi}'_1-\boldsymbol{\pi}'_2)\ge 0$, which implies that the policy $\boldsymbol{\pi}'_1$ is closer to the optimal policy in the neighborhood that the policy $\boldsymbol{\pi}'_2$ Therefore, any convex combination of these neighbourhood policy, $\boldsymbol{\lambda}^\top(\hat{\boldsymbol{\pi}}^\star-(a\boldsymbol{\pi}'_1+(1-a)\boldsymbol{\pi}'_2))=\boldsymbol{\lambda}^\top(\hat{\boldsymbol{\pi}}^\star-\boldsymbol{\pi}'_2)-a\boldsymbol{\lambda}^\top(\boldsymbol{\pi}'_1-\boldsymbol{\pi}'_2)\le c$. Therefore, the set $\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})$ is also bounded and convex in $\boldsymbol{\pi}$.
Also, since we are working with optimistic estimate of A, the set $\hat{\mathcal{F}}$ will always be non-empty, because we will find at least one $\tilde{\mathbf{A}}_0$ which is non-singular and it's inverse exists.

Step 2: Slater's condition. From step 1 of this proof we have the following properties

1. $\hat{\mathcal{F}}$ is non-empty, bounded and convex in $\boldsymbol{\pi}$.
2. The perturbed neighborhood $\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi})$ is convex for any $\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi})$
3. $\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})$ is also bounded and convex in $\boldsymbol{\pi}$.

Leveraging these three results we claim that there exists a $\hat{\boldsymbol{\pi}}^\star$ that uniquely solves the optimisation problem in Equation (7) and satisfy the constraints with strict inequality. Thus, we claim Slater's sufficient conditions hold for $\boldsymbol{\pi}$.

Step 3: Bound on the Lagrangian multiplier. Here, we try to bound the L-1 norm of the Lagrangian multiplier. Since, $\|\boldsymbol{l}\|_1$ cannot be less than 0, then we already have a lower bound.

Now we refer to lemma 5 for the upper bound. An immediate implication of this result is that for any dual optimal solution $\boldsymbol{l}^*$, we have $\|\boldsymbol{l}^*\|_1\le\frac{1}{\gamma}(f(\bar{x})-q^*)$. Since Slater's conditions hold in our case for $\boldsymbol{\pi}$, we can write that the optimal solution of the Lagrangian dual,

$$0 \leq ||l^\star||_1 \leq \frac{1}{\gamma}\mathcal{D}(\boldsymbol{\omega}^\star, \hat{\boldsymbol{\mu}}, \mathcal{F})$$
$$\text{where, } \gamma \triangleq \min_{i\in[1,d]}\{-\tilde{A}^i\hat{\boldsymbol{\pi}}^\star\}$$

Where, $\hat{\boldsymbol{\pi}}^\star$ is the r-optimal feasible policy.

Step 4: Establishing strong duality. Therefore the domain of the Lagrangian multiplier is also bounded and convex. So again we say that $l_t^\star$ uniquely minimises Equation 7. We define $\mathcal{L} \triangleq \{\boldsymbol{l} \in \mathbb{R}^d \mid 0 \leq \|\boldsymbol{l}\|_1 \leq \frac{1}{\gamma}T^{-1}_{\hat{\mathcal{F}},r}(\hat{\boldsymbol{\mu}})\}$, where $\gamma \triangleq \min_{i\in[1,d]}\{-\tilde{A}^i\boldsymbol{\omega}^*\}$ Then according to **Heine-Borel's theorem** (Theorem 9) we can say that these sets are compact as well. We can then conclude that Strong duality holds which means that it perfectly make sense of solving the Lagrangian dual formulation of the primal optimisation problem because there is no duality gap. We later on will consider this formulation as two player zero sum game. Due to strong duality we claim that the agent wile playing this game, Nash equilibrium will be eventually established.

Now that everything is put into place we can conclude with the very statement of the theorem that due to strong duality the following holds

$$\begin{aligned}
&\inf_{\boldsymbol{l}\in\mathbb{R}^d_+}\min_{A'\in\mathcal{C}}\sup_{\boldsymbol{\omega}\in\Delta_K}\max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}}\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})}\boldsymbol{\omega}^\top d\left(\boldsymbol{\mu},\boldsymbol{\lambda}\right) - \boldsymbol{l}^\top A'\boldsymbol{\omega} \\
&= \sup_{\boldsymbol{\omega}\in\Delta_K}\min_{\boldsymbol{l}\in\mathcal{L}}\max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}}\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})}\boldsymbol{\omega}^\top d\left(\boldsymbol{\mu},\boldsymbol{\lambda}\right) - \boldsymbol{l}^\top \tilde{A}\boldsymbol{\omega}\,.
\end{aligned}$$

□

D Lagrangian Relaxation of Projection Lemma: Proof of Theorem 2

Theorem 2. *For a sequence $\{\hat{\mathcal{F}}_t\}_{t\in\mathbb{N}}$ and $\{\hat{\boldsymbol{\lambda}}_t\}_{t\in\mathbb{N}}$, we first show that (a) $\lim_{t\to\infty}\hat{\mathcal{F}}_t \to \mathcal{F}$, (b) $\boldsymbol{\lambda}^\star$ is unique, and (c) $\lim_{t\to\infty}\hat{\boldsymbol{\lambda}}_t \to \boldsymbol{\lambda}^\star$. Thus, for any $\boldsymbol{\omega}\in\mathcal{F}$ and $\boldsymbol{\mu}$,*

$$\lim_{t\to\infty}\mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}}_t)\to\mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\mathcal{F}),$$

where $\boldsymbol{\lambda}^* \in \arg\min_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})}\boldsymbol{\omega}^\top d\left(\boldsymbol{\mu},\boldsymbol{\lambda}\right)$.

Proof. Here, we prove the three parts of the theorem consecutively.

Statement (a): Convergence of the limit $\lim_{t\to\infty}\hat{\mathcal{F}}$. To begin with the proof of the first statement of Theorem 2 we leverage the results stated in Theorem 10. Let $H(\tilde{A}) \triangleq \left\{\boldsymbol{\pi}\in\Delta_K : \tilde{A}\boldsymbol{\pi}\le 0\right\}$ and the set function $\tilde{A}\to H(\tilde{A})\cap C$ where $C=\hat{\mathcal{F}}$ is a non-empty compact (proven in Section C subset of Δ_K Then the set $H(\tilde{A})\cap C$ can be written as

$$H(\tilde{A})\cap C = \left\{\boldsymbol{\pi}\in\hat{\mathcal{F}} : \tilde{A}\boldsymbol{\pi}\le 0\right\}$$

To apply Theorem 10, $\{\tilde{A}^r, r\in\mathbb{N}\}$, must be a convergent sequence of affine function. It is evident that $\tilde{A}^r$ for any $r\in\mathbb{N}$ is an affine function since A is linear in A and the induced pessimism works as a translation. Then we can proceed to the next part of the proof of statement 1 where we prove that $\{\tilde{A}^r\}_{r\in\mathbb{N}}$ is a convergent sequence of functions. For ease of notation we will denote $\tilde{A}_t$ for the t-th element of the sequence $\{\tilde{A}^r\}_{r\in\mathbb{N}}$ for $t\in\mathbb{N}$.

The definition of the confidence radius for any constraint $i\in[d]$ follows from the Definition 2 as $f(\delta,t)\triangleq 1+\sqrt{\frac{1}{2}\log\frac{K}{\delta}+\frac{1}{4}\log\det\Sigma_t}$. It is evident from the definition that $f(t,\delta)$ is a non-decreasing function w.r.t time and it grows with order of at least $\mathcal{O}(\sqrt{\log t})$

We have from the definition of the confidence set, for all $i\in[d]$

$$\begin{aligned}
&\mathbb{P}\left(\hat{A}^i_t - f(t,\delta)||\omega_t||_{\Sigma_t^{-1}} \le A^i \le \hat{A}^i_t + f(t,\delta)||\omega_t||_{\Sigma_t^{-1}}\right)\ge 1-\delta\\
\implies &\mathbb{P}\left(-\frac{f(t,\delta)||\omega_t||_{\Sigma_t^{-1}}}{\sigma(\hat{A}^i_t)} \le \frac{\hat{A}^i_t - A^i}{\sigma(\hat{A}^i_t)} \le \frac{f(t,\delta)||\omega_t||_{\Sigma_t^{-1}}}{\sigma(\hat{A}^i_t)}\right)\ge 1-\delta\\
\implies &\mathbb{P}\left(-\frac{f(t,\delta)||\omega_t||_{\Sigma_t^{-1}}}{\sigma(\hat{A}^i_t)} \le Z \le \frac{f(t,\delta)||\omega_t||_{\Sigma_t^{-1}}}{\sigma(\hat{A}^i_t)}\right)\ge 1-\delta\\
\implies &2\Phi\left(\frac{f(t,\delta)||\omega_t||_{\Sigma_t^{-1}}}{\sigma(\hat{A}^i_t)}\right)\ge 2-\delta\\
\implies &\frac{f(t,\delta)||\omega_t||_{\Sigma_t^{-1}}}{\sigma(\hat{A}^i_t)} \ge \Phi^{-1}\left(1-\frac{\delta}{2}\right)\\
\implies &f(t,\delta)||\omega_t||_{\Sigma_t^{-1}} \ge \sigma(\hat{A}^i_t)\Phi^{-1}\left(1-\frac{\delta}{2}\right)
\end{aligned}$$

where $Z \triangleq \frac{\hat{A}^i_t - A^i}{\sigma(\hat{A}^i_t)}$ and $\Phi(.)$ is the cumulative distribution function of $\mathcal{N}(0,1)$ distribution. $\lim_{t\to\infty} f(t,\delta)\|\boldsymbol{\omega}_t\|_{\Sigma_t^{-1}}\to 0$ since $\sigma(\hat{A}^i_t)=\mathcal{O}(\sqrt{\frac{\log t}{t}})$. Leveraging CLT at this point we say

$$\hat{A}^i_t \xrightarrow{d} A^i, \forall i\in[d]$$

Then by Slutsky's theorem (Slutsky, 1925), we conclude $\hat{A}^i_t - f(t,\delta)||\omega_t||_{\Sigma_t^{-1}} \xrightarrow{d} A^i, \forall i\in[d]$.

It implies that $\{\tilde{A}_r\}_{r\in\mathbb{N}}$ is a convergent sequence of function for A. Now, we use Theorem 10 and get the following properties of the feasible set.

1. $H(\tilde{A}) \cap C \subset \lim_{r\to\infty} H(\tilde{A}^r) \cap C$
2. $\lim_{r\to\infty} H(\tilde{A}^r) \cap C$ is a closed convex superset of $H(\tilde{A}) \cap C$.
3. $H(\tilde{A}) \cap C$ has non-empty interior because of the feasibility condition and no component in $\tilde{A}$ is identically $\mathbf{0}$.

$$\lim_{r\to\infty} H(\tilde{A}^r) \cap C = H(\tilde{A}) \cap C$$

4. Even if the set $H(\tilde{A}) \cap C$ has empty interior or some component if $\tilde{A}$ is identically zero, by the last statement of the Theorem 10 we can say for any closed convex set Q of $H(\tilde{A}) \cap C$ we can design the function $\{\tilde{A}^r\}$ in such a way that $\lim_{r\to\infty} H(\tilde{A}^r) \cap C$ includes Q.

As the convergence of $\tilde{A}_t$ is guaranteed now asymptotically, we can guaranty convergence of the following limit $\lim_{t\to\infty} \hat{\mathcal{F}}_t \to \mathcal{F}$.

Statement (b) : Proof of Uniqueness of $\boldsymbol{\lambda}^\star$ Here, we prove if there exists a confusing instance $\boldsymbol{\lambda}^* \in \Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})$ which uniquely minimises the the function $\mathcal{D}(.)$ defined as

$$\mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \hat{\mathcal{F}}) \triangleq \inf_{\boldsymbol{l}\in\mathcal{L}} \inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})} \boldsymbol{\omega}^\top d\left(\boldsymbol{\mu}, \boldsymbol{\lambda}\right) - \boldsymbol{l}^\top \tilde{A}\boldsymbol{\omega}$$

We observe that only the leading quantity on the R.H.S associated with the KL is dependent on $\boldsymbol{\lambda}$. So, in this proof we will only show that $\boldsymbol{\lambda}^\star$ minimizes the KL divergence uniquely and since the KL is linearly dependent on the expression, proving this will be enough to ensure uniqueness of $\boldsymbol{\lambda}^\star$.

Now, from the properties of KL we know that $d(\boldsymbol{\mu}, \boldsymbol{\lambda})$ is convex on the pair $(\boldsymbol{\mu}, \boldsymbol{\lambda})$. But it is also strictly convex on $\boldsymbol{\lambda}$ if $\text{supp}(\boldsymbol{\lambda}) \subseteq \text{supp}(\boldsymbol{\mu})$ which is true in our case, since $\boldsymbol{\mu}, \boldsymbol{\lambda} \in \mathcal{D} \subseteq \mathbb{R}^k$.

Let us assume there are two local minima $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$, with the condition,

$$d(\boldsymbol{\mu}, \boldsymbol{\lambda}_1) \leq d(\boldsymbol{\mu}, \boldsymbol{\lambda}_2)$$

Then, we can write from the property of strict convexity, for some $\{h : 0 < h < 1\}$,

$$d(\boldsymbol{\mu}, h\boldsymbol{\lambda}_1 + (1-h)\boldsymbol{\lambda}_2) < hd(\boldsymbol{\mu}, \boldsymbol{\lambda}_1) + (1-h)d(\boldsymbol{\mu}, \boldsymbol{\lambda}_2)$$

Now, from the assumed condition on $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$, we can write —

$$\begin{aligned}
&d(\boldsymbol{\mu}, \boldsymbol{\lambda}_1) \leq d(\boldsymbol{\mu}, \boldsymbol{\lambda}_2) \\
\implies &hd(\boldsymbol{\mu}, \boldsymbol{\lambda}_1) \leq hd(\boldsymbol{\mu}, \boldsymbol{\lambda}_2) \text{ , since } h > 0 \\
\implies &hd(\boldsymbol{\mu}, \boldsymbol{\lambda}_1) + (1-h)d(\boldsymbol{\mu}, \boldsymbol{\lambda}_2) \leq hd(\boldsymbol{\mu}, \boldsymbol{\lambda}_2) + (1-h)d(\boldsymbol{\mu}, \boldsymbol{\lambda}_2) \\
\implies &hd(\boldsymbol{\mu}, \boldsymbol{\lambda}_1) + (1-h)d(\boldsymbol{\mu}, \boldsymbol{\lambda}_2) \leq d(\boldsymbol{\mu}, \boldsymbol{\lambda}_2)
\end{aligned}$$

Putting this result in the strict convexity condition we get

$$d(\boldsymbol{\mu}, h\boldsymbol{\lambda}_1 + (1-h)\boldsymbol{\lambda}_2) < d(\boldsymbol{\mu}, \boldsymbol{\lambda}_2)$$

which is a contradiction.

Thus, we conclude that for a strictly convex function f(x) with supp(x) being convex as well, the set of minimisers is either empty or singleton. Then, we can say $\boldsymbol{\lambda}^\star$ uniquely minimizes the KL, or say $\mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \hat{\mathcal{F}})$. Let us now once again remind the definition of perturbed alt-set $\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) \triangleq \{\boldsymbol{\lambda} \in \mathbb{D} : \max_{\boldsymbol{\pi}\in\hat{\mathcal{F}}} \boldsymbol{\lambda}^\top(\hat{\boldsymbol{\pi}}^\star - \boldsymbol{\pi}) < r\}$. Let us denote $\nu(\hat{\boldsymbol{\pi}}^\star)$ as the neighborhood of $\hat{\boldsymbol{\pi}}^\star$. Any $\boldsymbol{\pi} \in \hat{\mathcal{F}}$ is called a neighbor of $\hat{\boldsymbol{\pi}}^\star$, if it is an extreme point of $\hat{\mathcal{F}}$ and shares (K-1) active constraints with $\hat{\boldsymbol{\pi}}^\star$. Then, we can decompose the perturbed alt-set as $\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) = \bigcup_{\boldsymbol{\pi}\in\nu_{\hat{\mathcal{F}}}(\hat{\boldsymbol{\pi}}^\star)} \left\{\boldsymbol{\lambda} : \boldsymbol{\lambda}^\top(\hat{\boldsymbol{\pi}}^\star - \boldsymbol{\pi}) < 0\right\}$, which is a union of half-spaces for each neighbor. From this decomposition we can observe that $\hat{\boldsymbol{\pi}}^\star$ is not the r-optimal policy for $\boldsymbol{\lambda}$, i.e, $\{\exists \boldsymbol{\pi}' \in \Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) : \boldsymbol{\lambda}^\top(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}') < 0\}$. Then, it follows similar argument in Carlsson et al. (2024) to argue that the most confusing instance w.r.t $\boldsymbol{\mu}$ lies in the boundary of the normal cone, which lands us to Proposition 1.

For any $\boldsymbol{\omega} \in \hat{\mathcal{F}}$ and $\boldsymbol{\mu} \in \mathcal{D}$, the following projection lemma holds for the Lagrangian relaxation,

$$\begin{aligned}\mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \hat{\mathcal{F}}) &= \min_{\boldsymbol{l} \in \mathcal{L}} \max_{\boldsymbol{\pi} \in \Pi^r_{\hat{\mathcal{F}}}} \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi})} \min_{\boldsymbol{\lambda} : \boldsymbol{\lambda}^\top (\hat{\boldsymbol{\pi}}^\star - \boldsymbol{\pi}') = 0} \boldsymbol{\omega}^\top d(\boldsymbol{\mu}, \boldsymbol{\lambda}) - \boldsymbol{l}^\top \tilde{A} \boldsymbol{\omega} \\ &= \min_{\boldsymbol{l} \in \mathcal{L}} \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\hat{\boldsymbol{\pi}}^\star)} \min_{\boldsymbol{\lambda} : \boldsymbol{\lambda}^\top (\hat{\boldsymbol{\pi}}^\star - \boldsymbol{\pi}') = 0} \boldsymbol{\omega}^\top d(\boldsymbol{\mu}, \boldsymbol{\lambda}) - \boldsymbol{l}^\top \tilde{A} \boldsymbol{\omega}\end{aligned}$$

Statement (c): Convergence of the sequence $\{\hat{\boldsymbol{\lambda}}_n\}_{n\in\mathbb{N}}$. In known constraint setting the agent has access to $\mathcal{F}$. That means there is the actual sequence $\{\boldsymbol{\lambda}_n\}_{n\in\mathbb{N}}$ for which $\boldsymbol{\lambda}_n \to \boldsymbol{\lambda}^*$ as $n \to \infty$ since $\mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \hat{\mathcal{F}})$ is convex and continuous on $\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu}, \boldsymbol{\pi})$. But in this setting we try to estimate $\mathcal{F}$ as $\hat{\mathcal{F}}_n$ at each time step $n \in \mathbb{N}$. So there exists the $\{\hat{\boldsymbol{\lambda}}_n\}_{n\in\mathbb{N}}$ such that $\hat{\boldsymbol{\lambda}}_n \in \Lambda_{\hat{\mathcal{F}}_n}(\boldsymbol{\mu}, \boldsymbol{\pi})$ and we have to ensure it converges to the unique optimal $\boldsymbol{\lambda}^*$ i.e $\boldsymbol{\lambda}^\star \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu}) \subseteq \lim_{n\to\infty} \Lambda_{\hat{\mathcal{F}}_n}(\boldsymbol{\mu}, \boldsymbol{\pi})$ implies $\{\hat{\boldsymbol{\lambda}}_n\} \to \boldsymbol{\lambda}^*$ as $n \to \infty$

We use the fundamental theorem of limit to carry out this proof with the help of properties of the sets $\hat{\mathcal{F}}$ and Λ. The properties we have already proven for these sets are

1. $\hat{\mathcal{F}}_n$ for any $n \in \mathbb{N}$ is a superset of $\mathcal{F}$ due to the optimistic choice of A.
2. $\hat{\mathcal{F}}_n$ is a non-empty compact subset of Δ_K and $\lim_{n\to\infty} \hat{\mathcal{F}}_n = \mathcal{F}$.
3. $\Lambda_{\hat{\mathcal{F}}_n}(\boldsymbol{\mu}, \boldsymbol{\pi})$ is a closed convex set and it also is a superset of the real alt-set $\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})$.

Leveraging these properties we claim that for any $\boldsymbol{\mu} \in \mathcal{D}$, $\lim_{n\to\infty} \Lambda_{\hat{\mathcal{F}}_n}(\boldsymbol{\mu}, \boldsymbol{\pi}) = \Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})$. Since we have already proven uniqueness of $\boldsymbol{\lambda}$ in statement 2, we say $\hat{\boldsymbol{\lambda}}_n$ uniquely minimises $\Lambda_{\hat{\mathcal{F}}_n}(\boldsymbol{\mu}, \boldsymbol{\pi})$. Now from the (ϵ, δ)-definition of limits we say if $\Lambda_{\hat{\mathcal{F}}_n}(\boldsymbol{\mu}, \boldsymbol{\pi})$ is an ϵ-cover of $\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})$ for $\epsilon > 0$, then $|\hat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}^*| \leq \delta$ for $\delta > 0$ sufficiently small. It implies for a sequence of $\{\hat{\lambda_n}\}_{n\in\mathbb{N}}$ we claim $\lim_{n\to\infty} \hat{\boldsymbol{\lambda}}_n = \boldsymbol{\lambda}^*$ i.e the sequence convergence. Therefore we conclude by the statement itself

$$\{\hat{\boldsymbol{\lambda}}_n\}_{n\in\mathbb{N}} \to \boldsymbol{\lambda}^\star$$

Hence, proved. □

E Characterization of the Unique Optimal Policy: Proof of Theorem 3

Theorem 3. *For all $\boldsymbol{\mu} \in \mathcal{D}$, $\boldsymbol{\omega}^\star(\boldsymbol{\mu})$ satisfies the conditions*
1. Both the sets $\hat{\mathcal{F}}$ and $\boldsymbol{\omega}^\star(\boldsymbol{\mu})$ are closed and convex.
2. For all $\boldsymbol{\mu} \in \mathcal{D}$ and $\boldsymbol{\omega} \in \hat{\mathcal{F}}$, $\lim_{t\to\infty} \mathcal{D}(\boldsymbol{\omega}, \hat{\boldsymbol{\mu}}_t, \hat{\mathcal{F}}_t)$ is continuous.
3. Reciprocal of the characteristic time $\lim_{t\to\infty} T^{-1}_{\hat{\mathcal{F}}_t,r}(\boldsymbol{\mu})$ is continuous for all $\boldsymbol{\mu} \in \mathcal{D}$.
4. For all $\boldsymbol{\mu} \in \mathcal{D}$, $\boldsymbol{\mu} \to \boldsymbol{\omega}^\star(\boldsymbol{\mu})$ is upper hemi-continuous.
Thus, the optimization problem $\max_{\boldsymbol{\pi}\in\hat{\mathcal{F}}} \boldsymbol{\mu}^\top \boldsymbol{\pi}$ has a unique solution.

Proof. The theorem has four statements as the sufficient condition for the existence of unique optimal policy. So naturally we will dictate the proof structure in four steps and prove the statements one by one.

Statement 1: Convexity of feasible space and optimal set function. Let us first analyse the properties of $\hat{\mathcal{F}}$. For any two member of $\boldsymbol{\omega}_1, \boldsymbol{\omega}_2 \in \hat{\mathcal{F}}$ satisfying $\tilde{A}\boldsymbol{\omega}_1 \le 0$ and $\tilde{A}\boldsymbol{\omega}_2 \le 0$, their convex combination for any $\alpha \in [0,1]$,

$$\tilde{A}\left(\alpha\boldsymbol{\omega}_1 + (1-\alpha)\boldsymbol{\omega}_2\right) = \alpha\tilde{A}\boldsymbol{\omega}_1 + (1-\alpha)\tilde{A}\boldsymbol{\omega}_2 \le 0$$

Therefore we can say $\hat{\mathcal{F}}$ is convex because it is closed under convex operation. We claim $\hat{\mathcal{F}}$ is also closed since (a) The complement of $\hat{\mathcal{F}}$, $\hat{\mathcal{F}}^c \triangleq \{\boldsymbol{\pi} \in \Delta_K : \tilde{A}\boldsymbol{\pi} > 0\}$ is an open set. (b) we have already proven the limit of $\hat{\mathcal{F}}$ to be $\mathcal{F}$ which is always contained by $\hat{\mathcal{F}}$.

The elements in the domain of optimal allocation set function must be included in $\hat{\mathcal{F}}$. So compactness of $\boldsymbol{\omega}^*(\boldsymbol{\mu})$ is a direct consequence of compactness of $\hat{\mathcal{F}}$.

Statement 2: Continuity of limit. We have already proven in Section D that $\lim_{t\to\infty} \hat{\mathcal{F}}_t \to \mathcal{F}$. Also by convexity of KL and CLT we claim $\hat{\boldsymbol{\mu}}_t \to \boldsymbol{\mu}$ as $t \to \infty$ and since $\boldsymbol{\omega}$ is linear in $\mathcal{D}(\boldsymbol{\omega}_t, \boldsymbol{\mu}, \hat{\mathcal{F}}_t)$ it will converge to $\boldsymbol{\omega}^*(\boldsymbol{\mu})$ as $t \to \infty$, also due to convergence of $\hat{\boldsymbol{\mu}}_t$. Then we can say that the limiting value is same as the value if we plug in the limits in $\mathcal{D}$ i.e $\lim_{t\to\infty} \mathcal{D}(\boldsymbol{\omega}_t, \hat{\boldsymbol{\mu}}_t, \hat{\mathcal{F}}_t) = \mathcal{D}(\boldsymbol{\omega}^*(\boldsymbol{\mu}), \boldsymbol{\mu}, \mathcal{F})$. So we ensure the continuity of $\lim_{t\to\infty} \mathcal{D}(\boldsymbol{\omega}_t, \hat{\boldsymbol{\mu}}_t, \hat{\mathcal{F}})$.

Statement 3: Continuity of limit of inverse sampling complexity. This statement directly follows from the statement 2. Due to convexity of KL-divergence and convergence of $\hat{\mathcal{F}}$, the limiting value exists and it is equal to the inverse of characteristic time with the limiting value.

Statement 4: Upper hemi-continuity of optimal allocation function. We refer to (Magureanu et al., 2014) (see Lemma 12) for this proof. We denote $Q(\tilde{A}') \triangleq \lim_{\hat{\mathcal{F}}\to\mathcal{F}} \max_{\boldsymbol{\omega}\in\Delta_K} \left\{ \sum_{a=1}^K \boldsymbol{\omega}_a d\left(\boldsymbol{\mu}, \boldsymbol{\lambda}\right) - \boldsymbol{l}^\top \tilde{A}''\boldsymbol{\omega} \middle| \tilde{A}''\boldsymbol{\omega} \le 0, \boldsymbol{\omega}_a \ge 0 \forall i \in [K] \right\} = \boldsymbol{\omega}(\boldsymbol{\mu})$ where $\tilde{A}'' \in \mathbb{R}^{K\times K}$ is the rank-1 update of $\tilde{A}'$ which is a sub-matrix of $\tilde{A}$ with K number of active constraints. We define limiting set as

$$Q^\star(\tilde{A}'') = \left\{ \boldsymbol{\omega} : \lim_{\hat{\mathcal{F}}\to\mathcal{F}} \sum_{a=1}^K \boldsymbol{\omega}_a d\left(\boldsymbol{\mu}, \boldsymbol{\lambda}\right) = Q(\tilde{A}'') \middle| \tilde{A}''\boldsymbol{\omega} \le 0, \boldsymbol{\omega}_a \ge 0 \forall i \in [K] \right\} = \boldsymbol{\omega}^*(\boldsymbol{\mu})$$

As a direct consequence of Lemma 12 we get the following results

1. The function $\boldsymbol{\omega}^*(\boldsymbol{\mu})$ is continuous in $(\mathbb{R}^{K\times K}) \times \mathbb{R}^K$
2. $\boldsymbol{\omega}^*(\boldsymbol{\mu})$ is upper-hemicontinuous on $(\mathbb{R}^{K\times K}) \times \mathbb{R}^K$

Leveraging these four sufficient statements ensure that there exist unique solution for the optimization problem $\max_{\boldsymbol{\pi}\in\hat{\mathcal{F}}} \boldsymbol{\mu}^\top \boldsymbol{\pi}, \forall \boldsymbol{\mu} \in \mathcal{D}$ i.e the image set of the set-valued function $\boldsymbol{\omega}^*(.)$ is singleton. □

F Lagrangian Lower Bound for Gaussians: Proof of Theorem 4

Theorem 4. *Let $\{P_a\}_{a\in[K]}$ be Gaussian distributions with equal variance $\sigma^2 > 0$*

$$T^{-1}_{\hat{\mathcal{F}},r}(\boldsymbol{\mu}) = \max_{\boldsymbol{\omega}\in\hat{\mathcal{F}}} \min_{\boldsymbol{l}\in\mathcal{L}} \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi})} \left\{ \frac{(r-\boldsymbol{\mu}^\top(\boldsymbol{\pi}-\boldsymbol{\pi}'))^2}{2\sigma^2\|\boldsymbol{\pi}-\boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} - \boldsymbol{l}^\top\tilde{A}\boldsymbol{\omega} \right\}.$$

where $\mathrm{Diag}(1/\boldsymbol{\omega}_a)$ *is a K-dimensional diagonal matrix with a-th diagonal entry $1/\boldsymbol{\omega}_a$.*

Proof. We start the proof by the definition of $\mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}})$ as per Equation (8)

$$\begin{aligned}
\mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}}) &= \min_{\boldsymbol{l}\in\mathcal{L}} \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}} \min_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})} \left\{ \sum_{a=1}^{k} \boldsymbol{\omega}_a d(\boldsymbol{\mu}_a,\boldsymbol{\lambda}_a) - \boldsymbol{l}^\top\tilde{A}\boldsymbol{\omega} \right\} \\
&= \min_{\boldsymbol{l}\in\mathcal{L}} \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \min_{\boldsymbol{\lambda}:\boldsymbol{\lambda}^\top(\boldsymbol{\pi}-\boldsymbol{\pi}')=r} \left\{ \sum_{a=1}^{k} \boldsymbol{\omega}_a d(\boldsymbol{\mu}_a,\boldsymbol{\lambda}_a) - \boldsymbol{l}^\top\tilde{A}\boldsymbol{\omega} \right\} \rightsquigarrow \text{ via Proposition 1} \qquad (9)
\end{aligned}$$

The Lagrangian formulation of $\mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}})$ is written as

$$\mathcal{L}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}}) = \min_{\boldsymbol{l}\in\mathcal{L}} \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi})} \left\{ \sum_{a=1}^{K} \boldsymbol{\omega}_a d\left(\boldsymbol{\mu}_a,\boldsymbol{\lambda}_a\right) - \boldsymbol{l}^\top\tilde{A}\boldsymbol{\omega} - \gamma\left(\sum_{a=1}^{K}\boldsymbol{\lambda}_a\boldsymbol{v}_a - r\right)\right\}$$

where $\boldsymbol{v}_a \triangleq (\boldsymbol{\pi}-\boldsymbol{\pi}')_a$.

We assume both the instances $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ follow Gaussian distribution with same variance σ^2. Then, we can rewrite the Lagrangian putting the value of the KL as —

$$\mathcal{L}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}}) = \min_{\gamma\in\mathbb{R}_+} \min_{\boldsymbol{l}\in\mathcal{L}} \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi})} \left\{ \sum_{a=1}^{K} \boldsymbol{\omega}_a \frac{(\boldsymbol{\mu}_a-\boldsymbol{\lambda}_a)^2}{2\sigma^2} - \boldsymbol{l}^\top\tilde{A}\boldsymbol{\omega} - \gamma\left(\sum_{a=1}^{K}\boldsymbol{\lambda}_a\boldsymbol{v}_a - r\right)\right\} \qquad (10)$$

Differentiating the Lagrangian w.r.t $\boldsymbol{\lambda}_a$ and equating it to 0, we get

$$\begin{aligned}
\nabla_{\boldsymbol{\lambda}_a}\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\omega},\boldsymbol{\mu}) &= 0 \\
\text{or, } -\frac{\boldsymbol{\omega}_a(\boldsymbol{\mu}_a-\boldsymbol{\lambda}_a)}{\sigma^2} - \gamma\boldsymbol{v}_a &= 0 \\
\text{or, } \boldsymbol{\lambda}_a = \boldsymbol{\mu}_a + \frac{\gamma\boldsymbol{v}_a\sigma^2}{\boldsymbol{\omega}_a}
\end{aligned}$$

Then putting back the value of $\boldsymbol{\lambda}_a$ in Equation 10 we get

$$\mathcal{L}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}}) = \min_{\gamma\in\mathbb{R}_+} \min_{\boldsymbol{l}\in\mathcal{L}} \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi})} \left\{ -\sum_{a=1}^{K} \gamma^2\frac{\boldsymbol{v}_a^2\sigma^2}{2\boldsymbol{\omega}_a} - \boldsymbol{l}^\top\tilde{A}\boldsymbol{\omega} - \gamma\left(\sum_{a=1}^{K}\boldsymbol{\mu}_a\boldsymbol{v}_a - r\right)\right\} \qquad (11)$$

Again differentiating the Lagrangian w.r.t γ and equating it to 0, we get

$$\begin{aligned}
\nabla_\gamma\mathcal{L}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}}) = 0 &\implies r - \sum_{a=1}^{K}\boldsymbol{\mu}_a\boldsymbol{v}_a - \gamma\sum_{a=1}^{K}\frac{\sigma^2}{\boldsymbol{\omega}_a}\boldsymbol{v}_a = 0 \\
&\implies \gamma = \frac{(r-\boldsymbol{\mu}^\top v)}{\sum_{a=1}^{K}\frac{\sigma^2}{\boldsymbol{\omega}_a}\boldsymbol{v}_a^2}
\end{aligned}$$

Putting the value of γ in Equation 11, we get —

$$\mathcal{L}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}}) = \min_{\boldsymbol{l}\in\mathcal{L}} \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi})} \left\{ \frac{(r-\boldsymbol{\mu}^\top v)^2}{2\sigma^2\sum_{a=1}^{K}\frac{\boldsymbol{v}_a^2}{\boldsymbol{\omega}_a}} - \boldsymbol{l}^\top\tilde{A}\boldsymbol{\omega} \right\}$$

$$= \min_{\boldsymbol{l}\in\mathcal{L}} \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi})} \left\{ \frac{1}{2\sigma^2} \frac{(r-\boldsymbol{\mu}^\top(\boldsymbol{\pi}-\boldsymbol{\pi}'))^2}{\|\boldsymbol{\pi}-\boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} - \boldsymbol{l}^\top \tilde{A}\boldsymbol{\omega} \right\}$$

Therefore inverse characteristic time for Lagrangian relaxation with unknown constraints satisfies,

$$T^{-1}_{\hat{\mathcal{F}},r}(\boldsymbol{\mu}) = \max_{\boldsymbol{\omega}\in\hat{\mathcal{F}}} \min_{\boldsymbol{l}\in\mathcal{L}} \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi})} \left\{ \frac{1}{2\sigma^2} \frac{(r-\boldsymbol{\mu}^\top(\boldsymbol{\pi}-\boldsymbol{\pi}'))^2}{\|\boldsymbol{\pi}-\boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} - \boldsymbol{l}^\top \tilde{A}\boldsymbol{\omega} \right\}$$

□

F.1 Bounds on Sample complexity: Proof of Corollary 1 Part (a)

Corollary 1. ***Part (a)*** *Let* $d^2_{\boldsymbol{\pi}} \triangleq \frac{\|\boldsymbol{\pi}^\star_{\mathcal{F}}-\boldsymbol{\pi}\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^\top}}{\|\boldsymbol{\pi}^\star_{\mathcal{F}}-\boldsymbol{\pi}\|^2_2}$ *be the norm of the projection of* $\boldsymbol{\mu}$ *on the policy gap* $(\boldsymbol{\pi}^\star_{\mathcal{F}}-\boldsymbol{\pi})$.
Part i. *Then, we get* $\frac{2\sigma^2 K}{C_{\text{known}}}(1+\mathfrak{s}_{\tilde{A}}) \leq T_{\hat{\mathcal{F}},0}(\boldsymbol{\mu}) \leq \frac{2\sigma^2 K}{C_{\text{known}}}$*, where* $C_{\text{known}} = \min_{\boldsymbol{\pi}''\in\nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} d^2_{\boldsymbol{\pi}''}$.
Part ii. *Additionally,* $T_{\hat{\mathcal{F}},0}(\boldsymbol{\mu}) \geq \frac{H}{\kappa^2}(1+\mathfrak{s}_{\tilde{A}})$*, where* H *is inversely proportional to the sum of squares of gaps and* κ_{known} *is the condition number of a sub-matrix of* A *consisting* K *linearly independent active constraints for* $\boldsymbol{\pi}^\star$.

Proof. Here, we derive explicit expression for gaussian characterisation of the lower and upper bound on the characteristic time. We start the proof with the difference in sample complexity between unknown and known constraint setting

$$\begin{aligned}\mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}}) - \mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\mathcal{F}) &= \min_{\boldsymbol{l}\in\mathcal{L}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \left\{ \frac{1}{2\sigma^2} \frac{\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}-\boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^\top}}{\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}-\boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} - \boldsymbol{l}^\top\tilde{A}\boldsymbol{\omega} \right\} \\ &\quad - \min_{\boldsymbol{\pi}''\in\nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} \frac{1}{2\sigma^2} \frac{\|\boldsymbol{\pi}^\star_{\mathcal{F}}-\boldsymbol{\pi}''\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^\top}}{\|\boldsymbol{\pi}^\star_{\mathcal{F}}-\boldsymbol{\pi}''\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} \end{aligned} \tag{12}$$

Let us remind, due to pessimistic choice of $\tilde{A}$, $\mathcal{F} \subseteq \hat{\mathcal{F}}$. $\boldsymbol{\pi}'$ is a neighbor of $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$ if it is an extreme point in the polytope $\hat{\mathcal{F}}$ and shares (K-1) active constraints with $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$. Then $\boldsymbol{\pi}^\star_{\mathcal{F}}$ and $\boldsymbol{\pi}''$ lies in the interior of $\hat{\mathcal{F}}$ i.e, they can be expressed as a convex combination of $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$ and $\boldsymbol{\pi}'$. Let, $\exists 0 \leq t_1 \leq 1 : \boldsymbol{\pi}^\star_{\mathcal{F}} = t_1\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} + (1-t_1)\boldsymbol{\pi}'$ and $\exists 0 \leq t_1 \leq 1 : \boldsymbol{\pi}'' = t_2\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} + (1-t_2)\boldsymbol{\pi}'$ Then, $(\boldsymbol{\pi}^\star_{\mathcal{F}}-\boldsymbol{\pi}') = t_1(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}-\boldsymbol{\pi}')$ and $(\boldsymbol{\pi}''-\boldsymbol{\pi}') = t_2(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}-\boldsymbol{\pi}')$. Then,

$$\|\boldsymbol{\pi}^\star_{\mathcal{F}}-\boldsymbol{\pi}''\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^\top} = \|(\boldsymbol{\pi}^\star_{\mathcal{F}}-\boldsymbol{\pi}')-(\boldsymbol{\pi}''-\boldsymbol{\pi}')\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^\top} = \|t_1(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}-\boldsymbol{\pi}')-t_2(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}-\boldsymbol{\pi}')\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^\top} = (t_1-t_2)^2\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}-\boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^\top}$$

Applying this equality in Equation 12, we get —

$$\begin{aligned}&\mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\mathcal{F}) - \mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}}) \\ &= \min_{\boldsymbol{\pi}''\in\nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} \frac{1}{2\sigma^2} \frac{\|\boldsymbol{\pi}^\star_{\mathcal{F}}-\boldsymbol{\pi}''\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^\top}}{\|\boldsymbol{\pi}^\star_{\mathcal{F}}-\boldsymbol{\pi}''\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} - \min_{\boldsymbol{l}\in\mathcal{L}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \left\{ \frac{1}{2\sigma^2} \frac{\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}-\boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^\top}}{\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}-\boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} - \boldsymbol{l}^\top\tilde{A}\boldsymbol{\omega} \right\} \\ &= \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \frac{1}{2\sigma^2} \frac{(t_1-t_2)^2\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}-\boldsymbol{\pi}''\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^\top}}{(t_1-t_2)^2\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}-\boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} - \min_{\boldsymbol{l}\in\mathcal{L}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \left\{ \frac{1}{2\sigma^2} \frac{\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}-\boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^\top}}{\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}-\boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} - \boldsymbol{l}^\top\tilde{A}\boldsymbol{\omega} \right\} \\ &= \min_{\boldsymbol{l}\in\mathcal{L}} \boldsymbol{l}^\top\tilde{A}\boldsymbol{\omega} \\ &\leq \mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}})\mathfrak{s}_{\tilde{A}} \end{aligned} \tag{13}$$

The last inequality is a direct consequence of Lemma 6.
Part I. Since $\min_{\boldsymbol{l}\in\mathcal{L}} -\boldsymbol{l}^\top\tilde{A}\boldsymbol{\omega} \geq 0$ by the design of the Lagrangian and the constraints, we get

$$\begin{aligned}&\mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}}) \geq \mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\mathcal{F}) \\ \implies &T^{-1}_{\hat{\mathcal{F}},0}(\boldsymbol{\mu}) \geq T^{-1}_{\mathcal{F},0}(\boldsymbol{\mu})\end{aligned} \tag{14}$$

This leads us to the upper bound. For ease of comparison, we follow the same notation used in Carlsson et al. (2024) and define for any $\boldsymbol{\pi} \in \Delta_{K-1}$, $d^2_{\boldsymbol{\pi}} \triangleq \frac{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^\top}}{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}\|^2_2}$, which is the squared distance between $\boldsymbol{\mu}$ and the hyperplane $(\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}) = \mathbf{0}$. Therefore, using (Carlsson et al., 2024, Corollary 1) and Equation (14), we get

$$T_{\hat{\mathcal{F}},0}(\boldsymbol{\mu}) \le \frac{2\sigma^2 K}{\min_{\boldsymbol{\pi}'' \in \nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} d^2_{\boldsymbol{\pi}''}}$$

Along with Equation (13), this implies that

$$\begin{aligned} &\mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \hat{\mathcal{F}})\,(1 + \mathfrak{s}_{\tilde{A}}) \le \mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \mathcal{F}) \\ \implies &T^{-1}_{\hat{\mathcal{F}},0}(\boldsymbol{\mu})\,(1 + \mathfrak{s}_{\tilde{A}}) \le T^{-1}_{\mathcal{F},0}(\boldsymbol{\mu}) \end{aligned} \tag{15}$$

We again leverage (Carlsson et al., 2024, Corollary 1) to get lower bound on the characteristic time as,

$$T_{\hat{\mathcal{F}},0}(\boldsymbol{\mu}) \ge (1 + \mathfrak{s}_{\tilde{A}})\,\frac{2\sigma^2}{\min_{\boldsymbol{\pi}'' \in \nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} d^2_{\boldsymbol{\pi}''}}$$

Further, we define $C_{\text{Known}} \triangleq \min_{\boldsymbol{\pi}'' \in \nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} d^2_{\boldsymbol{\pi}''}$, which concludes the proof.

□

F.2 Impact of Unknown Linear Constraints: Proof of Corollary 1 Part (b)

Corollary 1. ***Part (b)*** *Let $d^2_{\boldsymbol{\pi}} \triangleq \frac{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^\top}}{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}\|^2_2}$ be the norm of the projection of $\boldsymbol{\mu}$ on the policy gap $(\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi})$. Then, the characteristic time $T_{\hat{\mathcal{F}},0}(\boldsymbol{\mu})$ satisfies $T_{\hat{\mathcal{F}},0}(\boldsymbol{\mu}) \ge \frac{H}{\kappa^2}\left(1 - \frac{\epsilon}{\gamma}\right)$. H is the sum of squares of gaps. κ is the condition numbers of a sub-matrix of A that consists K linearly independent active constraints for $\boldsymbol{\pi}^\star$.*

Proof. This result a direct implication of Corollary 2 in Carlsson et al. (2024), that states,

$$T^{-1}_{\mathcal{F},0}(\boldsymbol{\mu}) \le \frac{\kappa^2}{H}$$

Here, $H = \frac{2\sigma^2}{\Delta^2}$, where $\Delta^2 = \sum_{a=1}^{K}(\boldsymbol{\mu}^* - \boldsymbol{\mu}_a)^2$, i.e the sum of squares of sub-optimal gaps in the arms and κ is the condition number of a sub-matrix of A consisting at least K linearly independent active constraints for $\boldsymbol{\pi}^\star$. Referring to the proof structure of Corollary 2 in (Carlsson et al., 2024) is of independent interest.

Now we leverage equation Equation (15) to get

$$T_{\hat{\mathcal{F}},0}(\boldsymbol{\mu}) \ge \frac{H}{\kappa^2}\,(1 + \mathfrak{s}_{\tilde{A}})$$

□

G Sample Complexity Upper Bounds (Analysis of Algorithms)

G.1 Proof of Lemma 2 : Implication of $(1-\delta)$-correctness

Lemma 2. *If the recommended policy is $(1-\delta)$-correct then it is $(1-\delta)$-feasible.*

Proof. Let the recommended policy $\boldsymbol{\pi}$ is $(1-\delta)$-correct. It means

$$\begin{aligned}&\mathbb{P}(\boldsymbol{\pi}=\boldsymbol{\pi}^\star_{\mathcal{F}}) \geq 1-\delta\\ \implies &\mathbb{P}(\boldsymbol{\pi}=\boldsymbol{\pi}^\star_{\mathcal{F}} \wedge A\boldsymbol{\pi}^\star_{\mathcal{F}} \leq 0) \geq 1-\delta\\ \implies &\mathbb{P}(A\boldsymbol{\pi} \leq 0) \geq 1-\delta\end{aligned}$$

Hence $(1-\delta)$-correctness automatically implies $(1-\delta)$-feasibility. □

G.2 Stopping Criterion

Theorem 5. *The Chernoff* ***stopping rule to ensure*** $(1-\delta)$***-correctness and*** $(1-\delta)$***-feasibility*** *is*

$$\max_{\boldsymbol{\pi}_t\in\Pi^r_{\hat{\mathcal{F}}_t}} \inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^{K} N_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t}, \lambda_a) > \beta(t,\delta)\,,$$

where $\beta(t,\delta) \triangleq 3S_0\log(1+\log N_{a,t}) + S_0\mathcal{T}\left(\frac{(K\wedge d)+\log\frac{1}{\delta}}{S_0}\right)$ *and* $0\leq S_0\leq K$*.*

Proof. We dictate the proof in 2 steps. In the first step, we prove that the stopping time τ_δ is finite. Then, in next step, we give an explicit expression of the stopping threshold by upper bounding probability of the bad event for stopping time τ_δ.

Let us first go through some notations.

$\boldsymbol{\pi}_t \triangleq \arg\max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}_t}} \hat{\boldsymbol{\mu}}_t^\top\boldsymbol{\pi}$, where $\hat{\boldsymbol{\mu}}_t \in \arg\min_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^{K} N_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t},\boldsymbol{\lambda}_a) - \boldsymbol{l}_t^\top \tilde{A}_t N_{a,t}$.

Algorithm 2 and 3 stops at a finite $\tau_\delta\in\mathbb{N}$ if the events $\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\hat{\boldsymbol{\mu}}_t)}\{\exists t\in\mathbb{N}: \sum_{a=1}^{K} N_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t},\boldsymbol{\lambda}_a)\} > \beta(t,\delta)$ and $\{\exists t\in\mathbb{N}: \|(\tilde{A}_t - A)N_t\|_\infty \leq t\rho(t,\delta)\}$ jointly occurs.

Step 1: Finiteness of the stopping time. A stopping time is finite if the parameters in the system converges to their true values in finite time, in our case the means of arms and the constraint matrix. Let us define $\mathcal{A} \triangleq \{a\in[K] : \lim_{t\to\infty} N_{a,t} < \infty\}$ as a sampling rule i.e if an arm belongs to this set $\mathcal{A}$, it has been sampled finitely and otherwise the arm has been sampled enough number of times so that the mean of that arm has converged to it's true value and the column in the constraint matrix corresponding to that arm as also converged. For arms $a\in[K]$ and $a\in\mathcal{A}^c, \hat{\boldsymbol{\mu}}_{a,t}\to\tilde{\boldsymbol{\mu}}_a\neq\boldsymbol{\mu}_a$. When all parameters are concentrated $\mathcal{A}=\emptyset$, we say $\forall a\in[K] : \hat{\boldsymbol{\mu}}_a\to\boldsymbol{\mu}_a$. We also define the limit of this empirical sampling rule as $\boldsymbol{\omega}_\infty = \lim_{t\to\infty}\frac{N_{a,t}}{t}\forall a\in[K]$. We then write the stopping condition in a new way $\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\hat{\boldsymbol{\mu}}_t)}\{\exists t\in\mathbb{N} : \sum_{a=1}^{K}\frac{N_{a,t}}{t} d(\hat{\boldsymbol{\mu}}_{a,t},\boldsymbol{\lambda}_a) > \frac{\beta(t,\delta)}{t}\}$. By continuity properties and knowing $\beta(t,.)\to\log\log t$ and $\rho(t,.)\to 0$ as $t\to\infty$, we claim by taking taking asymptotic limits both sides $\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\hat{\boldsymbol{\mu}}_t)}\sum_{a=1}^{K}\frac{\boldsymbol{\omega}_{\infty,a}}{t} d(\hat{\boldsymbol{\mu}}_{a,t},\boldsymbol{\lambda}_a) > 0$. We get strict inequality for the both the cases by the virtue of construction of the set $\mathcal{A}$ such that for arms $a\in\mathcal{A}, \boldsymbol{\omega}_\infty\neq 0$ and the KL-divergence is non-zero as $\boldsymbol{\lambda}_a\neq\boldsymbol{\mu}$ since $\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)$.

Step 2: Probability of bad event to Stopping threshold. Let $\boldsymbol{\omega}_t$ is the allocation associated to N_t. Then we define the bad event as

$$U_t \triangleq \{\boldsymbol{\pi}_{\tau_\delta}\notin\Pi^r_{\mathcal{F}}\} = \bigcup_{\boldsymbol{\pi}\notin\Pi^r_{\mathcal{F}}}\left\{\exists t\in\mathbb{N} : \boldsymbol{\pi}_{t+1}=\boldsymbol{\pi}\wedge \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}_t}}\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)}\sum_{a=1}^{K} N_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t},\boldsymbol{\lambda}_a) > \beta(t,\delta)\right\}$$

Therefore probability of this bad event

$$\mathbb{P}(U_t) = \mathbb{P}\left(\bigcup_{\boldsymbol{\pi}\notin\Pi^r_{\mathcal{F}}}\left\{\exists t\in\mathbb{N} : \boldsymbol{\pi}_{t+1}=\boldsymbol{\pi}\wedge \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}_t}}\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)}\sum_{a=1}^{K} N_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t},\boldsymbol{\lambda}_a) > \beta(t,\delta)\right\}\right)$$

$$\leq \sum_{\boldsymbol{\pi}\notin\Pi^r_{\mathcal{F}}} \mathbb{P}\left\{\exists t\in\mathbb{N}: \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}_t}} \inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^{K} N_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t},\boldsymbol{\lambda}_a)\} > \beta(t,\delta)\right\}$$

$$\leq \sum_{\boldsymbol{\pi}\notin\Pi^r_{\mathcal{F}}} \mathbb{P}\left\{\exists t\in\mathbb{N}: \sum_{a=1}^{K} N_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t},\boldsymbol{\mu}_a) > \beta(t,\delta)\right\} \tag{16}$$

We define $I_{\boldsymbol{\pi}} \triangleq \text{Supp}(\hat{\boldsymbol{\pi}}^\star)\Delta\text{Supp}(\boldsymbol{\pi})$ and also $S_0 \triangleq \max_{\boldsymbol{\pi}} |I_{\boldsymbol{\pi}}|$. We note that $0 \leq S_0 \leq K$.

We get from Theorem 9 in (Kaufmann and Koolen, 2021) with the notation of $\mathcal{T}(.)$ follows from Lemma 9

$$\sum_{\boldsymbol{\pi}\notin\Pi^r_{\mathcal{F}}} \mathbb{P}\left\{\exists t\in\mathbb{N}: \sum_{a\in I_{\boldsymbol{\pi}}} N_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t},\boldsymbol{\mu}_a) \geq \sum_{a\in I_{\boldsymbol{\pi}}} 3\log(1+\log N_{a,t}) + |S_0|\mathcal{T}\left(\frac{\log\frac{|\boldsymbol{\pi}_t|-1}{\delta}}{S_0}\right)\right\} \leq \delta$$

where δ is chosen to be $\frac{\delta}{|\boldsymbol{\pi}_t|-1}$ such that $\log\frac{|\boldsymbol{\pi}_t|-1}{\delta} \leq \log(\frac{2^K}{\delta}) \leq (K\wedge d) + \log\frac{1}{\delta}$

Also $\sum_{a\in I_{\boldsymbol{\pi}}} 3\log(1+\log N_{a,t} \leq 3S_0\log(1+\log N_{a,t})$. Therefore the stopping threshold is given by

$$\beta(t,\delta) = 3S_0\log(1+\log N_{a,t}) + S_0\mathcal{T}\left(\frac{(K\wedge d)+\log\frac{1}{\delta}}{S_0}\right)$$

In practice, we use $S_0 = K$. □

G.3 Upper Bound of LATS

Theorem 6. *The sample complexity upper bound of LATS to yield a $(1-\delta)$-correct optimal policy is*

$$\lim_{\delta\to 0}\frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq \alpha(1+\mathfrak{s})T_{\mathcal{F},r}(\mu),$$

where $\mathfrak{s}$ is the shadow price of the true constraint A, $T_{\mathcal{F},r}(\mu)$ is the characteristic time under the true constraint (Equation (4)), and $\delta\in(0,1]$.

Proof. We will prove this theorem in 5 steps. In the first step, we define what is considered to be the good event in our unknown constraint setting, then we go on bounding the probability of the complement of this good event in step 2. Once the parameter concentrations are taken care of, we show how we can lower bound the instantaneous complexity of the algorithm in step 3. In step 4, we finally prove the upper bound on stopping time for both good and bad events. We conclude with the asymptotic upper bound on stopping time i.e when $\delta\to 0$ and $\epsilon\to 0$ in step 5.

Step 1: Defining the good event. Given an $\epsilon > 0$, for $h(T) = \max\sqrt{T}, f(T,\delta)$ we define the good event G_T as,

$$G_T \triangleq \bigcap_{t=h(T)}^{\top} \{||\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}||_\infty \leq \xi(\epsilon) \wedge ||(\tilde{A}^i_t - A^i)\boldsymbol{\omega}||_\infty \leq \phi(\epsilon)\forall\boldsymbol{\omega}\in\hat{\mathcal{F}}\}$$

where, $\xi(\epsilon) \leq \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}} \max_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi})} \frac{1}{5}\boldsymbol{\mu}^\top(\boldsymbol{\pi}-\boldsymbol{\pi}')$, and $\phi(\epsilon) \triangleq \max(1,\epsilon)$ for a given $\epsilon > 0$. The good event implies that the means and constraints are well concentrated in an ϵ-ball around their true values. Thus, we have to now bound the extra cost of their correctness and the number of samples required to reach the good events.

We also observe that $||\boldsymbol{\mu}'-\boldsymbol{\mu}||_\infty \leq \xi(\epsilon)$ and $||(\tilde{A}^i_t - A^i)\boldsymbol{\omega}||_\infty \leq \phi(\epsilon)$ implies that $\sup_{\boldsymbol{\omega}'\in\boldsymbol{\omega}^\star(\boldsymbol{\mu}')}\sup_{\boldsymbol{\omega}\in\boldsymbol{\omega}^\star(\boldsymbol{\mu})} ||\boldsymbol{\omega}'-\boldsymbol{\omega}|| \leq \epsilon$ due to upper hemicontinuity of $\boldsymbol{\omega}^*(\boldsymbol{\mu})$ (Theorem 3).

Step 2: Samples to Achieve the Good Event. Now, let us bound the probability of bad event,

$$\mathbb{P}(G^c_T) = \sum_{t=h(T)}^{\top}\left(\mathbb{P}\{||\hat{\boldsymbol{\mu}}_t-\boldsymbol{\mu}||_\infty > \xi(\epsilon)\} + \mathbb{P}\left\{||(\tilde{A}^i_t - A^i)\boldsymbol{\omega}||_\infty > \phi(\epsilon)\right\}\right)$$

$$\leq \sum_{t=h(T)}^{\top} \mathbb{P}\left\{||\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}||_\infty > \xi(\epsilon)\right\} + \sum_{t=h(T)}^{\top} \mathbb{P}\left\{||(\tilde{A}^i_t - A^i)\boldsymbol{\omega}||_\infty > \phi(\epsilon)\right\}$$

$$\leq BT \exp\left(-CT^{\frac{1}{8}}\right) + K \sum_{t=h(T)}^{\top} \frac{1}{t}$$

The first inequality is due to the union bound. The second inequality is due to the Lemma 7 (Lemma 19 of Garivier and Kaufmann (2016)), which states that

$$\sum_{t=h(T)}^{\top} \mathbb{P}\left\{\|\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}\|_\infty > \xi(\epsilon)\right\} \leq BT \exp\left(-CT^{\frac{1}{8}}\right),$$

and also due to Lemma 4 that proves concentration bound of the constraint matrix over time.

Step 3: Tracking argument. Now, we state how concentrating on means and constraints leads to good concentration on the allocations too. Since we use C-tracking, we can leverage the concentration in allocation by(Lemma 17 of Degenne et al. (2019b)). We use this lemma than D-tracking or the tracking argument in (Lemma 7 of Garivier and Kaufmann (2016)) because the optimal allocations might not be unique but the set $\boldsymbol{\omega}^*(\boldsymbol{\mu})$ is convex (Theorem 3).

Hence, there exists a T_ϵ such that under the good event and $t \geq \max(T_\epsilon, h(T))$, we have,

$$\begin{aligned} |(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_t)^\top \boldsymbol{\pi}^\star| &\leq |(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_t)^\top \hat{\boldsymbol{\pi}}^\star| + |(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_t)^\top (\hat{\boldsymbol{\pi}}^\star - \boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})| \leq 4\xi(\epsilon) \\ &\leq \max_{\boldsymbol{\pi} \in \Pi^r_{\hat{\mathcal{F}}}} \max_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi})} \boldsymbol{\mu}^\top (\boldsymbol{\pi} - \boldsymbol{\pi}') \end{aligned}$$

We have replaced the perturbed alt-set with the true one because for $t \geq \max(T_\epsilon, h(T))$, we can ensure the convergence of $\hat{\mathcal{F}}$ almost surely i.e $\hat{\mathcal{F}} \overset{\text{a.s}}{\to} \mathcal{F}$

Step 3: Complexity of identification under good event and constraint. Now, we want to understand how hard it is to hit the stopping rule even under the good event. First, we define

$$C_{\epsilon,\hat{\mathcal{F}}} \triangleq \inf_{\substack{\boldsymbol{\mu}' : ||\boldsymbol{\mu}' - \boldsymbol{\mu}||_\infty \leq \xi(\epsilon) \\ \boldsymbol{\omega}' : ||\boldsymbol{\omega}' - \boldsymbol{\omega}||_\infty \leq 3\epsilon \\ \tilde{A}' : ||(\tilde{A} - A)\boldsymbol{\omega}||_\infty \leq \phi(\epsilon)}} \mathcal{D}(\boldsymbol{\mu}', \boldsymbol{\omega}', \hat{\mathcal{F}}) - \boldsymbol{l}^\top \tilde{A}\boldsymbol{\omega}\,.$$

Now leveraging Lemma 6, we obtain

$$(1 + \psi) \inf_{\substack{\boldsymbol{\mu}' : ||\boldsymbol{\mu}' - \boldsymbol{\mu}||_\infty \leq \xi(\epsilon) \\ \boldsymbol{\omega}' : ||\boldsymbol{\omega}' - \boldsymbol{\omega}||_\infty \leq 3\epsilon \\ \tilde{A} : ||(\tilde{A} - A)\boldsymbol{\omega}||_\infty \leq \phi(\epsilon)}} \mathcal{D}(\boldsymbol{\mu}', \boldsymbol{\omega}', \hat{\mathcal{F}}) \geq \inf_{\substack{\boldsymbol{\mu}' : ||\boldsymbol{\mu}' - \boldsymbol{\mu}||_\infty \leq \xi(\epsilon) \\ \boldsymbol{\omega}' : ||\boldsymbol{\omega}' - \boldsymbol{\omega}||_\infty \leq 3\epsilon \\ \tilde{A}' : ||(\tilde{A} - A)\boldsymbol{\omega}||_\infty \leq \phi(\epsilon)}} \mathcal{D}(\boldsymbol{\mu}', \boldsymbol{\omega}', \hat{\mathcal{F}}) - \boldsymbol{l}^\top \tilde{A}\boldsymbol{\omega}\,,$$

where definition of ψ follows from Lemma 6. It quantifies how the Lagrangian lower bound relates with the Likelihood Ratio Test-based quantity in the stopping time.

Therefore by the C-tracking argument in Theorem 10, we can state

$$\mathcal{D}(\hat{\boldsymbol{\mu}}_t, N_t, \hat{\mathcal{F}}) \geq t \inf_{\substack{\boldsymbol{\mu}' : ||\boldsymbol{\mu}' - \boldsymbol{\mu}||_\infty \leq \xi(\epsilon) \\ \boldsymbol{\omega}' : ||\boldsymbol{\omega}' - \boldsymbol{\omega}||_\infty \leq 3\epsilon \\ \tilde{A} : ||(\tilde{A} - A)\boldsymbol{\omega}||_\infty \leq \phi(\epsilon)}} \mathcal{D}(\boldsymbol{\mu}', \boldsymbol{\omega}', \hat{\mathcal{F}}) \geq \frac{tC_{\epsilon,\hat{\mathcal{F}}}}{1 + \psi}\,. \tag{17}$$

Here, LHS is the quantity that we use to stop and yield a $(1 - \delta)$-correct policy.

Step 4: Bounding the stopping time with good and bad events. We denote τ_δ as the stopping time. So for $T \geq T_\epsilon$ we can write upper bound on this stopping time for both good and bad events as

$$\min(\tau_\delta, T) \leq \max(\sqrt{T}, f(T, \delta)) + \sum_{t=T_\epsilon}^{\top} \mathbb{1}_{\tau_\delta > t}$$

By the correctness of the stopping time, the event $\tau(\delta) > t$ happens if $\mathcal{D}(\hat{\boldsymbol{\mu}}_t, N_t, \hat{\mathcal{F}}) \leq \beta(t,\delta)$ for any $t \leq T$.

Now using the lower bound on $\mathcal{D}(\hat{\boldsymbol{\mu}}_t, N_t, \hat{\mathcal{F}})$ (Equation 17), we get

$$\begin{aligned}
\min(\tau_\delta, T) \leq \beta(t,\delta)) &\leq \max(\sqrt{T}, \beta(T,\delta)) + \sum_{t=T_\epsilon}^{\top} \mathbb{1}\left(t\frac{C_{\epsilon,\hat{\mathcal{F}}}}{1+\psi} \leq \beta(t,\delta)\right) \\
&\leq \max(\sqrt{T}, \beta(T,\delta)) + \sum_{t=T_\epsilon}^{\top} \mathbb{1}\left(t\frac{C_{\epsilon,\hat{\mathcal{F}}}}{1+\psi} \leq \beta(T,\delta)\right) \\
&\leq \max(\sqrt{T}, \beta(T,\delta)) + \frac{\beta(T,\delta)(1+\psi)}{C_{\epsilon,\hat{\mathcal{F}}}}
\end{aligned}$$

Let us define $T_\delta \triangleq \inf\{T \in \mathbb{N} : \max(\sqrt{T}, f(T,\delta)) + \frac{\beta(T,\delta)(1+\psi)}{C_{\epsilon,\hat{\mathcal{F}}}} \leq T\}$. To find a lower bound on T_δ, we refer to (Garivier and Kaufmann, 2016). Specifically, let us define $c(\eta) \triangleq \inf\{T : T - \max(\sqrt{T}, \beta(T,\delta)) \geq \frac{T}{1+\eta}\}$ for some $\eta > 0$. Therefore,

$$T_\delta \leq c(\eta) + \inf\{T \in \mathbb{N} : T\frac{C_{\epsilon,\hat{\mathcal{F}}}}{(1+\psi_{\hat{\mathcal{F}}})(1+\eta)} \geq \beta(T,\delta)\}$$

Thus, finally combining the results, we upper bound the stopping time as

$$\mathbb{E}[\tau_\delta] \leq T_\epsilon + T_\delta + T_{\text{bad}}\,.$$

Here, $T_{\text{bad}} = \sum_{t=1}^{\infty} BT\exp\left(-CT^{1/8}\right) + K\zeta(1) < \infty$ is the sum of probability of the bad events over time. $\zeta(.)$ denotes the Euler-Riemann Zeta function.

Step 5. Deriving the asymptotics. Now, we leverage the continuity properties of the Lagrangian characteristic time under approximate constraint to show that we converge to traditional hardness measures as ϵ and δ tends to zero.

First, we observe that for some $\alpha > 1$

$$\limsup_{\delta\to 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq \frac{\alpha(1+\psi_{\hat{\mathcal{F}}})(1+\eta)}{C_{\epsilon,\hat{\mathcal{F}}}}$$

Now, if also $\epsilon \to 0$, by the Equation 3, we get $\tilde{A} \to A$, and thus, $\hat{\mathcal{F}} \to F$.

Thus, by continuity properties in Theorem 3 and Theorem 2, we get that

$$\mathcal{D}(\hat{\mathcal{F}}, .) \to \mathcal{D}(\mathcal{F}, .) \text{ and } \psi_{\hat{\mathcal{F}}} \to \frac{\max_{i\in[1,N]}\Gamma}{\min_{i\in[1,N]}\Gamma} \triangleq \frac{\Gamma_{\max}}{\Gamma_{\min}} \triangleq \mathfrak{s}\,.$$

Here, $\mathfrak{s}$ is the shadow price of the true constraint matrix, and quantifies the change in the constraint values due to one unit change in the policy vector.

Hence, we conclude that

$$\lim_{\delta\to 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq \alpha T_{\mathcal{F},r}(\boldsymbol{\mu})(1+\mathfrak{s})\,.$$

□

G.4 Upper Bound for LAGEX

Theorem 7. *The expected sample complexity of LAGEX* $\lim_{\delta\to 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq T_{\mathcal{F},r}(\boldsymbol{\mu})$.

Proof. We will do this proof in two parts. In part (a) we will assume that the current recommended policy is the correct policy and try to find an upper bound on the sample complexity of LAGEX. In the next part (b) we break that assumption and try to get an upper bound on the number of steps the recommended policy is not the correct policy.

Part (a) : Current recommended policy is correct. Proof structure of this part involves several steps. We start with defining the good event where we introduce a new event associated with the concentration event of the constraint set, then proceeding to prove concentration on that good event. Third step starts with the stopping criterion explained in G.2. In step 4 we define LAGEX as an approximate saddle point algorithm. The next step further transforms the stopping criterion with the help of allocation and instance player's regret that play the zero-sum game. We conclude with the asymptotic upper on the sample complexity characterised by the additive effect of the novel quantity shadow price $\mathfrak{s}$.

Step 1: Defining the good event. We start the proof first by defining the good event as

$$G_t = \{\forall t \leq T, \forall a \in [K] : N_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\mu}_a) \leq g(t) \wedge \|(\tilde{A}_t - A)\boldsymbol{\omega}\|_\infty \leq \rho(t,\delta)\}$$

where, $g(t) = 3\log t + \log\log t$ and $\rho(t,\delta)$ is defined in Lemma 3. The choice of $g(t)$ is motivated from (Degenne et al., 2019a) which originates from the negative branch of the Lambert's W function. This eventually helps us upper bounding the cumulative probability of the bad event.

Step 2: Concentrating to the good event We denote G_t^c as the bad event where any one of the above events does not occur. Cumulative probability of this bad event

$$\sum_{s=1}^{\top} \mathbb{P}\left(G_t^c\right) = \sum_{s=1}^{\top} \mathbb{P}\left(\sum_{a=1}^{K} N_{a,s} d(\hat{\boldsymbol{\mu}}_{a,s}, \boldsymbol{\mu}_a) > g(s)\right) + \sum_{s=1}^{\top} \mathbb{P}\left(\|(\tilde{A}_s - A)\boldsymbol{\omega}\|_\infty > \rho(s,\delta) \forall \boldsymbol{\omega} \in \Delta_K\right)$$

We get the upper bound on

$$\sum_{s=1}^{\top} \mathbb{P}\left(\sum_{a=1}^{K} N_{a,s} d(\hat{\boldsymbol{\mu}}_{a,s}, \boldsymbol{\mu}_a) > g(s)\right) \leq 1 + \sum_{s=2}^{\infty} \frac{\exp(2)}{s^3 \log s}(g(s) + g^2(s)\log s) \leq 19.48$$

as a direct consequence of (Lemma 6 of Degenne et al. (2019a)). The second cumulative probability is bounded by $\zeta(1) < 0.578$ using Lemma 4, which is finite.

In the next step, we work with the stopping criterion where we do not have access to $\mathcal{F}$ rather a bigger feasible set $\hat{\mathcal{F}}_t$.

Step 3: Working with the stopping criterion. The stopping criterion implies that

$$\beta(t,\delta) \geq \max_{\boldsymbol{\pi} \in \Pi^r_{\hat{\mathcal{F}}_t}} \inf_{\boldsymbol{\lambda} \in \Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^{K} N_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\lambda}_a)\,,$$

where the exact expression of $\beta(t,\delta)$ is defined in Theorem 5.

We use the C-tracking lemma (Lemma 8) to express the stopping in terms of allocations

$$\beta(t,\delta) \geq \max_{\boldsymbol{\pi} \in \Pi^r_{\hat{\mathcal{F}}_t}} \inf_{\boldsymbol{\lambda} \in \Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)} \sum_{s=1}^{t} \sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\lambda}_a) - (1+\sqrt{t})K \tag{18}$$

L-Lipschitz property of KL divergence gives

$$|d(\boldsymbol{\mu}_a, \boldsymbol{\lambda}_a) - d(\hat{\boldsymbol{\mu}}_{a,s}, \boldsymbol{\lambda}_a)| \leq L\sqrt{2\sigma^2 \frac{g(s)}{N_{a,s}}} \tag{19}$$

Using this result in Equation (18) we get

$$\sum_{s=1}^{t} \sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} d(\hat{\boldsymbol{\mu}}_{a,s}, \boldsymbol{\lambda}_a) \geq \sum_{s=1}^{t} \sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} d(\hat{\boldsymbol{\mu}}_a, \boldsymbol{\lambda}_a) - L\sqrt{2\sigma^2 K t g(t)}$$

$$\geq \sum_{s=1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} d(\hat{\boldsymbol{\mu}}_{a,s}, \boldsymbol{\lambda}_a) - L\sqrt{2\sigma^2 Ktg(t)} - 2L\sqrt{2\sigma^2 g(t)}(K^2 + 2\sqrt{2Kt})$$
$$\geq \sum_{s=1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} d(\hat{\boldsymbol{\mu}}_{a,s}, \boldsymbol{\lambda}_a) - \mathcal{O}(\sqrt{t\log t})$$

the penultimate inequality yields from using the Equation (19). Using this result in Equation (18)

$$\beta(t,\delta) \geq \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}_t}} \inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\lambda}_a) - (1+\sqrt{t})K - \mathcal{O}(\sqrt{t\log t}) \tag{20}$$

Step 4: LAGEX (Algorithm 3) as an optimistic saddle point algorithm We follow the definition of approximate saddle point algorithm in (Degenne et al., 2019a). LAGEX acts as an approximate saddle point algorithm if

$$\max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}_t}} \inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)} \sum_{s=1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_{s,a} d(\hat{\boldsymbol{\mu}}_{a,s}, \boldsymbol{\lambda}_a)$$
$$\geq \max_{\boldsymbol{\omega}\in\hat{\mathcal{F}}_t} \sum_{a=1}^{K}\sum_{s=1}^{t} \boldsymbol{\omega}_a U_{a,s} - R_t^{\boldsymbol{\omega}} - \sum_{s=1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} C_{a,s} + \boldsymbol{l}_t^\top \tilde{A}_t \boldsymbol{\omega}_t\,, \tag{21}$$

where $U_{a,s} \triangleq \max\left\{\frac{g(s)}{N_{a,s}}, \max_{\xi\in[\alpha_{a,s},\beta_{a,s}]} d(\xi,\lambda_{a,s})\right\}$, $C_{a,s} \triangleq U_{a,s} - d(\hat{\boldsymbol{\mu}}_{a,s}, \lambda_{a,s})$, and $R_t^{\boldsymbol{\omega}}$ is defined in Equation (23).

Step 5: Bounds cumulative regret of players Algorithm 3 at each step solves a two player zero-sum game. First one is the allocation player who uses AdaGrad to maximize the inverse of characteristic time function to find the optimal allocation.

Step 5a. $\boldsymbol{\lambda}$-player's regret.

$$R^t_{\boldsymbol{\lambda}} \triangleq \sum_{s=1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} d(\hat{\boldsymbol{\mu}}_{a,s}, \lambda_{a,s}) - \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}_t}} \inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)} \sum_{s=1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} d(\hat{\boldsymbol{\mu}}_{a,s}, \boldsymbol{\lambda}_a) \leq 0$$

The last inequality holds because we take infimum over $\boldsymbol{\lambda}$ in the perturbed alt-set. now let us prove that LAGEX is a optimistic saddle point algorithm. From the definition of regret of the $\boldsymbol{\lambda}$-player we get

$$\max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}_t}} \inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)} \sum_{s=1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} d(\hat{\boldsymbol{\mu}}_{a,s}, \boldsymbol{\lambda}_a) \geq \sum_{s=1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} U_{a,s} - \sum_{s=1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} C_{a,s}\,.$$

Then we have from Equation (20) and Equation (21)

$$\beta(t,\delta) \geq \sum_{s=1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} U_{a,s} - \sum_{s=1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} C_{a,s} - (1+\sqrt{t})K - \mathcal{O}(\sqrt{t\log t})\,, \tag{22}$$

Step 5b. Allocation player's regret.

$$R_t^{\boldsymbol{\omega}} \triangleq \max_{\boldsymbol{\omega}\in\Delta_K} \sum_{s=1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_a U_{a,s} - \sum_{s=1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} U_{a,s} \tag{23}$$

We note that AdaGrad enjoys regret of order $R_t^{\boldsymbol{\omega}} \leq \mathcal{O}(\sqrt{Qt})$ where Q is an upper bound on the losses such that $Q \geq \max_{x,y\in[\boldsymbol{\mu}_{\min},\boldsymbol{\mu}_{\max}]} d(x,y)$.

Now, from the allocation player's regret, we have

$$\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})} \sum_{s=1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} d(\hat{\boldsymbol{\mu}}_{a,s}, \lambda_s) \geq \max_{\boldsymbol{\omega}\in\hat{\mathcal{F}}} \sum_{s=1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_a U_{a,s} - R_t^{\boldsymbol{\omega}} - \underbrace{\sum_{s=1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} C_{a,s}}_{T1} + \boldsymbol{l}_t^\top \tilde{A}_t \boldsymbol{\omega}_t$$

$$\geq \max_{\boldsymbol{\omega}\in\hat{\mathcal{F}}} \sum_{s=1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_a U_{a,s} - \mathcal{O}(\sqrt{Qt}) - \underbrace{\sum_{s=1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} C_{a,s}}_{T1} + \boldsymbol{l}_t^\top \tilde{A}_t \boldsymbol{\omega}_t$$

which shows that LAGEX is a approximate saddle point algorithm with a slack $R_t^{\boldsymbol{\omega}} + \sum_{s=1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} C_{a,s} - \boldsymbol{l}_t \tilde{A}_t \boldsymbol{\omega}_t$.

Step 5c. Bounding T1. Now we have to ensure that $T1$ in the slack is bounded.

$$\begin{aligned}
\sum_{s=K+1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} C_{a,s} &\leq \sum_{s=K+1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} \left(\frac{g(s)}{N_{a,s}} + \max_{\xi\in[\alpha_{a,s},\beta_{a,s}]} d(\xi, \lambda_{a,s}) - d(\hat{\boldsymbol{\mu}}_{a,s}, \lambda_{a,s}) \right) \\
&\leq \sum_{s=K+1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} \left\{ \frac{g(s)}{N_{a,s}} + L\sqrt{2\sigma^2 \frac{g(s)}{N_{a,s}}} \right\} \qquad (24) \\
&\leq g(t) \sum_{s=K+1}^{t}\sum_{a=1}^{K} \frac{\boldsymbol{\omega}_{a,s}}{N_{a,s}} + L\sqrt{2\sigma^2 g(t)} \sum_{s=K+1}^{t}\sum_{a=1}^{K} \frac{\boldsymbol{\omega}_{a,s}}{\sqrt{N_{a,s}}} \\
&\leq g(t) \left(K^2 + 2K \log \frac{t}{K} \right) + L\sqrt{2\sigma^2 g(t)} (K^2 + 2\sqrt{2Kt}) \\
&\leq \mathcal{O}(\sqrt{t \log t}) \,. \qquad (25)
\end{aligned}$$

Equation (24) holds due to L-Lipschitz property of KL under the good event G_T. Specifically,

$$|d(\boldsymbol{\mu}_a, \boldsymbol{\lambda}_a) - d(\hat{\boldsymbol{\mu}}_{a,s}, \boldsymbol{\lambda}_a)| \leq L\sqrt{2\sigma^2 \frac{g(s)}{N_{a,s}}} \,.$$

Step 6. From the two-player regrets to stopping time. Finally, combining the results in Step 5, i.e. Equation (22) and Equation (25), we get

$$\begin{aligned}
\beta(t,\delta) &\geq \max_{\boldsymbol{\omega}\in\Delta_K} \left(\sum_{a=1}^{K}\sum_{s=1}^{t} \boldsymbol{\omega}_a d(\boldsymbol{\mu}_a, \lambda_{a,s}) \right) - \mathcal{O}(\sqrt{Qt}) - \mathcal{O}(\sqrt{t \log t}) - (1+\sqrt{t})K + \boldsymbol{l}_t^\top \tilde{A}_t \boldsymbol{\omega}_t \\
&\geq \max_{\boldsymbol{\omega}\in\Delta_K} \sum_{a=1}^{K}\sum_{s=1}^{t} \boldsymbol{\omega}_a d(\boldsymbol{\mu}_a, \lambda_{a,s}) - \mathcal{O}(\sqrt{Qt}) - \mathcal{O}(\sqrt{t \log t}) - (1+\sqrt{t})K + \boldsymbol{l}_t^\top \tilde{A}_t \boldsymbol{\omega}_t \\
&\geq \max_{\boldsymbol{\omega}\in\Delta_K} \sum_{a=1}^{K}\sum_{s=1}^{t} \boldsymbol{\omega}_a d(\boldsymbol{\mu}_a, \lambda_{a,s}) - \mathcal{O}(\sqrt{Qt}) - \mathcal{O}(\sqrt{t \log t}) - (1+\sqrt{t})K - \frac{\psi_t}{T_{\hat{\mathcal{F}}_t}(\boldsymbol{\mu})} \,. \qquad (26)
\end{aligned}$$

The last inequality is true due to Lemma 6, i.e. $-\boldsymbol{l}_t^\top \tilde{A}_t \boldsymbol{\omega}_t \leq \frac{\psi_t}{T_{\hat{\mathcal{F}}_t}(\boldsymbol{\mu})}$.

Now, we observe that

$$\max_{\boldsymbol{\omega}\in\Delta_K} \sum_{s=1}^{t}\sum_{a=1}^{K} \boldsymbol{\omega}_a d(\boldsymbol{\mu}_a, \lambda_{a,s}) \geq t \max_{\boldsymbol{\omega}\in\Delta_K} \max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}_t}} \inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^{K} \boldsymbol{\omega}_a d(\boldsymbol{\mu}_a, \boldsymbol{\lambda}_a) = t T^{-1}_{\hat{\mathcal{F}}_t}(\boldsymbol{\mu})$$

Replacing this inequality in Equation (26), we finally obtain

$$t \leq T_{\hat{\mathcal{F}}_t}(\boldsymbol{\mu}) \left(\beta(t,\delta) + R_t^{\boldsymbol{\omega}} + \mathcal{O}(\sqrt{t \log t}) \right) + \psi_t$$

Part (b) : Current recommended policy is wrong (not r-good). To get on with the proof for this part we will use similar argument as (Carlsson et al., 2024). Though the argument was motivated by the work (Degenne et al., 2019b). We define the event

$$B_t \triangleq \left\{ \boldsymbol{\pi}^\star \triangleq \arg\max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}_t}} \hat{\boldsymbol{\mu}}_t^\top \boldsymbol{\pi} \notin \Pi^r_{\mathcal{F}} \right\}$$

i.e the current recommendation policy is not correct which implies that the mean estimate or the constraint estimate has not been concentrated yet. If we define Chernoff's information function as $\mathrm{ch}(u,v) \triangleq \inf_{z\in\mathcal{D}}(d(u,x)+d(v,x))$. Therefore the current mean estimate will yield positive Chernoff's information since it has not been converged yet i.e $\exists\epsilon>0 : \mathrm{ch}(\hat{\boldsymbol{\mu}}_{a,t},\boldsymbol{\mu}_a) > \epsilon$. Consequently under the good event G_T defined earlier

$$\frac{g(t)}{N_{a,t}} \leq \epsilon$$

since $\mathrm{ch}(\hat{\boldsymbol{\mu}}_{a,t},\boldsymbol{\mu}_a) \leq d(\hat{\boldsymbol{\mu}}_{a,t},\boldsymbol{\mu}_a)$. At time $s\in\mathbb{N}$, let $\boldsymbol{\pi'}$ be an extreme point in $\hat{\mathcal{F}}_s$ that is not the r-good optimal policy. But since it is an extreme point in $\hat{\mathcal{F}}_s$ that shares $(K-1)$ active constraints with $\boldsymbol{\pi}^\star$, it has to be an r-good optimal policy w.r.t some $\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_s}(\hat{\boldsymbol{\mu}}_s,\boldsymbol{\pi}) : \boldsymbol{\pi'} = \arg\max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}_s}} \boldsymbol{\lambda}^\top\boldsymbol{\pi}$. So we again define the event B_t as

$$B_t \triangleq \left\{ \boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t) : \boldsymbol{\pi'} = \arg\max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}_t}} \boldsymbol{\lambda}^\top \boldsymbol{\pi} \notin \Pi^r_{\mathcal{F}} \right\}$$

We again define $n_{\boldsymbol{\pi'}}(t)$ be the number of steps when $\boldsymbol{\pi}_s = \boldsymbol{\pi'}, s\in[t]$. Therefore

$$\epsilon = \sum_{s=1,B_s}\sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} d(\hat{\boldsymbol{\mu}}_{a,s},\mu_a) \geq \sum_{\boldsymbol{\pi'}\notin\Pi^r_{\mathcal{F}}} \inf_{\boldsymbol{\lambda}\in B_s} \sum_{s=1,\boldsymbol{\pi}_s=\boldsymbol{\pi'}}^{t} \sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} d(\hat{\boldsymbol{\mu}}_{a,s},\mu_a) \tag{27}$$

Now to break the RHS of the above inequality we go back to step 5 of the proof of part (a) where we showed LAGEX is an approximate saddle point algorithm. In this case the slack will be $x_t = R^{\boldsymbol{\omega}}_{n_{\boldsymbol{\pi'}}(t)} + \sum_{s=1,\boldsymbol{\pi}_s=\boldsymbol{\pi'}}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}C_{a,s} - \boldsymbol{l}_t\tilde{A}_t\boldsymbol{\omega}_t$. Therefore we can write the RHS of Equation (27) as

$$\begin{aligned}
&\sum_{\boldsymbol{\pi'}\notin\Pi^r_{\mathcal{F}}} \inf_{\boldsymbol{\lambda}\in B_s} \sum_{s=1,\boldsymbol{\pi}_s=\boldsymbol{\pi'}}^{t} \sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} d(\hat{\boldsymbol{\mu}}_{a,s},\mu_a) \\
&\geq \max_{\boldsymbol{\pi}\in\hat{\mathcal{F}}_s} \underbrace{\sum_{s=1,\boldsymbol{\pi}_s=\boldsymbol{\pi'}}^{t} \sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} U_{a,s} - R^{\boldsymbol{\omega}}_{n_{\boldsymbol{\pi'}}(t)}}_{T1} - \underbrace{\sum_{s=1,\boldsymbol{\pi}_s=\boldsymbol{\pi'}}^{t} \sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} C_{a,s}}_{T2} + \boldsymbol{l}_t^\top \tilde{A}_t \boldsymbol{\omega}_t \,.
\end{aligned} \tag{28}$$

We apply the same logic as in (Carlsson et al., 2024) and (Degenne et al., 2019b) that $\exists a'\in[K]$ for which $U_{a',s}\geq\epsilon$. That means the term $T1$ grows at most linear with $n_{\boldsymbol{\pi'}}(t)$. From the proof of part (a) it is clear that the term $T2 = \mathcal{O}\left(\sqrt{n_{\boldsymbol{\pi'}}(t)\log n_{\boldsymbol{\pi'}}(t)}\right) \leq \mathcal{O}(\sqrt{t\log t})$ and the allocation player regret is bounded by $R_{n_{\boldsymbol{\pi'}}(t)} = \mathcal{O}(\sqrt{Qn_{\boldsymbol{\pi'}}(t)}) \leq \mathcal{O}(\sqrt{Qt})$. That means the number of times the event B_t occurs is upper bounded by $\mathcal{O}(\sqrt{t\log t})$. Now the extra term in Equation (28) appearing with term $T1$ and $T2$ induces same implication as part (a).

Then incorporating part (a) and (b) to get the upper bound on the expected stopping time asymptotically

$$\mathbb{E}[\tau_\delta] \leq T_{\hat{\mathcal{F}}_t}(\boldsymbol{\mu})\left(\beta(t,\delta) + \mathcal{O}\left(\sqrt{t\log t}\right) + \mathcal{O}\left(\sqrt{Qt}\right)\right) + 2\psi_t$$

as $\hat{\mathcal{F}}_t \to \mathcal{F}$ and $\psi_t \to \mathfrak{s}$, we have the explicit upper bound

$$\mathbb{E}[\tau] \leq T_0(\delta) + 2\mathfrak{s} + CK \tag{29}$$

where $T_0(\delta) \triangleq \max\left\{t\in\mathbb{N} : t \leq T_{\mathcal{F}}(\boldsymbol{\mu})\left(\beta(t,\delta) + \mathcal{O}\left(\sqrt{t\log t}\right) + \mathcal{O}\left(\sqrt{Qt}\right)\right)\right\}$ and C is a problem independent constant that appears due to the concentration guaranty to good event G_t i.e concentration of mean vector and constraint matrix. Thus, dividing both sides by $\log(\frac{1}{\delta})$ and taking the limit $\delta\to 0$ we conclude the proof. Note that, we do not see the effect of not knowing the constraints in the asymptotic guaranty, though in non-asymptotic sense in Equation (29), we see an additive effect of the shadow price on the sample complexity. □

G.5 Applications to Existing Problems

End-of-time Knapsack. We can model the BAI problem with end-of-time knapsack constraints as discussed in Section B.2. In such a setting the shadow price comes out to be $\mathfrak{s} \leq c$ i.e the maximum consumable resource. So if we were to implement LATS for this the asymptotic sample complexity upper bound will translate to $\alpha T_{\hat{\mathcal{F}}}(\boldsymbol{\mu})(1+c)$, the multiplicative part being the effect of the end of time knapsack constraint. In case of LAGEX the unknown knapsack constraint will leave a additive effect quantified by $2c$. Recently people have deviated from only devising a no-regret learner in BwK rather people are interested to also give good sub-optimal guaranties on constraint violation as well. We think algorithms like LAGEX will perform well if we translate this model to our setting since it has shown not only good sample complexity but also better constraint violation guaranies as well.

Fair BAI across subpopulations. This problem is a direct consequence of our setting. The shadow price in this setting $\mathfrak{s} = \frac{\max_{i \in K} \boldsymbol{\pi}^\star_i}{\min_{i \in K} \boldsymbol{\pi}^\star_i} > 1$ i.e the ratio between maximum and minimum non-zero weight in the recommended policy. Similar to the knapsack scenario, here also this ratio will appear as a extra cost of not knowing the fairness constraint in multiplicative way in case of LATS and gets added to the sample complexity upper bound of LAGEX.

Pure exploration with Fairness of exposure. We can think of a problem where we want to select a pool of employees from different sub-sections of a whole population for a task. As we want to maximise the reward or utility of this selected group we also must also give fair exposure to all race or say gender. As discussed earlier in Section B.2, a direct application of our algorithms LATS and LAGEX to use them in the problem of pure exploration with unknown constraints on fairness of exposure. The shadow price in such a setting would be $\mathfrak{s} = \frac{\max_{i,j \in [K]} \left(\frac{1}{\mu_i} - \frac{1}{\mu_j}\right)}{\min_{i,j \in [K]} \left(\frac{1}{\mu_i} - \frac{1}{\mu_j}\right)} = \frac{\max_{i,j \in [K]} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{\min_{i,j \in [K]} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)} \geq 1$.

Thresholding bandits. The problem of Thresholding bandit is motivated from the safe dose finding problem in clinical trials, where one wants to identify the highest dose of a drug that is below a known safety level. From the translated optimisation problem in Section B.2 we easily find out the shadow price for this setting to be $\mathfrak{s} = \frac{\max_{i \in [K]} (\boldsymbol{\pi} - \boldsymbol{\theta})^i}{\min_{i \in [K]} (\boldsymbol{\pi} - \boldsymbol{\theta})^i} \geq 1$. This shadow price is similar to ours because the constraint structure is very similar. Our setting generalises thresholding bandit problem by giving the liberty of choosing different threshold levels for different support index of $\boldsymbol{\pi}$. Similarly to other settings this shadow price will come as a price of handling different unknown thresholds for every arm as addition in case of LAGEX and as multiplication for LATS.

Feasible arm selection. Feasible arm selection problem is motivated by the spirit of recommending an optimal arm which should satisfy a performance threshold. For example one might be interested to find a combination of food among a plethora of options which maximises the nutrient intake, rather the nutrient value of the food combination should exceed a threshold value. The structure of the optimisation problem for such a setting is discussed in detail in Section B.2. Then the shadow price comes out as $\mathfrak{s} = \frac{\tau - f_{\min}}{\tau - f_{\max}} \geq 1$ where $f \in \mathbb{R}^{\mathrm{Supp}(\boldsymbol{\pi})}$ can be compared to a utility function. In our setting we will not have access to the true utility function rather we have to track it per step. This shadow price again get multiplied to the LATS sample complexity upper bound as a cost of not knowing the true utility of the arms, whereas we see a additive cost incurred in case of LAGEX.

H Constraint Violations During Exploration

H.1 Upper Bound on Constraint Violation

In a linear programming problem we say constraint is violated if the chosen allocation fails to satisfy any of the true linear constraints. In other words when the event $A\boldsymbol{\omega}_t \geq 0$. We start with the optimization problem relaxed with slack if the constraints were known,

$$\max_{\boldsymbol{\pi}\in\Pi^r_{\hat{\mathcal{F}}}} \boldsymbol{\mu}^\top\boldsymbol{\pi}$$
$$\text{such that, } A\boldsymbol{\pi} + \Gamma \leq 0$$

where, Γ is the slack. Cumulative violation of constraints can be expressed as $\mathcal{V}_t = \sum_{s=1}^{t} \max_{i\in[K]} \left[A^i\boldsymbol{\omega}_t\right]_+$ where $[z]_+ = \max\{z, 0\}$. Then, at any time step $t \in [T]$, instantaneous violation is given by, $v_t = \max_{i\in[k]} \left[A^i\boldsymbol{\omega}_t\right]_+$. Since, A is feasible, we define the game value as, $\eta = \max_{\boldsymbol{\omega}\in\mathcal{F}} \max_{i\in[K]} A^i\boldsymbol{\omega} \leq -\Gamma$ and $\eta = \max_{\boldsymbol{\omega}\in\mathcal{F}} \max_{i\in[K]} A^i\boldsymbol{\omega} \geq \min_{\tilde{A}\in\mathcal{C}^t_A} \max_{\boldsymbol{\omega}\in\hat{\mathcal{F}}_t} \max_{i\in[K]} \tilde{A}^i_t\boldsymbol{\omega}_t = \tilde{A}^{i_t}_t\boldsymbol{\omega}_t$, holds due to optimistic estimate of A.

Again, we define, $i_{\min}(\boldsymbol{\omega}) = \arg\min_{i\in[K]} \tilde{A}^i\boldsymbol{\omega}$.

$$\begin{aligned}
\max_{i\in[K]} \left[A^i\boldsymbol{\omega}_t\right] &= \max_{i\in[K]}(A^i - \tilde{A}^i_t)\boldsymbol{\omega}_t + \max_{i\in[K]} \tilde{A}^i_t\boldsymbol{\omega}_t \\
&\leq \max_{i\in[K]} ||(A^i - \tilde{A}^i_t)||_{\Sigma_t} ||\boldsymbol{\omega}_t||_{\Sigma_t^{-1}} + \max_{i\in[K]} \tilde{A}^i_t\boldsymbol{\omega}_t \\
&\leq f(t,\delta)||\boldsymbol{\omega}_t||_{\Sigma_t^{-1}} + \max_{i\in[K]} \tilde{A}^i_t\boldsymbol{\omega}_t \leq \rho(t,\delta) - \Gamma
\end{aligned}$$

where the last inequality holds because $\max_{i\in[K]} \tilde{A}^i_t\boldsymbol{\omega}_t \leq \max_{i\in[K]} A^i\boldsymbol{\omega}_t \leq -\Gamma$

Let stopping time is denoted by $\tau_\delta < T$ following the expression from the Stopping criterion section. Then expected cumulative constraint violation is denoted by,

$$\begin{aligned}
\mathbf{E}\Big[\frac{\mathcal{V}_{\tau_\delta}}{\tau_\delta}\Big] = \sum_{t\leq\mathbf{E}[\tau_\delta]} s_t \leq \sum_{t\leq\mathbf{E}[\tau_\delta]} [\rho(t,\delta) - \Gamma]_+ &= \sum_{t\leq\mathbf{E}[\tau_\delta]} \{\rho(t,\delta) - \Gamma\} \quad \text{because } \rho(t,\delta) - \Gamma \text{ is positive.} \\
&\leq \sqrt{2df^2(\mathbf{E}[\tau_\delta],\delta)\log\left(1 + \frac{1+\mathbf{E}[\tau_\delta]}{d}\right)} - \Gamma
\end{aligned}$$

The last inequality is a direct consequence of Lemma 3. Thus the expected constraint violation is of order $\tilde{\mathcal{O}}\left(t\sqrt{d}\right)$, ignoring the lower order dependence on t.

H.2 Experimental Results on Constraint Violation

We compare the constraint violation i.e the number of times $A\boldsymbol{\omega}_t > 0$, where A is the true constraint matrix.

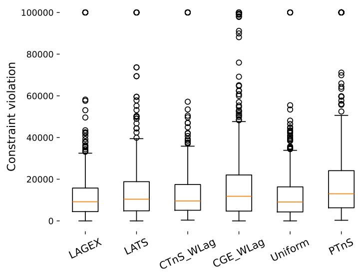

Figure 8: Constraint violation (median±std.) algorithms for **hard environment in setup 1**.

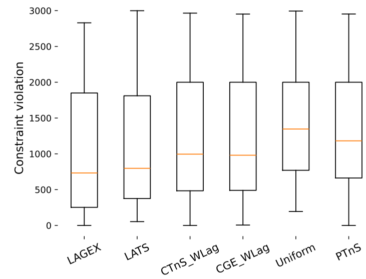

Figure 9: Constraint violation (median±std.) of algorithms for **easy environment in setup 1**.

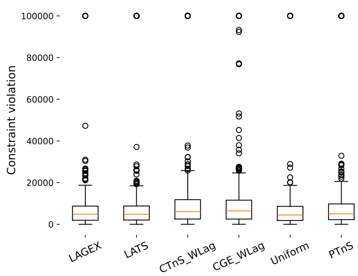

Figure 10: Constraint violation (median±std.) algorithms for **hard environment in setup 2**.

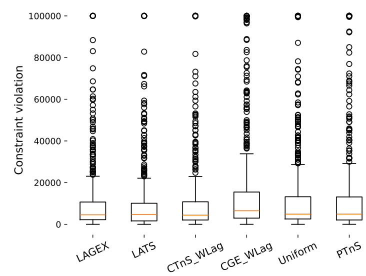

Figure 11: Constraint violation (median±std.) of algorithms for **easy environment in setup 2**.

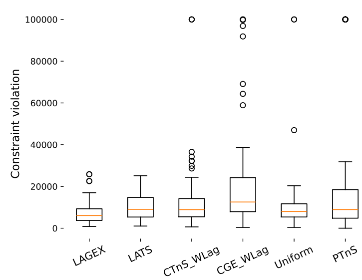

Figure 12: Constraint violation (median±std.) of algorithms for **IMDB environment**.

Observation: LAGEX is the "safest" among the existing baselines.

Setup 1. From Figure 8 unlike LAGEX, we see LATS shows comparable amount of constraint violation with Uniform sampling, whereas in Figure 9 we prominently observe LATS and LAGEX outperforming other algorithms in safety.

Setup 2. In the **hard env** as per Carlsson et al. (2024) in Figure 10 we observe both LAGEX and LATS show constraints violation as low as Uniform sampling outperforming other algorithms, but in the **easy env** in Figure 11 we see all the algorithms except CGE without Lagrangian relaxation shows more or less same amount of safety violation while choosing allocation.

IMDB env. During the experiment with IMDB-50K dataset wee see that LAGEX proves to be the safest among all other algorithms with a distinguishably low safety violations. Though again LATS shows comparable violation with Uniform sampling.

Therefore, we see that across multiple environments LAGEX is the "safest" out of all the competing algorithms.

I Experimental Details

Computation resources. We run the algorithms on a 64-bit 13th Gen Intel® Core™ i7-1370P × 20 processor machine with 32GB ram with a disc capability of 1TB and graphics Mesa Intel® Graphics (RPL-P). The hardware model is HP EliteBook 840 14 inch G10 Notebook PC.

Throughout this paper we have set $r = 0.01$ and $\beta(t, \delta) = \log \frac{1+\log \log t}{\delta}$ for all the experiments. For CGE we have used $g(t) = \log t$. Each plot has been generated over 500 random seeds.

Baseline algorithms. We compare LAGEX and LATS with the two algorithms under known constraints, i.e. CTnS and CGE (Carlsson et al., 2024). Also, to understand the utility of Lagrangian relaxation, we implement versions of CTnS and CGE with estimated constraints. In these variants, we solve the constrained optimisation problems without Lagrangian relaxation but with estimated constraints. We also compare with PTnS (Projected Track and Stop), a variant of TnS, where the algorithm computes the allocation by solving the unconstrained BAI optimisation problem and projects this allocation in the feasible set as necessary.

J Discussion on Impossibility Result

Suppose that we are given a bandit instance $\boldsymbol{\nu} \triangleq (P_a, C_a)_{a=1}^K$ with $\mathbb{E}_{R\sim P_a}[R] \in \mathbb{R}$ and $A_a \triangleq \mathbb{E}_{C\sim C_a}[C] \in \mathbb{R}^d$. Note, P_a and C_a denote reward and cost distributions for arm $a \in [K]$, respectively. Here, we follow our paper for the definition of true optimal policy $\boldsymbol{\pi}^\star_{\mathcal{F}}$. In bandit literature, we call another bandit instance $\boldsymbol{\lambda} \triangleq (P'_a, C'_a)_{a=1}^K$ an alternative instance for $\boldsymbol{\nu}$, if its optimal policy is different than $\boldsymbol{\pi}^\star_{\mathcal{F}}$. Thus, complexity of the inner optimisation problem, in most cases translates to indistinguishability of $\boldsymbol{\nu}$ and $\boldsymbol{\lambda}$, i.e.,

$$\mathcal{D}(\omega, \boldsymbol{\mu}, \mathcal{F}) \propto \mathrm{KL}\left((P_a, C_a) \mid (P'_a, C'_a)\right).$$

In the true environment $\boldsymbol{\nu}$, let $\exists$ at least one tight constraint at the optimal policy. That means $\exists i \in [d]$:

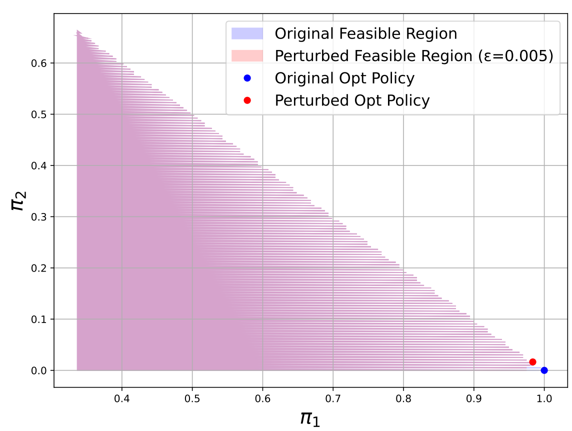

Figure 13: Effect of rank-one update of active constraint

$A_i^\top \boldsymbol{\pi}^\star_{\mathcal{F}} = 0$. Now, if we perturb each element of this tight constraint by some $\epsilon > 0$ as $A'_i \leftarrow A_i + \epsilon \mathbf{1}_K$. It produces an alternative instance $(\boldsymbol{\mu}_a, A'_a)_{a=1}^K$, since under this perturbed constraint set, $\boldsymbol{\pi}^\star_{\mathcal{F}}$ is no longer feasible, thus cannot serve as an candidate optimal policy. Hence, for even simple multivariate Gaussian cost vectors i.e., $C_a \triangleq \mathcal{N}_d\left(A_a, \text{ Variance }\right)$ and $C'_a \triangleq \mathcal{N}_d\left(A'_a, \text{Variance}\right)$, we have

$$\mathcal{D}(\omega, \boldsymbol{\mu}, \mathcal{F}) \propto \mathrm{KL}\left(C_a \mid C'_a\right) = \frac{\left(A'_a - A_a\right)^\top \left(A'_a - A_a\right)}{2 * \text{ Variance}} \propto \epsilon^2.$$

Finally, then the expected number of interaction satisfies: $\mathbb{E}\left[\tau_\delta\right] \propto \frac{1}{\epsilon^2}$

Now, ϵ can be arbitrarily small, any positive value of ϵ results in infeasibility of the true optimal policy $\boldsymbol{\pi}^\star_{\mathcal{F}}$. If we plug in $\epsilon = 0.005$, we get the complexity mentioned in Example of the paper. We chose 0.005 as the perturbation specifically, so that visually it is easy for a reader to be able to distinguish optimal policies (Figure 1).

K Technical Results: New and Known

In this section, we will devise some technical lemma using the help of standard text on online linear regression to ensure the convergence of unknown constraints. We specifically give the expression of the radius of confidence ellipsoid mentioned in the main text in Equation 2. We then prove an upper bound on the *bad event* i.e when the constraint matrix is not concentrated around the true matrix. We also acknowledge some known theoretical results from BAI and pure exploration literature that are used in this work.

K.1 Concentration Lemma for Constraints

Here, we want to get concentration on the deviation of the pessimistic estimate of the constraint matrix from the actual one quantified by $\|(\tilde{A} - A)\boldsymbol{\omega}\|_\infty$, it becomes very crucial to prove upper bounds on sample complexity of our proposed algorithms. The following lemma ensures the concentration of the constraint matrix.

Lemma 3. *For the pessimistic estimate $\tilde{A}$ of A, the following holds*

1. $|(\tilde{A}^i - A^i)\boldsymbol{\omega}| \leq \rho(t,\delta)$ *where* $\rho(t,\delta) \triangleq f(t,\delta)\|\boldsymbol{\omega}\|_{\Sigma_t^{-1}}$
2. $\sum_{s=1}^t \|\boldsymbol{\omega}_s\|^2_{\Sigma_s^{-1}} \leq 2d \log\left(1 + \frac{1+t}{d}\right)$
3. $\sum_{s=1}^t \rho(s,\delta) \leq t\sqrt{2df^2(t,\delta)\log\left(1+\frac{1+t}{d}\right)}$

Proof. The first result gives control on the deviations $(\tilde{A} - A)\boldsymbol{\omega}$ for $A \in \mathcal{C}_t^A(\delta)$. Then $\forall i \in [d]$

$$|(\tilde{A}^i - A^i)\boldsymbol{\omega}| \leq |(\tilde{A}^i_t - A^i)\boldsymbol{\omega}| \leq 2 \sup_{A \in \mathcal{C}_t^A(\delta)} \|\tilde{A}^i_t - A^i\|_{\Sigma_t} \|\boldsymbol{\omega}\|_{\Sigma_t^{-1}} \leq f(t,\delta)\|\boldsymbol{\omega}\|_{\Sigma_t^{-1}} \leq \rho(t,\delta)$$

here, we define $\rho(t,\delta) \triangleq f(t,\delta)\|\boldsymbol{\omega}\|_{\Sigma_t^{-1}}$. The penultimate inequality follows from the definition of the confidence set defined in Equation 2. Now we want to derive an explicit expression of this upper bound. It is natural to use the concentration of the gram matrix Σ_t over time. We refer to Abbasi-yadkori et al. (2011) for the control over the behaviour of Σ_t and we directly get the second as

$$\sum_{s=1}^t \|\boldsymbol{\omega}_s\|^2_{\Sigma_t^{-1}} \leq 2\log \mathsf{det}\Sigma_{t+1} \leq 2d\log\left(1 + \frac{1+t}{d}\right)$$

Refer Abbasi-yadkori et al. (2011) for the context. Now we have to control the cumulative deviation because later on when we define the bad event based on this concentration we will need to know the cumulative behavior of $\rho(t,\delta)$.

Then for an arbitrary sequence of actions $\{\boldsymbol{\omega}_s\}_{s\in[T]}$

$$\sum_{s=1}^t \rho(s,\delta) \leq t\sqrt{\sum_{s=1}^t \rho^2(s,\delta)} \leq \sqrt{2dtf^2(t,\delta)\log\left(1+\frac{1+t}{d}\right)}$$

where, $\sum_{s=1}^t \rho^2(s,\delta) \leq 2dtf^2(t,\delta)\left(1+\frac{1+t}{d}\right)$ using result 2 of this lemma. This holds because as we have already stated $\{f(s,\delta)\}_{s\in[T]}$ is a non-decreasing sequence of function and $f(t,\delta)$ is the maximum possible value in the set i.e $\sum_{s=1}^t f^2(s,\delta) \leq tf^2(t,\delta)$

□

Now we proceed to state an upper bound on the cumulative probability of the *bad event* i.e the event $|(\tilde{A}_t - A)\boldsymbol{\omega}|_\infty > |(\tilde{A}^i_t - A^i)\boldsymbol{\omega}| > \rho(t,\delta)$.

Lemma 4. *The cumulative probability of the bad event till time t¡T,*

$$\sum_{s=1}^t \mathbb{P}\left(\|(\tilde{A}_t - A)\boldsymbol{\omega}\|_\infty > \rho(t,\delta)\right) \leq \zeta(1)$$

where $\zeta(.)$ is the Euler-Riemann zeta function.

Proof. We already have stated in the main paper

$$f(t,\delta) = 1 + \sqrt{\frac{1}{2}\log\frac{K}{\delta} + \frac{1}{4}\log\det\Sigma_t} \leq 1 + \sqrt{\frac{1}{2}\log\frac{K}{\delta} + \frac{d}{4}\log\left(1+\frac{t}{d}\right)} \triangleq f'(t,\delta)$$

by Lemma 3 and we define $\rho'(t,\delta) \triangleq f'(t,\delta)\|\boldsymbol{\omega}\|_{\Sigma_t^{-1}}$. It implies $\mathbb{P}\left(\exists t\in[T]\|(\tilde{A}_t - A)\boldsymbol{\omega}\|_\infty > \rho'(t,\delta)\right) \leq \mathbb{P}\left(\exists t\in[T]\|(\tilde{A}_t - A)\boldsymbol{\omega}\|_\infty > \rho(t,\delta)\right) \leq \delta$. Now if we replace $\log\frac{1}{\delta}$ by u, we can write

$$\mathbb{P}\left(\exists t\in[T], \forall i\in[d]\|\|\tilde{A}_t^i - A^i\|_{\Sigma_t} > 1 + \sqrt{\frac{1}{2}\log K + \frac{u}{2} + \frac{d}{4}\log\left(1+\frac{t}{d}\right)}\right) \leq \exp(-u).$$

We can directly assign $\log t$ as the simplest and natural choice for u, since $\sum_{s=1}^\infty \frac{1}{t} = \zeta(1)$, $\zeta(.)$ being the Euler-Riemann zeta function. Though this integral is improper, it has a Cauchy principal value as Euler–Mascheroni constant which means $\sum_{s=1}^\infty \frac{1}{t} \approx \gamma = 0.577$. So we assign $u = \log t$

$$\sum_{s=1}^t \mathbb{P}\left(\|(\tilde{A}_t - A)\boldsymbol{\omega}\|_\infty > \rho(t,\delta)\right) \leq \sum_{s=1}^t \frac{1}{t} \leq \sum_{s=1}^\infty \frac{1}{t} \leq \zeta(1) \approx 0.577$$

□

Lemma 5. *Let $\bar{\boldsymbol{\mu}} \geq 0$ be a vector, and consider the set $Q_{\bar{\boldsymbol{\mu}}} = \{\boldsymbol{\mu} \geq 0 \mid q(\boldsymbol{\mu}) \geq q(\bar{\boldsymbol{\mu}})\}$. Let Slater condition hold. Then, the set $Q_{\bar{\boldsymbol{\mu}}}$ is bounded and, in particular, we have $\|\boldsymbol{\mu}\|_1 \leq \frac{1}{\gamma}(f(\bar{x}) - q(\bar{\boldsymbol{\mu}})), \forall \boldsymbol{\mu} \in Q_{\bar{\boldsymbol{\mu}}}$ where, $\gamma = \min_{1\leq j\leq m}\{-g_j(\bar{x})\}$ and $\bar{x}$ is a Slater vector. f(.) and q(.) respectively denotes the primal and the dual function of the optimization problem.*

Using the aforementioned lemmas we give a bound for the part in the inverse of the characteristic time function that gets added for Lagrangian relaxation which eventually helps up landing on a unique formulation of sample complexity upper bounds of our proposed algorithm.

Lemma 6. *For any $\boldsymbol{l} \in \mathcal{L}$ and $\boldsymbol{\omega} \in \Delta_K$, $-\boldsymbol{l}^\top \tilde{A}\boldsymbol{\omega} \leq \mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\omega}, \hat{\mathcal{F}})\psi$*

where, $\psi = \frac{\|(\tilde{A}-A)\boldsymbol{\omega}\|_\infty + \max_{i\in[1,N]}(-A\boldsymbol{\omega})}{\min_{i\in[1,N]}(-\tilde{A}\boldsymbol{\omega})}$

Proof. For any $\boldsymbol{l} \in \mathcal{L}$ and $\boldsymbol{\omega} \in \Delta_K$ we write

$$\begin{aligned}(-\boldsymbol{l}^\top \tilde{A}\boldsymbol{\omega}) = \boldsymbol{l}^\top(-\tilde{A} + A - A)\boldsymbol{\omega} &\leq \|\boldsymbol{l}\|_1\|(A-\tilde{A})\boldsymbol{\omega}\|_\infty + \|\boldsymbol{l}\|_1 \max_{i\in[1,N]}(-A^i\boldsymbol{\omega}) \\ &\leq \|\boldsymbol{l}\|_1\left((\|A-\tilde{A})\boldsymbol{\omega}\|_\infty + \max_{i\in[1,N]}(-A\boldsymbol{\omega})\right) \\ &\leq \mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \mathcal{F})\frac{\|(\tilde{A}-A)\boldsymbol{\omega}\|_\infty + \max_{i\in[1,N]}(-A\boldsymbol{\omega})}{\min_{i\in[1,N]}(-\tilde{A}\boldsymbol{\omega})}\end{aligned}$$

Plugging in the definition of ψ mentioned in the statement of the lemma concludes the proof. □

K.2 Useful Results from BAI and Pure Exploration Literature

Lemma 7 (Lemma 19 in (Garivier and Kaufmann, 2016))**.** *There exists two constants B and C (depends on $\boldsymbol{\mu}$ and ϵ), such that—*

$$\sum_{t=h(T)}^{\top} \mathbb{P}\left\{\|\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}\|_\infty > \xi(\epsilon)\right\} \leq BT\exp(-CT^{\frac{1}{8}})$$

Lemma 8 (Lemma 7, Garivier and Kaufmann (2016))**.** *For all $t \geq 1$ and $\forall a \in [K]$, C-Tracking ensures $N_{a,t} \geq \sqrt{t+K^2} - K$ and*

$$\max_{a\in[K]} |N_{a,t} - \sum_{s=1}^t \boldsymbol{\omega}_{a,s}| \leq K(1+\sqrt{t})$$

Lemma 9 (Theorem 9 in (Kaufmann and Koolen, 2021))**.** *Let $\delta > 0, \nu$ be independent one-parameter exponential families with mean $\boldsymbol{\mu}$ and $S \subset [d]$. Then we have,*

$$\mathbb{P}_\nu\left[\exists t \in \mathbb{N} : \sum_{a \in S} \tilde{N}_{t,a} d_{KL}\left(\mu_{t,a}, \mu_a\right) \geq \sum_{a \in S} 3 \ln\left(1 + \ln\left(\tilde{N}_{t,a}\right)\right) + |S| \mathcal{T}\left(\frac{\ln\left(\frac{1}{\delta}\right)}{|S|}\right)\right] \leq \delta \,.$$

Here, $\mathcal{T} : \mathbb{R}^+ \to \mathbb{R}^+$ is such that $\mathcal{T}(x) = 2\tilde{h}_{3/2}\left(\frac{h^{-1}(1+x) + \ln\left(\frac{\pi^2}{3}\right)}{2}\right)$ with

$$\forall u \geq 1, \quad h(u) = u - \ln(u) \tag{30}$$

$$\forall z \in [1, e], \forall x \geq 0, \quad \tilde{h}_z(x) = \begin{cases} \exp\left(\frac{1}{h^{-1}(x)}\right) h^{-1}(x) & \text{if } x \geq h^{-1}\left(\frac{1}{\ln(z)}\right) \\ z(x - \ln(\ln(z))) & \text{else} \end{cases} . \tag{31}$$

Lemma 10 (Lemma 17 in (Degenne and Koolen, 2019))**.** *Under the good event G_T, there exists a T_ϵ such that for T where $h(T) \geqslant T_\epsilon$ C-tracking will satisfy*

$$\inf_{\boldsymbol{w} \in w^*(\boldsymbol{\mu})} \left\| \frac{N_t}{t} - \boldsymbol{w} \right\|_\infty \leqslant 3\epsilon, \forall t \geqslant 4\frac{K^2}{\epsilon^2} + 3\frac{h(T)}{\epsilon}$$

Lemma 11 (Theorem 2 in (Degenne et al., 2019b))**.** *The sample complexity of GE is $\mathbb{E}[\tau] \leqslant T_0(\delta) + \frac{eK}{a}$ where $T_0(\delta) = \max\left\{t \in \mathbb{N} : t \leqslant T(\boldsymbol{\mu}) c(t, \delta) + C_{\boldsymbol{\mu}}\left(R_t^{\boldsymbol{\lambda}} + R_t^{\boldsymbol{w}} + O(\sqrt{t \log t})\right)\right\}$ where $R_t^{\boldsymbol{\lambda}}$ is the regret of the instance player, $R_t^{\boldsymbol{w}}$ the regret of the allocation player and $C_{\boldsymbol{\mu}}$ an instance-dependent constant.*

K.3 Useful Results for Continuity of Convex Functions

Definition 2 (Definition of Upper Hemicontinuity)**.** *We say that a set-valued function $C : \Theta \to \boldsymbol{\omega}$ is upper hemicontinuous at the point $\theta \in \Theta$ if for any open set $S \subset \boldsymbol{\omega}$ with $C(\theta) \in S$ there exists a neighborhood U around θ, such that $\forall x \in U, C(x)$ is a subset of S.*

Theorem 8 (Berge's maximum theorem, (Berge, 1963))**.** *Let X and Θ be topological spaces. Let $f : X \times \Theta \to \mathbb{R}$ be a continuous function and let $C : \Theta \to \bar{X}$ be a compact-valued correspondence such that $C(\theta) \neq \varnothing \forall \theta \in \Theta$. If C is continuous at θ then $f^*(\theta) = \sup_{x \in C(\theta)} f(x, \theta)$ is continuous and $C^* = \{x \in C(\theta) : f(x, \theta) = f^*(\theta)\}$ is upper hemicontinuous.*

Theorem 9 (Heine-Borel theorem, due to Eduard Heine and Émile Borel)**.** *For a subset S in $\mathbb{R}^n$, the following two statements are equivalent*

1. *S is closed and bounded.*
2. *S is compact, means every open cover of S has a finite sub-cover.*

Theorem 10. *Let C be a closed convex set with nonempty (topological) interior. Let f and $\{f^r\}$ be affine functions from E^n to E^m with $f^r \to f$. Then*

(II.1.2) $\overline{\lim}_{r \to \infty}\left(H\left(f^r\right) \cap C\right) \subset H(f) \cap C$

(II.1.3). $\underline{\lim}_{r \to \infty}\left(H\left(f^r\right) \cap C\right)$ *is a closed convex subset of $H(f) \cap C$,*

(II.1.4). If $H(f) \cap C$ has nonempty interior and no component of f is identically zero, then $\lim_{r \to \infty}\left(H\left(f^r\right) \cap C\right) = H(f) \cap C$.

Lemma 12 (Lemma 13, Magureanu et al. (2014))**.** *Consider $A \in \left(\mathbb{R}^+\right)^{k \times k}, c \in \left(\mathbb{R}^+\right)^k$, and $\mathcal{T} \subset \left(\mathbb{R}^+\right)^{k \times k} \times \left(\mathbb{R}^+\right)^k$. Define $t = (A, c)$. Consider the function Q and the set-valued map $Q^\star$*

$$Q(t) = \inf_{x \in \mathbb{R}^k} \{cx \mid Ax \geq 1, x \geq 0\}$$

$$Q^\star(t) = \{x : cx \leq Q(t) \mid Ax \geq 1, x \geq 0\}.$$

Assume that: For all $t \in \mathcal{T}$, all rows and columns of A are non-identically 0 and $\min_{t \in \mathcal{T}} \min_k c_k > 0$. Then, 1. Q is continuous on $\mathcal{T}$, 2. $Q^\star$ is upper-hemicontinuous on $\mathcal{T}$.