Learning to Explore with Lagrangians for Bandits under Unknown Constraints

Udvas Das and Debabrota Basu
Équipe Scool, Univ. Lille, Inria,
CNRS, Centrale Lille, UMR 9189- CRIStAL, F-59000 Lille, France
udvas.das@inria.fr debabrota.basu@inria.fr

Abstract

Pure exploration in bandits can model eclectic real-world decision making problems, such as tuning hyper-parameters or conducting user studies, where sample frugality is desired. Thus, considering different safety, resource, and fairness constraints on the decision space has gained increasing attention. In this paper, we study generalisation of these problems as pure exploration in multi-armed bandits with unknown linear constraints. First, we propose a Lagrangian relaxation of the sample complexity lower bound for pure exploration. We further derive how this lower bound converges to the existing lower bound for pure exploration under known constraints, and how the hardness of the problem changes with the geometry induced by the constraint estimation procedure. We further leverage the Lagrangian lower bound and properties of convex optimisation to propose two computationally efficient extensions of Track-and-Stop and Gamified Explorations, namely LATS and LAGEX. Designing these algorithms require us to propose a new constraint-adaptive stopping rule, and also at each step, using pessimistic estimates of constraints in the Lagrangian lower bound. We show that these algorithms asymptotically achieve the desired sample complexity bounds. Finally, we conduct numerical experiments with different reward distributions and constraints that validate efficient performance of LAGEX and LATS with respect to baselines.

1 Introduction

Decision-making under uncertainty is a ubiquitous challenge encountered across various domains, including clinical trials [VBW15], recommendation systems [ZY24] and more. Multi-Armed Bandit (MAB) serves as an archetypal framework for sequential decision-making under uncertainty and allows to study the involved information-utility trade-offs [LS20]. In MAB, at each step, an agent interacts with an environment consisting of K decisions (aka arms) corresponding to K noisy feedback distribution (aka reward distribution). At each step, the agent takes a decision, and obtains a reward from the corresponding reward distribution. The goal of the agent is to compute a *policy*, i.e. a distribution over the decisions, that maximises a certain utility metric (e.g. accumulated rewards [ACBF02], probability of identifying the best arm [KCG16] etc.) over time.

In this paper, we focus on the *pure exploration* problem of MABs, where the agent interacts by sequentially realising a sequence of policies (or experiments) with the goal of *answering a query as correctly as possible*. A well-studied pure exploration problem is Best-Arm Identification (BAI), where the agent aims to identify the arm with the highest expected reward [EDMM02a, BMS09, JN14, KCG16]. BAI has been increasingly applied for hyper-parameter tuning [LJD$^+$17], communication networks [LPJ22], influenza mitigation [LVR$^+$19], finding the optimal dose of a drug [AKR21a] etc. However, real-world scenarios often involve constraints on the arms that must be satisfied [CBJD24].

Example 1 (Optimal treatment plan [KCJ23, CGS22a])**.** *We want to identify the optimal treatment of a patient with rheumatoid arthritis, when the first and second-line of treatments have failed [KCJ23].*

17th European Workshop on Reinforcement Learning (EWRL 2024).

There is large variability in the choice of next treatment, and several drugs are a priori considered to be equally good but might not work equally well together [KCJ23]. Let us assume that we have K drugs for efficacies and side-effects are unknown. We assume that efficacy of each drug comes from a reward distribution with unknown mean $\boldsymbol{\mu}_a$ and also d side-effects (e.g. drop in heart- and liver-function scores) are due to unknown drug-specific constraints $\mathbf{A}_a$. The constraint $\mathbf{A}_a$ represent scores that are deemed to be unsafe, and thus, we want $\mathbf{A}\boldsymbol{\pi} \leq \mathbf{0}$. Thus, given a treatment plan $\boldsymbol{\pi}$ of mixing the drugs, the mean efficacy is $\langle \boldsymbol{\mu}, \boldsymbol{\pi} \rangle$, and the extent of side-effects is $\mathbf{A}\boldsymbol{\pi}$. These scores and the efficacyserves can be measured after each application of a drug, with the stochasticity due to inter-patient variability. Additionally, changing a treatment has additional human involvement and cost, which is better to be minimised. We aim to reliably find the most effective drug cocktail $\boldsymbol{\pi}^*$ with minimum number of interactions with patients, while retaining the side-effects in the safe zone.

Pure exploration under constraints. In recent years, real-life applications like above naturally motivated study of pure exploration under a set of known and unknown constraints [KSS18, WWJ21, LZYL23, WZZ23, CBJD24]. Specifically, we aim to find the optimal policy that maximises the expected rewards over the set of arms and also satisfies the true constraints. The agent, at each step t, selects an action according to a chosen policy observes associated reward and the cost incurred with respect to the constraints. The agent uses the feedbacks to update the estimates of the expected rewards and constraints. Using these estimates, the agent chooses the next policy and action till the optimal policy satisfying constraints is identified with confidence $1 - \delta$. This is known as the *fixed-confidence setting* of pure exploration [WWJ21, CBJD24], while there is also fixed-budget setting which is of independent interest [KSS18, LZYL23, FN22]. Existing literature has studied either the general linear constraints when they are known [WWJ21, CBJD24, CWM$^+$22a], or very specific type of unknown constraints, e.g. safety [WWJ21], knapsack [LZYL23], fairness [WZZ23], preferences [LTHK22] 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). Further discussions on related works is in Appendix B.1.

Recently, [CBJD23] show if the constraints are known, i.e the feasible set is deterministic, a bandit instance may become harder or easier depending on the geometry of the constraints. They state that *studying similar phenomenon for unknown constraints as an open problem* as the feasible policy space is not deterministic anymore. As we have to construct an estimate of the feasible policy set, we have to simultaneously control concentration of the means of the unknown reward distributions, and those of the feasible sets, simultaneously. This leads us to two questions

- *How does the hardness of the pure exploration under unknown constraints change if the constraints are estimated sequentially?*
- *How can we design a generic algorithmic scheme to track both the constraints and the optimal policy with sample- and computational-efficiency?*

Our contributions address these questions as follows:

1. *Lagrangian relaxation of the lower bound.* The minimum number of samples required to conduct pure exploration with fixed confidence and the corresponding interactive policy are expressed through lower bound, which is an optimisation problem under known constraints. For unknown constraints, we propose a novel Lagrangian relaxation of this optimisation problem in the lower bound (Sec. 3). At every step, we use pessimistic estimates of the constraints, while the Lagrangian multipliers help us to track the interaction between the objective function and the structure of the estimated constraints. We use multiple results from convex analysis to show that the Lagrangian relaxation with pessimistic constraints preserves all the continuity properties of the lower bound under known constraints. Additionally, it satisfies strong duality to yield a unique optimal policy for interactions, and also bounds on the Lagrangian multipliers at every step. Further, we characterise the Lagrangian lower bound for Gaussian rewards, which connects with the lower bound for known constraints.

2. *A generic algorithm design.* Now, we leverage this Lagrangian lower bound with pessimistic estimates of constraints to propose two algorithms, namely LATS (Lagrangian Track and Stop) and LAGEX (Lagrangian Gamified EXplorer). First, we develop a new stopping rule for the unknown constraint setting as we have to ensure that both the mean estimates and pessimistic estimates of constraints concentrate close to their true values before reliably recommend an optimal policy using them. Then, we extend the Track-and-Stop [GK16] and gamified explorer [DKM19b] approaches for the Lagrangian lower bound to design LATS and LAGEX, respectively (Sec. 4).

3. *Upper bound on sample complexities.* We provide upper bounds on sample complexities of LATS and LAGEX (Sec. 3). This requires proving a novel concentration of the constraints, and also consequent concentration of optimal policies under constraints. We show that due to constraint LATS achieve an upper bound, which is $(1 + \mathfrak{s})$ times the upper bound of TS under known constraints. $\mathfrak{s}$ is the shadow price of the true constraint and quantifies its stability under perturbation. In contrast, LAGEX leads to an upper bound that has only an additive $\mathfrak{s}$ factor with the known constraint lower bound. This suggests that LAGEX should be more sample-efficient than LATS. Our experimental results (Sec. 5) across multiple settings validate that LAGEX requires the least samples among competing algorithms and it can exactly follow the hardness due to constraints across environments.

2 Pure exploration under unknown constraints

Notation. x, $\mathbf{x}$, $\mathbf{X}$, adn $\mathcal{X}$ denote a scalar, a vector, a matrix, and a set respectively. For a positive semi-definite matrix $\mathbf{A}$ and vector $\mathbf{z}$, $\|\mathbf{z}\|^2_{\mathbf{A}} = \langle \mathbf{z}, \mathbf{A}\mathbf{z} \rangle$. Also, in $\mathbb{R}^d_+$, we include $\mathbf{0}^d$. $[K]$ refers to $\{1, \dots, K\}$. $\mathrm{Supp}(P)$ denotes the support of a distribution P. Δ_K is the simplex over $[K]$.

Problem formulation. We work with a MAB instance with $K \in \mathbb{N}$ arms. Each arm $a \in [K]$ has a reward distribution P_a with unknown means $\boldsymbol{\mu}_a \in \mathbb{R}$. The agent, at each time step $t \in \mathbb{N}$, 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 $\mathbf{A}\boldsymbol{\pi} \leq \mathbf{0}$ with respect to the set of d linear constraints $\mathbf{A} \in \mathbb{R}^{d \times K}$[1].

If we have known $\mathbf{A}$, the agent would have access to the non-empty and compact set of feasible policies $\mathcal{F} \triangleq \{\boldsymbol{\pi} \in \Delta_K : \mathbf{A}\boldsymbol{\pi} \leq \mathbf{0}\}$ and the agent would aim to identify the optimal feasible policy, i.e. the one yielding the highest expected reward while maintaining the constraints,

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

In our setting, we do not have access to the true set of constraints. At any time $t \in \mathbb{N}$, we rather produce $\hat{\mathbf{A}}_t$ as an estimate of $\mathbf{A}$. Then, the agent has access to an estimated feasible set $\hat{\mathcal{F}} \triangleq \{\boldsymbol{\pi} \in \Delta_K : \hat{\mathbf{A}}\boldsymbol{\pi} \leq \mathbf{0}\}$ and can identify the optimal feasible policy to be $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} \triangleq \arg\max_{\boldsymbol{\pi} \in \hat{\mathcal{F}}} \boldsymbol{\mu}^T \boldsymbol{\pi}$.

Using the estimated constraints, we want to design an algorithm that finally recommends policy that is $(1-\delta)$-correct and $(1-\delta)$-feasible.

Definition 1 ($(1-\delta)$-correct and $(1-\delta)$-feasible recommended policy)**.** *For $\delta \in [0, 1)$, a policy recommended by a pure exploration algorithm is $(1-\delta)$-correct and $(1-\delta)$-feasible if* $\Pr[\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} \neq \boldsymbol{\pi}^\star_{\mathcal{F}}] \leq \delta$ *and* $\Pr[\mathbf{A}\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} \geq \mathbf{0}] \leq \delta$.

For the algorithm to yield correct and feasible recommendations at any time it stops, we want such an estimate of $\mathbf{A}$ that $\hat{\mathcal{F}}$ is a superset of $\mathcal{F}$. Otherwise, we cannot ensure that the true optimal and feasible policy $\boldsymbol{\pi}^\star_{\mathcal{F}} \in \hat{\mathcal{F}}$, i.e. the set where we are searching for the recommended policy. Additionally, we might hit a degenerate case where $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$ might not exist. To ensure these properties, in Section 3, we design pessimistic estimates of $\mathbf{A}$ and use them further. This property echoes the spirit of the optimistic-pessimistic algorithms designed for regret-minimisation setting of bandits under unknown constraints [PGBJ20, MAAT20, LLSY21].

Goal. In addition to recommending a $(1-\delta)$-correct and $(1-\delta)$-feasible policy, we want to use minimum number of interactions to identify it. For this, pure exploration algorithms use a stopping rule which stops at a random stopping time τ_δ. Here, we aim to design an algorithm that recommends a $(1-\delta)$-correct and $(1-\delta)$-feasible policy while keeping $\mathbb{E}[\tau_\delta]$ as small as possible.

Motivation: Extension of multiple prior problems. First, we clarify our motivation by showing how different interesting problems are special cases of our setting.

a. Thresholding Bandits. Thresholding bandits [AKR21a] are motivated from the safe dose finding problem, where one wants to identify the highest dose of a drug below a known safety level. This has also motivated the safe arm identification problem [WWJ21]. Our setting generalises it further to detect the optimal combination of doses of a set of drugs yielding highest efficacy while staying below the safety threshold. Formally, we want to identify $\boldsymbol{\pi}^\star = \arg\max \boldsymbol{\mu}^T \boldsymbol{\pi}$, such that $\mathbf{I}\boldsymbol{\pi} \leq \mathbf{I}\boldsymbol{\theta}$. Additionally, we allow different different thresholds for different drugs.

[1]We assume that simplex constraints are augmented in $\mathbf{A}$ for ease of notation.

b. Optimal policy under knapsack. Bandits under knapsack constraints are studied both in BAI [LZYL23, TTCRJ12, LSY21] and regret minimisation literature [BKS18, AD16, ISSS22, ADL16, SS18]. Detecting an optimal arm might have additional resource constraints than the number of required samples. This led to study of BAI with knapsacks under fixed-budget setting [LZYL23]. But as in regret-minimisation [SS18], one might finally want a policy that maximises utility and satisfies knapsack constraints. For example, we want to manage caches where the recommended memory allocation should satisfy a certain resource budget but can violate them during exploration. Formally, $\boldsymbol{\pi}^\star_\tau = \arg\max_{\boldsymbol{\pi} \in C_{\mathbf{A}}} \hat{\boldsymbol{\mu}}^T_{\tau_\delta} \boldsymbol{\pi}$, where $C_{\mathbf{A}} \triangleq \{\mathbf{A}\boldsymbol{\pi}_{\tau_\delta} \leq c\}$.

c. BAI with fairness across sub-populations. BAI with fairness constraints on sub-populations (BAICS) [WZZ23] aims to select an arm that must be fair across all the $\boldsymbol{l}$ sub-populations rather than the whole population as in standard BAI. 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 $\boldsymbol{\mu}_k$. Then we are looking for a combination of drugs rather than a single drug to administer as $\boldsymbol{\pi}^\star = \arg\max_{\boldsymbol{\pi} \in \Delta_K} \boldsymbol{\mu}^T \pi$, such that $\mathbb{1}^T_{\boldsymbol{\mu}_m \geq 0} \boldsymbol{\pi} = 1, \forall m \in [l]$. Hence, our setting solve the BAICS as a special case.

3 Lagrangian Relaxation of the Lower Bound and its Properties

Now, we discuss the Lagrangian relaxation of the lower bound and its properties that are necessary to ensure design of a correct and feasible pure exploration algorithm under unknown constraints. We require two structural assumptions to establish our approach.

Assumption 1 (Structural assumptions on means, policy, and constraints)**.** *(a) The mean vector $\boldsymbol{\mu}$ belongs to a bounded subset $\mathcal{D}$ of $\mathbb{R}^K$. (b) There exists a unique optimal feasibly policy (Equation* (1)*). (c) For the true constraint $\mathbf{A}$, there exists a non-zero slack vector Γ, such that* $\max_{\boldsymbol{\pi} \in \Delta_K}(-\mathbf{A}\boldsymbol{\pi}) = \Gamma$.

The unique optimal and feasible policy assumption is used following [CBJD23] to ensure that solution of Equation (1) is an extreme point of the polytope $\mathcal{F}$. The assumption on slack is analogous of using the assumption of existence of a safe-arm [PGBJ20], or existence of Slater's condition for the constraint optimisation problem [LLSY21]. Standing on these assumptions, we further prove that $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$ is unique, i.e $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$ is an extreme point in the polytope $\hat{\mathcal{F}}$.

3.1 Information acquisition and estimates of constraints

The agent acquires new information at every step $t \in \mathbb{N}$ by sampling an action $A_t \sim \boldsymbol{\omega}_t$. $\boldsymbol{\omega}_t \in \Delta_K$ is called the allocation policy, and is used for interaction at step t. This yields a noisy reward $R_t \in \mathbb{R}$ and cost vector $\mathbf{A}_t \in \mathbb{R}^d$. We remind that in MAB, due to independence of arms, we can represent the a-th arm as the a-th basis of $\mathbb{R}^K$. Thus, using the observations obtained till t, we estimate the mean vector as $\hat{\boldsymbol{\mu}}_t \triangleq \Sigma_t^{-1} \left(\sum_{s=1}^{t-1} R_s A_s \right)$. Here, $\Sigma_t \triangleq \sum_{s=1}^{t} A_s A_s^\top$ is the Gram matrix or the design matrix at time t. Similarly, the estimate of the i-th row of the constraint matrix is $\hat{\mathbf{A}}^i_t \triangleq \Sigma_t^{-1} \left(\sum_{s=1}^{t-1} \mathbf{A}^{i,A_s} A_s \right)$. But we observe that using $\hat{\mathbf{A}}_t$ to define the feasible policy set does not ensure that for any t, it is a superset of $\mathcal{F}$.

Thus, we define a confidence ellipsoid around $\hat{\mathbf{A}}_t$ that always includes $\mathbf{A}$ with probability $1 - \delta$, and further allows us to define a pessimistic estimate of $\mathbf{A}$. Formally, the confidence ellipsoid for each row $i \in [d]$ of $\hat{\mathbf{A}}^i$ is

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

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

Lemma 1 (Pessimistic estimate of constraints)**.** *If we use the pessimistic estimate of $\mathbf{A}$ from the confidence ellipsoid $\mathcal{C}_t$ to define the feasible policy set at time t as*

$$\hat{\mathcal{F}}_t \triangleq \{ \boldsymbol{\pi} \in \Delta_K : \min_{\mathbf{A}' \in \mathcal{C}_t} \mathbf{A}' \boldsymbol{\pi} \leq \mathbf{0} \}, \tag{3}$$

we observe that $\mathcal{F} \subseteq \hat{\mathcal{F}}_t$ for all $t \in \mathbb{N}$.

In Figure 1, we visualise this result using the numerical values obtained from our algorithms. We observe that as we acquire more samples, our estimated feasible policy set $\hat{\mathcal{F}}_t \to \mathcal{F}$.

Remark. Our design principle of the estimators echos the spirit of the optimistic-pessimistic algorithms from the regret-minimisation under constraints literature [PGBJ20, LLSY21, CGS22b,

PGB24], which has led to successful and efficient algorithms. The idea there is also to use a pessimistic constraint so that we can always find the 'true' optimal policy in it. Similarly, here having a pessimistic choice of $\mathbf{A}$ results in a bigger alternative set for $\boldsymbol{\mu}$ than the alternative set we get using $\mathbf{A}$, and a bigger feasible policy set that always includes the true optimal policy.

3.2 Lagrangian relaxation of the lower bound with estimated constraints

When we have a bandit instance with mean vector $\boldsymbol{\mu}$ and a constraint matrix $\mathbf{A}$, we try to find the most confusing instance of $\boldsymbol{\mu}$, so that we can minimise the KL-divergence between these two instances to make sure we have gathered enough statistical evidence to rule out all the confusing instances. Extending the BAI lower bound of [GK16, CBJD24] prove that if $\mathbf{A}$ is known, expected stopping time of any $(1-\delta)$-correct and always-feasible algorithm satisfies

$$\mathbb{E}[\tau_\delta] \geq T_{\mathcal{F}}(\boldsymbol{\mu}) \ln \frac{1}{2.4\delta} \,. \tag{4}$$

Here, the reciprocal of the characteristic time is defined by an optimisation problem over the set of alternative instances $\Lambda_{\mathcal{F}}(\boldsymbol{\mu}) \triangleq \left\{\boldsymbol{\lambda} \in \mathcal{D} : \max_{\boldsymbol{\pi}\in\mathcal{F}} \boldsymbol{\lambda}^T\boldsymbol{\pi} > \boldsymbol{\lambda}^T\boldsymbol{\pi}^\star_{\mathcal{F}}\right\}$:

$$T^{-1}_{\mathcal{F}}(\boldsymbol{\mu}) \triangleq \sup_{\boldsymbol{\omega}\in\mathcal{F}} \inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{a=1}^{K} \boldsymbol{\omega}_a d\left(\boldsymbol{\mu}_a, \boldsymbol{\lambda}_a\right) \triangleq \boldsymbol{\omega}^\top d\left(\boldsymbol{\mu}, \boldsymbol{\lambda}\right) \,. \tag{5}$$

$\Lambda_{\mathcal{F}}(\boldsymbol{\mu})$, aka the Alt-set, is the set of all bandit instances that have mean vectors in a bounded subset $\mathcal{D} \in \mathbb{R}^K$ but a different optimal policy than that of $\boldsymbol{\mu} \in \mathcal{D}$.

Now, given an estimate $\hat{\mathcal{F}}_t$ of the feasible policies at any $t > 0$ (Eq. (3)), the corresponding Alt-set is

$$\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) \triangleq \left\{\boldsymbol{\lambda} \in \mathcal{D} : \max_{\boldsymbol{\pi}\in\hat{\mathcal{F}}_t} \boldsymbol{\lambda}^T\boldsymbol{\pi} > \boldsymbol{\lambda}^T\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}\right\} \tag{6}$$

Since the estimated feasible policy set $\hat{\mathcal{F}}_t$ is superset of the original feasible policy set $\mathcal{F}$, we observe that $\Lambda_{\mathcal{F}}(\boldsymbol{\mu}) \subseteq \Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})$. Enabled by these definitions, we are ready define to Lagrangian relaxation of the lower bound. Specifically, we observe that

$$\begin{aligned}
T^{-1}_{\mathcal{F}}(\boldsymbol{\mu}) &\triangleq \sup_{\boldsymbol{\omega}\in\mathcal{F}} \inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \boldsymbol{\omega}^\top d\left(\boldsymbol{\mu}, \boldsymbol{\lambda}\right) \\
&\leq \inf_{\boldsymbol{l}\in\mathbb{R}^d_+} \sup_{\boldsymbol{\omega}\in\mathcal{F}} \inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \boldsymbol{\omega}^\top d\left(\boldsymbol{\mu}, \boldsymbol{\lambda}\right) + \boldsymbol{l}^T\Gamma \\
&\leq \inf_{\boldsymbol{l}\in\mathbb{R}^d_+} \min_{\mathbf{A}'\in\mathcal{C}_t} \sup_{\boldsymbol{\omega}\in\hat{\mathcal{F}}} \inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})} \boldsymbol{\omega}^\top d\left(\boldsymbol{\mu}, \boldsymbol{\lambda}\right) - \boldsymbol{l}^\top\mathbf{A}'\boldsymbol{\omega}
\end{aligned} \tag{7}$$

Equation (7) defines the Lagrangian relaxation of the characteristic time under unknown constraints, i.e. denoted by $T^{-1}_{\hat{\mathcal{F}}_t}(\boldsymbol{\mu})$. For non-negative Lagrange multipliers $\boldsymbol{l} \in \mathbb{R}^d_+$, the first inequality is true due to the existence of a slack for the true constraints $\mathbf{A}$. The second inequality is due to the pessimistic choice of the estimated constraint. Equation (7) shows that the Lagrangian relaxation of the $T_{\hat{\mathcal{F}}_t}(\boldsymbol{\mu})$ always serves as a lower bound of the characteristic time $T^{-1}_{\mathcal{F}}(\boldsymbol{\mu})$ for known constraints [CBJD23], and thus, leads to a valid lower bound to optimise for the expected stopping time $\mathbb{E}[\tau_\delta]$. For brevity, we denote $\mathbf{A}'$ achieving the minimum as $\tilde{\mathbf{A}}$, and omit t from $\hat{\mathcal{F}}_t$ if it is true for any $t \in \mathbb{N}$.

The Lagrangian relaxation leads to a natural question:

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

Theorem 1 (Strong Duality and Range of Lagrange Multipliers)**.** *For a bounded sequence of $\{l_t\}_{t\in\mathbb{N}}$, strong-duality holds for the optimisation problem stated in Equation* (7)*, i.e.*

$$\begin{aligned}
&\inf_{\boldsymbol{l}\in\mathbb{R}^d_+} \min_{\mathbf{A}'\in\mathcal{C}} \sup_{\boldsymbol{\omega}\in\hat{\mathcal{F}}} \inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})} \boldsymbol{\omega}^\top d\left(\boldsymbol{\mu}, \boldsymbol{\lambda}\right) - \boldsymbol{l}^T\tilde{\mathbf{A}}\boldsymbol{\omega} \\
&= \sup_{\boldsymbol{\omega}\in\hat{\mathcal{F}}} \min_{\boldsymbol{l}\in\mathcal{L}} \inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})} \boldsymbol{\omega}^\top d\left(\boldsymbol{\mu}, \boldsymbol{\lambda}\right) - \boldsymbol{l}^T\tilde{\mathbf{A}}\boldsymbol{\omega} \,.
\end{aligned} \tag{8}$$

Here, $\mathcal{L} \triangleq \{\boldsymbol{l} \in \mathbb{R}^d \mid 0 \leq \|\boldsymbol{l}\|_1 \leq \frac{1}{\gamma} T^{-1}_{\hat{\mathcal{F}}}(\hat{\boldsymbol{\mu}})\}$, where $\gamma \triangleq \min_{i\in[1,d]}\{-\tilde{\mathbf{A}}^i\boldsymbol{\omega}^*\}$, i.e. the minimum slack for pessimistic constraints w.r.t. the optimal allocation.

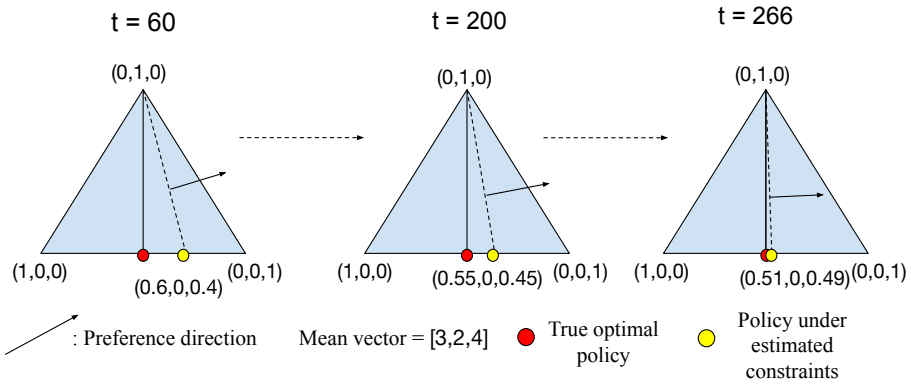

Figure 1: Convergence of the feasible set and optimal policy.

Hereafter, we use the RHS of Eq. (8) as $T^{-1}_{\hat{\mathcal{F}}}(\boldsymbol{\mu})$. Theorem 1 provides us a hypercube to search the minimising Lagrangian multipliers, and thus, turns into a linear programming problem.

Connections with Lagrangian-based methods in bandits. In regret minimisation literature Lagrangian-based optimistic-pessimistic methods [TPRL20, SSF23] are usually used to not only devise a no-regret learner but to control and get a sub-linear constraint violation guarantees [LLSY21, BCC24]. Our proposed algorithm LAGEX is a prime example where the "self-boundedness" of the dual variables results in better constraint violation guarantees (Figure 7 and 6). It would be interesting to see how LAGEX performs in regret minimisation setting.

Inner optimisation problem. Now, we peel the layers of the optimisation problem in Eq. (8). For known constraints, [CBJD23] has leveraged results from convex analysis [BV04] to show that the most confusing instance for $\boldsymbol{\mu}$ lie in the boundary of the normal cone $\Lambda_{\mathcal{F}}(\boldsymbol{\mu})^C$ (solid cone in Figure 2) spanned by the active constraints $\mathbf{A}_{\boldsymbol{\pi}^\star_{\mathcal{F}}}$ for $\boldsymbol{\pi}^\star_{\mathcal{F}}$. $\mathbf{A}_{\boldsymbol{\pi}^\star_{\mathcal{F}}}$ is a sub-matrix of $\mathbf{A}$ consisting at least K linearly independent rows. Specifically, they show that $\mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \mathcal{F}) \triangleq \inf_{\boldsymbol{\lambda} \in \Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})} \boldsymbol{\omega}^\top d\left(\boldsymbol{\mu}, \boldsymbol{\lambda}\right) = \min_{\boldsymbol{\pi}' \in \nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} \min_{\boldsymbol{\lambda}: \boldsymbol{\lambda}^T(\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}') = 0} \boldsymbol{\omega}^\top d\left(\boldsymbol{\mu}, \boldsymbol{\lambda}\right)$.

In our setting, we are estimating both the mean vectors and the constraints. Hence, the normal cone is also estimated at every step. Let $\tilde{\mathbf{A}}_{\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}}$ be the sub-matrix spanned by the active constraints for $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$. Since we are working with the pessimistic estimate of $\mathbf{A}$, the vector space spanned by the linearly independent rows of $\mathbf{A}_{\boldsymbol{\pi}^\star_{\mathcal{F}}}$ is a subset of the vector space spanned by those of $\tilde{\mathbf{A}}_{\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}}$. Since the estimated Alt-set $\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})$ is always a superset of the true Alt-set $\Lambda_{\mathcal{F}}(\boldsymbol{\mu})$, the normal cone around $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$ is always a subset of the true normal cone $\boldsymbol{\pi}^\star_{\mathcal{F}}$. We illustrate them in Figure 2. We now extend the projection lemma for known constraints to the Lagrangian formulation with unknown constraints.

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,*

$$\mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \hat{\mathcal{F}}) = \min_{\boldsymbol{l} \in \mathcal{L}} \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \min_{\boldsymbol{\lambda}: \boldsymbol{\lambda}^T(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}') = 0} \boldsymbol{\omega}^\top d\left(\boldsymbol{\mu}, \boldsymbol{\lambda}\right) - \boldsymbol{l}^T \tilde{\mathbf{A}} \boldsymbol{\omega}\,. \tag{9}$$

This 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 continuously track the lower bound defined by projection lemma for unknown constraints?

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) $\lim_{t \to \infty} \hat{\boldsymbol{\lambda}}_t \to \boldsymbol{\lambda}^\star$. Thus, for any $\boldsymbol{\omega} \in \mathcal{F}$ and $\boldsymbol{\mu}$, $\lim_{t \to \infty} \mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \hat{\mathcal{F}}) \to \mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \mathcal{F})$ where $\boldsymbol{\lambda}^*$ is such that for any $\boldsymbol{\lambda}^* \in \arg\min_{\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \boldsymbol{\omega}^T d\left(\boldsymbol{\mu}, \boldsymbol{\lambda}\right)$.*

Outer optimisation problem. As the convergence of $\hat{\mathcal{F}}_t$, we are left with the outer optimisation problem in Equation (8). Since it is a linear problem in $\boldsymbol{\omega}$, we can use a linear programming method which would lead to one of the vertices of $\hat{\mathcal{F}}$. But to be sure of an existence of a solution at each $t \in \mathbb{N}$, we try and derive some 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 $\lim_{t\to\infty} \mathcal{D}(\boldsymbol{\omega}, \hat{\boldsymbol{\mu}}_t, \hat{\mathcal{F}}) \to \mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \mathcal{F})$. Now, we ensure well-behavedness and existence of an optimal allocation for all $t > 0$.

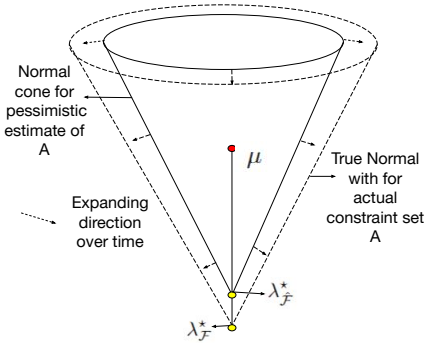

Figure 2: Evolution of the normal cone.

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

Characterising the 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.

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

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

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

Corollary 1 gives explicit upper and lower bound on characteristic bound for Gaussian rewards. It also shows explicit connection to the lower bounds under known constraints. Finally, it tells that the hardness of the pure exploration under constraints depends inversely on the condition number of the estimated and true constraints, which quantify their invertibility.

Corollary 1. *Let $d^2_{\boldsymbol{\pi}} \triangleq \frac{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}\|^2_2}$ be the norm of the projection of $\boldsymbol{\mu}$ on the policy gap $(\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi})$. The characteristic time $T_{\hat{\mathcal{F}}}(\boldsymbol{\mu})$ satisfies two bounds. (a) $\frac{2\sigma^2}{C_{\mathrm{known}}+2C_{\mathrm{unknown}}} \le T_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) \le \frac{2\sigma^2 K}{C_{\mathrm{known}}}$, where $C_{\mathrm{unknown}} \triangleq \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} d^2_{\boldsymbol{\pi}'}$ and $C_{\mathrm{known}} = \min_{\boldsymbol{\pi}''\in\nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} d^2_{\boldsymbol{\pi}''}$. (b) $T_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) \ge \frac{H}{\kappa^2_{\mathrm{known}}+2\kappa^2_{\mathrm{unknown}}}$. H is the sum of squares of gaps. κ_{known} and $\kappa_{\mathrm{unknown}}$ are condition numbers of $\mathbf{A}$ and $\tilde{\mathbf{A}}$.*

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 $C_{\mathrm{unknown}} = 0$. *(b) BAI without constraints.* In BAI, we consider only deterministic policies playing an arm. Thus, $d_{\boldsymbol{\pi}_a} = \frac{\boldsymbol{\mu}^T(\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}_a)}{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}_a\|_2} = \boldsymbol{\mu}^\star - \boldsymbol{\mu}_a$, i.e. the sub-optimality gap for arm a. Here, $\boldsymbol{\mu}^\star$ is the mean of the best arm. Then, the 'known' term in the denominator is the minimum squared sub-optimality gap and matches the BAI bound [CBJD24]. But the 'unknown' term in the lower bound is due to the unknown constraints. The term is the squared sub-optimality gap under the estimated feasible set $\hat{\mathcal{F}}$. Thus, our upper bound on characteristic time matches that of BAI [KCG16], while the lower bound has an added term due to the extra cost of exploring under unknown constraints. *(c) Safe linear BAI.* [WWJ21] stated a lower bound in for BAI under safety constraints but when the reward generation has a linear structure, and while the rewards and constraints are jointly generated for an arm. This is a bit different setting. Similar to our lower bound, they also show an added cost for constraints which is not directly comparable.

4 LATS and LAGEX: Algorithm design and analysis

Now, we propose two algorithms to conduct pure exploration with the Lagrangian relaxation of the 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 $\mathbf{A}^i$ for $i \in [d]$.*

These assumptions are standard in bandits under constraints [CBJD23, DK19, PGBJ20, PGB24].

Algorithm design. Any algorithm in pure exploration setting comprises of three main components.

Component 1 : Stopping rule. Stopping rule of an exploring algorithm decides when to stop sampling. While exploring, once we gather enough statistical information about the parameters in the system, the test statistic crosses the stopping threshold with the chosen confidence level δ and exploration stops to recommend the best policy.

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

$$\inf_{\boldsymbol{\lambda} \in \Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^{K} N_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t}, \lambda_a) > \beta(t,\delta) \wedge \|(\tilde{\mathbf{A}}_t - \mathbf{A})\boldsymbol{\omega}_t\|_\infty \leq \rho(t,\delta)\,,$$

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 $\rho(t,\delta)$ is defined in Lemma 4.

First, we need to check the Chernoff condition under the estimated alternative set as we do not know the true one. Secondly, it seems we also need to ensure that the constraint matrix has also concentrated around the true matrix. Though we show that to stop the algorithm we only need to ensure $(1-\delta)$-correctness. This phenomenon is stated formally in Lemma 2.

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

Thus, while implementing Algorithm 1 and 2 we just need to check the first condition to stop sampling.

Algorithm 1 LATS - **LA**grangian **T**rack and **S**top

Input : Time Horizon T, Confidence level $\delta > 0$
Initialization : $\hat{\mathbf{A}}_0 = \mathbf{0}_{d\times K}$, $\hat{\boldsymbol{\mu}}_0 = \mathbf{0}_K$, $\Sigma_0 = \boldsymbol{\lambda}\mathbf{1}_K$, l_0
Play each arm once to set μ_1 and $\hat{\mathbf{A}}_1$.
while $\beta(t,\delta) > \mathcal{D}(\boldsymbol{\omega}_t, \hat{\boldsymbol{\mu}}_t, \hat{\mathcal{F}}_t, \boldsymbol{l}_t^\star) \mid \|(\tilde{\mathbf{A}}_t - \mathbf{A})\boldsymbol{\omega}_t\|_\infty > \rho(t,\delta)$ **do** ⇝ Proposition 1)
  Optimal Policy: $\boldsymbol{\omega}_t^\star \in \arg\max_{\boldsymbol{\omega}\in\hat{\mathcal{F}}_t} \mathcal{D}(\boldsymbol{\omega}, \hat{\boldsymbol{\mu}}_t, \hat{\mathcal{F}}_t)$
  Optimize Lagrange Multiplier : $\boldsymbol{l}_t^\star \in \arg\min_{\boldsymbol{l}\in\mathcal{L}_t} \mathcal{D}(\omega_t^\star, \hat{\boldsymbol{\mu}}_t, \hat{\mathcal{F}}_t)$
  C-Tracking: Play $a_t \in \arg\min_{a\in[1,K]} N_{a,t} - \sum_{s=1}^t \omega_{a,s}^\star$
  Feedback : Observe reward r_t and cost $\mathbf{A}_{a_t}$, and update Σ_{t+1}, $\hat{\boldsymbol{\mu}}_{t+1}$ and $\hat{\mathbf{A}}_{t+1}$
end while
Recommended policy: $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}_t} = \arg\max_{\boldsymbol{\pi}\in\hat{\mathcal{F}}_t} \hat{\boldsymbol{\mu}}_t^T \boldsymbol{\pi}$

Algorithm 2 LAGEX - **LA**grangian **G**ame **EX**plorer

Input and Initialisation as same as LATS
Play each arm once to set μ_1 and $\hat{\mathbf{A}}_1$.
while $g(t,\delta) > \mathcal{D}(\boldsymbol{\omega}_t, \hat{\boldsymbol{\mu}}_t, \hat{\mathcal{F}}_t, \boldsymbol{l}_t^*) \mid \|(\tilde{\mathbf{A}}_t - \mathbf{A})\boldsymbol{\omega}_t\|_\infty > \rho(t,\delta)$ **do** ⇝ Proposition 1)
  Optimal allocation $\boldsymbol{\omega}_t$ ⇝ Using AdaGrad via Theorem 4
  Optimize Lagrange Multiplier: $\boldsymbol{l}_t^\star \in \arg\min_{\boldsymbol{l}\in\mathcal{L}_t} \mathcal{D}(\omega_t, \hat{\boldsymbol{\mu}}_t, \hat{\mathcal{F}}_t)$
  Compute confusing instance : $\boldsymbol{\lambda}_t$ ⇝ Via Proposition 1 plugging in $\boldsymbol{\omega}_t, \boldsymbol{l}_t^\star$
  Confidence intervals: for all $a \in [K]$ $[\alpha_{t,a}, \beta_{t,a}] : \{\zeta : N_{a,t} d(\mu_{t,a}, \zeta) \leq g(t)\}$ $U_t^a = \max\left\{\frac{g(t)}{N_{a,t}}, d(\alpha_{t,a}, \boldsymbol{\lambda}_{t,a}), d(\beta_{t,a}, \boldsymbol{\lambda}_{t,a})\right\}$
  Update loss for regret minimizer: $l(\boldsymbol{\omega}_t) = \langle \omega_t, U_t\rangle - \boldsymbol{l}_t^{\star T} \tilde{\mathbf{A}}_t \boldsymbol{\omega}_t$
  C-Tracking : Play $a_t \in \arg\min_{a\in[1,K]} N_{a,t} - \sum_{s=1}^t \omega_{a,s}^\star$
  Feedback : Observe reward r_t and cost $\mathbf{A}_{a_t}$, and update Σ_{t+1}, $\hat{\boldsymbol{\mu}}_{t+1}$ and $\hat{\mathbf{A}}_{t+1}$
end while
Recommended: $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}_t} = \arg\max_{\boldsymbol{\pi}\in\hat{\mathcal{F}}_t} \hat{\boldsymbol{\mu}}_t^T \boldsymbol{\pi}$

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 $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}_{\tau_\delta}} = \arg\max_{\boldsymbol{\pi}\in\hat{\mathcal{F}}_{\tau_\delta}} \hat{\boldsymbol{\mu}}_t^T \boldsymbol{\pi}$.

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

a. LATS. The algorithm LATS (Algorithm 1) uses a Track and Stop strategy adapted to the unknown constraint setting. We use markers in the algorithm which are novel approaches used to handle the challenge of estimating the feasible space per step. The algorithm first warms up the parameter estimates by playing each arm once. Then until the test statistic jumps the threshold, first in line 5, it calculates the optimal policy under the estimated feasible space at the current step solving the Lagrangian relaxed optimization problem in Proposition 1 by plugging in the best choice of Lagrangian multiplier optimized in previous step. Using this current optimal policy we optimize the Lagrangian multiplier maintaining the bounds of it's 1-norm stated in Theorem 1. It uses C-tracking [GK16] to track the action taken per step. At line 9 and 10, we basically observe the instantaneous reward and cost feedback and update the parameter estimates based on them.

Theorem 6. *Let $\mathfrak{s}$ be the shadow price $\mathfrak{s} \triangleq \frac{\Gamma_{\max}}{\Gamma_{\min}}$ of the slack Γ. For any $\alpha > 1$, the expected stopping time of LATS satisfies* $\lim_{\delta \to 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq \alpha T_{\mathcal{F}}(\boldsymbol{\mu})(1+\mathfrak{s})$.

b. LAGEX. Historically, algorithms base on track and stop mechanism seems to fail in case of larger problems where efficient optimization becomes a challenge due to the use of a max-min oracle per step (Line 5, Algorithm 2). To improve on this we land on the two-player zero sum game approach introduced in [DKM19b]. The second algorithm 2 we introduce in this work also starts by playing each arm once to initially start the estimation of the parameter in the system. Then it uses a **allocation player** (We have used AdaGrad) to optimize the allocation $\boldsymbol{\omega}_t$ in line 5 against the most confusing instance w.r.t current estimate of $\hat{\boldsymbol{\mu}}_t$ optimized by a **instance player** which minimizes $\sum_{a=1}^{K} \omega_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\lambda}_a) - \boldsymbol{l}_t^{\star T} \tilde{\mathbf{A}}_t \boldsymbol{\omega}_t$ w.r.t $\boldsymbol{\lambda} \in \lambda_{\hat{\mathcal{F}}_t}(\boldsymbol{\mu})$ by plugging in the chosen allocation, estimated feasible set $\tilde{\mathbf{A}}_t$ and the optimized Lagrangian multiplier $\boldsymbol{l}_t^{\star}$. Since our search space in closed and convex, the allocation player enjoys sub-linear regret of order $\mathcal{O}(\sqrt{t \log t})$, whereas the instance player computes the best confusing instance using the Lagrangian formulation of the weighted projection lemma stated in Proposition 1. Then, in line 11, Adagrad loss function is updated with a loss by introducing optimism as U_t defined in line 10 with an extra term that is novel in the literature to track the unknown constraint set. LAGEX also uses C-tracking similar to LATS to track the actions taken per step. Then it goes on to observe the instantaneous reward and cost to update the estimates for the next optimization step.

Theorem 7. *The expected sample complexity of LAGEX satisfies* $\lim_{\delta \to 0} \frac{\mathbb{E}(\tau_\delta)}{\ln(1/\delta)} \leq T_{\mathcal{F}}(\boldsymbol{\mu}) + 2\mathfrak{s}$.

Implications. Theorem 6 and 7 show both LATS and LAGEX are asymptotically stable. Also, for LATS, the effect of the unknown constraints in the form of **shadow price** arises in a multiplicative way, whereas for LAGEX, it is additive. Thus, LAGEX should show a lower sample complexity than LATS, which is later validated in the experiments.

5 Experimental analysis

Now, we experimentally evaluate performances of LAGEX and LAGTS across environments and against baseline algorithms. Code available in this Link.

Setup. We run the algorithms on a 64-bit 13th Gen Intel® Core™ i7-1370P × 20 processor machine with 32GB ram. We evaluate using a set of environments with mean vectors $[1.5, 1.0, \boldsymbol{\mu}_3, 0.4, 0.3, 0.2, 0.1]$. We impose two linear constraints $\boldsymbol{\pi}_1 + \boldsymbol{\pi}_2 + \boldsymbol{\pi}_3 \leq 0.5$ and $\boldsymbol{\pi}_4 + \boldsymbol{\pi}_5 \leq 0.5$. We conduct two experiments to validate universality and efficiency of LAGEX and LATS. We set $\delta = 0.01$ for all the experiments.

Experiment 1: Universality. We vary $\boldsymbol{\mu}_3$ from 0.5 to 2.5. For each environment, we plot the corresponding BAI lower bounds (in red) and lower bounds under constraints (in blue) in Figure 3. We observe that the constraint problem gets easier with increasing $\boldsymbol{\mu}_3$. In contrast, the BAI problem changes non-monotonically. BAI problem gets harder when $\boldsymbol{\mu}_3$ is around 1.5 as the suboptimality gap gets very small. But the constraint problem stays easier than BAI. In Figure 3, we also plot the median sample complexity of LAGEX across these environments over 500 runs. We observe that LAGEX grows parallel to the lower bound under constraints and can track it across environments.

Experiment 2: Efficiency against existing algorithms. We compare LAGEX and LATS with the two algorithms under known constraints, i.e. CTnS and CGE [CBJD23]. 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

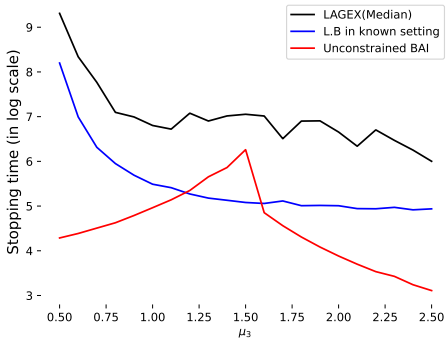

Figure 3: Lower bounds with and without constraints, and 500 runs of LAGEX for $\boldsymbol{\mu}_3 \in [0.5, 2.5]$.

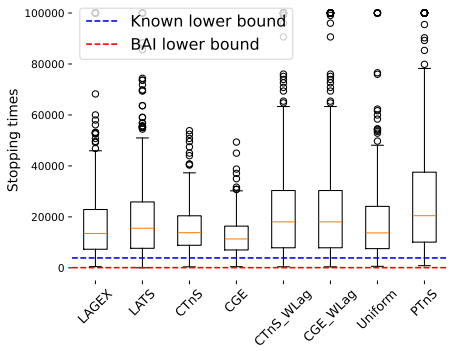

Figure 4: Sample complexity (median±std.) of algorithms over 500 runs for **hard env.**

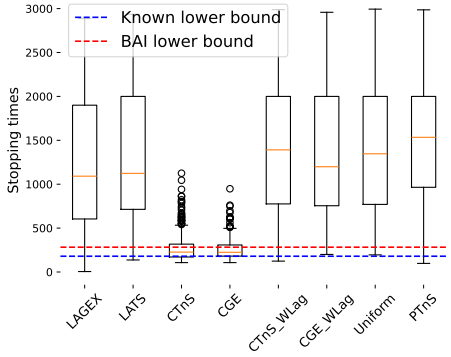

Figure 5: Sample complexity (median±std.) of algorithms over 500 runs for **easy env**.

TnS, where the algorithm solves a standard BAI problem and projects the allocation to the estimated feasible space. We run all these algorithms in two environment: (i) **hard** with $\boldsymbol{\mu}_3 = 0.5$ and (ii) **easy** with $\boldsymbol{\mu}_3 = 1.3$. We call the first environment hard as it is harder than BAI and similarly, the second environment easy. We plot three quantiles of the sample complexities of these algorithms over 500 runs in Figure 4 and 5. We observe that (i) among the algorithms with unknown constraints 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**.

6 Conclusion and Future Works

We study the problem of pure exploration under unknown linear constraints. This problem requires tracking both mean vectors and constraints to recommend a correct and feasible policy. We encompass this effect with a Lagrangian relaxation of the lower bound for known constraints. We further design an pessimistic estimate of constraints to ensure identification of the optimal feasible policy. These tools allows us to propose two algorithms LATS and LAGEX. We prove their sample complexity upper bounds, and conduct numerical experiments to find that LAGEX is the most efficient among baselines. In reality, constraints can be non-linear. Our concentration bounds are tailored for linearity. It would be interesting to extend our Lagrangian-based technique to nonlinear constraints.

Acknowledgments and Disclosure of Funding

We acknowledge the ANR JCJC project REPUBLIC (ANR-22-CE23-0003-01), the PEPR project FOUNDRY (ANR23-PEIA-0003), and the Inria-Japan associate team RELIANT for supporting the project.

References

[AAT19] Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints, 2019.

[ACBF02] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[AD14] Shipra Agrawal and Nikhil R. Devanur. Bandits with concave rewards and convex knapsacks. EC '14, page 989–1006, New York, NY, USA, 2014. Association for Computing Machinery.

[AD16] Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[ADL16] Shipra Agrawal, Nikhil R. Devanur, and Lihong Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, 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, 23–26 Jun 2016. PMLR.

[AKR21a] Maryam Aziz, Emilie Kaufmann, and Marie-Karelle Riviere. On multi-armed bandit designs for dose-finding clinical trials. *The Journal of Machine Learning Research*, 22(1):686–723, 2021.

[AKR21b] Maryam Aziz, Emilie Kaufmann, and Marie-Karelle Riviere. On multi-armed bandit designs for dose-finding trials. *Journal of Machine Learning Research*, 22(14):1–38, 2021.

[AyPS11] Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[BB62] Robert E. Bechhofer and Saul Blumenthal. 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, 1962.

[BCC24] Martino Bernasconi, Matteo Castiglioni, and Andrea Celli. No-regret is not enough! bandits with general constraints through adaptive regret minimization, 2024.

[Ber63] C. Berge. *Topological Spaces: Including a Treatment of Multi-valued Functions, Vector Spaces and Convexity*. Macmillan, 1963.

[BKS18] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *J. ACM*, 65(3), mar 2018.

[BMS09] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. 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, 2009.

[BMS10] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration for multi-armed bandit problems, 2010.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[CBJD23] Emil Carlsson, Debabrota Basu, Fredrik D. Johansson, and Devdatt Dubhashi. Pure Exploration in Bandits with Linear Constraints. In *EWRL 2023 – European Workshop on Reinforcement Learning*, Brussels, Belgium, September 2023.

[CBJD24] Emil Carlsson, Debabrota Basu, Fredrik Johansson, and Devdatt Dubhashi. Pure exploration in bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 334–342. PMLR, 2024.

[CGS22a] Tianrui Chen, Aditya Gangrade, and Venkatesh Saligrama. Doubly-optimistic play for safe linear bandits. *arXiv preprint arXiv:2209.13694*, 2022.

[CGS22b] Tianrui Chen, Aditya Gangrade, and Venkatesh Saligrama. Strategies for safe multi-armed bandits with logarithmic regret and risk. In *International Conference on Machine Learning*, pages 3123–3148. PMLR, 2022.

[CLK+14] Shouyuan Chen, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[CMC21] James Cheshire, Pierre Menard, and Alexandra Carpentier. The influence of shape constraints on the thresholding bandit problem, 2021.

[CWM+22a] Romain Camilleri, Andrew Wagenmaker, Jamie H Morgenstern, Lalit Jain, and Kevin G Jamieson. Active learning with safety constraints. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 33201–33214. Curran Associates, Inc., 2022.

[CWM+22b] Romain Camilleri, Andrew Wagenmaker, Jamie H Morgenstern, Lalit Jain, and Kevin G Jamieson. Active learning with safety constraints. *Advances in Neural Information Processing Systems*, 35:33201–33214, 2022.

[DK19] Rémy Degenne and Wouter M. Koolen. Pure exploration with multiple correct answers. In *Neural Information Processing Systems*, 2019.

[DKM19a] Rémy Degenne, Wouter M Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[DKM19b] Rémy Degenne, Wouter M. Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games, 2019.

[EDMM02a] Eyal Even-Dar, Shie Mannor, and Y. Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *Annual Conference Computational Learning Theory*, 2002.

[EDMM02b] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. 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, 2002.

[FJJR19] Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems*, 32, 2019.

[FN22] Fathima Zarin Faizal and Jayakrishnan Nair. Constrained pure exploration multi-armed bandits with a fixed budget. *arXiv preprint arXiv:2211.14768*, 2022.

[GK16] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, 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, 23–26 Jun 2016. PMLR.

[GK21a] Aurélien Garivier and Tomáš Kocák. Epsilon Best Arm Identification in Spectral Bandits. In *Thirtieth International Joint Conference on Artificial Intelligence {IJCAI-21}*, pages 2636–2642, Montreal, Canada, August 2021. International Joint Conferences on Artificial Intelligence Organization.

[GK21b] Aurélien Garivier and Emilie Kaufmann. Non-asymptotic sequential tests for overlapping hypotheses and application to near optimal arm identification in bandit models, 2021.

[GMRM18] Aurélien Garivier, Pierre Ménard, Laurent Rossi, and Pierre Menard. Thresholding bandit for dose-ranging: The impact of monotonicity, 2018.

[HTA24] Spencer Hutchinson, Berkay Turan, and Mahnoosh Alizadeh. Directional optimism for safe linear bandits, 2024.

[ISSS22] Nicole Immorlica, Karthik Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. *J. ACM*, 69(6), nov 2022.

[JMKK21] Marc Jourdan, Mojmír Mutný, Johannes Kirschner, and Andreas Krause. Efficient pure exploration for combinatorial bandits with semi-bandit feedback. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132, 2021.

[JN14] Kevin Jamieson and Robert Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2014.

[KCG16] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models, 2016.

[KCJ23] Newton Mwai Kinyanjui, Emil Carlsson, and Fredrik D. Johansson. Fast treatment personalization with latent bandits in fixed-confidence pure exploration. *Trans. Mach. Learn. Res.*, 2023, 2023.

[KGAYR17] Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi-Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits, 2017.

[KK21] Emilie Kaufmann and Wouter M. Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021.

[KSS18] Julian Katz-Samuels and Clay Scott. Feasible arm identification. In *International Conference on Machine Learning*, pages 2535–2543. PMLR, 2018.

[LJD+17] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1):6765–6816, 2017.

[LLSY21] Xin Liu, Bin Li, Pengyi Shi, and Lei Ying. An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints, 2021.

[LPJ22] Simon Lindståhl, Alexandre Proutiere, and Andreas Johnsson. Measurement-based admission control in sliced networks: A best arm identification approach. In *GLOBECOM 2022-2022 IEEE Global Communications Conference*, pages 1484–1490. IEEE, 2022.

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[LSY21] Xiaocheng Li, Chunlin Sun, and Yinyu Ye. The symmetry between arms and knapsacks: A primal-dual approach for bandits with knapsacks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6483–6492. PMLR, 18–24 Jul 2021.

[LTHK22] David Lindner, Sebastian Tschiatschek, Katja Hofmann, and Andreas Krause. Interactively learning preference constraints in linear bandits. In *International Conference on Machine Learning*, pages 13505–13527. PMLR, 2022.

[LVR+19] Pieter JK Libin, Timothy Verstraeten, Diederik M Roijers, Jelena Grujic, Kristof Theys, Philippe Lemey, and Ann Nowé. 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, 2019.

[LZYL23] Shaoang Li, Lan Zhang, Yingqi Yu, and Xiangyang Li. Optimal arms identification with knapsacks. In *International Conference on Machine Learning*, pages 20529–20555. PMLR, 2023.

[Ma] Will Ma. *Improvements and Generalizations of Stochastic Knapsack and Multi-Armed Bandit Approximation Algorithms: Extended Abstract*, pages 1154–1163.

[MAAT20] Ahmadreza Moradipari, Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Safe linear thompson sampling with side information, 2020.

[MCP14] Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz bandits: Regret lower bounds and optimal algorithms, 2014.

[MDP+11] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[MJTN20] Blake Mason, Lalit Jain, Ardhendu Tripathy, and Robert Nowak. Finding all \epsilon-good arms in stochastic bandits. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20707–20718. Curran Associates, Inc., 2020.

[MPK21] Arnab Maiti, Vishakha Patil, and Arindam Khan. Multi-armed bandits with bounded arm-memory: Near-optimal guarantees for best-arm identification and regret minimization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 19553–19565. Curran Associates, Inc., 2021.

[Pau64] Edward Paulson. A sequential procedure for selecting the population with the largest mean from k normal populations. *The Annals of Mathematical Statistics*, 35(1):174–180, 1964.

[PGB24] Aldo Pacchiano, Mohammad Ghavamzadeh, and Peter Bartlett. Contextual bandits with stage-wise constraints. *arXiv preprint arXiv:2401.08016*, 2024.

[PGBJ20] Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints, 2020.

[SCBC23] Xuedong Shang, Igor Colin, Merwan Barlier, and Hamza Cherkaoui. Price of safety in linear best arm identification, 2023.

[SJ19] Ashudeep Singh and Thorsten Joachims. Policy learning for fairness in ranking. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[Slu25] E. Slutsky. Über stochastische Asymptoten und Grenzwerte. Metron 5, Nr. 3, 3-89 (1925)., 1925.

[SS18] Karthik Abinav Sankararaman and Aleksandrs Slivkins. Combinatorial semi-bandits with knapsacks. In Amos Storkey and Fernando Perez-Cruz, 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, 09–11 Apr 2018.

[SSF23] Aleksandrs Slivkins, Karthik Abinav Sankararaman, and Dylan J Foster. Contextual bandits with packing and covering constraints: A modular lagrangian approach via regression. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 4633–4656. PMLR, 12–15 Jul 2023.

[TPRL20] Andrea Tirinzoni, Matteo Pirotta, Marcello Restelli, and Alessandro Lazaric. An asymptotically optimal primal-dual incremental algorithm for contextual linear bandits. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1417–1427. Curran Associates, Inc., 2020.

[TTCRJ12] Long Tran-Thanh, Archie Chapman, Alex Rogers, and Nicholas R. Jennings. 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, April 2012.

[VBW15] Sofia Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statistical Science*, 30:199–215, 05 2015.

[WBSJ21] Lequn Wang, Yiwei Bai, Wen Sun, and Thorsten Joachims. Fairness of exposure in stochastic bandits. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10686–10696. PMLR, 18–24 Jul 2021.

[WWJ21] Zhenlin Wang, Andrew Wagenmaker, and Kevin Jamieson. Best arm identification with safety constraints, 2021.

[WZZ23] Yuhang Wu, Zeyu Zheng, and Tingyu Zhu. Best arm identification with fairness constraints on subpopulations. In *2023 Winter Simulation Conference (WSC)*, pages 540–551. IEEE, 2023.

[ZY24] Yafei Zhao and Long Yang. Constrained contextual bandit algorithm for limited-budget recommendation system. *Engineering Applications of Artificial Intelligence*, 128:107558, 2024.

Appendix

Table of Contents

A Notations

Notation	Definition
Δ_K	K-simplex
T	Time Horizon
K	Number of Arms
$\mathbf{A}$	True constraint set
d	Number of constraints
$\mathcal{F}$	True feasible set w.r.t $\mathbf{A}$, $\mathcal{F} = \{\mathbf{A} \in \mathbb{R}^{N \times K} : \mathbf{A}\boldsymbol{\pi} \leq 0\}$
$\tilde{\mathbf{A}}_t$	Pessimistic estimate of constraint set at time t, $\tilde{\mathbf{A}}_t = \hat{\mathbf{A}} - f(t,\delta)\|\|\omega_t\|\|_{\Sigma_t^{-1}}$
$\hat{\mathcal{F}}_t$	Estimated feasible set w.r.t pessimistic estimate $\tilde{\mathbf{A}}$ at time t,, $\mathcal{F} = \{\tilde{\mathbf{A}}_t \in \mathbb{R}^{N \times K} : \tilde{\mathbf{A}}_t\boldsymbol{\pi} \leq 0\}$
$\mathbf{A}$	The action set of K possible choices
ω_t	Policy chosen at time t
a_t	Action at time t among K possible actions
N	Number of Constraints
Γ	Slack
σ^2	Variance of the reward distribution (Gaussian) of arms
T	Time Horizon
r_t, c_t	Reward and cost observed at time t
δ	Chosen confidence level
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 $\boldsymbol{\mu}$
$\nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})$	Neighbourhood set around $\boldsymbol{\pi}^\star_{\mathcal{F}}$
$\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})$	Neighbourhood set around $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$
τ_δ	Stopping time
$\boldsymbol{\pi}^\star_{\mathcal{F}}$	True optimal policy w.r.t actual constraint set $\mathbf{A}$
$\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$	Optimal Policy for the estimated feasible set

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 [BB62] and [Pau64] with the assumption of the populations being normally distributed. To talk about pure exploration setting, [EDMM02b], [BMS10] 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. [WWJ21] 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 [CBJD23] is closer as it tracks an optimal policy w.r.t to a known set of known constraints. On the other hand, [LLSY21] (Improvement over [PGBJ20] 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 1 and 2. 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. [LTHK22] 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 [FJJR19] [CWM$^+$22b] 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. [MAAT20] studies regret minimization using Linear Thompson Sampling (LTS) imposing known safety constraints on the chosen action. [AAT19] studies contextual bandits under unknown and unobserved linear constraint, whereas [KGAYR17] [PGBJ20] 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 [HTA24] [PGB24] [SCBC23] improved on regret guarantees and the first one relaxed the assumption of existence of a pessimistic safe space. [CGS22a] introduces doubly-optimistic setting to study safe linear bandit(SLB). [LLSY21] generalised the setting of [PGBJ20] 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 [CBJD23]. [LLSY21] 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 [WZZ23] 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 [WBSJ21] [SJ19] 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 [BKS18] [AD16] [ISSS22], [AD14] [ADL16] [SS18] [Ma] focused on mainly regret minimization, our setting aligns more as a special case of the Optimal Arm identification with Knapsack setting in [LZYL23,

TTCRJ12, LSY21]. 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 1 is an extension of the Track and Stop(TnS) strategy from [GK16], while the motivation for Algorithm 2 comes from the Gamified Explorer strategy from [DKM19b] where the lower bound is treated as a zero-sum game between the allocation and the instance player. We refer to [GK21b],[GK16],[KCG16],[DK19], [BV04],[JMKK21] 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 [CLK$^+$14] generalized the problem, then a line of work [AKR21b] [GMRM18] [CMC21] 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 existing 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 [AKR21a]. 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 [WWJ21]. 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 $\boldsymbol{\pi}^\star = \arg\max \boldsymbol{\mu}^T\boldsymbol{\pi}$, such that $\mathbf{I}\boldsymbol{\pi} \leq \mathbf{I}\boldsymbol{\theta}$. 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 [LZYL23, TTCRJ12, LSY21] and regret minimisation [BKS18, AD16, ISSS22, AD14, ADL16, SS18, Ma] 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 [LZYL23]. But as in regret-minimisation literature [SS18, Ma], 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 $\boldsymbol{\pi}^\star_\tau = \arg\max_{\boldsymbol{\pi}\in C_{\mathbf{A}}} \hat{\boldsymbol{\mu}}^T_{\tau_\delta}\boldsymbol{\pi}$, where $C_{\mathbf{A}} \triangleq \{\mathbf{A}\boldsymbol{\pi}_{\tau_\delta} \leq c\}$. 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 [KSS18]. 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 $t \in \mathbb{N}$, $\max_{\boldsymbol{\pi}\in\Delta_{K-1}} \boldsymbol{\mu}_t^T\boldsymbol{\pi}$ such that $f^T\boldsymbol{\pi} \geq \tau$ 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 [WZZ23] 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 $\boldsymbol{\pi}^\star = \arg\max_{\pi \in \Delta_K} \boldsymbol{\mu}^T \pi$ such that $\mathbb{1}^T_{\boldsymbol{\mu}_m \geq 0} \boldsymbol{\pi} = 1, \forall m \in [l]$. Thus, BAICS is a special case of ours.

Fairness of exposure in bandits. [WBSJ21] introduced positive merit based exposure of fairness constraints [SJ19] 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 [WBSJ21] 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 $\boldsymbol{\pi}^\star = \arg\max \boldsymbol{\mu}^T \boldsymbol{\pi}$ such that $\boldsymbol{\pi}^\star$ satisfies $\mathbf{A}^T_{f(\boldsymbol{\mu})} \boldsymbol{\pi} = 0$ where $\mathbf{A}_{f(\boldsymbol{\mu})}$ is of order $\frac{K(K-1)}{2} \times K$ and $\mathbf{A}_{f(\boldsymbol{\mu})}$ is expressed as,

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

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

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* (7) *i.e.*

$$\inf_{\boldsymbol{l}\in\mathbb{R}^d_+}\min_{\mathbf{A}'\in\mathcal{C}}\sup_{\boldsymbol{\omega}\in\hat{\mathcal{F}}}\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})}\boldsymbol{\omega}^\top d\left(\boldsymbol{\mu},\boldsymbol{\lambda}\right)-\boldsymbol{l}^T\tilde{\mathbf{A}}\boldsymbol{\omega}=\sup_{\boldsymbol{\omega}\in\hat{\mathcal{F}}}\min_{\boldsymbol{l}\in\mathcal{L}}\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})}\boldsymbol{\omega}^\top d\left(\boldsymbol{\mu},\boldsymbol{\lambda}\right)-\boldsymbol{l}^T\tilde{\mathbf{A}}\boldsymbol{\omega}\,. \quad (10)$$

Here, $\mathcal{L}\triangleq\{\boldsymbol{l}\in\mathbb{R}^d\mid 0\leq\|\boldsymbol{l}\|_1\leq\frac{1}{\gamma}T^{-1}_{\hat{\mathcal{F}}}(\hat{\boldsymbol{\mu}})\}$, where $\gamma\triangleq\min_{i\in[1,d]}\{-\tilde{\mathbf{A}}^i\boldsymbol{\omega}^*\}$, i.e. the minimum slack for pessimistic constraints w.r.t. the optimal allocation.

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 8. Then 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 $\nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})$. For that, let us remind the definitions of these sets. The estimated feasible set is defined as $\hat{\mathcal{F}}\triangleq\left\{\boldsymbol{\pi}\in\Delta_K:\tilde{\mathbf{A}}\boldsymbol{\pi}\leq 0\right\}$. The set of alternative (confusing) instances for the optimal policy $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$ is $\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})\triangleq\left\{\boldsymbol{\lambda}\in\mathbb{D}:\max_{\boldsymbol{\pi}\in\hat{\mathcal{F}}}\boldsymbol{\lambda}^T\boldsymbol{\pi}>\boldsymbol{\lambda}^T\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}\right\}$. For $\boldsymbol{\pi}'$ being a neighbour of $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$ or in other words, an extreme point in $\hat{\mathcal{F}}$, we decompose the alternative set as the union of half-spaces as,

$$\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})=\bigcup_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})}\left\{\boldsymbol{\lambda}:\boldsymbol{\lambda}^T(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}-\boldsymbol{\pi}')<0\right\}$$

We should note, $\boldsymbol{\pi}'$ shares at least $(K-1)$ active constraints with $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$. 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 neighbourhood of the optimal policy $\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})$ also shares $K-1$ active constraints with $\boldsymbol{\pi}^\star$, so $\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})$ is convex in $\boldsymbol{\pi}'$.

Let, $\boldsymbol{\pi}'_1$ and $\boldsymbol{\pi}'_2$ are two policies in the neighbourhood of $\boldsymbol{\pi}^\star$, such that for any alternative instance $\boldsymbol{\lambda}$, $\boldsymbol{\lambda}^T(\boldsymbol{\pi}'_1-\boldsymbol{\pi}'_2)\geq 0$ Above equation implies that the policy $\boldsymbol{\pi}'_1$ is closer to the optimal policy in the neighbourhood that the policy $\boldsymbol{\pi}'_2$
Therefore, any convex combination of these neighbourhood policy, $\boldsymbol{\lambda}^T\left(\boldsymbol{\pi}^\star-(a\boldsymbol{\pi}'_1+(1-a)\boldsymbol{\pi}'_2)\right)=\boldsymbol{\lambda}^T(\boldsymbol{\pi}^\star-\boldsymbol{\pi}'_2)-a\boldsymbol{\lambda}^T(\boldsymbol{\pi}'_1-\boldsymbol{\pi}'_2)\leq c$. Therefore, the set $\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})$ is also bounded and convex in $\boldsymbol{\pi}$
Also, since we are working with pessimistic estimate of $\mathbf{A}$, 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 neighbourhood $\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})$ is convex for any $\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})$
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 $\boldsymbol{\pi}^\star$ that uniquely solve the optimisation problem in Equation (8) and satisfy the constraints with strict inequality. As a consequence of this we claim Slater's sufficient conditions hold.

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 $\boldsymbol{\mu}^*$, we have $\|\boldsymbol{\mu}^*\|_1\leq\frac{1}{\gamma}\left(f(\bar{x})-q^*\right)$. Since Slater's conditions hold in our case for $\boldsymbol{\pi}$, we can write that the optimal solution of the Lagrangian dual,

$$0 \leq ||l_t^\star||_1 \leq \frac{1}{\gamma}\left(\mathcal{D}(\omega_t^\star, \hat{\boldsymbol{\mu}}_t, \hat{\mathcal{F}}_t) - \underbrace{\mathcal{D}(\boldsymbol{\pi}^\star, \hat{\boldsymbol{\mu}}_t, \hat{\mathcal{F}}_t)}_{\text{not tractable}}\right) \leq \frac{1}{\gamma}\mathcal{D}(\omega_t^\star, \hat{\boldsymbol{\mu}}_t, \hat{\mathcal{F}}_t, l_{t-1}^\star) = \frac{1}{\gamma}T_{\hat{\mathcal{F}}}^{-1}(\hat{\boldsymbol{\mu}})$$

where, $\gamma \triangleq \min_{i\in[1,d]}\{-\tilde{\mathbf{A}}^i\boldsymbol{\omega}^*\}$

Where, $\boldsymbol{\pi}^\star$ is the pure-exploration solution. We can replace the dual function with primal $\mathcal{D}(\omega_t^\star, \hat{\boldsymbol{\mu}}_t, \hat{\mathcal{F}}_t, l_{t-1}^\star)$ because it is always upper bounded by the dual function, so we don't have to explicitly calculate the dual function. Though it is not tractable anyway due to not knowing the pure exploration solution.

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 8. We define $\mathcal{L} \triangleq \{\boldsymbol{l} \in \mathbb{R}^d \mid 0 \leq \|\boldsymbol{l}\|_1 \leq \frac{1}{\gamma}T_{\hat{\mathcal{F}}}^{-1}(\hat{\boldsymbol{\mu}})\}$, where $\gamma \triangleq \min_{i\in[1,d]}\{-\tilde{\mathbf{A}}^i\boldsymbol{\omega}^*\}$ 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

$$\inf_{\boldsymbol{l}\in\mathbb{R}_+^d}\min_{\mathbf{A}'\in\mathcal{C}}\sup_{\boldsymbol{\omega}\in\hat{\mathcal{F}}}\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})}\boldsymbol{\omega}^\top d\left(\boldsymbol{\mu},\boldsymbol{\lambda}\right) - \boldsymbol{l}^T\tilde{\mathbf{A}}\boldsymbol{\omega} = \sup_{\boldsymbol{\omega}\in\hat{\mathcal{F}}}\min_{\boldsymbol{l}\in\mathcal{L}}\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})}\boldsymbol{\omega}^\top d\left(\boldsymbol{\mu},\boldsymbol{\lambda}\right) - \boldsymbol{l}^T\tilde{\mathbf{A}}\boldsymbol{\omega}\,.$$

□

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 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 $\boldsymbol{\mu}$, $\lim_{t\to\infty}\mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}}) \to \mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\mathcal{F})$ where $\boldsymbol{\lambda}^*$ is such that for any $\boldsymbol{\lambda}\in\mathcal{D}$: $\inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})}\boldsymbol{\omega}^T d\left(\boldsymbol{\mu},\boldsymbol{\lambda}\right) \geq \boldsymbol{\omega}^T d(\boldsymbol{\mu},\boldsymbol{\lambda}^*)$.*

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 $H(\tilde{\mathbf{A}}) \triangleq \left\{\boldsymbol{\pi}\in\Delta_K : \tilde{\mathbf{A}}\boldsymbol{\pi}\leq 0\right\}$ and the set function $\tilde{\mathbf{A}} \to H(\tilde{\mathbf{A}})\cap C$ where $C=\hat{\mathcal{F}}$ is a non-empty compact (proven in Section C subset of Δ_K Then the set $H(\tilde{\mathbf{A}})\cap C$ can be written as

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

To apply Theorem 10, $\{\tilde{\mathbf{A}}^r, r\in\mathbb{N}\}$, must be a convergent sequence of affine function. It is evident that $\tilde{\mathbf{A}}^r$ for any $r\in\mathbb{N}$ is an affine function since $\mathbf{A}$ is linear in $\mathbf{A}$ 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 $\{\tilde{\mathbf{A}}^r\}_{r\in\mathbb{N}}$ is a convergent sequence of functions. For ease of notation we will denote $\tilde{\mathbf{A}}_t$ for the t-th element of the sequence $\{\tilde{\mathbf{A}}^r\}_{r\in\mathbb{N}}$ for $t\in\mathbb{N}$.

The definition of the confidence radius for any constraint $i\in[d]$ follows from the Definition 2 as $f(\delta,t) \triangleq 1+\sqrt{\frac{1}{2}\log\frac{K}{\delta}+\frac{1}{4}\log\mathsf{det}\Sigma_t}$. 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 $i\in[d]$

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

where $Z \triangleq \frac{\hat{\mathbf{A}}_t^i-\mathbf{A}^i}{\sigma(\hat{\mathbf{A}}_t^i)}$ 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 $\sigma(\hat{\mathbf{A}}_t^i)=\mathcal{O}(\sqrt{\frac{\log t}{t}})$. Leveraging CLT at this point we say

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

Then by Slutsky's theorem [Slu25], we conclude $\hat{\mathbf{A}}_t^i - f(t,\delta)||\omega_t||_{\Sigma_t^{-1}} \xrightarrow{d} \mathbf{A}^i, \forall i\in[d]$.

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

1. $H(\tilde{\mathbf{A}})\cap C \subset \lim_{r\to\infty} H(\tilde{\mathbf{A}}^r)\cap C$

2. $\lim_{r\to\infty} H(\tilde{\mathbf{A}}^r) \cap C$ is a closed convex superset of $H(\tilde{\mathbf{A}}) \cap C$.

3. $H(\tilde{\mathbf{A}}) \cap C$ has non-empty interior because of the feasibility condition and no component in $\tilde{\mathbf{A}}$ is identically $\mathbf{0}$.

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

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

As the convergence of $\tilde{\mathbf{A}}_t$ 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 try to 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

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

We can 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 can 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 $\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) \triangleq \{\boldsymbol{\lambda} \in \mathbb{D} : \max_{\boldsymbol{\pi}\in\hat{\mathcal{F}}} \boldsymbol{\lambda}^T(\boldsymbol{\pi} - \boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}) > 0\}$. Let us denote $\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})$ as the neighbourhood of $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$. Any $\boldsymbol{\pi}' \in \hat{\mathcal{F}}$ is called a neighbour of $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$, if it is an extreme point of $\hat{\mathcal{F}}$ and shares (K-1) active constraints with $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$. Then, we can decompose the perturbed alt-set as $\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) = \bigcup_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \left\{\boldsymbol{\lambda} : \boldsymbol{\lambda}^T(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}') < 0\right\}$, which is a union of half-spaces for each neighbour. From this decomposition we can observe that $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$ is not the optimal policy for $\boldsymbol{\lambda}$, i.e, $\{\exists\boldsymbol{\pi}' \in \Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) : \boldsymbol{\lambda}^T(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}') < 0\}$. Then, it follows similar argument in [CBJD23] 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,

$$\mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \hat{\mathcal{F}}) = \min_{\boldsymbol{l}\in\mathcal{L}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \min_{\boldsymbol{\lambda}:\boldsymbol{\lambda}^T(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}')=0} \boldsymbol{\omega}^\top d\left(\boldsymbol{\mu}, \boldsymbol{\lambda}\right) - \boldsymbol{l}^T \tilde{\mathbf{A}}\boldsymbol{\omega}\,. \tag{11}$$

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 $\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})$. 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 $\hat{\boldsymbol{\lambda}}_n \in \Lambda_{\hat{\mathcal{F}}_n}(\boldsymbol{\mu})$ and we have to ensure it converges to the unique optimal $\boldsymbol{\lambda}^*$ i.e $\boldsymbol{\lambda}^\star \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu}) \subseteq \lim_{n\to\infty} \Lambda_{\hat{\mathcal{F}}_n}(\boldsymbol{\mu})$ 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 pessimistic choice of $\mathbf{A}$.
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. $\Lambda_{\hat{\mathcal{F}}_n}(\boldsymbol{\mu})$ 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 $\boldsymbol{\mu} \in \mathcal{D}$, $\lim_{n\to\infty} \Lambda_{\hat{\mathcal{F}}_n}(\boldsymbol{\mu}) = \Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})$. Since we have already proven uniqueness of $\boldsymbol{\lambda}$ in statement 2, we say $\hat{\boldsymbol{\lambda}}_n$ uniquely minimises $\Lambda_{\hat{\mathcal{F}}_n}(\boldsymbol{\mu})$. Now from the (ϵ, δ)-definition of limits we say if $\Lambda_{\hat{\mathcal{F}}_n}(\boldsymbol{\mu})$ 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 any $\boldsymbol{\mu} \in \mathcal{D}$, the optimization problem $\max_{\boldsymbol{\pi} \in \hat{\mathcal{F}}} \boldsymbol{\mu}^T \boldsymbol{\pi}$ has a unique solution if $\boldsymbol{\omega}^\star(\boldsymbol{\mu})$ satisfies the following conditions:*

1. *Both the sets $\hat{\mathcal{F}}$ and $\boldsymbol{\omega}^\star(\boldsymbol{\mu})$ are closed and convex.*
2. *$\forall \boldsymbol{\mu} \in \mathcal{D}$ and $\boldsymbol{\omega} \in \hat{\mathcal{F}}$, the function $\lim_{t\to\infty} \mathcal{D}(\boldsymbol{\omega}, \hat{\boldsymbol{\mu}}_t, \hat{\mathcal{F}})$ is continuous.*
3. *Reciprocal of the characteristic time function $\lim_{t\to\infty} T^{-1}_{\hat{\mathcal{F}}}(\boldsymbol{\mu})$ is continuous $\forall \boldsymbol{\mu} \in \mathcal{D}$.*
4. *$\boldsymbol{\mu} \in \mathcal{D} : \boldsymbol{\mu} \to \boldsymbol{\omega}^\star(\boldsymbol{\mu})$ is upper hemi-continuous.*

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 $\tilde{\mathbf{A}}\boldsymbol{\omega}_1 \leq 0$ and $\tilde{\mathbf{A}}\boldsymbol{\omega}_2 \leq 0$, their convex combination for any $\alpha \in [0, 1]$,

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

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

1. The complement of $\hat{\mathcal{F}}$, $\hat{\mathcal{F}}^c \triangleq \{\boldsymbol{\pi} \in \Delta_K : \tilde{\mathbf{A}}\boldsymbol{\pi} > 0\}$ is an open set.
2. 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 $\lim_{t\to\infty} \hat{\mathcal{F}} \to \mathcal{F}$. Also by convexity of KL and CLT we claim $\hat{\mu}_t \to \mu$ as $t \to \infty$ and since $\boldsymbol{\omega}$ is linear in $\mathcal{D}(\boldsymbol{\omega}, \mu, \hat{\mathcal{F}})$ it will converge to $\boldsymbol{\omega}^*(\mu)$ as $t \to \infty$, also due to convergence of $\hat{\mu}_t$. Then we can say that the limiting value is same as the value if we plug in the limits in $\mathcal{D}$ i.e $\lim_{t\to\infty} \mathcal{D}(\boldsymbol{\omega}, \hat{\mu}_t, \hat{\mathcal{F}}_t) = \mathcal{D}(\boldsymbol{\omega}^*(\mu), \mu, \mathcal{F})$. So we ensure the continuity of $\lim_{t\to\infty} \mathcal{D}(\boldsymbol{\omega}, \hat{\boldsymbol{\mu}}_t, \hat{\mathcal{F}})$.

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 [MCP14] (see Lemma 12) for this proof. We denote $Q(\tilde{\mathbf{A}}') \triangleq \lim_{\hat{\mathcal{F}}\to\mathcal{F}} \max_{\boldsymbol{\omega}\in\Delta_K} \left\{ \sum_{a=1}^K \boldsymbol{\omega}_a d\left(\boldsymbol{\mu}, \boldsymbol{\lambda}\right) - \boldsymbol{l}^T\tilde{\mathbf{A}}''\boldsymbol{\omega} \middle| \tilde{\mathbf{A}}''\boldsymbol{\omega} \leq 0, \boldsymbol{\omega}_a \geq 0 \forall i \in [K] \right\} = \boldsymbol{\omega}(\boldsymbol{\mu})$ where $\tilde{\mathbf{A}}'' \in \mathbb{R}^{K\times K}$ is the rank-1 update of $\tilde{\mathbf{A}}'$ which is a sub-matrix of $\tilde{\mathbf{A}}$ with K number of active constraints. We define limiting set as

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

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

1. The function $\boldsymbol{\omega}^*(\mu)$ is continuous in $(\mathbb{R}^{K\times K}) \times \mathbb{R}^K$
2. $\boldsymbol{\omega}^*(\mu)$ 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 $\max_{\boldsymbol{\pi}\in\hat{\mathcal{F}}} \boldsymbol{\mu}^T\boldsymbol{\pi}, \forall \mu \in \mathcal{D}$ 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 $\sigma^2 > 0$, and $\mathrm{Diag}(1/\boldsymbol{\omega}_a)$ be a K-dimensional diagonal matrix with a-th diagonal entry $1/\boldsymbol{\omega}_a$. Then, we get*

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

Proof. We start the proof by the definition of $\mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \hat{\mathcal{F}})$ as per Equation (9)

$$\begin{aligned}\mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \hat{\mathcal{F}}) &= \min_{\boldsymbol{l}\in\mathcal{L}} \min_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})} \left\{ \sum_{a=1}^{k} \boldsymbol{\omega}_a d(\boldsymbol{\mu}_a, \boldsymbol{\lambda}_a) - \boldsymbol{l}^T\tilde{\mathbf{A}}\boldsymbol{\omega} \right\} \\ &= \min_{\boldsymbol{l}\in\mathcal{L}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \min_{\boldsymbol{\lambda}:\boldsymbol{\lambda}^T(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}-\boldsymbol{\pi}')=0} \left\{ \sum_{a=1}^{k} \boldsymbol{\omega}_a d(\boldsymbol{\mu}_a, \boldsymbol{\lambda}_a) - \boldsymbol{l}^T\tilde{\mathbf{A}}\boldsymbol{\omega} \right\} \rightsquigarrow \text{via Proposition 1}\end{aligned} \tag{12}$$

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

$$\mathcal{L}(\boldsymbol{\omega}, \boldsymbol{\mu}, \hat{\mathcal{F}}) = \min_{\boldsymbol{l}\in\mathcal{L}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \left\{ \sum_{a=1}^{K} \boldsymbol{\omega}_a d\left(\boldsymbol{\mu}_a, \boldsymbol{\lambda}_a\right) - \boldsymbol{l}^T\tilde{\mathbf{A}}\boldsymbol{\omega} - \gamma \sum_{a=1}^{K} \boldsymbol{\lambda}_a \boldsymbol{v}_a \right\}$$

where $\boldsymbol{v}_a \triangleq (\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}')_a$.

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 —

$$\mathcal{L}(\boldsymbol{\omega}, \boldsymbol{\mu}, \hat{\mathcal{F}}) = \min_{\gamma\in\mathbb{R}_+} \min_{\boldsymbol{l}\in\mathcal{L}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \left\{ \sum_{a=1}^{K} \boldsymbol{\omega}_a \frac{(\boldsymbol{\mu}_a - \boldsymbol{\lambda}_a)^2}{2\sigma^2} - \boldsymbol{l}^T\tilde{\mathbf{A}}\boldsymbol{\omega} - \gamma \sum_{a=1}^{K} \boldsymbol{\lambda}_a \boldsymbol{v}_a \right\} \tag{13}$$

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 13 we get

$$\mathcal{L}(\boldsymbol{\omega}, \boldsymbol{\mu}, \hat{\mathcal{F}}) = \min_{\gamma\in\mathbb{R}_+} \min_{\boldsymbol{l}\in\mathcal{L}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \left\{ \sum_{a=1}^{K} \gamma^2 \frac{\boldsymbol{v}_a^2\sigma^2}{2\boldsymbol{\omega}_a} - \boldsymbol{l}^T\tilde{\mathbf{A}}\boldsymbol{\omega} - \gamma \sum_{a=1}^{K} \boldsymbol{\mu}_a \boldsymbol{v}_a \right\} \tag{14}$$

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

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

Putting the value of γ in Equation 14, we get —

$$\mathcal{L}(\boldsymbol{\omega}, \boldsymbol{\mu}, \hat{\mathcal{F}}) = \min_{\boldsymbol{l}\in\mathcal{L}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \left\{ \frac{(\boldsymbol{\mu}^T v)^2}{2\sigma^2 \sum_{a=1}^{K} \frac{\boldsymbol{v}_a^2}{\boldsymbol{\omega}_a}} - \boldsymbol{l}^T\tilde{\mathbf{A}}\boldsymbol{\omega} \right\}$$

$$= \min_{\boldsymbol{l}\in\mathcal{L}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \left\{ \frac{1}{2\sigma^2} \frac{[\boldsymbol{\mu}^T(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}')]^2}{\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} - \boldsymbol{l}^T\tilde{\mathbf{A}}\boldsymbol{\omega} \right\}$$

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

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

□

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

Corollary 1. ***Part (a)*** *Let* $d^2_{\boldsymbol{\pi}} \triangleq \frac{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}\|^2_2}$ *be the norm of the projection of* $\boldsymbol{\mu}$ *on the policy gap* $(\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi})$. *Then, the characteristic time* $T_{\hat{\mathcal{F}}}(\boldsymbol{\mu})$ *satisfies* $\frac{2\sigma^2}{C_{\mathrm{known}}+2C_{\mathrm{unknown}}} \leq T_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) \leq \frac{2\sigma^2 K}{C_{\mathrm{known}}}$, *where* $C_{\mathrm{unknown}} \triangleq \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} d^2_{\boldsymbol{\pi}'}$ *and* $C_{\mathrm{known}} = \min_{\boldsymbol{\pi}''\in\nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} d^2_{\boldsymbol{\pi}''}$.

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

$$\begin{aligned}\mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}}) - \mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\mathcal{F}) = &\min_{\boldsymbol{l}\in\mathcal{L}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \left\{ \frac{1}{2\sigma^2} \frac{\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} - \boldsymbol{l}^T\tilde{\mathbf{A}}\boldsymbol{\omega} \right\} \\ &- \min_{\boldsymbol{\pi}''\in\nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} \frac{1}{2\sigma^2} \frac{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}''\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}''\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} \end{aligned} \tag{15}$$

Let us remind, due to pessimistic choice of $\tilde{\mathbf{A}}$, $\mathcal{F} \subseteq \hat{\mathcal{F}}$. $\boldsymbol{\pi}'$ is a neighbour 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,

$$\begin{aligned}\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}''\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)} &= \|(\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}') - (\boldsymbol{\pi}'' - \boldsymbol{\pi}')\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)} \\ &\leq 2\left\{ \|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)} + \|\boldsymbol{\pi}'' - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)} \right\} \\ &= 2\{t_1^2 + t_2^2\}\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}\end{aligned}$$

Putting the above inequality in Equation 15, we get —

$$\begin{aligned}&\mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}}) - \mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\mathcal{F}) \\ \leq &\min_{\boldsymbol{l}\in\mathcal{L}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \left\{ \frac{1}{2\sigma^2} \frac{\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} - \boldsymbol{l}^T\tilde{\mathbf{A}}\boldsymbol{\omega} \right\} \\ \leq &\frac{1}{2\sigma^2} \min_{\boldsymbol{l}\in\mathcal{L}} \min_{\boldsymbol{\pi}'\in\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \min_{\boldsymbol{\pi}''\in\nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} \left[\frac{2\{t_1^2+t_2^2\}\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T} - \|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}''\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{2\{t_1^2+t_2^2\}\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} \right. \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \left. - \boldsymbol{l}^T\tilde{\mathbf{A}}\boldsymbol{\omega} \right]\end{aligned} \tag{16}$$

Now, we have already proven in Lemma 6 that $(-\boldsymbol{l}^T\tilde{\mathbf{A}}\boldsymbol{\omega}) \leq \mathcal{D}(\boldsymbol{\omega},\boldsymbol{\mu},\hat{\mathcal{F}})\psi$ where, $\psi \triangleq \frac{\|(\tilde{\mathbf{A}}-\mathbf{A})\boldsymbol{\omega}\|_\infty + + \max_{i\in[1,N]}\Gamma}{\gamma}$. Also,

$$\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}''\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T} = \|(\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}') - (\boldsymbol{\pi}'' - \boldsymbol{\pi}')\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}$$

$$
\begin{aligned}
&\geq 2\left\{ \|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T} - \|(\boldsymbol{\pi}'' - \boldsymbol{\pi}')\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T} \right\} \\
&= 2\left\{ t_1^2 \|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T} - t_2^2 \|(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}')\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T} \right\} \\
&= 2(t_1^2 - t_2^2)\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}
\end{aligned}
$$

Therefore Equation 16 gives us

$$
\begin{aligned}
&(1+\psi)\mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \hat{\mathcal{F}}) \\
&\leq \mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \mathcal{F}) + \frac{1}{2\sigma^2} \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \min_{\boldsymbol{\pi}'' \in \nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} \frac{2\{t_1^2 + t_2^2\} - 2(t_1^2 - t_2^2)}{2\{t_1^2 + t_2^2\}} \frac{\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} \\
&= \mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \mathcal{F}) + \frac{1}{2\sigma^2} \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \min_{\boldsymbol{\pi}'' \in \nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} \frac{2t_2^2}{\{t_1^2 + t_2^2\}} \frac{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} \qquad (17)
\end{aligned}
$$

Now to get the lower bound on characteristic time in unknown constraint setting, we take maximum over $\boldsymbol{\omega}$ in Equation 17,

$$
\begin{aligned}
T^{-1}_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) &\leq \frac{1}{(1+\psi)} \left\{ T^{-1}_{\mathcal{F}}(\boldsymbol{\mu}) + \frac{1}{2\sigma^2} \max_{\boldsymbol{\omega}} \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \min_{\boldsymbol{\pi}'' \in \nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} \frac{2t_2^2}{\{t_1^2 + t_2^2\}} \frac{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} \right\} \\
&< T^{-1}_{\mathcal{F}}(\boldsymbol{\mu}) + \frac{1}{\sigma^2} \max_{\boldsymbol{\omega}} \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \frac{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} \qquad (18)
\end{aligned}
$$

To get the lower bound on characteristic time we minimize $\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}$ with the simplex constraint $\sum_{a=1}^K \boldsymbol{\omega}_a = 1$ for each and every neighbour in $\nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})$ and $\nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})$. the lagrangian formulation gives us the solution, $\boldsymbol{\omega}_a = \frac{|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'|_a}{\sum_{a=1}^K |\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'|_a}$. Therefore, $\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)} \geq \|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'\|_1^2 \geq \|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'\|_2^2$, since $\forall a \in [1, K] : \|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'\|_a \leq 1$.Then, putting the value of $\boldsymbol{\omega}_a$ in Equation 18,

$$
T^{-1}_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) \leq T^{-1}_{\mathcal{F}}(\boldsymbol{\mu}) + \frac{1}{\sigma^2} \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \frac{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'\|_2^2}
$$

For ease of comparison, we follow the same notation used in [CBJD23] and define for any $\boldsymbol{\pi} \in \Delta_{K-1}$, $d^2_{\boldsymbol{\pi}} = \frac{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}\|_2^2}$, which is the squared distance between $\boldsymbol{\mu}$ and the hyper-plane $(\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}) = \mathbf{0}$. Therefore, using [CBJD23, Corollary 1], we get

$$
T^{-1}_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) \leq \frac{1}{2\sigma^2} \left(\min_{\boldsymbol{\pi}'' \in \nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} d^2_{\boldsymbol{\pi}''} + 2 \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} d^2_{\boldsymbol{\pi}'} \right)
$$

Therefore lower bound on the characteristic time is given by,

$$
T_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) \geq \left(\frac{2\sigma^2}{\min_{\boldsymbol{\pi}'' \in \nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} d^2_{\boldsymbol{\pi}''} + 2 \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} d^2_{\boldsymbol{\pi}'}} \right)
$$

Now, let us find an upper bound on the characteristic time. We have already shown that, $\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}''\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T} \geq 2(t_1^2 - t_2^2)\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}$, which also implies $\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}''\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)} \geq 2(t_1^2 - t_2^2)\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}$.

Using this result in Equation 15,

$$
\begin{aligned}
&\mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \hat{\mathcal{F}}) - \mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \mathcal{F}) \\
&\geq \frac{1}{2\sigma^2} \left\{ \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \frac{\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} - \min_{\boldsymbol{\pi}'' \in \nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} \frac{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}''\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{2(t_1^2 - t_2^2)\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}''\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} \right\}
\end{aligned}
$$

$$= \frac{1}{2\sigma^2}\left\{ \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \min_{\boldsymbol{\pi}'' \in \nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} \frac{2(t_1^2 - t_2^2)\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T} - \|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}''\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{2(t_1^2 - t_2^2)\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} \right\} \tag{19}$$

We have already shown, $\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}''\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)} \leq 2(t_1^2 + t_2^2)\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}$. Therefore $\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}''\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T} \leq 2(t_1^2 + t_2^2)\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}$. Consequently Equation 19 gives,

$$\begin{aligned}
&\mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \hat{\mathcal{F}}) - \mathcal{D}(\boldsymbol{\omega}, \boldsymbol{\mu}, \mathcal{F}) \\
&\geq \frac{1}{2\sigma^2}\left\{ \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \min_{\boldsymbol{\pi}'' \in \nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} \frac{2(t_1^2 - t_2^2)\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T} - 2(t_1^2 + t_2^2)\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{2(t_1^2 - t_2^2)\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} \right\} \\
&= \frac{1}{2\sigma^2}\left\{ \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \min_{\boldsymbol{\pi}'' \in \nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} \left(\frac{2t_2^2}{t_2^2 - t_1^2}\right) \frac{\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/\boldsymbol{\omega}_a)}} \right\} \geq 0
\end{aligned}$$

$$\implies T^{-1}_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) \geq T^{-1}_{\mathcal{F}}(\boldsymbol{\mu}) \geq \frac{1}{2\sigma^2 K}\left\{ \min_{\boldsymbol{\pi}'' \in \nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} d^2_{\boldsymbol{\pi}''} \right\}$$

Therefore, upper bound on the characteristic time is given by,

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

Let, $C_{\mathrm{unknown}} \triangleq \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} d^2_{\boldsymbol{\pi}'}$ and $C_{\mathrm{known}} \triangleq \min_{\boldsymbol{\pi}'' \in \nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} d^2_{\boldsymbol{\pi}''}$ and the result follows. □

F.2 Impact of unknown linear constraints: Proof of Corollary 1 Part (b)

Corollary 1. ***Part (b)*** *Let $d^2_{\boldsymbol{\pi}} \triangleq \frac{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}\|^2_2}$ be the norm of the projection of $\boldsymbol{\mu}$ on the policy gap $(\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi})$. Then, the characteristic time $T_{\hat{\mathcal{F}}}(\boldsymbol{\mu})$ satisfies $T_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) \geq \frac{H}{\kappa^2_{\mathrm{known}} + 2\kappa^2_{\mathrm{unknown}}}$. H is the sum of squares of gaps. κ_{known} and $\kappa_{\mathrm{unknown}}$ are condition numbers of $\mathbf{A}$ and $\tilde{\mathbf{A}}$.*

Proof. We dictate this proof in total four steps. The first step mainly deals with a sub-matrix $\tilde{\mathbf{A}}'$of $\tilde{\mathbf{A}}$ with K number if of active constraints where it will be shown that we can get any neighbouring policy of $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$ just by a rank-1 update of $\tilde{\mathbf{A}}'$ which means we just one of the constraints inactive for $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$. Using this result we will find an upper bound in the next step on the deviation between $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$ and any of it's neighbouring policy sharing $K-1$ number of active constraints. Once we find the upper bound on this deviation, in the third step we give a new expression for the characteristic time and we conclude the fourth step by reducing a new lower bound characterised the condition number of $\tilde{\mathbf{A}}'$ and $\hat{\mathbf{A}}$, where $\hat{\mathbf{A}}$ is the sub-matrix of the actual constraint matrix $\mathbf{A}$ with at least K number of active constraints.

Step 1 : Optimal policy to neighbouring policy via rank-1 update. We have the pessimistic estimate of $\mathbf{A}$ as $\tilde{\mathbf{A}}$, which gives us the perturbed feasible space $\hat{\mathcal{F}}$. Let, $\tilde{\mathbf{A}}'$ be a sub-matrix of $\tilde{\mathbf{A}}$ such that it consists of K linearly independent rows of $\tilde{\mathbf{A}}$ active at $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$. We can then say $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} \in \mathrm{Null}(\tilde{\mathbf{A}}')$, where, $\mathrm{Null}(\tilde{\mathbf{A}}')$ is the null space of $\tilde{\mathbf{A}}'$. Now for some neighbour of $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$, $\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})$ belongs to the null space of some $\tilde{\mathbf{A}}''$, where this $\tilde{\mathbf{A}}''$ can be expressed as a rank-1 update of $\tilde{\mathbf{A}}'$. Specifically, $\tilde{\mathbf{A}}'' = \tilde{\mathbf{A}}' + e_r(a''_r - a'_r)^T$. Here, a'_r is the column corresponding to the r-th constraint of $\tilde{\mathbf{A}}'$. We just want to replace this column with a new column a''_r, so we set e_r as a vector which has 1 at the r-th position and 0 everywhere else.

Step 2: Bounding distance between neighbor and optimal policy. From Equation 18, we have —

$$T^{-1}_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) < T^{-1}_{\mathcal{F}}(\boldsymbol{\mu}) + \frac{1}{2\sigma^2} \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} 2\frac{(\boldsymbol{\mu}^T(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'))^2}{\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_2} \tag{20}$$

So, it becomes evident that we need to get an upper bound on the quantity $\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|$. Let us start with,

$$\tilde{\mathbf{A}}'(\boldsymbol{\pi}' - \boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}) = \tilde{\mathbf{A}}'\boldsymbol{\pi}' \text{ since, } \boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} \in \mathrm{Null}(\tilde{\mathbf{A}}')$$

$$
\begin{aligned}
&= \left\{ \tilde{\mathbf{A}}'' + e_r({a'}_r - {a''}_r)^T \right\} \boldsymbol{\pi}' \\
&= e_r({a'}_r - {a''}_r)^T \boldsymbol{\pi}' \\
&= e_r {a'}_r^T \boldsymbol{\pi}' \text{ , since } {a''}_r \in \text{ column space of } \tilde{\mathbf{A}}'' \\
\implies (\boldsymbol{\pi}' - \boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}) &= {\tilde{\mathbf{A}}'}^{-1} (e_r {a'}_r^T) \boldsymbol{\pi}'
\end{aligned}
$$

We denote $\xi \triangleq {a'}_r^T \boldsymbol{\pi}'$ as it is the slack for the new r-th row/constraint. We also define $\sigma_{\min}(\tilde{\mathbf{A}}')$ and $\sigma_{\max}(\tilde{\mathbf{A}}')$ be respectively the minimum and maximum singular value of $\tilde{\mathbf{A}}'$. Also, let $\kappa_{\text{unknown}} \triangleq \frac{\sigma_{\max}(\tilde{\mathbf{A}}')}{\sigma_{\min}(\tilde{\mathbf{A}}')}$ be the minimum condition number for $\tilde{\mathbf{A}}'$. Then by property of singular value of a matrix, it follows —

$$
\begin{aligned}
\frac{1}{\sigma_{\min}(\tilde{\mathbf{A}}')} = \sigma_{\min}({\tilde{\mathbf{A}}'}^{-1}) &\leq \min_{v:\|v\|_2=1} \|{\tilde{\mathbf{A}}'}^{-1} v\| \leq \|{\tilde{\mathbf{A}}'}^{-1} e_r\| \leq \max_{v:\|v\|_2=1} \|{\tilde{\mathbf{A}}'}^{-1} v\| \\
&\leq \sigma_{\max}({\tilde{\mathbf{A}}'}^{-1}) = \frac{1}{\sigma_{\max}(\tilde{\mathbf{A}}')} \\
\implies \frac{|\xi|}{\sigma_{\min}(\tilde{\mathbf{A}}')} \leq \|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|_2 &\leq \frac{\xi}{\sigma_{\max}(\tilde{\mathbf{A}}')}
\end{aligned}
$$

Step 3 : A new expression for Characteristic time. Plugging in the above obtained bound on $\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|_2$ in the expression of inverse of characteristic time in Theorem 4, we get a new expression for the characteristic time.

$$
\begin{aligned}
\frac{1}{2\sigma^2} \frac{\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{\|\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}'\|^2_{\text{Diag}(1/\boldsymbol{\omega}_a)}} &= \frac{1}{2\sigma^2} \frac{\|{\tilde{\mathbf{A}}'}^{-1}(e_r {a'}_r^T)\boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{\|{\tilde{\mathbf{A}}'}^{-1}(e_r {a'}_r^T)\boldsymbol{\pi}'\|^2_{\text{Diag}(1/\boldsymbol{\omega}_a)}} \\
&= \frac{1}{2\sigma^2} \frac{\|{\tilde{\mathbf{A}}'}^{-1} e_r \xi\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{\|{\tilde{\mathbf{A}}'}^{-1} e_r \xi\|^2_{\text{Diag}(1/\boldsymbol{\omega}_a)}} \\
&= \frac{1}{2\sigma^2} \frac{(\Delta^T {\tilde{\mathbf{A}}'}^{-1} e_r)^2}{\|{\tilde{\mathbf{A}}'}^{-1} e_r\|^2_{\text{Diag}(1/\boldsymbol{\omega}_a)}}
\end{aligned}
$$

where Δ is the vector of sub-optimality gaps of arms, i.e $\Delta_a \triangleq \boldsymbol{\mu}^\star - \boldsymbol{\mu}_a$. Then, the new expression for the inverse of characteristic time is as follows,

$$
T^{-1}_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) = \max_{\boldsymbol{\omega} \in \hat{\mathcal{F}}} \min_{\boldsymbol{l} \in \mathcal{L}} \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \left\{ \frac{1}{2\sigma^2} \frac{(\Delta^T {\tilde{\mathbf{A}}'}^{-1} e_r)^2}{\|{\tilde{\mathbf{A}}'}^{-1} e_r\|^2_{\text{Diag}(1/\boldsymbol{\omega}_a)}} - \boldsymbol{l}^T \tilde{\mathbf{A}} \boldsymbol{\omega} \right\} \tag{21}
$$

Step 4 : New expression for the lower bound.

Combining Equation 18 and the new expression in Equation 21, we get —

$$
\begin{aligned}
T^{-1}_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) &< T^{-1}_{\mathcal{F}}(\boldsymbol{\mu}) + \frac{1}{\sigma^2} \max_{\boldsymbol{\omega}} \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \frac{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'\|^2_{\text{Diag}(1/\boldsymbol{\omega}_a)}} \\
&\leq T^{-1}_{\mathcal{F}}(\boldsymbol{\mu}) + \frac{1}{\sigma^2} \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \frac{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^T}}{\|\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}'\|^2_2} \\
&= T^{-1}_{\mathcal{F}}(\boldsymbol{\mu}) + \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \frac{1}{2\sigma^2} \frac{(\Delta^T {\tilde{\mathbf{A}}'}^{-1} e_r)^2}{\|{\tilde{\mathbf{A}}'}^{-1} e_r\|^2_2} \\
&\leq T^{-1}_{\mathcal{F}}(\boldsymbol{\mu}) + \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \frac{\Delta^2}{2\sigma^2} \frac{\sigma_{\max}(\tilde{\mathbf{A}}')}{\sigma_{\min}(\tilde{\mathbf{A}}')}
\end{aligned}
$$

For ease of comparing, we denote $H \triangleq \frac{2\sigma^2}{\Delta^2}$. Therefore, the new expression of characteristic time satisfies,

$$T^{-1}_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) \leq \min_{\boldsymbol{\pi}'' \in \nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})} \frac{\kappa^2_{\text{known}}}{H} + \min_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \frac{\kappa^2_{\text{unknown}}}{H} = \frac{1}{H}\left(\kappa^2_{\text{known}} + \kappa^2_{\text{unknown}}\right)$$

where, $\kappa_{\text{known}} \triangleq \frac{\sigma_{\max}(\hat{\mathbf{A}})}{\sigma_{\min}(\hat{\mathbf{A}})}$ and $\kappa_{\text{unknown}} \triangleq \frac{\sigma_{\max}(\tilde{\mathbf{A}}')}{\sigma_{\min}(\tilde{\mathbf{A}}')}$, $\hat{\mathbf{A}}$ and $\tilde{\mathbf{A}}'$ being the sub-matrix of $\mathbf{A}$ and $\tilde{\mathbf{A}}$ having Klinearly independent rows.

Consequently expected stopping time is then lower bounded by

$$\mathbb{E}[\tau_\delta] \geq \frac{H}{\kappa^2_{\text{known}} + \kappa^2_{\text{unknown}}} \text{kl}(\delta \| 1-\delta)$$

□

G Sample Complexity upper bounds (Analysis of algorithms)

G.1 Stopping Criterion

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

$$\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^{K} N_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t}, \lambda_a) > \beta(t,\delta) \wedge \|(\tilde{\mathbf{A}}_t - \mathbf{A})\boldsymbol{\omega}_t\|_\infty \leq \rho(t,\delta)\,,$$

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* $\rho(t,\delta)$ *is in Lemma 4.*

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.

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

Algorithm 1 and 2 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 $\{\exists t \in \mathbb{N} : \|(\tilde{\mathbf{A}}_t - \mathbf{A})N_t\|_\infty \leq t\rho(t,\delta)\}$ jointly occurs.

Step 1: Finiteness of the stopping time. $\mathbf{A}$ 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 $\mathbf{A} \triangleq \{a \in [K] : \lim_{t\to\infty} N_{a,t} < \infty\}$ as a sampling rule i.e if an arm belongs to this set $\mathbf{A}$, 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 $a \in \mathbf{A}^c, \hat{\boldsymbol{\mu}}_{a,t} \to \tilde{\boldsymbol{\mu}}_a \neq \boldsymbol{\mu}_a$ and $(\tilde{\mathbf{A}})_{a,t} \to (\mathbf{A}')_a \neq (\mathbf{A})_a$. When all parameters are concentrated $\mathbf{A} = \emptyset$, we say $\forall a \in [K] : \hat{\boldsymbol{\mu}}_a \to \boldsymbol{\mu}_a$ and $\tilde{\mathbf{A}} \to \mathbf{A}$. 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 $\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\hat{\boldsymbol{\mu}}_t)}\{\exists t \in \mathbb{N} : \sum_{a=1}^{K} \frac{N_{a,t}}{t} d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\lambda}_a) > \frac{\beta(t,\delta)}{t}\}$ and $\{\exists t \in \mathbb{N} : \|(\tilde{\mathbf{A}}_t - \mathbf{A})\frac{N_t}{t}\|_\infty \leq \rho(t,\delta)\}$. 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 $\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^{K} \frac{\boldsymbol{\omega}_{\infty,a}}{t} d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\lambda}_a) > 0$ and also $\|(\tilde{\mathbf{A}}_t - \mathbf{A})\boldsymbol{\omega}_{\infty,a}\|_\infty < 0$. We get strict inequality for the both the cases by the virtue of construction of the set $\mathbf{A}$ such that for arms $a \in \mathbf{A}, \boldsymbol{\omega}_\infty \neq 0$ and the KL-divergence is non-zero as $\boldsymbol{\lambda}_a \neq \boldsymbol{\mu}$ since $\boldsymbol{\lambda} \in \Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)$. Also for the second strict inequality, since $\tilde{\mathbf{A}}_t$ is the pessimistic estimate of $\mathbf{A}$ at time t the condition will hold.

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

$$\begin{aligned}
&U_t \triangleq \{\boldsymbol{\pi}_{\tau_\delta} \neq \boldsymbol{\pi}^\star_{\mathcal{F}}\} \\
&= \bigcup_{\boldsymbol{\pi}\neq\boldsymbol{\pi}^\star_{\mathcal{F}}} \Bigg\{ \exists t \in \mathbb{N} : \boldsymbol{\pi}_{t+1} = \boldsymbol{\pi} \\
&\qquad\qquad \wedge \left\{ \inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^{K} N_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\lambda}_a) > \beta(t,\delta) \wedge \mathbf{A}\boldsymbol{\omega}_t \leq 0 \right\} \Bigg\} \\
&\overset{(a)}{\subseteq} \bigcup_{\boldsymbol{\pi}\neq\boldsymbol{\pi}^\star_{\mathcal{F}}} \Bigg\{ \exists t \in \mathbb{N} : \boldsymbol{\pi}_{t+1} = \boldsymbol{\pi} \\
&\qquad\qquad \wedge \left\{ \inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^{K} N_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\lambda}_a) > \beta(t,\delta) \wedge \|(\tilde{\mathbf{A}}_t - \mathbf{A})\boldsymbol{\omega}_t\|_\infty \leq \rho(t,\delta) \right\} \Bigg\}
\end{aligned}$$

The argument (a) holds because

$$\mathbb{P}\{0 \geq \mathbf{A}\boldsymbol{\omega}_t\} = \mathbb{P}\{-\tilde{\mathbf{A}}_t\boldsymbol{\omega}_t \leq (\tilde{\mathbf{A}}_t - \mathbf{A})\boldsymbol{\omega}_t\} \leq \mathbb{P}\{\|\tilde{\mathbf{A}}_t\boldsymbol{\omega}_t\|_1 > \|(\tilde{\mathbf{A}}_t - \mathbf{A})\boldsymbol{\omega}_t\|_1 \geq \|(\tilde{\mathbf{A}}_t - \mathbf{A})\boldsymbol{\omega}_t\|_\infty\}$$

$$\leq \mathbb{P}\{\|\tilde{\mathbf{A}}_t\boldsymbol{\omega}_t\|_1 \leq \rho(t,\delta)\}\mathbb{P}\{\|(\tilde{\mathbf{A}}_t - \mathbf{A})\boldsymbol{\omega}_t\|_\infty \leq \rho(t,\delta)\}$$
$$= \mathbb{P}\{\|(\tilde{\mathbf{A}}_t - \mathbf{A})\boldsymbol{\omega}_t\|_\infty \leq \rho(t,\delta)\}$$

since the event $\{\|\tilde{\mathbf{A}}_t\boldsymbol{\omega}_t\|_1 \leq \rho(t,\delta)\}$ is a sure event.

Therefore probability of this bad event

$$\mathbb{P}(U_t) \leq \bigcup_{\boldsymbol{\pi}\neq\boldsymbol{\pi}^\star_{\mathcal{F}}} \mathbb{P}\left\{\exists t\in\mathbb{N} : \boldsymbol{\pi}_{t+1} = \boldsymbol{\pi} \wedge \inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^K N_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\lambda}_a)\} > \beta(t,\delta)\right\}$$
$$\mathbb{P}\left\{\exists t\in\mathbb{N} : \boldsymbol{\pi}_{t+1} = \boldsymbol{\pi} \wedge \|(\tilde{\mathbf{A}}_t - \mathbf{A})\boldsymbol{\omega}_t\|_\infty > \rho(t,\delta)\right\}$$
$$= \sum_{\boldsymbol{\pi}\neq\boldsymbol{\pi}^\star_{\mathcal{F}}} \mathbb{P}\left\{\exists t\in\mathbb{N} : \inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^K N_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\lambda}_a)\} > \beta(t,\delta)\right\}$$
$$+ \sum_{\boldsymbol{\pi}\neq\boldsymbol{\pi}^\star_{\mathcal{F}}} \mathbb{P}\left\{\exists t\in\mathbb{N} : \|(\tilde{\mathbf{A}}_t - \mathbf{A})\boldsymbol{\omega}_t\|_\infty > \rho(t,\delta)\right\} \tag{22}$$

The second cumulative probability can be bound using Lemma 4, i.e

$$\sum_{\boldsymbol{\pi}\neq\boldsymbol{\pi}^\star_{\mathcal{F}}} \mathbb{P}\left\{\exists t\in\mathbb{N} : \|(\tilde{\mathbf{A}}_t - \mathbf{A})\boldsymbol{\omega}_t\|_\infty > \rho(t,\delta)\right\} \leq \frac{1}{t}$$

for the choice of $\rho(t,\delta)$ given in Lemma 4. We work with the first term in R.H.S of Equation (22).

$$\sum_{\boldsymbol{\pi}\neq\boldsymbol{\pi}^\star_{\mathcal{F}}} \mathbb{P}\left\{\exists t\in\mathbb{N} : \inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^K N_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\lambda}_a)\} > \beta(t,\delta)\right\}$$
$$= \sum_{\boldsymbol{\pi}\neq\boldsymbol{\pi}^\star_{\mathcal{F}}} \mathbb{P}\Bigg\{\exists t\in\mathbb{N} : \inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)} \inf_{\boldsymbol{\lambda}'\in\Lambda_{\mathcal{F}}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^K N_{a,t}\left(d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\lambda}_a) - d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\lambda}'_a)\right)$$
$$+ \inf_{\boldsymbol{\lambda}'\in\Lambda_{\mathcal{F}}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^K N_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\lambda}'_a) > \beta(t,\delta)\Bigg\}$$
$$\leq \sum_{\boldsymbol{\pi}\neq\boldsymbol{\pi}^\star_{\mathcal{F}}} \mathbb{P}\left\{\exists t\in\mathbb{N} : \inf_{\boldsymbol{\lambda}'\in\Lambda_{\mathcal{F}}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^K N_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\lambda}'_a) \leq \beta(t,\delta)\right\}$$

The last inequality holds because

$$\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)} \inf_{\boldsymbol{\lambda}'\in\Lambda_{\mathcal{F}}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^K N_{a,t}\left(d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\lambda}_a) - d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\lambda}'_a)\right) \leq 0$$

since under $\hat{\mathcal{F}}_t$ and the bad event we are assuming that the estimated alt-set is still bigger than the actual alt-set. So any $\boldsymbol{\lambda} \in \Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)$ will have a bigger distance with the estimate of $\boldsymbol{\mu}$ than any $\boldsymbol{\lambda}' \in \Lambda_{\mathcal{F}}(\hat{\boldsymbol{\mu}}_t)$. We define $I_{\boldsymbol{\pi}} \triangleq \text{Supp}(\boldsymbol{\pi}^\star_{\mathcal{F}})\Delta\text{Supp}(\boldsymbol{\pi})$ and also $S_0 \triangleq \max_{\boldsymbol{\pi}} |I_{\boldsymbol{\pi}}|$. We note that $0 \leq S_0 \leq K$.

We get from Lemma 9 in [KK21] with the notation of $\mathcal{T}(.)$ follows from Lemma 9

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

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

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.2 Proof of Lemma 2

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

$$\begin{aligned}
&\mathbb{P}(\boldsymbol{\pi} \neq \boldsymbol{\pi}^\star_{\mathcal{F}}) \leq \delta \\
\implies &\mathbb{P}(\boldsymbol{\pi} \neq \boldsymbol{\pi}^\star_{\mathcal{F}} \wedge \mathbf{A}\boldsymbol{\pi}^\star_{\mathcal{F}} \leq 0) \leq \delta \\
\implies &\mathbb{P}(\{\mathbf{A}\boldsymbol{\pi} \leq \mathbf{A}\boldsymbol{\pi}^\star_{\mathcal{F}} \vee \mathbf{A}\boldsymbol{\pi} > \mathbf{A}\boldsymbol{\pi}^\star_{\mathcal{F}}\} \wedge \mathbf{A}\boldsymbol{\pi}^\star_{\mathcal{F}} \leq 0) \leq \delta \\
\implies &\mathbb{P}(\mathbf{A}\boldsymbol{\pi} \leq 0 \vee -\Gamma^\star < \mathbf{A}\boldsymbol{\pi}) \leq \delta \quad \text{,where } \Gamma^\star \text{ is the slack value at stopping time} \\
\implies &\mathbb{P}(\mathbf{A}\boldsymbol{\pi} \leq 0) \leq \delta
\end{aligned}$$

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

G.3 Upper Bound of LATS

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

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

where $\mathfrak{s}$ is the shadow price of the true constraint $\mathbf{A}$, $T_{\mathcal{F}}(\mu)$ is the characteristic time under the true constraint (Equation (5)*), 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 $\epsilon > 0$, we define the good event G_T as,

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

where, $\xi(\epsilon) \leq \max_{\boldsymbol{\pi}' \in \nu_{\hat{\mathcal{F}}}(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})} \frac{1}{5}\boldsymbol{\mu}^T(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}')$, and $\phi(\epsilon) \triangleq \max(1,\epsilon)$ for a given $\epsilon > 0$. The good event implies that the means and the 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 $||(\tilde{\mathbf{A}}^i_t - \mathbf{A}^i)\boldsymbol{\omega}||_\infty \leq \phi(\epsilon)$ 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 complement of the good event,

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

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

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

The first inequality is due to the union bound. The second inequality is due to the Lemma 7 (Lemma 19 of [GK16]), which states that

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

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 [DKM19b, Lemma 17]. We use this lemma than D-tracking or the tracking argument in [GK16, Lemma 7] 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,

$$|(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_t)^T\boldsymbol{\pi}^\star_{\mathcal{F}}| \leq |(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_t)^T\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}| + |(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_t)^T(\boldsymbol{\pi}^\star_{\mathcal{F}} - \boldsymbol{\pi}^\star_{\hat{\mathcal{F}}})| \leq 4\xi(\epsilon)$$
$$\leq \max_{\boldsymbol{\pi}' \in \nu_{\mathcal{F}}(\boldsymbol{\pi}^\star_{\mathcal{F}})'} \boldsymbol{\mu}^T(\boldsymbol{\pi}^\star_{\mathcal{F}'} - \boldsymbol{\pi}')$$

We have replaced the perturbed alt-set with the true alt-set 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

$$C_{\epsilon,\hat{\mathcal{F}}} \triangleq \inf_{\substack{\boldsymbol{\mu}':||\boldsymbol{\mu}'-\boldsymbol{\mu}||_\infty \leq \xi(\epsilon) \\ \boldsymbol{\omega}':||\boldsymbol{\omega}'-\boldsymbol{\omega}||_\infty \leq 3\epsilon \\ \tilde{\mathbf{A}}':||(\tilde{\mathbf{A}}-\mathbf{A})\boldsymbol{\omega}||_\infty \leq \phi(\epsilon)}} \mathcal{D}(\boldsymbol{\mu}', \boldsymbol{\omega}', \hat{\mathcal{F}}) - \boldsymbol{l}^T\tilde{\mathbf{A}}\boldsymbol{\omega}\,.$$

Now leveraging Lemma 6, we obtain

$$(1+\psi)\inf_{\substack{\boldsymbol{\mu}':||\boldsymbol{\mu}'-\boldsymbol{\mu}||_\infty \leq \xi(\epsilon) \\ \boldsymbol{\omega}':||\boldsymbol{\omega}'-\boldsymbol{\omega}||_\infty \leq 3\epsilon \\ \tilde{\mathbf{A}}:||(\tilde{\mathbf{A}}-\mathbf{A})\boldsymbol{\omega}||_\infty \leq \phi(\epsilon)}} \mathcal{D}(\boldsymbol{\mu}', \boldsymbol{\omega}', \hat{\mathcal{F}}) \geq \inf_{\substack{\boldsymbol{\mu}':||\boldsymbol{\mu}'-\boldsymbol{\mu}||_\infty \leq \xi(\epsilon) \\ \boldsymbol{\omega}':||\boldsymbol{\omega}'-\boldsymbol{\omega}||_\infty \leq 3\epsilon \\ \tilde{\mathbf{A}}':||(\tilde{\mathbf{A}}-\mathbf{A})\boldsymbol{\omega}||_\infty \leq \phi(\epsilon)}} \mathcal{D}(\boldsymbol{\mu}', \boldsymbol{\omega}', \hat{\mathcal{F}}) - \boldsymbol{l}^T\tilde{\mathbf{A}}\boldsymbol{\omega}\,,$$

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 10, we can state

$$\mathcal{D}(\hat{\boldsymbol{\mu}}_t, N_t, \hat{\mathcal{F}}) \geq t \inf_{\substack{\boldsymbol{\mu}':||\boldsymbol{\mu}'-\boldsymbol{\mu}||_\infty \leq \xi(\epsilon) \\ \boldsymbol{\omega}':||\boldsymbol{\omega}'-\boldsymbol{\omega}||_\infty \leq 3\epsilon \\ \tilde{\mathbf{A}}:||(\tilde{\mathbf{A}}-\mathbf{A})\boldsymbol{\omega}||_\infty \leq \phi(\epsilon)}} \mathcal{D}(\boldsymbol{\mu}', \boldsymbol{\omega}', \hat{\mathcal{F}}) \geq \frac{tC_{\epsilon,\hat{\mathcal{F}}}}{1+\psi}\,. \tag{23}$$

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 we can write upper bound on this stopping time for both good and bad events as

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

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 23), we get

$$T_\epsilon + \sum_{t=T_\epsilon}^{T} \mathbb{1}(\mathcal{D}(\hat{\boldsymbol{\mu}}_t, N_t, \hat{\mathcal{F}}) \leq \beta(t,\delta)) \leq \max(\sqrt{T}, \beta(T,\delta)) + \sum_{t=T_\epsilon}^{T} \mathbb{1}\Big(t\frac{C_{\epsilon,\hat{\mathcal{F}}}}{1+\psi} \leq \beta(T,\delta)\Big)$$
$$\leq \max(\sqrt{T}, \beta(T,\delta)) + \frac{\beta(T,\delta)(1+\psi)}{C_{\epsilon,\hat{\mathcal{F}}}}$$

Let us define a $T_\delta \triangleq \inf\{T \in \mathbb{N} : \max(\sqrt{T}, \beta(T,\delta)) + \frac{\beta(T,\delta)(1+\psi)}{C_{\epsilon,\hat{\mathcal{F}}}} \leq T\}$. To find a lower bound on T_δ, we refer to [GK16]. Specifically, let us define $\beta(\eta) \triangleq \{\inf T : T - \max(\sqrt{T}, \beta(T,\delta)) \geq \frac{T}{1+\eta}\}$ for some $\eta > 0$. Therefore,

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

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 4, we get $\tilde{\mathbf{A}} \to \mathbf{A}$, and thus, $\hat{\mathcal{F}} \to F$.

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

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

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

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

□

G.4 Upper Bound for LAGEX

Theorem 7. *The expected sample complexity of LAGEX satisfies* $\lim_{\delta\to 0} \frac{\mathbb{E}(\tau_\delta)}{\ln(1/\delta)} \leq T_{\mathcal{F}}(\boldsymbol{\mu}) + 2\mathfrak{s}$.

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.1. 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

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

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 [DKM19a] 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

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

We get the upper bound on $\sum_{s=1}^{T} \mathbb{P}\left(\sum_{a=1}^{K} N_{a,s} d(\hat{\boldsymbol{\mu}}_{a,s}, \boldsymbol{\mu}_a) > g(s)\right) \leq \frac{\exp(2)}{t^3 \log t}(g(t) + g^2(t)\log t) \leq \infty$ as a direct consequence of [DKM19a, Lemma 6]. The second cumulative probability is bounded by $\zeta(1)$ 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

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

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

$$\beta(t,\delta) \geq \inf_{\boldsymbol{\lambda} \in \Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)} \sum_{s=1}^{t} \sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\lambda}_a) - (1+\sqrt{t})K \tag{24}$$

L-Lipschitz property of KL divergence gives

$$|d(\boldsymbol{\mu}_a, \boldsymbol{\lambda}_a) - d(\hat{\boldsymbol{\mu}}_{a,s}, \boldsymbol{\lambda}_a)| \leq L\sqrt{2\sigma^2 \frac{g(s)}{N_{a,s}}} \tag{25}$$

Using this result in Equation (24) we get

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

the penultimate inequality yields from using the Equation (25). Using this result in Equation (24)

$$\beta(t,\delta) \geq \inf_{\boldsymbol{\lambda} \in \Lambda_{\hat{\mathcal{F}}}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^{K} \boldsymbol{\omega}_{a,s} d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\lambda}_a) - (1+\sqrt{t})K - \mathcal{O}(\sqrt{t\log t}) \tag{26}$$

Step 4: LAGEX (Algorithm 2) as an optimistic saddle point algorithm We follow the definition of approximate saddle point algorithm in [DKM19a]. LAGEX acts as an approximate saddle point algorithm if

$$\inf_{\boldsymbol{l}_t\in\mathbb{R}^d_+}\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\hat{\boldsymbol{\mu}}_t)}\sum_{s=1}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{s,a}d(\hat{\boldsymbol{\mu}}_{a,s},\boldsymbol{\lambda}_a)-\boldsymbol{l}_t^T\tilde{\mathbf{A}}_t\boldsymbol{\omega}_t\geq\max_{\boldsymbol{\omega}\in\hat{\mathcal{F}}_t}\sum_{a=1}^{K}\sum_{s=1}^{t}\boldsymbol{\omega}_a U_{a,s}-x_t \tag{27}$$

where, $U_{a,s}=\max\left\{\frac{g(t)}{N_{a,s}},\max_{\xi\in[\alpha_{a,s},\beta_{a,s}]}d(\xi,\lambda_{a,s})\right\}$ and $x_t=R_t^{\boldsymbol{\omega}}+\sum_{s=1}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}C_{a,s}$. The quantities $R_t^{\boldsymbol{\omega}}$ and $C_{a,s}$ will be defined later on.

Step 5: Bounds cumulative regret of players Algorithm 2 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. The regret of the AdaGrad player is defined at time $t\in\mathbb{N}$ as

Allocation player's regret.

$$R_t^{\boldsymbol{\omega}}=\max_{\boldsymbol{\omega}\in\hat{\mathcal{F}}}\sum_{s=1}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_a U_{a,s}-\sum_{s=1}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}U_{a,s}$$

We should 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 $Q\geq\max_{x,y\in[\boldsymbol{\mu}_{\min},\boldsymbol{\mu}_{\max}]}d(x,y)$.

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

$$R_{\boldsymbol{\lambda}}^t=\sum_{s=1}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}d(\hat{\boldsymbol{\mu}}_{a,s},\lambda_{a,s})-\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})}\sum_{s=1}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}d(\hat{\boldsymbol{\mu}}_{a,s},\boldsymbol{\lambda}_a)\leq 0$$

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. We define $C_{a,s}\triangleq U_{a,s}-d(\hat{\boldsymbol{\mu}}_{a,s},\lambda_{a,s})$. From the definition of regret of the $\boldsymbol{\lambda}$-player we get

$$\begin{aligned}&\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})}\sum_{s=1}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}d(\hat{\boldsymbol{\mu}}_{a,s},\boldsymbol{\lambda}_a)\\&\geq\inf_{\boldsymbol{l}\in\mathbb{R}^d_+}\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})}\left\{\sum_{s=1}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}U_{a,s}-\sum_{s=1}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}C_{a,s}+\boldsymbol{l}^T\tilde{\mathbf{A}}_t\boldsymbol{\omega}_t\right\}\end{aligned}$$

Then we have from Equation (26) and Equation (27)

$$\begin{aligned}\beta(t,\delta)&\geq\inf_{\boldsymbol{l}\in\mathbb{R}^d_+}\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})}\left\{\sum_{s=1}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}U_{a,s}-\sum_{s=1}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}C_{a,s}+\boldsymbol{l}^T\tilde{\mathbf{A}}_t\boldsymbol{\omega}_t\right\} &(28)\\&\quad-(1+\sqrt{t})K-\mathcal{O}(\sqrt{t\log t}) &(29)\\&\geq\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})}\sum_{s=1}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}U_{a,s}-\sum_{s=1}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}C_{a,s} &(30)\\&\quad-(1+\sqrt{t})K-\mathcal{O}(\sqrt{t\log t})+\boldsymbol{l}_t^T\tilde{\mathbf{A}}_t\boldsymbol{\omega}_t &(31)\\&\geq\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})}\sum_{s=1}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}U_{a,s}-\sum_{s=1}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}C_{a,s}-(1+\sqrt{t})K &(32)\\&\quad-\mathcal{O}(\sqrt{t\log t})-\mathcal{D}(\boldsymbol{\omega}_t,\hat{\boldsymbol{\mu}}_t,\hat{\mathcal{F}}_t)\psi_t &(33)\end{aligned}$$

where ψ_t is defined as Lemma 6. The penultimate inequality holds as we have replaced $\inf_{\boldsymbol{l}\in\mathcal{L}}\boldsymbol{l}$ with $\boldsymbol{l}_t$ i.e the optimised Lagrangian multiplier at the last step. Whereas the last inequality follows from Lemma 6. From the definition of the allocation player regret we have

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

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

Now we have to ensure that $T1$ in the slack is bounded.

$$\begin{aligned}\sum_{K+1}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}C_{a,s} &\leq \sum_{K+1}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}\left\{\frac{g(s)}{N_{a,s}} + 2L\sqrt{2\sigma^2\frac{g(s)}{N_{a,s}}}\right\}\\ &\leq g(t)\sum_{K+1}^{t}\sum_{a=1}^{K}\frac{\boldsymbol{\omega}_{a,s}}{N_{a,s}} + 2L\sqrt{2\sigma^2 g(t)}\sum_{K+1}^{t}\sum_{a=1}^{K}\frac{\boldsymbol{\omega}_{a,s}}{\sqrt{N_{a,s}}}\\ &\leq g(t)\left(K^2 + 2K\log\frac{t}{K}\right) + 2L\sqrt{2\sigma^2 g(t)}(K^2 + 2\sqrt{2Kt})\\ &\leq \mathcal{O}(\sqrt{t\log t})\,. \end{aligned} \tag{34}$$

The inequalities hold due to the good event G_T and L-Lipschitz property of KL

$$\begin{aligned}&|d(\boldsymbol{\mu}_a, \boldsymbol{\lambda}_a) - d(\hat{\boldsymbol{\mu}}_{a,s}, \boldsymbol{\lambda}_a)| \leq L\sqrt{2\sigma^2\frac{g(s)}{N_{a,s}}}\\ \implies &\sup_{\xi\in[\alpha_{a,s},\beta_{a,s}]} U_{a,s} - d(\xi, \lambda_{a,s}) \leq \max\left\{2L\sqrt{2\sigma^2\frac{g(s)}{N_{a,s}}}, \frac{g(s)}{N_{a,s}}\right\}\end{aligned}$$

Now the stopping time expression changes to

$$\begin{aligned}\beta(t,\delta) &\geq \max_{\boldsymbol{\omega}\in\hat{\mathcal{F}}}\left(\sum_{a=1}^{K}\sum_{s=1}^{t}\boldsymbol{\omega}_a d(\boldsymbol{\mu}_a, \lambda_{a,s}) - \mathcal{D}(\boldsymbol{\omega}_t, \hat{\boldsymbol{\mu}}_t, \hat{\mathcal{F}}_t)\psi_t\right) - R_t^{\boldsymbol{\omega}} - \mathcal{O}(\sqrt{t\log t}) - (1+\sqrt{t})K\\ &\geq \max_{\boldsymbol{\omega}\in\hat{\mathcal{F}}}\sum_{a=1}^{K}\sum_{s=1}^{t}\boldsymbol{\omega}_a d(\boldsymbol{\mu}_a, \lambda_{a,s}) - T^{-1}_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t)\psi_t - R_t^{\boldsymbol{\omega}} - \mathcal{O}(\sqrt{t\log t}) - (1+\sqrt{t})K\end{aligned}$$

Step 6: Characteristic time Accumulating Equation (28) and Equation (34)

$$\begin{aligned}\max_{\boldsymbol{\omega}\in\hat{\mathcal{F}}_t}\sum_{s=1}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_a d(\boldsymbol{\mu}_a, \lambda_{a,s}) &\geq t\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})}\max_{\boldsymbol{\omega}\in\hat{\mathcal{F}}_t}\sum_{a=1}^{K}\boldsymbol{\omega}_a d(\boldsymbol{\mu}_a, \boldsymbol{\lambda}_a)\\ &\geq t\max_{\boldsymbol{\omega}\in\hat{\mathcal{F}}_t}\inf_{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu})}\sum_{a=1}^{K}\boldsymbol{\omega}_a d(\boldsymbol{\mu}_a, \boldsymbol{\lambda}_a) = tT^{-1}_{\hat{\mathcal{F}}_t}(\boldsymbol{\mu})\end{aligned}$$

Then the sample complexity is upper bounded by

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

Now asymptotically when $\hat{\mathcal{F}}_t \to \mathcal{F}$ and $\psi_t \to \mathfrak{s}$, the expression for the characteristic time is given by

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

Part (b) : Current recommended policy is wrong. To get on with the proof for this part we will use similar argument as [CBJD23]. Though the argument was motivated by the work [DKM19b]. We define the event

$$B_t \triangleq \left\{\boldsymbol{\pi}^\star \neq \arg\max_{\boldsymbol{\pi}\in\hat{\mathcal{F}}_t}\hat{\boldsymbol{\mu}}_t^T\boldsymbol{\pi}\right\}$$

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

$$B_t \triangleq \left\{\boldsymbol{\lambda}\in\Lambda_{\hat{\mathcal{F}}_t}(\hat{\boldsymbol{\mu}}_t) : \boldsymbol{\pi}' = \arg\max_{\boldsymbol{\pi}\in\hat{\mathcal{F}}_t}\boldsymbol{\lambda}^T\boldsymbol{\pi} \neq \boldsymbol{\pi}^\star_{\hat{\mathcal{F}}_s}\right\}$$

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

$$\epsilon = \min_{\boldsymbol{l}\in\mathcal{L}}\sum_{s=1,B_s}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}d(\hat{\boldsymbol{\mu}}_{a,s},\mu_a) \geq \min_{\boldsymbol{l}\in\mathcal{L}}\sum_{\boldsymbol{\pi}'\neq\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}_s}}\inf_{\boldsymbol{\lambda}\in B_s}\sum_{s=1,\boldsymbol{\pi}_s=\boldsymbol{\pi}'}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}d(\hat{\boldsymbol{\mu}}_{a,s},\mu_a) - \boldsymbol{l}^T\tilde{\mathbf{A}}_t\boldsymbol{\omega}_t \tag{35}$$

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 $x_t = R^{\boldsymbol{\omega}}_{n_{\boldsymbol{\pi}'}(t)} + \sum_{s=1,\boldsymbol{\pi}_s=\boldsymbol{\pi}'}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}C_{a,s}$. Therefore we can write the RHS of Equation (35) as

$$\sum_{\boldsymbol{\pi}'\neq\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}_s}}\inf_{\boldsymbol{\lambda}\in B_s}\sum_{s=1,\boldsymbol{\pi}_s=\boldsymbol{\pi}'}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}d(\hat{\boldsymbol{\mu}}_{a,s},\mu_a) \tag{36}$$

$$\geq \max_{\boldsymbol{\pi}\in\hat{\mathcal{F}}_s}\min_{\boldsymbol{l}\in\mathcal{L}}\underbrace{\sum_{s=1,\boldsymbol{\pi}_s=\boldsymbol{\pi}'}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}U_{a,s} - R^{\boldsymbol{\omega}}_{n_{\boldsymbol{\pi}'}(t)}}_{T1} - \underbrace{\sum_{s=1,\boldsymbol{\pi}_s=\boldsymbol{\pi}'}^{t}\sum_{a=1}^{K}\boldsymbol{\omega}_{a,s}C_{a,s}}_{T2} + \boldsymbol{l}^T\tilde{\mathbf{A}}_t\boldsymbol{\omega}_t \tag{37}$$

We apply the same logic as in [CBJD23] and [DKM19b] 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 (36) 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

$$\mathbb{E}[\tau_\delta] \leq T_{\hat{\mathcal{F}}_t}(\boldsymbol{\mu})\left(\beta(t,\delta) + R^{\boldsymbol{\omega}}_t + \mathcal{O}\left(\sqrt{t\log t}\right)\right) + 2\psi_t \implies \lim_{\delta\to 0}\frac{\mathbb{E}[\tau_\delta]}{\log\frac{1}{\delta}} \leq T_{\mathcal{F}} + 2\mathfrak{s}\,.$$

□

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 $\mathfrak{s} = \frac{\max_{i \in K} \boldsymbol{\pi}_i^\star}{\min_{i \in K} \boldsymbol{\pi}_i^\star} > 1$ 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 $\mathbf{A}\boldsymbol{\omega}_t \geq 0$. We start with the optimization problem relaxed with slack if the constraints were known,

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

where, Γ is the slack. Cumulative violation of constraints can be expressed as,

$$\mathcal{V}_t = \sum_{s=1}^{t} \max_{i\in[K]} \left[\mathbf{A}^i\omega_t\right]_+$$

where, $[z]_+ = \max\{z, 0\}$. Then, at any time step $t \in [T]$, instantaneous violation is given by, $v_t = \max_{i\in[k]} \left[\mathbf{A}^i\omega_t\right]_+$. Since, $\mathbf{A}$ is feasible, we define the game value as, $\eta = \max_{\boldsymbol{\omega}\in\mathcal{F}} \max_{i\in[K]} \mathbf{A}^i\boldsymbol{\omega} \leq \Gamma$ and $\eta = \max_{\boldsymbol{\omega}\in\mathcal{F}} \max_{i\in[K]} \mathbf{A}^i\boldsymbol{\omega} \geq \min_{\tilde{\mathbf{A}}\in\mathcal{C}_A^t} \max_{\boldsymbol{\omega}\in\hat{\mathcal{F}}_t} \max_{i\in[K]} \tilde{\mathbf{A}}_t^i\omega_t = \tilde{\mathbf{A}}_t^{i_t}\omega_t$, holds because of pessimistic estimate of $\mathbf{A}$.

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

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

Again, $\max_{i\in[K]} \tilde{\mathbf{A}}_t^i\omega_t \leq \max_{i\in[K]} \mathbf{A}^i\omega_t \leq \Gamma$

Then, the instantaneous violation becomes,

$$\begin{aligned}
v_t &\leq [f(t,\delta)||\omega_t||_{\Sigma_t^{-1}} + \Gamma]_+ \\
&\leq \underbrace{f(t,\delta)||\omega_t||_{\Sigma_t^{-1}}}_{\rho(t,\delta)} + |\Gamma|
\end{aligned}$$

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

$$\begin{aligned}
\mathcal{V}_{\tau_\delta} &= \sum_{t\leq\tau_\delta} s_t \\
&= \sum_{t\leq\tau_\delta} \rho(t,\delta) + |\Gamma| \\
&\leq 2\sum_{t\leq\tau_\delta} |\Gamma| \mathbb{1}\{\rho(t,\delta) \leq |\Gamma|\} + 2\sum_{t\leq\tau_\delta} \rho(t,\delta)\mathbb{1}\{|\Gamma| \leq \rho(t,\delta)\} \\
&\leq 2\tau_\delta|\Gamma| + 2\sum_{t\leq\tau_\delta} \frac{(\rho(t,\delta)^2}{|\Gamma|}, \text{ where, } \mathbb{1}(u \leq v) \leq \frac{v}{u} \\
&\leq 2\tau_\delta|\Gamma| + \frac{6d^2\log^2(1+\frac{\tau_\delta+1}{d}) + 12d\log(1+\frac{\tau_\delta+1}{d})(1+\log\frac{K}{\delta})}{|\Gamma|}
\end{aligned}$$

The last inequality is a direct consequence of Lemma 3.

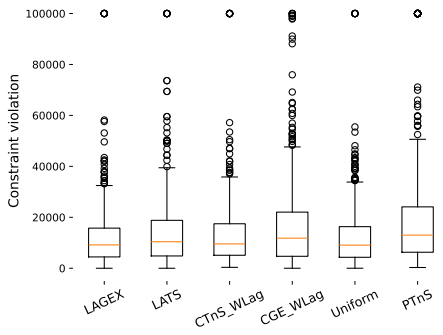

Figure 6: Constraint violation (median±std.) algorithms over 500 runs for **hard environment**.

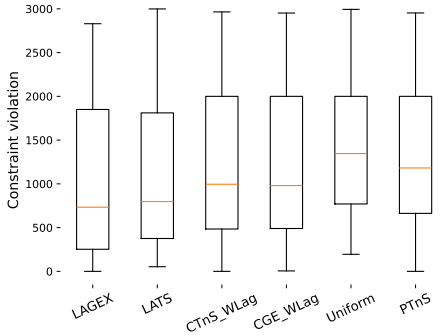

Figure 7: Constraint violation (median±std.) of algorithms over 500 runs for **easy environment**.

H.2 Experimental results

In all the experiments we have taken $\delta = 0.01$. Here, we will also explain what the abbreviations used in the plots. "CTnS-WLag" and "CGE-WLag" is basically the CTnS and CGE algorithm ([CBJD23]) under unknown constraints without Lagrangian relaxation. For the experiments in **easy environment** are clipped at 3000 for better visualisation. For CGE we have used $g(t) = \log t$. Each plot has been generated over 500 random seeds.

Observation 2. LAGEX least violates constraints. We compare the constraint violation i.e the number of times $\mathbf{A}\boldsymbol{\omega}_t > 0$, where $\mathbf{A}$ is the true constraint matrix.

Observations in Environment 1. From Fig. 6 we see that LAGEX shows least number of constraint violation across all the algorithms compared, followed by LATS and CTnS without Lagrangian relaxation shows the worst performance by having highest number of constraint violations. Though performance of Uniform explorer in terms of constraint violation is at par with LAGEX and LATS.

Observations in Environment 2. From Fig. 7 we can say our proposed algorithms LATS and LAGEX also better in terms of showing least number of constraint violations for the **easy environment**. An interesting application of these algorithm would be to explore BAI with end-of-time Knapsack constraint since LAGEX and LATS work "safer" than other algorithms in the unknown constraint setting with better constraint violation guaranties.

H.3 Experiment on IMDB dataset

We evaluate our proposed algorithms Algorithm 1 and 2 with other algorithms using the publicly available and often used IMDB 50K dataset [MDP+11]. For ease of comparison we use the same bandit environment as [CBJD23] using 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. The true optimal policy is $[0.3, 0.3, 0, 0, 0.4, 0, 0, 0, 0, 0, 0, 0]$. We assume $\delta = 0.1$. We compare the same set of algorithms as before.

Observations 1. LAGEX shows better sample complexity From figure 8 we can observe that the LAGEX (Algorithm 2) performs better any other algorithm in the unknown constraint setting. The algorithm LATS (Algorithm 1) performs also well on the IMDB environment but notably we cannot distinguish it's performance from the performance of the Uniform explorer as well. The diagram 8 also properly demonstrates the extra cost we had to pay due to not knowing the constraints in the allocation on the different genres of movies.

Observation 2. LAGEX shows least constraint violation Interestingly not only LAGEX performs as the most efficient algorithm among the all other algorithms in the unknown constraint setting but also it shows least constraint violation. It means LAGEX performs as the most safe algorithm even during exploration.

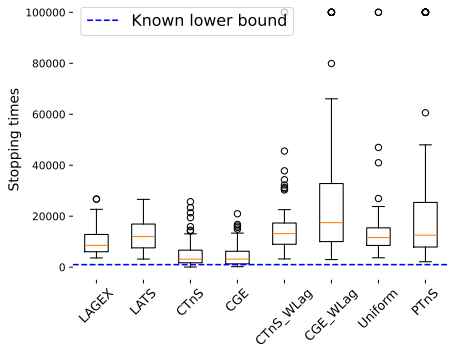

Figure 8: Sample complexity (median±std.) of algorithms over 500 runs for **IMDB environment**.

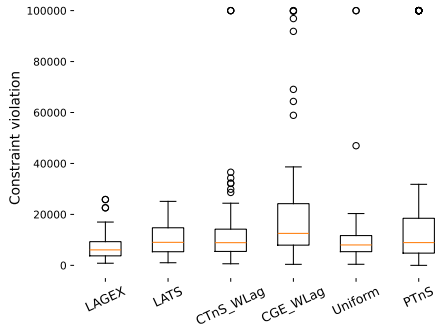

Figure 9: Constraint violation (median±std.) of algorithms over 500 runs for **IMDB environment**.

I ϵ-good policies under unknown linear constraints

Throughout this paper we have talked about converging to the optimal policy while tracking the estimates of the unknown linear constraints and unknown means of the reward distributions of K arm. One might not want to land on the exact optimal policy, rather may be interested in finding an ϵ-good policy i.e the recommended policy is in a ϵ-ball of the actual true optimal policy. This approach may get us to a much lower sample complexity lower bound [[GK21b],[MPK21], [GK21a], [MJTN20]]. We can extend our proposed Algorithms 1 and 2 can be extended to this setting by changing the definition of the set of alternative instances as $\Lambda_{\hat{\mathcal{F}}}(\boldsymbol{\mu}) \triangleq \left\{\boldsymbol{\lambda} \in \mathbb{D} : \boldsymbol{\lambda}^T(\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}} - \boldsymbol{\pi}) > \epsilon\right\}$ for a pre-specified alpha as an input in the algorithms. So we will say the recommended policy in ϵ-good for $\boldsymbol{\pi}^\star_{\mathcal{F}}$ after the stopping rule fires, if $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$ has converged to $\boldsymbol{\pi}^\star_{\mathcal{F}}$ with concentration in means and constraint matrix. But there are two main problem that appears. Since $\epsilon > 0$, the most confusing instance for $\boldsymbol{\mu}$ will not lie on the boundary of the normal cone. So the projection lemma for lagrangian formulation (Proposition 1) is no more sufficient. Also, the algorithm may start oscillating when the allocation comes inside the epsilon ball of $\boldsymbol{\pi}^\star_{\hat{\mathcal{F}}}$ among near-optimal policies since convexity of $\omega^\star(\boldsymbol{\mu})$ is no more ensured. To handle the first problem we can use the approximation error ϵ as an *added pessimism* in the system. That means we are interested in the optimization problem

$$\max_{\boldsymbol{\pi} \in \mathcal{F}} \boldsymbol{\mu}^T \boldsymbol{\pi}$$
$$\text{such that } \mathbf{A}\boldsymbol{\pi} \leq \epsilon \mathbf{A}\mathbf{1}.$$

Then to build a new superset for the new feasible set we use pessimistic estimate of $\mathbf{A}$ i.e $(\mathbf{A}\epsilon)\boldsymbol{\pi} \leq \mathbf{A}\boldsymbol{\pi} \leq 0$. We would also want to track the sequence $\{\epsilon_t\}_{t\in\mathbb{N}}$ and use it's concentration properties to find introduce new quantities in the lower bound that will capture the effect of the approximation. For the second hurdle we can add the notion of *sticky* approach from [DK19] that can help the agent stick to a specific ϵ-good policy rather than oscillating.

J Technical results and known tools in BAI and pure exploration

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.

J.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 $\|(\tilde{\mathbf{A}} - \mathbf{A})\boldsymbol{\omega}\|_\infty$, 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 $\tilde{\mathbf{A}}$ of $\mathbf{A}$, the following holds*

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

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

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

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 [AyPS11] for the control over the behaviour of Σ_t and we directly get the second as

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

Refer [AyPS11] 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 behaviour of $\rho(t,\delta)$.

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

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

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

□

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

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

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

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 $\mathbb{P}\left(\exists t \in [T] \|(\tilde{\mathbf{A}}_t - \mathbf{A})\boldsymbol{\omega}\|_\infty > \rho'(t,\delta)\right) \leq \mathbb{P}\left(\exists t \in [T] \|(\tilde{\mathbf{A}}_t - \mathbf{A})\boldsymbol{\omega}\|_\infty > \rho(t,\delta)\right) \leq \delta$. Now if we replace $\log\frac{1}{\delta}$ by u, we can write $\mathbb{P}\left(\exists t \in [T], \forall i \in [d] \|\tilde{\mathbf{A}}_t^i - \mathbf{A}^i\|_{\Sigma_t} > 1 + \sqrt{\frac{1}{2}\log K + \frac{u}{2} + \frac{d}{4}\log\left(1+\frac{t}{d}\right)}\right) \leq \exp(-u)$. 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 $u = \log t$

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

□

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 $\boldsymbol{\omega} \in \Delta_K$*

$$-\boldsymbol{l}^T\tilde{\mathbf{A}}\boldsymbol{\omega} \leq \mathcal{D}(\boldsymbol{\mu},\boldsymbol{\omega},\hat{\mathcal{F}})\psi$$

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

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

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

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

J.2 Useful results from BAI and pure exploration literature

Lemma 7. *(Lemma 19 in [GK16]) There exists two constants B and C (depends on $\boldsymbol{\mu}$ and ϵ), such that—*

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

Lemma 8. *[GK16, Lemma 7] 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 14 in [KK21]) 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

$$\forall u \geq 1, \quad h(u) = u - \ln(u) \tag{38}$$

$$\forall z \in [1, e], \forall x \geq 0, \quad \tilde{h}_z(x) = \begin{cases} \exp\left(\frac{1}{h^{-1}(x)}\right) h^{-1}(x) & \text{if } x \geq h^{-1}\left(\frac{1}{\ln(z)}\right) \\ z(x - \ln(\ln(z))) & \text{else} \end{cases} . \tag{39}$$

Lemma 10. *(Lemma 17 in [DK19]) 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 [DKM19b]) 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.

J.3 Useful definitions and theorems from literature on 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, [Ber63]) 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, 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 $\lim_{r \to \infty} \left(H\left(f^r\right) \cap C = H(f) \cap C\right.$

Lemma 12. *[MCP14, Lemma 13] 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 $\mathcal{T}$.