Pure Exploration in Bandits with Linear Constraints

Anonymous Author(s)
Affiliation
Address
email

Abstract

We address the problem of identifying the optimal policy with a fixed confidence level in a multi-armed bandit setup, when *the arms are subject to linear constraints*. Unlike the standard best-arm identification problem which is well studied, the optimal policy in this case may not be deterministic and could mix between several arms. This changes the geometry of the problem which we characterize via an information-theoretic lower bound. We introduce two asymptotically optimal algorithms for this setting, one based on the Track-and-Stop method and the other based on a game-theoretic approach. Both these algorithms try to track an optimal allocation based on the lower bound and computed by a weighted projection onto the boundary of a normal cone. Finally, we provide empirical results that validate our bounds and visualize how constraints change the hardness of the problem.

1 Introduction

A classical problem in the multi-armed bandit framework is *pure exploration* (Lattimore and Szepesvári, 2020), where the task of a learner is to answer some query about a set of actions, also known as arms, by iteratively choosing between the actions and receiving an immediate reward sampled from a distribution associated with the action. A very well-studied problem in this context is Best-Arm Identification (BAI), where a learner is trying to identify the arm with the highest expected reward (Even-Dar et al., 2002; Bubeck et al., 2009; Kalyanakrishnan et al., 2012). The BAI problem has many applications such as hyper-parameter tuning (Li et al., 2017), clinical trials (Aziz et al., 2021), communication networks (Lindståhl et al., 2022) and user studies (Losada et al., 2022). However, many real-world scenarios often involve *constraints on the arms* that must be satisfied. For example, in recommender systems, one may need to ensure diversity and genre constraints (Kunaver and Požrl, 2017), or fairness of exposure (Wang et al., 2021). In clinical trials, one may need to account for toxicity constraints of the available treatments (Brannath et al., 2009; Chen, 2021; Demirel et al., 2022). As a result, standard BAI algorithms are not perfectly fitted in these settings.

In this paper, we introduce the problem of pure exploration in bandits with linear constraints where the goal is to identify, with a fixed confidence, a policy that maximizes the expected rewards over arms while satisfying some given constraints. A set of constraints may change the nature of the pure exploration problem fundamentally. In particular, the optimal policy *may not be deterministic*, and *finding the best arm may not be sufficient*. Let us consider the following example.

Example 1 (Optimal meal plan)**.** *Two people, A and B, are searching for a meal plan $\boldsymbol{\pi}$ that maximizes taste, i.e. expected reward $\boldsymbol{\mu}^\top \boldsymbol{\pi}$, while satisfying some nutrition constraints. Without any constraints this setting reduces to BAI and can be viewed as searching for the optimal policy over the probability simplex. However, as illustrated in Figure 1, the nutrition constraints alter the set of feasible sets and a person might have to mix between several dishes to satisfy the constraints while maximizing the reward. In Figure 1, the red arrow indicates the preference direction and the red dot corresponds to the optimal policy for each case. The dotted arrows, $\mathbf{b}_i$, corresponds to the normal of that boundary, i.e. the constraint causing the boundary, and as we will see later, Figure 2,*

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

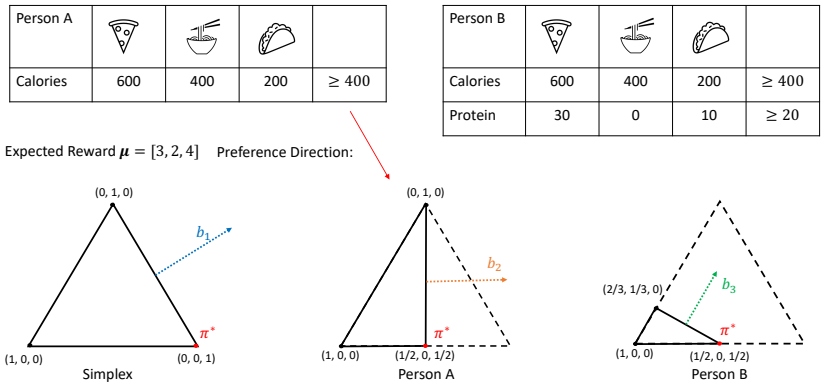


Figure 1: A Visual Representation of Example 1. Left figure with the full simplex represents the unconstrained problem. While the constraints of person A (middle) and person B (right) modify the problem to be harder and easier than the unconstrained one.

the distance between $\boldsymbol{\mu}$ and $\mathbf{b}_i$ controls the hardness of the problem. For person A, the distance between $\mathbf{b}_2$ and $\boldsymbol{\mu}$ decreases compared to the unconstrained case, while it increases for person B. Thus, the problem of finding the optimal pure exploration policy gets easier for person B while harder for person A. This is quantified by the minimum number of samples required to identify the optimal policies for person A, B, and the unconstrained case (ref. Fig. 2).

As illustrated in Figure 1, a learner may need to search for a *stochastic policy* that allocates positive probabilities to multiple arms and this influences how an efficient learner should explore. Depending on the constraints, the learner's task may become easier or harder, e.g. because the learner may need to explore several arms more extensively, or the constraints may remove several near-optimal policies, which makes the problem easier. These observations yield two fundamental questions:

1. *How do a specific set of constraints impact a pure exploration problem in terms of the minimum number samples required to identify the optimal policy?*

2. *How can we modify existing algorithms to adapt to the constraint setting?*

Our Contributions. We define the problem of pure exploration in bandits with linear constraints and derive a generic lower bound on the sample complexity of any algorithm in this setting. We further derive an explicit lower bound for arms corresponding to Gaussian distributions, which shows that the hardness depends on the projection of $\boldsymbol{\mu}$ onto boundary of a normal cone, and that the lower bound diminishes with the increasing condition number of the constraints defining the optimal policy. These insights allow us to modify the classical pure exploration algorithms, such as Track-and-Stop (Garivier and Kaufmann, 2016) and the game-theoretic algorithm (Degenne et al., 2019), and extend them to the constraint setting. We prove that our proposed algorithms CTnS and CGE are optimal in the asymptotic regime for the pure exploration problem with known linear constraints. Finally, we empirically evaluate the algorithms, both on synthetic data and scenarios based on the IMDB dataset (Maas et al., 2011). Experimental results validate our theoretical findings.

1.1 Related Work

Now, we review some works on policy learning, a classical problem in decision-making (Bechhofer, 1958), that deal with known or learned constraints on decisions and/or constraint exploration due to safety, fairness, or other preferences.

Adapting to Known Constraints. Constraints are often used to ensure safety in reinforcement learning, online learning and control (Moldovan and Abbeel, 2012; Gillulay and Tomlin, 2011; Wan et al., 2022; Vaswani et al., 2022). In the bandit literature, some variants of the best-arm identification (BAI) problem impose constraints on the chosen arm, or on the exploration process. Wang et al. (2022) study the setting with unknown linear rewards and known constraints but only allow single coordinate actions. Faizal and Nair (2022) consider BAI under fixed budget with known constraints on the arms. Their setting differs from ours in that *we look for a best "policy" over arms with linear constraints rather than a single best arm.*

Learning Unknown Constraints. Sui et al. (2015, 2018) study online optimization of an unknown function f with constraints on f, but without formal analysis. In the bandit literature, constraints are mostly studied in the regret-minimization setting. Moradipari et al. (2021) and Pacchiano et al. (2021) consider regret minimization in linear bandits under linear constraints from Bayesian and Frequentist

perspectives, respectively. Amani et al. (2019) study regret minimization in linear contextual bandits with unknown and unobserved linear constraints. Wang et al. (2021) aims to minimize the fairness regret to ensure proportional exposure for each arm, which implies a known structure for the policies. Unlike these works, we focus on the pure exploration setting. Lindner et al. (2022) considers constrained linear best-arm identification where arm are vectors with *known* rewards and a single *unknown* constraint (representing preferences) on the actions.

Pure Exploration Algorithms. Our Constrained Track-and-Stop algorithm, (CTnS, Section 4), follows the Track-and-Stop TnS) meta-scheme proposed by Garivier and Kaufmann (2016). In TnS, one tracks an optimal allocation with respect to a lower bound and assumes that the current estimate is the true environment. This approach has been applied to various bandits, e.g., linear bandits (Jedra and Proutiere, 2020), spectral bandits (Kocák and Garivier, 2021), heavy-tailed bandits (Agrawal et al., 2020), bandits with multiple correct answers (Degenne and Koolen, 2019), and latent bandits (Kinyanjui et al., 2023). The Constrained Game Explorer, (CGE, Section 4), follows the gamification approach to pure-exploration, which treats the lower bound as a zero-sum game between an allocation player and instance player. This approach was first introduced by Degenne et al. (2019), and later used for best-arm identification in linear bandits (Degenne et al., 2020) and combinatorial bandits (Degenne et al., 2020). In particular, CGE is an extension of the sampling rule of (Degenne et al., 2019) to the case of known linear constraints.

2 Problem Formulation: Pure Exploration with Linear Constraints

We consider a multi-armed bandit problem with K arms that corresponds to reward distributions, $\{P_a\}_{a=1}^K$, with unknown means $\{\mu_a\}_{a=1}^K$ and support $\mathbb{R}$. At each time step t, a learner chooses to play one of the arms, $A_t \in [K]$, and observes an immediate reward R_t, drawn from the reward distribution P_{A_t}. The learner has access to a non-empty and compact set of feasible policies

$$\mathcal{F} \triangleq \{\boldsymbol{\pi} \in \Delta_{K-1} : B\boldsymbol{\pi} \leqslant c\}, \tag{1}$$

where Δ_{K-1} is the K-simplex and $B \in \mathbb{R}^{N\times K}$ and $\boldsymbol{c} \in \mathbb{R}^N$, are known parameters of the linear constraints. For the ease of the presentation, we absorb the simplex constraints in B and $\boldsymbol{c}$. Hereafter, these variables refer to both the simplex constraints, and the additional linear constraints of the problem. *The goal of the learner is to recommend, with probability at least* $1-\delta$*, the* unique *optimal policy* $\boldsymbol{\pi}^*_{\boldsymbol{\mu},\mathcal{F}}$ satisfying

$$\boldsymbol{\pi}^*_{\boldsymbol{\mu},\mathcal{F}} \triangleq \underset{\boldsymbol{\pi}\in\mathcal{F}}{\arg\max} \quad \boldsymbol{\mu}^\top \boldsymbol{\pi}. \tag{2}$$

When it is clear from the context, we denote $\boldsymbol{\pi}^*_{\boldsymbol{\mu},\mathcal{F}}$ as $\boldsymbol{\pi}^*$. We refer to such a learner as a δ*-PAC learner*. As $1-\delta$ quantifies the correctness of the learner, we also want it to be efficient, i.e. to detect the optimal policy *fast*. Let τ_δ denotes the random *stopping time* at which the learner stops interacting with the bandit and makes a recommendation with confidence $1-\delta$. *We aim to design a* δ*-PAC learner that minimizes the expected stopping time* $\mathbb{E}[\tau_\delta]$*, a.k.a. sample complexity, needed to find the optimal policy.*

Depending on the application, a learner can abide by the constraints of Equation (1) in two ways:

- **Scenario 1: End-of-Time Constraint:** The learner does not have to take the constraints into account during exploration. Only the final recommended policy needs to satisfy the constraints.
- **Scenario 2: Anytime Constraint:** The exploration policy needs to satisfy the constraints *in expectation* during exploration, i.e. the exploration policy $\boldsymbol{w}_t$ needs to satisfy $\boldsymbol{w}_t \in \mathcal{F}$.

For example, Scenario 1 arises while using a more sophisticated hardware to search for an optimal policy, that should satisfy some energy-constraints, before deploying it on a low-energy hardware. In contrast, Scenario 2 can be thought of as performing the search directly on the low-energy hardware. Now, we explicitly state the assumptions used in this study:

- **Assumption 1:** The reward of each arm $i \in [K]$ is distributed according to a sub-Gaussian single-parameter exponential family parameterized by its unknown mean μ_i.
- **Assumption 2:** The vector of arm means, $\boldsymbol{\mu}$, lies in a bounded domain $\mathcal{D} = [\mu_{\min}, \mu_{\max}]^K$.
- **Assumption 3:** The optimal solution $\boldsymbol{\pi}^*_{\boldsymbol{\mu},\mathcal{F}}$ to the linear program in Equation (2) is unique.

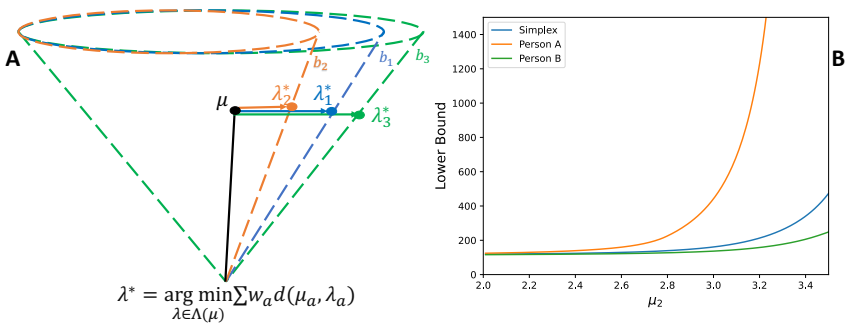


Figure 2: Computing the $\boldsymbol{\lambda}$ satisfying Equation 6, i.e. the *most confusing instance*, can be viewed as an information-theoretic projection onto the boundary of the normal cone spanned by the active constraints at $\boldsymbol{\pi}_{\boldsymbol{\mu}}$. In A) we see the different normal cones for the three different examples in Figure 1. In B) we have fixed μ_1 and μ_3, as in Figure 1, and plot the lower bound, assuming $N(0,1)$ noise and with $\delta = 0.1$, for increasing μ_2 which mean that we are moving $\boldsymbol{\mu}$ closer to the boundaries in A). We observe an inverse relationship between the distance to the boundary and the lower bound, properly characterized in Proposition 2. As a result, any δ-PAC learner needs to gather enough samples from $\boldsymbol{\mu}$ such that it's confidence set is fully contained within the normal cone.

Assumptions 1 and 2 are standard in the literature (Degenne and Koolen, 2019; Degenne et al., 2020). Assumption 3 is the analogue of assuming a unique best arm in the BAI problem, and it ensures that the optimum of Equation (2) is an extreme point. Hence, the optimal policy $\boldsymbol{\pi}^*_{\boldsymbol{\mu},\mathcal{F}}$ always corresponds to an extreme point in the polytope $\mathcal{F}$. In Appendix D, we discuss about relaxation to ϵ-good policies.

Notations. Let Π denote the set of feasible exploration policies. Thus, for Scenario 1, $\Pi = \Delta_{K-1}$, and $\Pi = \mathcal{F}$ for Scenario 2. We denote the KL-divergence between two single-parameter exponential family distributions with mean x and y as $d(x, y)$. Additionally, if the random variables are Bernoulli, we denote the KL-divergence as $\mathbf{kl}(x||y)$.

3 Lower Bound on The Sample Complexity of any δ-Correct Algorithm

Lower bounds on the sample complexity of a δ-correct algorithm, i.e. $\mathbb{E}[\tau_\delta]$, is a driving force in designing good algorithms in the BAI literature (Garivier and Kaufmann, 2016; Degenne and Koolen, 2019; Agrawal et al., 2020).

Given a problem instance $\boldsymbol{\mu}$, a learner needs to collect enough information about the problem to be able to rule out all alternative instances, $\boldsymbol{\lambda}$, for which we have $\max_{\boldsymbol{\pi}\in\mathcal{F}} \boldsymbol{\lambda}^\top \boldsymbol{\pi} > \boldsymbol{\lambda}^\top \boldsymbol{\pi}^*$ with confidence at least $1-\delta$. We refer to this set of instances as the Alt-set and denote it as

$$\Lambda_{\mathcal{F}}(\boldsymbol{\mu}) \triangleq \{\boldsymbol{\lambda} \in \mathcal{D} : \max_{\boldsymbol{\pi}\in\mathcal{F}} \boldsymbol{\lambda}^\top \boldsymbol{\pi} > \boldsymbol{\lambda}^\top \boldsymbol{\pi}^*\}. \tag{3}$$

Garivier and Kaufmann (2016) introduced general techniques for deriving lower bounds on the sample complexity of any δ-PAC learner, which depends on the the distance from $\boldsymbol{\mu}$ to the closest $\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})$ in an information-theoretic sense. We extend these general proof techniques and show that the expected stopping time of any δ-PAC algorithm ϕ for BAI with linear constraints satisfies

$$\mathbb{E}_{\boldsymbol{\mu},\phi}[\tau_\delta] \geqslant T_{\mathcal{F}}(\boldsymbol{\mu})\mathbf{kl}(\delta||1-\delta). \tag{4}$$

where $T_{\mathcal{F}}(\boldsymbol{\mu})$ is the *characteristic time*, defined as

$$T_{\mathcal{F}}^{-1}(\boldsymbol{\mu}) = \sup_{\boldsymbol{w}\in\Pi} \inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a). \tag{5}$$

The supremum in Equation (5) hints towards the existence of some optimal exploration policy $\boldsymbol{w}$, which any optimal algorithm should try to track. This is exactly the idea behind the Track-and-Stop meta-scheme (Garivier and Kaufmann, 2016) (details in Section 4). In order to design algorithms achieving the lower bound in Equation (4), we need to solve the optimization problem in Equation (5). This requires a more explicit characterization of $\Lambda_{\mathcal{F}}(\mu)$, continuity properties of of the function $D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F}) \triangleq \inf_{\boldsymbol{\lambda}\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a)$, and the set of optimal allocations $w^*(\boldsymbol{\mu})$.

To derive an explicit expression for $\Lambda_{\mathcal{F}}(\mu)$, let M be the number of active constraints for $\boldsymbol{\pi}^*$, $B_{\boldsymbol{\pi}*} \in \mathbb{R}^{M\times K}$ be a submatrix of B consisting of all these active constraints, and $\boldsymbol{c}_{\boldsymbol{\pi}*} \in \mathbb{R}^M$ the corresponding bounds in $\boldsymbol{c}$. Hence, there exists *at least* K linearly independent rows in $B_{\boldsymbol{\pi}*}$, i.e.

a matrix $\hat{B}_{\boldsymbol{\pi}*} \in \mathbb{R}^{K\times K}$ and vector $\hat{\boldsymbol{c}}_{\boldsymbol{\pi}*} \in \mathbb{R}^K$, such that $\boldsymbol{\pi}^* = \hat{B}_{\boldsymbol{\pi}*}^{-1}\hat{\boldsymbol{c}}_{\boldsymbol{\pi}*}$. Since our objective (Equation (2)) is a linear program, we can leverage the optimality condition stating that $\boldsymbol{\mu}$ must be in the normal cone of the optimal solution (Boyd and Vandenberghe, 2004). Hence, we express the Alt-set as $\Lambda_{\mathcal{F}}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} : \boldsymbol{\lambda} \notin \mathcal{N}(\boldsymbol{\pi}^*)\}$. Here, $\mathcal{N}(\boldsymbol{\pi}^*) := \left\{\boldsymbol{\lambda} : \boldsymbol{\lambda} = B_{\boldsymbol{\pi}*}^\top \boldsymbol{v}, \boldsymbol{v} \in \mathbb{R}^M_{\geqslant 0}\right\}$ is the normal cone spanned by the active constraints for $\boldsymbol{\pi}^*$.

Further, we say that $\boldsymbol{\pi}'$ is a neighbor of $\boldsymbol{\pi}^*$ if it is an extreme point in $\mathcal{F}$ and shares $K-1$ active constraints with $\boldsymbol{\pi}^*$. We denote the set of all neighbors of $\boldsymbol{\pi}^*$ as $\mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)$. Hence, we can decompose the Alt-set into a union of a finite number of half-spaces $\Lambda_{\mathcal{F}}(\boldsymbol{\mu}) = \bigcup_{\boldsymbol{\pi}'\in\mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \left\{\boldsymbol{\lambda} : \boldsymbol{\lambda}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}') < 0\right\}$. This formulation implies that if $\boldsymbol{\pi}^*$ is not an optimal policy for the instance $\boldsymbol{\lambda}$, there must exist an direction for the simplex algorithm to follow to increase the expected reward, i.e. $\exists \boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*) : \boldsymbol{\lambda}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}') < 0$. This formulation of Alt-sets lead us to the observation that the most confusing instances in the Alt-set w.r.t. $\boldsymbol{\mu}$ lay on the boundary of the normal cone. Specifically, Lemma 1 shows that the function $D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$ is a weighted projection onto the plane $\boldsymbol{\lambda}^\top(\boldsymbol{\pi}' - \boldsymbol{\pi}^*) = 0$ for some $\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)$, as shown in Figure 2.

Lemma 1 (Projection lemma)**.** *For any $\boldsymbol{w} \in \Pi$ and $\boldsymbol{\mu}$ it holds that*

$$D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F}) = \min_{\boldsymbol{\pi}'\in\mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \min_{\boldsymbol{\lambda}\in\mathcal{D}:\boldsymbol{\lambda}^\top(\boldsymbol{\pi}^*-\boldsymbol{\pi}')=0} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) \tag{6}$$

To compute $D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$ from Equation (6), we need to have access to the true instance $\boldsymbol{\mu}$, which we do not have in reality. Rather, we sequentially obtain samples from the arms yielding an estimate $\hat{\boldsymbol{\mu}}_t$. Thus, we need $D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$ and $\boldsymbol{w}^*(\boldsymbol{\mu})$ to satisfy continuity properties (Theorem 1) w.r.t $\boldsymbol{\mu}$, that ensures as the estimates $\hat{\boldsymbol{\mu}}_t$ converge to $\boldsymbol{\mu}$, $D(\boldsymbol{w}, \hat{\boldsymbol{\mu}}_t, \mathcal{F}) \to D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$ and our empirical distribution of plays gets closer to some $\boldsymbol{w} \in w^*(\mu)$.

Theorem 1. *Following properties are true for all $\boldsymbol{\mu}$ and $\mathcal{F} = \{\boldsymbol{\pi} \in \Delta_{K-1} : B\boldsymbol{\pi} \leqslant c\}$ such that the problem $\max_{\boldsymbol{\pi}\in\mathcal{F}} \boldsymbol{\mu}^\top\boldsymbol{\pi}$ has a unique solution.*

- *The function $(\boldsymbol{w}, \boldsymbol{\mu}) \mapsto D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$ is continuous.*
- *The function $\boldsymbol{\mu} \mapsto T_{\mathcal{F}}(\boldsymbol{\mu})$ is continuous.*
- *The set-valued function $\boldsymbol{\mu} \mapsto w^*(\boldsymbol{\mu})$ is upper hemicontinuous (definition in Appendix F).*
- *The set $w^*(\boldsymbol{\mu})$ is convex.*

3.1 Lower Bound for Gaussian Distributions

To gain further insights on how the constraints alter the lower bound in Equation (4), we consider the special case where all arms are Gaussian distributions with equal variance σ^2. This leads us to a close-form of the projection in Lemma 1 as in Proposition 1.

Proposition 1. *If the arms follow Gaussian distributions with identical variance σ^2 and $w_a > 0$ $\forall a$, we have that the projection $\min_{\boldsymbol{\lambda}\in\mathcal{D}:\boldsymbol{\lambda}^\top(\boldsymbol{\pi}^*-\boldsymbol{\pi}')\leqslant 0} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a)$ for any $\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)$ is satisfied by*

$$\lambda_{a,\boldsymbol{\pi}'} = \mu_a - \gamma\frac{(\boldsymbol{\pi}^* - \boldsymbol{\pi}')_a}{w_a},$$

for $\gamma = \frac{\boldsymbol{\mu}^\top(\boldsymbol{\pi}^*-\boldsymbol{\pi}')}{\sum_a \frac{(\boldsymbol{\pi}^*-\boldsymbol{\pi}')^2}{w_a}}$, and the characteristic time is

$$T_{\mathcal{F}}(\boldsymbol{\mu}) = \max_{\boldsymbol{w}\in\Pi} \min_{\boldsymbol{\pi}'\in\mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} 2\sigma^2 \frac{\sum_a \frac{(\boldsymbol{\pi}^*-\boldsymbol{\pi}')_a^2}{w_a}}{\left(\boldsymbol{\mu}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}')\right)^2} = \max_{\boldsymbol{w}\in\Pi} \min_{\boldsymbol{\pi}'\in\mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} 2\sigma^2 \frac{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|^2_{\mathrm{Diag}(1/w_a)}}{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|^2_{\boldsymbol{\mu}\boldsymbol{\mu}^\top}} \tag{7}$$

Here, $\mathrm{Diag}(1/w_a)$ is a diagonal matrix with a-th entry of the diagonal as $1/w_a$.

From Proposition 1, we further derive a lower and an upper bound on the characteristic time. Let us define $d_{\boldsymbol{\pi}'} \triangleq \min_{\boldsymbol{\lambda}:\boldsymbol{\lambda}^\top(\boldsymbol{\pi}^*-\boldsymbol{\pi}')=0} ||\boldsymbol{\mu} - \boldsymbol{\lambda}||_2$ and note that this is the distance between $\boldsymbol{\mu}$ and the hyperplane $\boldsymbol{\pi}^* - \boldsymbol{\pi}' = 0$, see Figure 2 for illustration.

Proposition 2. *The characteristic time $T_{\mathcal{F}}(\boldsymbol{\mu})$ satisfies the following bounds:*

$$\min_{\boldsymbol{\pi}'\in\mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \frac{4\sigma^2}{d^2_{\boldsymbol{\pi}'}} \leqslant T_{\mathcal{F}}(\boldsymbol{\mu}) \leqslant \min_{\boldsymbol{\pi}'\in\mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \frac{4\sigma^2 K}{d^2_{\boldsymbol{\pi}'}}. \tag{8}$$

Proposition 2 implies a lower bound of $\mathbb{E}[\tau] \geqslant \min_{\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \frac{4\sigma^2}{d^2_{\boldsymbol{\pi}'}} \mathbf{kl}(\delta || 1-\delta)$.

Impact of Constraints: Geometric View. We first observe that, since the distance-to-projection $d_{\boldsymbol{\pi}'} = \frac{\boldsymbol{\mu}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}')}{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|_2}$, the problem becomes easier when the direction of the reward vector $\boldsymbol{\mu}$ is aligned with the deviation in policy $\boldsymbol{\pi}^* - \boldsymbol{\pi}'$. Especially, if we only consider deterministic policies, i.e. BAI problem, $d_{\boldsymbol{\pi}'} = \mu_1 - \mu_a$ where μ_1 is the best arm, a is the arm played by $\boldsymbol{\pi}'$ and we retrieve the lower bound of Kaufmann et al. (2016).

Impact of Constraints: Constrained Optimization View. We relate the lower bound more explicitly to the constraint matrix B by using the fact that any neighbor $\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)$ can be reached from $\boldsymbol{\pi}^*$ via an 1-rank update on a matrix $\hat{B}_{\boldsymbol{\pi}^*} \in \mathbb{R}^{K \times K}$ consisting of K active constraints at $\boldsymbol{\pi}^*$ that are linearly independent. Thus, we only need to change one row in $\hat{B}_{\boldsymbol{\pi}^*}$ and one element in the corresponding $\hat{\boldsymbol{c}}_{\boldsymbol{\pi}^*}$ to get B' and $\boldsymbol{c}'$ such that $\boldsymbol{\pi}' = B'^{-1}\boldsymbol{c}'$. This results in the lower bound on the sample complexity presented in Corollary 1. This bound is weaker than the bound implied by Proposition 2.

Corollary 1. *Given a feasible set of policies $\mathcal{F}$ defined by linear constraints $B\boldsymbol{\pi} \leqslant \boldsymbol{c}$, the sample complexity of any δ-PAC learner is lower bounded as*

$$\mathbb{E}[\tau] = \Omega\left(\frac{H}{\kappa^2}\mathbf{kl}(\delta||1-\delta)\right) \tag{9}$$

where κ^2 is the minimum condition number of any submatrix $\hat{B}_{\boldsymbol{\pi}^*} \in \mathbb{R}^{K \times K}$ consisting of M linearly independent active constraints at $\boldsymbol{\pi}^*$, and $H = \frac{2\sigma^2}{\sum_{a \neq a^*} \Delta_a^2}$ is the complexity of the bandit instance.

Corollary 1 illustrates the relation between constraints and complexity of identifying the optimal policy. The minimum condition number κ is a measure of how sensitive the active constraints are to changes in the policy space. A large value indicates that whenever we apply a small perturbation to the optimal policy, $\boldsymbol{\pi}^*$, there will be a large change in the value of the active constraints, i.e. there will be a large change in the corresponding slack variables. As a result, the set of possible policies, i.e. extreme points in $\mathcal{F}$, will structurally behave very different and the exploration problem should become easier. In contrast, if κ is small, perturbations to a policy $\boldsymbol{\pi}$ has smaller impact on $B\boldsymbol{\pi}$ rendering neighboring policies less distinguishable from the optimal policy.

4 Algorithms: Constraint Adaptive Track-n-Stop and Gamified Explorer

In this section, we focus on extending the classical pure exploration algorithms to the setting of pure exploration with linear constraints.

Algorithm Design. We begin by observing that any pure exploration algorithm consists of three components: *a stopping rule*, *a recommendation rule*, and *a sampling strategy*. The stopping rule consists of a condition deciding when to halt sampling further. The recommendation rule decides what policy to recommend as the optimal policy. The sampling rule decides which arm to sample next given the history of arms sampled and intermediate policies computed.

Component 1: Chernoff's Stopping Rule with Constraints. As a stopping rule, we extend the Chernoff's stopping rule (Garivier and Kaufmann, 2016). We first introduce the *confidence set* $\mathcal{C}_t(\delta) := \left\{\boldsymbol{\lambda} : \sum_{a=1}^K N_{a,t} d(\hat{\mu}_{a,t}, \lambda_a) \leqslant c(t,\delta)\right\}$, where $c(t,\delta)$ is a threshold defined in Lemma 2.

Lemma 2 (Garivier and Kaufmann (2016))**.** *For any $\alpha > 1$ there exists a constant $C(\alpha, K)$ such that for $c(t,\delta) = \log \frac{t^\alpha C(\alpha,K)}{\delta}$ we have for any $t \in \mathbb{N}$ $P\left(\boldsymbol{\mu} \notin \mathcal{C}_t(\delta)\right) \leqslant \delta$.*

Lemma 2 implies that Chernoff's stopping rule is a δ-PAC stopping rule, and we stop when

$$\inf_{\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\hat{\boldsymbol{\mu}}_t)} \sum_{a=1}^K N_{a,t} d(\hat{\mu}_{a,t}, \lambda_a) > c(t,\delta). \tag{10}$$

This means that the confidence set is a subset of the normal cone spanned by the active constraints at $\boldsymbol{\pi}^*_{\hat{\boldsymbol{\mu}}_t}$. The details of the constant in Lemma 2 are deferred to Appendix C.

Component 2: Recommendation Rule. We recommend the solution of the linear programming (Equation (2)) with the empirical means of the arms at the stopping time, $\boldsymbol{\pi}^*_{\hat{\boldsymbol{\mu}}_t} = \arg\max_{\boldsymbol{\pi} \in \mathcal{F}} \hat{\boldsymbol{\mu}}_t^\top \boldsymbol{\pi}$.

Algorithm 1 Constrained Track-and-Stop (CTnS)

Require: Confidence level δ, constraints (B, c), exploration set Π

1: Play each arm once.
2: **while** $c(t,\delta) > D(\boldsymbol{N}/t, \hat{\boldsymbol{\mu}}_t, \mathcal{F})$ **do** ▷ Weighted projection via Lemma 1
3: Compute $\boldsymbol{w}_t^* \in \arg\max_{\boldsymbol{w}\in\Pi} D(\boldsymbol{w}, \hat{\boldsymbol{\mu}}_t, \mathcal{F})$ ▷ Solve for optimal $\boldsymbol{w}$ w.r.t. the constraints
4: Play $A_t \in \arg\min_a N_{a,t} - \sum_{s=1}^t w^*_{a,s,\epsilon_s}$ and observe reward R_t
5: **end while**
6: Recommend $\boldsymbol{\pi}^*_{\hat{\boldsymbol{\mu}}_t} = \arg\max_{\boldsymbol{\pi}\in\mathcal{F}} \hat{\boldsymbol{\mu}}_t^\top \boldsymbol{\pi}$

Algorithm 2 Constrained Game Explorer (CGE)

Require: Confidence level δ, constraints (B, c), exploration set Π

1: **while** $c(t,\delta) > D(\boldsymbol{N}/t, \hat{\boldsymbol{\mu}}_t, \mathcal{F})$ **do** ▷ Weighted projection via Lemma 1
2: Get allocation $\boldsymbol{w}_t$ from regret minimizer ▷ Running Adagrad over Π
3: Compute best-response $\boldsymbol{\lambda}_t$ w.r.t. $\boldsymbol{w}_t$ and $\hat{\boldsymbol{\mu}}_t$ ▷ Weighted projection via Lemma 1
4: Compute confidence intervals $\forall a\ [\alpha_{t,a}, \beta_{t,a}] = \{\xi : N_{a,t} d(\hat{\mu}_{a,t}, \xi) \leqslant f(t)\}$
5: $\forall a$ let $U_t^a := \max\left\{\frac{f(t)}{N_{a,t}}, \max_{\xi\in\{\alpha_t^a, \beta_t^a\}} d(\xi, \lambda_{a,t})\right\}$
6: Update regret minimizer with loss $l(w_t) = \sum_{a=1}^K w_a U_{a,t}$
7: Play $A_t \in \arg\min_a N_{a,t} - \sum_{s=1}^t w^*_{a,s,\epsilon_s}$ and observe reward R_t
8: **end while**
9: Recommend $\boldsymbol{\pi}^*_{\hat{\boldsymbol{\mu}}_t} = \arg\max_{\boldsymbol{\pi}\in\mathcal{F}} \hat{\boldsymbol{\mu}}_t^\top \boldsymbol{\pi}$

Since the empirical means might not always be within the pre-specified range $\mathcal{D}$, we let $\hat{\boldsymbol{\mu}}_t$ denote the Euclidean projection of the empirical means onto $\mathcal{D}$.

Component 3a: CTnS (A Track-and-Stop Approach).

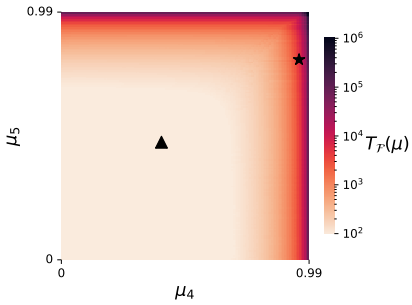
First, we present our Constrained Track-and-Stop Algorithm (CTnS, Algorithm 1), which is an adaptation of the Track-and-Stop (TnS) framework (Garivier and Kaufmann, 2016) to the linear constraint setting with aforementioned stopping and recommendation rules. In Algorithm 1, we highlight, in red, the computations that we modify from the original schematic to account for the linear constraints. The algorithm starts by playing each arm once. Then, until the stopping rule in Equation (10) fires, it performs *C-tracking* (Garivier and Kaufmann, 2016). This means that we perform a $\max - \min$ oracle call (Line 3), and solve the problem in Equation (4) w.r.t our current estimate of the means $\hat{\boldsymbol{\mu}}_t$ to get an optimal allocation $\boldsymbol{w}_t^*$. We track the optimal allocation via $A_t \in \arg\min_a N_{a,t} - \sum_{s=1}^t w^*_{a,s,\epsilon_s}$, where w^*_{a,t,ϵ_t} is the projection of $\boldsymbol{w}_t^*$ onto $\Pi \bigcap \{\boldsymbol{w} : w_a > \epsilon_t \forall a\}$, and $\epsilon_t = \frac{1}{2\sqrt{K^2+t}}$. Note that $\frac{1}{t}\sum_{s=1}^t w^*_{a,s,\epsilon_s} \in \Pi$ due to the convexity of the set feasible exploration policies/allocations.

Theorem 2 (Upper bound for CTnS)**.** *For any $\alpha > 1$ and $c(t,\delta)$ be defined as in Lemma 2, we have that the expected stopping time of CTnS satisfies* $\lim_{\delta\to 0} \frac{\mathbb{E}[\tau]}{\log\frac{1}{\delta}} \leqslant T_{\mathcal{F}}(\boldsymbol{\mu}),\ \forall \boldsymbol{\mu} \in \mathcal{D}$.

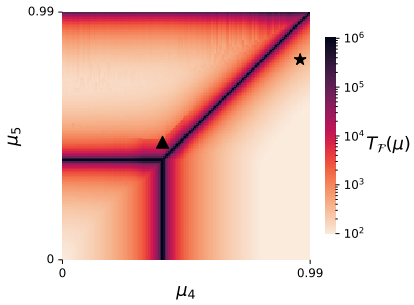
The proof of Theorem 2 can be found in Appendix C.2 and follows the same structure as the sample complexity proof of the original TnS in Garivier and Kaufmann (2016). However, the optimal allocation does not necessarily have to be unique. rather, we use the upper hemicontinuity and convexity of $w^*(\boldsymbol{\mu})$, while modifying the tracking lemma originally used by Garivier and Kaufmann (2016) with the tracking result of Degenne and Koolen (2019). This change allows to track a set of optimal solutions in absence of a unique optimum.

Component 3b: CGE (A Game-Theoretic Approach).

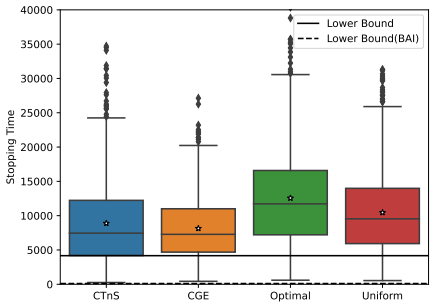
Track-and-Stop algorithms, like CTnS, tend to be computationally inefficient for larger problems since they requires a $\max - \min$ call at each iteration. To mitigate this, we adopt the approach of Degenne et al. (2019), and treat the optimization problem in Equation (5) as a two player zero-sum game. This results in the Constrained Game Explorer (CGE), in Algorithm 2. Instead of solving for an optimal $\boldsymbol{w}_t$ at each t, as in CTnS, we play one game between an allocation player, who plays $\boldsymbol{w}$ to maximize $\sum_{a=1}^K w_a d(\hat{\mu}_{a,t}, \lambda_a)$, and an instance player, who plays the confusing instance $\boldsymbol{\lambda}$



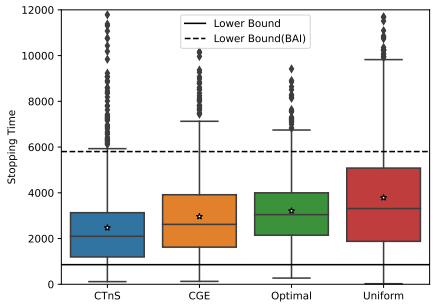
(a) Characteristic time of the BAI problem as we vary μ_4 and μ_5.



(b) Characteristic time of the constraint pure-exploration problem as we vary μ_4 and μ_5



(c) Results on the instance highlighted as a triangle in Figure 3a and Figure 3b, $\boldsymbol{\mu} = (1, 0.5, 0.4, 0.4, 0.5)$, with constraints and $\delta = 0.1$. Results are for 1000 random seeds.



(d) Results on the instance highlighted as a star in Figure 3a and Figure 3b, $\boldsymbol{\mu} = (1, 0.5, 0.4, 0.95, 0.8)$, with constraints and $\delta = 0.1$. Results are for 1000 random seeds.

Figure 3: Figure 3a and Figure 3b illustrate of the hardness of the problem, i.e. the Characteristic time, changes in the 5 arm instance $\boldsymbol{\mu} = (1.0, 0.5, 0.4, \mu_4, \mu_5)$ as we vary μ_4 and μ_5. Figure 3a corresponds to the hardness in the BAI while Figure 3b is the constraint setting with constraints $\boldsymbol{\pi}_1 + \boldsymbol{\pi}_2 \leqslant 0.5$ and $\boldsymbol{\pi}_3 + \boldsymbol{\pi}_4 \leqslant 0.5$. We clip the characteristic time at 10^6 for visual purposes.

to minimize $\sum_{a=1}^{K} w_a d(\hat{\mu}_{a,t}, \lambda_a)$. We deploy an instance of AdaGrad (Duchi et al., 2011) as the allocation player is taken to be, which enjoys sub-linear regret over any bounded domain when losses are convex, and the instance player is taken to be a best-response w.r.t. the allocation $\boldsymbol{w}_t$. The loss provided to AdaGrad at each time step is $\sum_{a=1}^{K} w_{a,t} U_{a,t}$, where $U_{a,t}$ induces optimism as $U_{a,t} \triangleq \max_{\xi \in \{\alpha_{a,t}, \beta_{a,t}\}} N_{a,t} d(\xi, \lambda_{a,t})$. Here, $(\alpha_{a,t}, \beta_{a,t})$ are the endpoints of the confidence interval around $\hat{\mu}_{a,t}$, i.e. $[\alpha_{t,a}, \beta_{t,a}] = \{\xi : N_{a,t} d(\mu_{a,t}, \xi) < f(t)\}$, and $f(t) = 3 \log t + \log \log t$. We apply the same tracking as in CTnS.

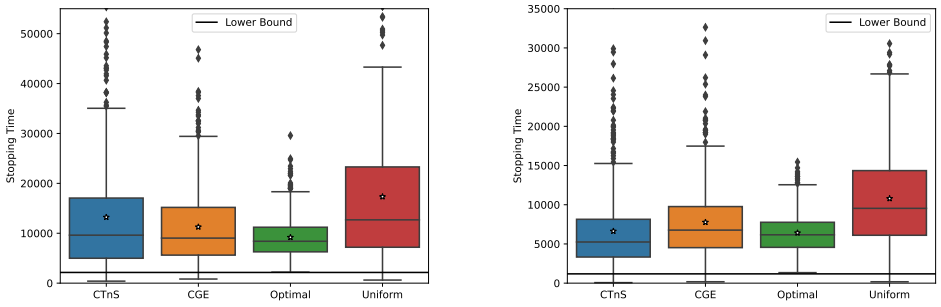
Theorem 3 (Upper bound for CGE)**.** *The expected sample complexity of CGE satisfies* $\mathbb{E}[\tau] \leqslant T_0(\delta) + CK$, *where* $T_0(\delta) := \max\left\{t \in \mathbb{N} : t \leqslant T_{\mathcal{F}}(\boldsymbol{\mu}) c(t, \delta) + O(\sqrt{tQ}) + O(\sqrt{t \log t})\right\}$. C_μ *is problem-dependent constant, C is a universal constant and Q is an upper bound on the losses provided to Adagrad.*

The full proof of Theorem 3 can be found in Appendix C.3. We simply follow the steps of the proof of Theorem 2 in Degenne et al. (2019) and apply specifics of our setting when applicable.

Theorem 2 and 3 show that CTnS and CGE are asymptotically optimally, i.e. upper bound on their sample complexities match the lower bound of constrained pure exploration for small enough δ.

5 Experimental Analysis

We evaluate our algorithms using the threshold $c(t, \delta) = \log \frac{1 + \log \log t}{\delta}$, which is commonly done in the literature (Garivier and Kaufmann, 2016), and we set $f(t) = \log t$ in CGE. As benchmarks we will use the lower bound, Equation 4, as well as a learner that samples from the optimal allocation, given by the lower bound, at all time steps. We also consider a learner that draws arms from the uniform distribution and in scenarios where the uniform distribution is not in the set of feasible exploration policies we project it onto the set and sample from the resulting distribution. Appendix E contains additional experiments.



(a) Anytime Constraints

(b) End-of-Time Constraints

Figure 4: Experiments on IMDB dataset with 12 movies and $\delta = 0.1$. Each experiment was performed over 500 random seeds.

Synthetic Dataset. In Figures 3a and 3b we illustrate how the hardness of a bandit instance $\boldsymbol{\mu}$ may differ once we introduce constraints, assuming anytime constraints. We consider the instance $\boldsymbol{\mu} = (1.0, 0.5, 0.4, \mu_4, \mu_5)$ and plot how the characteristic time $T_{\mathcal{F}}(\boldsymbol{\mu})$ changes as we vary μ_4 and μ_5, Figure 3a corresponds to the classical BAI, i.e. no constraints, and in Figure 3b we have introduced the two constraints $\boldsymbol{\pi}_1 + \boldsymbol{\pi}_2 \leqslant 0.5$ and $\boldsymbol{\pi}_3 + \boldsymbol{\pi}_4 \leqslant 0.5$. We have highlighted two instances, one where the BAI problem is easy but the constraint problem is hard (black triangle) and one where the reverse is true (black star). We run the algorithms on these two instances in Figure 3c and 3d, assuming anytime constraints, and observe that both algorithms operate close to the lower bound and outperforms the uniform allocation strategy. We also observe that the algorithms perform equally or better than the optimal learner, this is an interesting phenomena and have been observed earlier in other pure exploration scenarios (Degenne et al., 2019).

IMDB Movie Recommendation Environment. We construct a semi-synthetic task based on the widely used IMDB 50K Movie Dataset (Maas et al., 2011) which contains metadata on $k_0 = 50000$ movies including association with one or more of $d = 23$ genres, as indicated by a binary matrix $X \in \{0,1\}^{m \times d}$. In our setting, actions correspond to recommending one out of a subset of $k \leqslant k_0$ movies. To create reward distributions for each movie, we simulate a population of $n_u = 600$ users, each assigned $n_f = 5$ favorite genres f_i with weights $w_{if_i} = [20, 10, 5, 2, 2]$ and let $w_{ia} = 0$ for $a \notin f_i$. A score s_{ia} for user i and movie a is created as follows, $s_{ij} = \text{clip}(\lfloor \tilde{s}_{ia} / \sum_{a \in f_i} w_{ia} \cdot \sigma_0 + \sigma_1 \epsilon_{ia} \rceil ; 1, 5)$ where $\tilde{s}_{i\cdot} = w_i X^\top + w_0$, $\epsilon_{ia} \sim U(0,1)$, $\sigma_0 = 5, \sigma_1 = 3$, and $\lfloor x \rceil$ indicates rounding of x to the nearest integer. We construct the bandit environment by letting each movie a be represented by an arm with reward $R_a \sim \mathcal{N}(\hat{\mu}_{s\cdot a}, \hat{\sigma}^2_{s\cdot a})$ determined by the mean and standard deviation of user reviews for the movie. We sample a subset of movies and search for the optimal policy that allocates at most 0.3 to action movies, at least 0.3 to drama movies and at least 0.3 on family movies. Note that one movie might belong to more than one category. We present the result in Figure 4 for both the anytime scenario and the end of time scenario. We observe that CTnS and CGE outperform the uniform allocation strategy, which has a very high variance. We also observe a bigger difference between the algorithms under end of time constraints, this is reasonable since the set of plausible exploration policies is larger for that scenario. If the set of exploration policies is limited, there is little room for an algorithm to be adaptive. This is also captured in the fact that the lower bound for anytime constraints is always higher or equal to the bound for end-of-time constraints.

6 Conclusion and Future Directions

In this paper, we study the problem of pure exploration in bandits with linear constraints. We provide a generic lower bound for this setting that depends on an information-theoretic projection onto the boundary of the normal cone spanned by the active constraints at the optimal policy. We derive a closed-form lower bound for the case of Gaussian distributions and provide geometric insights into how constraints can make a problem easier or harder. Furthermore, we leverage the projection-based computation of the confusing instances to modify TnS (Garivier and Kaufmann, 2016) and GE (Degenne et al., 2019) to corresponding CTnS and CGE versions for pure exploration in constraint bandits. We empirically evaluate the algorithms on synthetic and real data to assess the impact of constraints on the hardness of the problem.

One interesting future direction is learning when both reward and constraints are unknown or partially unknown. Another future direction we deem very interesting is bandit with non-linear constraints, which will fundamentally change the structure of the normal cone and resulting projections.

References

Shubhada Agrawal, Sandeep Juneja, and Peter Glynn. Optimal δ-correct best-arm selection for heavy-tailed distributions. In *Algorithmic Learning Theory*, pages 61–110. PMLR, 2020.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

Robert E Bechhofer. A sequential multiple-decision procedure for selecting the best one of several normal populations with a common unknown variance, and its use with various experimental designs. *Biometrics*, 14(3):408–429, 1958.

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

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004.

Werner Brannath, Emmanuel Zuber, Michael Branson, Frank Bretz, Paul Gallo, Martin Posch, and Amy Racine-Poon. Confirmatory adaptive designs with bayesian decision tools for a targeted therapy in oncology. *Statistics in medicine*, 28(10):1445–1463, 2009.

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.

Ningyuan Chen. Multi-armed bandit requiring monotone arm sequences. *Advances in Neural Information Processing Systems*, 34:16093–16103, 2021.

Rémy Degenne and Wouter M Koolen. Pure exploration with multiple correct answers. 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, 2019.

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, 2019.

Rémy Degenne, Pierre Ménard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR, 2020.

Ilker Demirel, Ahmet Alparslan Celik, and Cem Tekin. Escada: Efficient safety and context aware dose allocation for precision medicine. *Advances in Neural Information Processing Systems*, 35: 27441–27454, 2022.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL `http://jmlr.org/papers/v12/duchi11a.html`.

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.

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

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

Aurélien Garivier and Emilie Kaufmann. Nonasymptotic sequential tests for overlapping hypotheses applied to near-optimal arm identification in bandit models. *Sequential Analysis*, 40(1):61–96, 2021. doi: 10.1080/07474946.2021.1847965. URL `https://doi.org/10.1080/07474946.2021.1847965`.

Jeremy H Gillulay and Claire J Tomlin. Guaranteed safe online learning of a bounded system. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2979–2984. IEEE, 2011.

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

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, 1, 01 2012.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *J. Mach. Learn. Res.*, 17(1):1–42, jan 2016. ISSN 1532-4435.

Newton Mwai Kinyanjui, Emil Carlsson, and Fredrik D. Johansson. Fast treatment personalization with latent bandits in fixed-confidence pure exploration. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Tomáš Kocák and Aurélien Garivier. Epsilon best arm identification in spectral bandits. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2636–2642. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/363. URL `https://doi.org/10.24963/ijcai.2021/363`. Main Track.

Matevž Kunaver and Tomaž Požrl. Diversity in recommender systems – a survey. *Knowledge-Based Systems*, 123:154–162, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

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.

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.

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.

David E Losada, David Elsweiler, Morgan Harvey, and Christoph Trattner. A day at the races: using best arm identification algorithms to reduce the cost of information retrieval user studies. *Applied Intelligence*, 52(5):5617–5632, 2022.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *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. URL `https://aclanthology.org/P11-1015`.

Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz bandits: Regret lower bound and optimal algorithms. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 975–999, Barcelona, Spain, 13–15 Jun 2014. PMLR.

Teodor Mihai Moldovan and Pieter Abbeel. Safe exploration in markov decision processes. *arXiv preprint arXiv:1205.4810*, 2012.

Ahmadreza Moradipari, Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Safe linear thompson sampling with side information. *IEEE Transactions on Signal Processing*, 69: 3755–3767, 2021.

Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. In *International conference on artificial intelligence and statistics*, pages 2827–2835. PMLR, 2021.

Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with gaussian processes. In *International conference on machine learning*, pages 997–1005. PMLR, 2015.

Yanan Sui, Vincent Zhuang, Joel Burdick, and Yisong Yue. Stagewise safe bayesian optimization with gaussian processes. In *International conference on machine learning*, pages 4781–4789. PMLR, 2018.

Sharan Vaswani, Lin F. Yang, and Csaba Szepesvári. Near-optimal sample complexity bounds for constrained mdps. In *NeurIPS*, November 2022.

Runzhe Wan, Branislav Kveton, and Rui Song. Safe exploration for efficient policy evaluation and comparison. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 22491–22511. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/wan22b.html`.

Lequn Wang, Yiwei Bai, Wen Sun, and Thorsten Joachims. Fairness of exposure in stochastic bandits. In *International Conference on Machine Learning*, pages 10686–10696. PMLR, 2021.

Zhenlin Wang, Andrew J Wagenmaker, and Kevin Jamieson. Best arm identification with safety constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 9114–9146. PMLR, 2022.

Appendix

A Notations

Table 1: Notations

K	:	Number of arms.
δ	:	Confidence parameter.
$d(x, y)$	:	KL-divergence between two random variables with means x and y.
$\mathbf{kl}(x\|\|y)$	:	KL-divergence between two Bernoulli random variables with means x and y.
$\mathcal{D}$	$\triangleq$	$[\boldsymbol{\mu}_{\min}, \boldsymbol{\mu}_{\max}]^K$, i.e. the range of expected rewards
$\boldsymbol{\mu}$	:	True reward vector, $\boldsymbol{\mu} \in \mathcal{D}$.
$\hat{\boldsymbol{\mu}}_t$	:	Empirical means at time t projected onto $\mathcal{D}$.
B	:	Matrix defining the linear constraints, i.e. $B\boldsymbol{\pi} \leqslant \boldsymbol{c}$.
$\boldsymbol{c}$	:	Vector defining the upper bound in the linear constraints, $B\boldsymbol{\pi} \leqslant \boldsymbol{c}$.
Δ_{K-1}	:	Simplex in K dimensions.
$\mathcal{F}$	$\triangleq$	$\{\boldsymbol{\pi} \in \Delta_{K-1} : B\boldsymbol{\pi} \leqslant \boldsymbol{c}\}$, i.e. the constrained policy space.
$\boldsymbol{\pi}$	:	A feasible policy over K arms, i.e. $\boldsymbol{\pi} \in \mathcal{F}$.
$\boldsymbol{\pi}^*$ or $\boldsymbol{\pi}^*_{\boldsymbol{\mu}}$	:	Unique optimal policy for bandit instance $\boldsymbol{\mu}$, defined as $\boldsymbol{\pi}^*_{\boldsymbol{\mu}} \triangleq \boldsymbol{\pi}^* \triangleq \arg\max_{\boldsymbol{\pi} \in \mathcal{F}} \boldsymbol{\mu}^\top \boldsymbol{\pi}$.
$\mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)$	:	Set of extreme points for $\boldsymbol{\pi}'$, which share $K-1$ linearly independent constraints with $\boldsymbol{\pi}^*$.
$\mathcal{N}(\boldsymbol{\pi}^*)$	:	Normal cone spanned by the active constraints at $\boldsymbol{\pi}^*$.
$\Lambda_{\mathcal{F}}(\boldsymbol{\mu})$	$\triangleq$	$\{\boldsymbol{\lambda} \in \mathcal{D} : \max_{\boldsymbol{\pi} \in \mathcal{F}} \boldsymbol{\lambda}^\top \boldsymbol{\pi} > \boldsymbol{\lambda}^\top \boldsymbol{\pi}^*_{\boldsymbol{\mu}}\}$, i.e. the set of alternative bandit instances.
τ	:	Random stopping time of a pure exploration algorithm.
Π	:	Set of possible exploration policies/allocations.
$T_{\mathcal{F}}(\boldsymbol{\mu})^{-1}$	$\triangleq$	$\sup_{\boldsymbol{w} \in \Pi} \inf_{\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{a=1}^K w_a d(\mu_a, \lambda_a)$, the characteristic time for the constrained policy space
$D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$	:	Shorthand for $\inf_{\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{a=1}^K w_a d(\mu_a, \lambda_a)$.
$D(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\lambda})$	:	Shorthand for $\sum_{a=1}^K w_a d(\mu_a, \lambda_a)$.
$w^*(\boldsymbol{\mu})$	:	Set of optimal allocations for bandit instance $\boldsymbol{\mu}$.
H	$\triangleq$	$\frac{2\sigma^2}{\|\boldsymbol{\Delta}\|_2^2}$ quantifies complexity of bandit instance $\boldsymbol{\mu}$

B Lower bound on sample complexity

The following lemma by Kaufmann et al. (2016) provides a general information-theoretic inequality that applies to any bandit model.

Lemma 3 (Kaufmann et al. (2016))**.** *Let $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ be two bandit models with K arms such that μ_a and λ_a are mutually continuous. For any almost surely finite stopping time τ we have*

$$\sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_{a,\tau}] d(\mu_a, \lambda_a) \geqslant \mathbf{kl}(P_{\boldsymbol{\mu}}(\mathcal{E}) || P_{\boldsymbol{\lambda}}(\mathcal{E})) \tag{11}$$

where $\mathcal{E}$ is any measurable event with respect to the filtration generated by the observed history.

From Lemma 3 we can directly derive a lower bound on the expected stopping time of any δ-PAC algorithm in the constraint multi-armed bandit setting. We present this lower bound in Theorem 4 and the proof is virtually the same as the proof for the lower bound in Garivier and Kaufmann (2016). We present it here for completeness.

Theorem 4 (Lower bound on sample complexity under constraints)**.** *The stopping time τ of any δ-PAC learner satisfy*

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geqslant T_{\mathcal{F}}(\boldsymbol{\mu}) \mathbf{kl}(\delta || 1-\delta). \tag{12}$$

Proof. Let $\boldsymbol{\mu}$ and $\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})$ be two bandit models with K arms such that they do not share optimal policy, i.e. $\boldsymbol{\pi}^*_{\boldsymbol{\mu}} \neq \boldsymbol{\pi}^*_{\boldsymbol{\lambda}}$.

Let $\mathcal{E}$ denote the event of recommending $\boldsymbol{\pi}^*_{\boldsymbol{\mu}}$ for any bandit instance at stopping using some δ-PAC algorithm. Then using Lemma 3, and δ-correctness of $\boldsymbol{\pi}^*_{\boldsymbol{\mu}}$ for $\boldsymbol{\mu}$, we have

$$\sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_{a,\tau}] d(\mu_a, \lambda_a) \geqslant \mathbf{kl}(1-\delta || \delta) = \mathbf{kl}(\delta || 1-\delta).$$

Further, we multiple and divide by $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$ which yields

$$\begin{aligned}\sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_{a,\tau}] d(\mu_a, \lambda_a) &= \mathbb{E}_{\boldsymbol{\mu}}[\tau] \sum_{a=1}^{K} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_{a,\tau}]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau]} d(\mu_a, \lambda_a) \\ &= \mathbb{E}_{\boldsymbol{\mu}}[\tau] \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) \geqslant \mathbf{kl}(\delta || 1-\delta)\ ,\end{aligned}$$

where $w_a \triangleq \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_{a,\tau}]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}$, and $\sum_{a=1}^{K} w_a = 1$.

Since the above inequality is true for any $\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})$, we have

$$\inf_{\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \mathbb{E}_{\boldsymbol{\mu}}[\tau] \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) = \mathbb{E}_{\boldsymbol{\mu}}[\tau] \inf_{\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) \geqslant \mathbf{kl}(\delta || 1-\delta)\ .$$

The equality is due to the fact that $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$ is independent of $\boldsymbol{\lambda}$.

Now, we further maximise over w_a to get

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \sup_{\boldsymbol{w} \in \Pi} \inf_{\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) \geqslant \mathbf{kl}(\delta || 1-\delta).$$

Finally, using the definition of the characteristic time $T_{\mathcal{F}}(\boldsymbol{\mu})$ yields

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geqslant T_{\mathcal{F}}(\boldsymbol{\mu}) \mathbf{kl}(\delta || 1-\delta)\ .$$

□

B.1 Projection lemma for $D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$: Proof of Lemma 1

To derive the key properties of the optimal solution and the set of optimal allocations, as presented in Lemma 1, we first explicate the set of optimal solutions, and then, use Berge's theorem (Theorem 5).

Step 1: Recall that

$$\Lambda_{\mathcal{F}}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} \in \mathcal{D} : \boldsymbol{\lambda} \notin \mathcal{N}(\boldsymbol{\pi}^*)\} \ ,$$

where the normal cone is expressed as

$$\mathcal{N}(\boldsymbol{\pi}^*) = \bigcap_{\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \left\{\boldsymbol{\lambda} \in \mathcal{D} : \boldsymbol{\lambda}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}') \geqslant 0\right\} \ .$$

This is due to the fact that if $\boldsymbol{\pi}^*$ is not the optimal policy under the environment $\boldsymbol{\lambda}$, there exists an improving direction in the simplex algorithm, i.e. a neighbor $\boldsymbol{\pi}'$, such that $\boldsymbol{\lambda}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}') < 0$.

Now, since the set of alternative hypotheses is the compliment of the normal cone, we write

$$\Lambda_{\mathcal{F}}(\boldsymbol{\mu}) = \bigcup_{\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \left\{\boldsymbol{\lambda} : \boldsymbol{\lambda}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}') < 0\right\}. \tag{13}$$

Applying Equation (13) in $D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$ leads to,

$$D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F}) = \inf_{\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) = \min_{\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \inf_{\boldsymbol{\lambda} : \boldsymbol{\lambda}^\top(\boldsymbol{\pi}' - \boldsymbol{\pi}^*) < 0} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) \ .$$

Step 2: What remains to be shown is that the $\inf$ is attained by some λ on $\boldsymbol{\lambda}^\top(\boldsymbol{\pi}' - \boldsymbol{\pi}^*) = 0$.

For some $\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)$ take an arbitrary $\boldsymbol{\lambda}' \in \left\{\boldsymbol{\lambda} : \boldsymbol{\lambda}^\top(\boldsymbol{\pi}' - \boldsymbol{\pi}^*) < 0\right\}$. There exists an $\boldsymbol{\lambda}'' \in \left\{\boldsymbol{\lambda} \in \mathcal{D} : \boldsymbol{\lambda}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}') = 0\right\}$ such that $|\mu_a - \lambda'_a| \geqslant |\mu_a - \lambda''_a| \ \forall a$ due to the convexity of $\mathcal{D}$. The mapping $y \to d(x, y)$ is an increasing function on the domain $y > x$ and a decreasing function on $y < x$ which implies that

$$\sum_{a=1}^{K} w_a d(\mu_a, \lambda'_a) \geqslant \sum_{a=1}^{K} w_a d(\mu_a, \lambda''_a). \tag{14}$$

There exists a sequence $\{\boldsymbol{\lambda}_t\}_{t=1}^{\infty} \subset \left\{\boldsymbol{\lambda} : \boldsymbol{\lambda}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}') < 0\right\}$ such that $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}'$ and $\lim_{t\to\infty} \boldsymbol{\lambda}_t = \boldsymbol{\lambda}''$. Hence, we can for any $\boldsymbol{\lambda}'$ get arbitrary close to some $\boldsymbol{\lambda}''$ such that Equation (14) holds.

Due to continuity of $d(x, .)$, the $\inf$ is attained by some $\boldsymbol{\lambda}'' \in \left\{\boldsymbol{\lambda} \in \mathcal{D} : \boldsymbol{\lambda}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}') = 0\right\}$. Hence, we conclude the proof.

B.2 Properties of $D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$: Proof of Theorem 1

<u>Property (a-b).</u> We first note that the function $D(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \triangleq \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a)$ is continuous in all elements. Take any $(\boldsymbol{w}, \boldsymbol{\mu})$ such that the optimal policy in $\mathcal{F}$ is unique. Let $(\boldsymbol{w}_t, \boldsymbol{\mu}_t)_{t\geqslant 1}$ be a sequence in $\Pi \times \mathcal{D}$ such that

$$(\boldsymbol{w}_t, \boldsymbol{\mu}_t) \xrightarrow{t\to\infty} (\boldsymbol{w}, \boldsymbol{\mu}).$$

Further, for any $\epsilon > 0$ there exists a $t' \geqslant 1$ such that $||(\boldsymbol{w}, \boldsymbol{\mu}) - (\boldsymbol{w}_t, \boldsymbol{\mu}_t)||_2 < \epsilon$ and $\Lambda_{\mathcal{F}}(\mu) = \Lambda_{\mathcal{F}}(\mu_t)$ $\forall t \geqslant t'$. By continuity of $D(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ we have that for any $\epsilon' > 0$ there exists exists an $t'' \geqslant 1$ such that for $t \geqslant t''$, we have

$$|D(\boldsymbol{w}_t, \boldsymbol{\mu}_t, \boldsymbol{\lambda}) - D(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\lambda})| \leqslant \epsilon', \forall \lambda \in \mathbb{R}^K.$$

Thus, by taking $t \geqslant t', t''$ leads to

$$|D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F}) - D(\boldsymbol{w}_t, \boldsymbol{\mu}_t, \mathcal{F})| = \left|\inf_{\lambda \in \Lambda_{\mathcal{F}}(\mu)} D(\boldsymbol{w}, \boldsymbol{\mu}, \lambda) - \inf_{\lambda \in \Lambda_{\mathcal{F}}(\mu_t)} D(\boldsymbol{w}_t, \boldsymbol{\mu}_t, \lambda)\right|$$

$$\leqslant \left| \inf_{\lambda \in \Lambda_{\mathcal{F}}(\mu)} \left(D(\boldsymbol{w}, \boldsymbol{\mu}, \lambda) - D(\boldsymbol{w}_t, \boldsymbol{\mu}_t, \lambda) \right) \right|$$
$$\leqslant \epsilon' ,$$

which establishes the continuity properties.

Property (c). The upper hemicontinuity of $w^*(\boldsymbol{\mu})$ and continuity of $D(\boldsymbol{\mu}, \mathcal{F})$ follows from Berge's maximum theorem, see Theorem 5, by letting $f(x, \theta) = D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$ and $C(\theta) = \Pi$. As a consequence of Berge's theorem (Theorem 5), we substitute the $\sup_w$ with $\max_w$.

Property (d). The convexity of the set $w^*(\boldsymbol{\mu})$ follows from the fact that it is the set of optimal solutions to $\max_{w \in \Pi} D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$ and $D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F})$ is concave (Specifically, it is linear in $\boldsymbol{w}$).

B.3 Projective representation of characteristic time for Gaussians: Proof of Proposition 1

For two bandit instances $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ consisting of Gaussian distributions with same variance σ^2, we have

$$D(\boldsymbol{w}, \boldsymbol{\mu}, \mathcal{F}) = \min_{\boldsymbol{\lambda} : \boldsymbol{\lambda}^\top (\boldsymbol{\pi}^* - \boldsymbol{\pi}') = 0} \sum_{a=1}^{K} w_a \frac{1}{2\sigma^2} (\mu_a - \lambda_a)^2 .$$

Now, by introducing the Lagrange multiplier γ, we obtain

$$L(\gamma, \boldsymbol{\lambda}) \triangleq \frac{1}{2\sigma^2} \sum_{a=1}^{K} w_a (\mu_a - \lambda_a)^2 - \gamma \boldsymbol{\lambda}^\top (\boldsymbol{\pi}^* - \boldsymbol{\pi}'). \tag{15}$$

For brevity, we denote $v \triangleq (\boldsymbol{\pi}^* - \boldsymbol{\pi}')$.

Computing the gradient $\nabla_\lambda L(\gamma, \boldsymbol{\lambda})$ and equating it to 0 yields

$$\lambda_a = \mu_a + \frac{\gamma \sigma^2}{w_a} v_a .$$

Substituting λ_a in Equation (15) yields

$$\begin{aligned} L(\gamma) = \min_{\lambda} L(\gamma, \boldsymbol{\lambda}) &= \frac{\sigma^2 \gamma^2}{2} \sum_{a=1}^{K} \frac{v_a^2}{w_a} - \gamma \boldsymbol{\mu}^\top v - \sum_{a=1}^{K} \frac{\gamma^2 \sigma^2}{w_a} v_a^2 \\ &= -\frac{\sigma^2 \gamma^2}{2} \sum_{a=1}^{K} \frac{v_a^2}{w_a} - \gamma \boldsymbol{\mu}^\top v . \end{aligned} \tag{16}$$

Minimizing over γ yields

$$\gamma = \frac{-\boldsymbol{\mu}^\top v}{\sigma^2 \sum_a \frac{v_a^2}{w_a}} ,$$

and putting it back in Equation (16) gives the final expression of λ_a

$$\lambda_a = \mu_a - \frac{v_a}{w_a} \left(\frac{\boldsymbol{\mu}^\top v}{\sum_a \frac{v_a^2}{w_a}} \right) . \tag{17}$$

B.4 Upper and lower bounding the characteristic time: Proof of Proposition 2

Lower bound on the characteristic time: To lower bound $T_{\mathcal{F}}(\boldsymbol{\mu})$, we need to upper bound the RHS in Equation (5), i.e. $T_{\mathcal{F}}(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{w}} \min_{\lambda} \sum_a w_a d(\mu_a, \lambda_a)$.

Step 1: We first observe that

$$\sup_{\boldsymbol{w}} \min_{\lambda} \sum_a w_a d(\mu_a, \lambda_a) = \max_{\boldsymbol{w}} \min_{\lambda} \sum_a w_a d(\mu_a, \lambda_a),$$

due to Berge's theorem. Further, the max-min inequality gives

$$\max_{\boldsymbol{w}} \min_{\lambda} \sum_a w_a d(\mu_a, \lambda_a) \leqslant \min_{\lambda} \max_{\boldsymbol{w}} \sum_a w_a d(\mu_a, \lambda_a).$$

Step 2: We proceed to upper bound $\max_{\boldsymbol{w}} \sum_a w_a d(\mu_a, \lambda_a)$ for each neighbor $\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)$ independently.

For a fixed $\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)$, Proposition 1 tells us that

$$\begin{aligned}
\min_{\lambda : \lambda^\top (\boldsymbol{\pi}^* - \boldsymbol{\pi}') = 0} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) &= \frac{\gamma^2}{2\sigma^2} \sum_{a=1}^{K} \frac{(\boldsymbol{\pi}^* - \boldsymbol{\pi}')_a^2}{w_a} \\
&= \left(\frac{\boldsymbol{\mu}^\top (\boldsymbol{\pi}^* - \boldsymbol{\pi}')}{\sum_a \frac{(\boldsymbol{\pi}^* - \boldsymbol{\pi}')^2}{w_a}} \right)^2 \frac{1}{2\sigma^2} \sum_{a=1}^{K} \frac{(\boldsymbol{\pi}^* - \boldsymbol{\pi}')_a^2}{w_a} \\
&= \frac{1}{2\sigma^2} \frac{\left(\boldsymbol{\mu}^\top (\boldsymbol{\pi}^* - \boldsymbol{\pi}')\right)^2}{\sum_{a=1}^{K} \frac{(\boldsymbol{\pi}^* - \boldsymbol{\pi}')_a^2}{w_a}}.
\end{aligned}$$

Step 3: We further minimize the expression $\frac{(\boldsymbol{\pi}^* - \boldsymbol{\pi}')_a^2}{w_a}$ under the constraint $\sum_a w_a = 1$.

Using Langrange multiplier technique, we get

$$w_a = \frac{|(\boldsymbol{\pi}^* - \boldsymbol{\pi}')|_a}{\sum_{a=1}^{K} |(\boldsymbol{\pi}^* - \boldsymbol{\pi}')|_a}$$

which yields that $\frac{(\boldsymbol{\pi}^* - \boldsymbol{\pi}')_a^2}{w_a} \geqslant \|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|_1^2$. Hence,

$$\frac{1}{2\sigma^2} \frac{\left(\boldsymbol{\mu}^\top (\boldsymbol{\pi}^* - \boldsymbol{\pi}')\right)^2}{\sum_{a=1}^{K} \frac{(\boldsymbol{\pi}^* - \boldsymbol{\pi}')_a^2}{w_a}} \leqslant \frac{1}{2\sigma^2} \frac{\left(\boldsymbol{\mu}^\top (\boldsymbol{\pi}^* - \boldsymbol{\pi}')\right)^2}{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|_1^2} \leqslant \frac{1}{2\sigma^2} \frac{\left(\boldsymbol{\mu}^\top (\boldsymbol{\pi}^* - \boldsymbol{\pi}')\right)^2}{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|_2^2}.$$

Here, the last part is exactly , $\frac{1}{2} d_{\boldsymbol{\pi}'}^2$, i.e. the squared distance between $\boldsymbol{\mu}$ and the hyperplane $\boldsymbol{\pi}^* - \boldsymbol{\pi} = 0$.

Thus, we conclude the lower bound.

Upper bound on the characteristic time: To obtain the upper bound, we aim to lower bound the inverse $T_{\mathcal{F}}(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{w}} \min_{\lambda} \sum_a w_a d(\mu_a, \lambda_a)$.

We let $w_a = \frac{1}{K}, \forall a$, and observe that

$$\max_{w} \min_{\lambda} \sum_a w_a d(\mu_a, \lambda_a) \geqslant \min_{\lambda} \frac{1}{K} \sum_a d(\mu_a, \lambda_a).$$

For some $\boldsymbol{\pi}' \in \mathcal{V}(\boldsymbol{\pi}^*)$ and using Proposition 1 with $w_a = \frac{1}{K}, \forall a$, we get

$$\frac{1}{K} \sum_a d(\mu_a, \lambda_a) = \frac{1}{2\sigma^2 K} \frac{\left(\boldsymbol{\mu}^\top (\boldsymbol{\pi}^* - \boldsymbol{\pi}')\right)^2}{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|_2^2} = d_{\boldsymbol{\pi}'}^2 \frac{1}{4\sigma^2 K}$$

This concludes the upper bound on the characteristic time.

B.5 Impact of linear constraints: Proof of Corollary 1

Step 1: Neighboring policies and rank-1 update. For any $\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)$, let $\hat{B} \in \mathbb{R}^{K\times K}$ be a set of linearly independent constraints at $\boldsymbol{\pi}^*$ and $\hat{\boldsymbol{c}}$ be the corresponding values in $\boldsymbol{c}$ such that $\boldsymbol{\pi}' = B'^{-1}\boldsymbol{c}'$. Specifically, B' and $\boldsymbol{c}'$ can be retrieved from the following rank-1 updates

$$B' = \hat{B} + \boldsymbol{e}_r(\boldsymbol{b}'_r - \hat{\boldsymbol{b}}_r)^\top \ ,$$
$$\boldsymbol{c}' = \widehat{\boldsymbol{c}} + (c'_r - c_r)\boldsymbol{e}_r \ ,$$

where $\hat{\boldsymbol{b}}_r$ a column vector corresponding to the constraint on the r-th row of $\hat{B}$ that we swap with $\boldsymbol{b}'_r$ in order to get B' and $\boldsymbol{e}_r$ a column vector with all elements equal to 0 except the r-th element which is equal to 0. Similarly, $(c'_r - c_r) \neq 0$ is the change that we perform on the r-th element in $\hat{\boldsymbol{c}}$ to get $\boldsymbol{c}'$.

Step 2: From perturbation in constraints to perturbations in policies. Now, we observe that

$$B'\boldsymbol{\pi}' - \hat{B}\boldsymbol{\pi}^* = (c'_r - c_r)\boldsymbol{e_r} \ .$$

Since $\hat{B}$ is invertible, further rearrangement yields

$$\begin{aligned}
\boldsymbol{\pi}' - \boldsymbol{\pi}^* &= \hat{B}^{-1}\left((c'_r - c_r)\boldsymbol{e}_r + \boldsymbol{e}_r(\hat{\boldsymbol{b}}_r - \boldsymbol{b}'_r)^\top\boldsymbol{\pi}'\right) \\
&= \hat{B}^{-1}\left((c'_r - c_r)\boldsymbol{e}_r + \boldsymbol{e}_r\hat{\boldsymbol{b}}_r^\top\boldsymbol{\pi}' - \boldsymbol{e}_r\boldsymbol{b}_r'^\top\boldsymbol{\pi}'\right) \\
&= \hat{B}^{-1}\left((c'_r - c_r)\boldsymbol{e}_r + \boldsymbol{e}_r\hat{\boldsymbol{b}}_r^\top\boldsymbol{\pi}' - c'_r\boldsymbol{e}_r\right) \\
&= \hat{B}^{-1}\left((\hat{\boldsymbol{b}}_r^\top\boldsymbol{\pi}' - c_r)\boldsymbol{e}_r\right)
\end{aligned}$$

The last part is the slack of $\boldsymbol{\pi}'$ at the r-th constraint in $\hat{B}$, hereby referred to as ξ.

We bound the norm of $\hat{B}^{-1}\boldsymbol{e}_r$ as follows

$$\sigma_{\min}(\hat{B}^{-1}) = \inf_{\boldsymbol{v}:\|v\|_2=1} \|\hat{B}^{-1}v\|_2 \leqslant \|\hat{B}^{-1}\boldsymbol{e}_r\|_2 \leqslant \sup_{\boldsymbol{v}:\|v\|_2=1} \|\hat{B}^{-1}v\|_2 = \sigma_{\max}(\hat{B}^{-1})$$

where $\sigma_{\min}(\hat{B}^{-1})$ and $\sigma_{\max}(\hat{B}^{-1})$ denote the smallest and largest singular value of $\hat{B}^{-1}$. From the properties of the inverse, we get

$$\frac{1}{\sigma_{\max(\hat{B})}} \leqslant \|\hat{B}^{-1}\boldsymbol{e}_r\|_2 \leqslant \frac{1}{\sigma_{\min(\hat{B})}}.$$

Thus, we obtain a lower and upper bound on the perturbation in policies

$$\frac{|\xi|}{\sigma_{\max(\hat{B})}} \leqslant \|\boldsymbol{\pi}' - \boldsymbol{\pi}^*\|_2 \leqslant \frac{|\xi|}{\sigma_{\min(\hat{B})}}. \tag{18}$$

Step 3: Bounds on characteristic time from perturbation in policies. From Proposition 2 we have

$$\begin{aligned}
T_{\mathcal{F}}(\boldsymbol{\mu})^{-1} = \frac{1}{2\sigma^2}\frac{\left(\boldsymbol{\mu}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}')\right)^2}{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|_2^2} &= \frac{1}{2\sigma^2}\frac{\left(\boldsymbol{\mu}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}') - \mu^*\mathbf{1}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}')\right)^2}{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|_2^2} \\
&= \frac{1}{2\sigma^2}\frac{(\boldsymbol{\mu} - \mu^*\mathbf{1})^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}')^2}{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}'\|_2^2} \\
&= \frac{1}{2\sigma^2}\frac{\left(\boldsymbol{\Delta}^\top\hat{B}^{-1}\boldsymbol{e}_r\right)^2}{\|\hat{B}^{-1}\boldsymbol{e}_r\|_2^2} \\
&\leqslant \frac{\|\boldsymbol{\Delta}\|_2^2}{2\sigma^2}\frac{\sigma_{\max}^2(\hat{B})}{\sigma_{\min}^2(\hat{B})} \ .
\end{aligned}$$

Step 4: Concluding with complexity of bandit instance and constraints. By referring to $\kappa(\hat{B}) \triangleq \frac{\sigma_{\max}(\hat{B})}{\sigma_{\min}(\hat{B})}$ as the condition number of $\hat{B}$, and $H \triangleq \frac{2\sigma^2}{\|\boldsymbol{\Delta}\|_2^2}$ as the quantifier complexity of bandit instance $\boldsymbol{\mu}$, we get

$$T_{\mathcal{F}}(\boldsymbol{\mu})^{-1} \leqslant \min_{\boldsymbol{\pi}'\in\mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \frac{\kappa^2(\hat{B})}{H} \ .$$

Hence, for any $\boldsymbol{\mu}$, we have that $T_{\mathcal{F}}(\boldsymbol{\mu}) \geqslant \frac{H}{\kappa^2}$, where κ^2 is the minimum condition number of any sub-matrix $\hat{B} \in \mathbb{R}^{K \times K}$ of B consisting of K linearly independent active constraints at $\boldsymbol{\pi}^*$. This leads to a lower bound

$$\mathbb{E}[\tau] \geqslant \Omega\left(\frac{H}{\kappa^2}\mathbf{kl}(\delta||1-\delta)\right).$$

C Upper bounds on sample complexity

C.1 Stopping criterion

Lemma 4 (Magureanu et al. (2014))**.** *$\forall \gamma > K+1$ and $t \in \mathbb{N}$ it holds*

$$P\left(\sum_{a=1}^{K} N_{a,t} d(\hat{\mu}_a, \mu_a) \geqslant \gamma\right) \leqslant e^{-\gamma} \left(\frac{\lceil \gamma \log t \rceil \gamma}{K}\right)^K e^{K+1}$$

The correctness of our stopping rule in Equation (10) follows easily from Lemma 4. Let $\boldsymbol{\pi}_\tau$ be our recommendation at stopping

$$\begin{aligned} P(\boldsymbol{\pi}_\tau \neq \boldsymbol{\pi}^*) &\leqslant P\left(\exists t \in \mathbb{N} : \sum_{a=1}^{K} N_{a,t} d(\hat{\boldsymbol{\mu}}_{a,t}, \boldsymbol{\mu}_a) \geqslant c(t,\delta)\right) \\ &\leqslant \sum_{t=1}^{\infty} e^{-c(t,\delta)} \left(\frac{\lceil c(t,\delta) \log t \rceil c(t,\delta)}{K}\right)^K e^{K+1}. \end{aligned}$$

We plug in $c(t,\delta) = \log \frac{t^\alpha C}{\delta}$ and choose C such that

$$\sum_{t=1}^{\infty} \left(\frac{\lceil c(t,\delta) \log t \rceil c(t,\delta)}{K}\right)^K e^{K+1} \leqslant C$$

which yields

$$P(\boldsymbol{\pi}_\tau \neq \boldsymbol{\pi}^*) \leqslant \delta.$$

C.2 Upper bound for CTnS

Proof of Theorem 2.
Step 1: Defining Good Event. Let $T \in \mathbb{N}$. For $\epsilon > 0$ and $h(T) = \sqrt{T}$, let $\mathcal{E}_T$ be the event

$$\mathcal{E}_T \triangleq \bigcap_{t=h(T)}^{T} \left\{ \|\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}\|_\infty \leqslant \xi(\epsilon) \right\},$$

where $\xi(\epsilon) < \max_{\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \frac{1}{4\sqrt{K}} \boldsymbol{\mu}^\top \left(\boldsymbol{\pi}^* - \boldsymbol{\pi}'\right)$ is such that

$$\|\boldsymbol{\mu}' - \boldsymbol{\mu}\|_\infty \leqslant \xi(\epsilon) \implies \sup_{\boldsymbol{w}' \in w^*(\boldsymbol{\mu}')} \sup_{\boldsymbol{w} \in w^*(\boldsymbol{\mu})} \|\boldsymbol{w}' - \boldsymbol{w}\| \leqslant \epsilon$$

This $\xi(\epsilon)$ exists due to the upper hemicontinuity of $w^*(\boldsymbol{\mu})$, Theorem 1.

Step 2: Concentrating to Good Event. We will make use of the following Lemma from Garivier and Kaufmann (2016) which bounds the probability of the compliment $\mathcal{E}_T^c$.

Lemma 5 (Concentration around means (Garivier and Kaufmann, 2016))**.** *There exist two constants B, C such that*

$$P(\mathcal{E}_T^c) \leqslant BT \exp\left(-CT^{\frac{1}{8}}\right)$$

This Lemma is due to the fact that C-tracking ensure that each arm has been played at least $\sqrt{t}$ times at each time t, see next Lemma.

Lemma 6 (Garivier and Kaufmann (2016))**.** *For all $t \geqslant 1$ and $\forall a$, C-Tracking ensures $N_{a,t} \geqslant \sqrt{t + K^2} - K$ and*

$$\max_a \left| N_{a,t} - \sum_{s=1}^{t} \boldsymbol{w}_{a,s} \right| \leqslant K(1 + \sqrt{t}) \tag{19}$$

We now leverage to following tracking Lemma of Degenne and Koolen (2019) which holds whenever we are tracking a set of optimal weights.

Lemma 7 (Concentration in allocations (Degenne and Koolen, 2019))**.** *Under $\mathcal{E}_T$, there exists a T_ϵ such that for T where $h(T) \geqslant T_\epsilon$ C-tracking will satisfy*

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

This shows that C-tracking is eventually going to produce an empirical distribution of plays that is close to an optimal allocation and the empirical distribution will converge to a point in $w^*(\boldsymbol{\mu})$ as $t \to \infty$. We need Lemma 7 instead of the original tracking result in Garivier and Kaufmann (2016) since the optimal allocation does not need to be unique. However, we know from Theorem 1 that the set of optimal allocations $w^*(\boldsymbol{\mu})$ is convex and we can thus apply Lemma 7.

There exists a T_ϵ such that under $\mathcal{E}_T$ and $t \geqslant \max(T_\epsilon, h(T))$ we have

$$|(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_t)^\top \boldsymbol{\pi}^*| \leqslant \sqrt{K}\xi < \frac{1}{4} \max_{\boldsymbol{\pi}' \in \mathcal{V}_{\mathcal{F}}(\boldsymbol{\pi}^*)} \boldsymbol{\mu}^\top \left(\boldsymbol{\pi}^* - \boldsymbol{\pi}'\right)$$

which implies that $\boldsymbol{\pi}^* = \arg\max_{\boldsymbol{\pi} \in \mathcal{F}} \hat{\boldsymbol{\mu}}_t^\top \boldsymbol{\pi}$. This ensures that we will be computing the stopping criterion w.r.t. to the correct Alt-set $\Lambda_{\mathcal{F}}(\boldsymbol{\mu})$.

Step 3: Complexity given the Good Event. Assume $T \geqslant T_\epsilon$ and let

$$C_{\epsilon,\mathcal{F}}(\boldsymbol{\mu}) \triangleq \inf_{\substack{\boldsymbol{\mu}' : \|\boldsymbol{\mu}' - \boldsymbol{\mu}\|_\infty \leqslant \xi(\epsilon) \\ \boldsymbol{w}' : \|\boldsymbol{w}' - \boldsymbol{w}\|_\infty \leqslant 3\epsilon, \forall \boldsymbol{w} \in w^*(\boldsymbol{\mu})}} D(\boldsymbol{w}', \boldsymbol{\mu}', \mathcal{F}).$$

This $C_{\epsilon,\mathcal{F}}(\boldsymbol{\mu})$ gives the worst-case characteristic time we might compute in the algorithm due to the fact that our estimates are not exact.

Assuming $\mathcal{E}_T$, Lemma 7 gives for $t \geqslant T_\epsilon$

$$D(\boldsymbol{N}_t, \hat{\boldsymbol{\mu}}_t, \mathcal{F}) \geqslant t C_{\epsilon,\mathcal{F}}(\boldsymbol{\mu}).$$

Step 4: Bounding the Stopping Time for Good and Bad Events. Let τ_δ be the stopping time, then

$$\min(\tau_\delta, T) \leqslant \sqrt{T} + \sum_{t=T_\epsilon}^{T} \mathbb{I}_{\tau_\delta > t}$$

and plugging in our stopping rule, i.e. $D(\boldsymbol{N}_t, \hat{\boldsymbol{\mu}}_t, \mathcal{F}) > c(t,\delta)$ yields

$$\begin{aligned} T_\epsilon + \sum_{t=T_\epsilon}^{T} \mathbb{I}(D(\boldsymbol{N}_t, \hat{\boldsymbol{\mu}}_t, \mathcal{F}) \leqslant c(t,\delta)) &\leqslant \sqrt{T} + \sum_{t=T_\epsilon}^{T} \mathbb{I}(t C_{\epsilon,\mathcal{F}}(\boldsymbol{\mu}) \leqslant c(T,\delta)) \\ &\leqslant \sqrt{T} + \frac{c(T,\delta)}{C_{\epsilon,\mathcal{F}}(\boldsymbol{\mu})}. \end{aligned}$$

We define $T_\delta := \inf\left\{T \in \mathbb{N} : \sqrt{T} + \frac{c(T,\delta)}{C_{\epsilon,\mathcal{F}}(\boldsymbol{\mu})} \leqslant T\right\}$. Hence,

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

where $\sum_{t=1}^{\infty} BT \exp\left(-Ct^{\frac{1}{8}}\right) \leqslant T' < \infty$. We bound T_δ in the same way as Garivier and Kaufmann (2016). Let $C(\eta) = \inf\{T : T - \sqrt{T} \geqslant T\frac{1}{1+\eta}\}$ for some $\eta > 0$. Then

$$T_\delta \leqslant C(\eta) + \inf\left\{T \in \mathbb{N} : T\frac{C_{\epsilon,\mathcal{F}}(\boldsymbol{\mu})}{1+\eta} \geqslant c(T,\delta)\right\}.$$

Step 5: Obtaining the Asymptotic Bound. Dividing Equation 20 with $\log\frac{1}{\delta}$ and taking the limit yields

$$\lim_{\delta\to 0} \inf \frac{\mathbb{E}[\tau_\delta]}{\log\frac{1}{\delta}} \leqslant \frac{\alpha(1+\eta)}{C_{\epsilon,\mathcal{F}}(\boldsymbol{\mu})}.$$

$C_{\epsilon,\mathcal{F}}(\boldsymbol{\mu})$ is continuous due to Theorem 1 and taking the limits $\eta, \epsilon \to 0$ yields

$$\lim_{\delta\to 0} \inf \frac{\mathbb{E}[\tau_\delta]}{\log\frac{1}{\delta}} \leqslant \alpha T_{\mathcal{F}}(\boldsymbol{\mu}), \forall \alpha > 1.$$

□

C.3 Upper bound for CGE

The proof follows the same structure as the proof of Theorem 2 in Degenne et al. (2019) and we use the same concentration analysis. The main difference is that we have to adjust the definition of approximate optimistic saddle point algorithm.

Proof of Theorem 3.
Step 1: Defining Good Event. We start by defining the good event

$$\mathcal{E}_T \triangleq \{\forall t \leqslant T\, \forall a, N_{a,t} d(\hat{\mu}_{a,t}, \mu_t) \leqslant f(t)\}$$

where $f(t) = 3\log t + \log\log t$.

Step 2: Concentration of Good Event.

We can bound $\sum_{t=1}^{\infty} P(\mathcal{E}_T^c)$ using Lemma 4. Hence, for any $t \in \mathbb{N}$ and arm a

$$\begin{aligned} P(N_{a,t} d(\hat{\mu}_{a,t}, \mu_t) \geqslant f(t)) &\leqslant e^{-f(t)}(1 + f(t)\log t) f(t) \\ &= \frac{e^2}{t^3 \log t}(f(t) + f(t)^2 \log t). \end{aligned}$$

Summing yields

$$\sum_{a=1}^{K}\sum_{t=1}^{\infty} P(\mathcal{E}_T^c) \leqslant K + K\sum_{t=2}^{\infty} \frac{e^2}{t^3 \log t}\left(f(t) + f(t)^2 \log t\right) \leqslant KC < \infty. \tag{20}$$

Here a constant $C = 21$ is sufficient.

Step 3: Starting from the Stopping Criterion The main idea of the proof is to work with the stopping criterion

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

and show that if we have the event $\mathcal{E}_T$, our current recommendation at some t is the correct policy $\boldsymbol{\pi}^*$ and we haven't stopped yet, we can lower bound $c(t,\delta)$ in a way that depends on the characteristic time and properties of the no-regret learners. We start with assuming our current recommendation at some t is the correct policy $\boldsymbol{\pi}^*$ and we have the event $\mathcal{E}_T$,

$$c(t,\delta) \geqslant \inf_{\lambda \in \Lambda_{\mathcal{F}}(\hat{\boldsymbol{\mu}}_t)} \sum_{s=1}^{t}\sum_{a=1}^{K} w_{a,s} d(\hat{\mu}_{a,t}, \lambda_a) - (1 + \sqrt{t})K$$

which follows from Tracking Lemma 6. We now use a concentration result, originally in Appendix D.1 of Degenne et al. (2019),

$$c(t,\delta) \geqslant \inf_{\lambda \in \Lambda_{\mathcal{F}}(\hat{\boldsymbol{\mu}}_t)} \sum_{s=1}^{t}\sum_{a=1}^{K} w_{a,s} d(\hat{\mu}_{a,s}, \lambda_a) - (1 + \sqrt{t})K - O(\sqrt{t\log t}). \tag{21}$$

This steps follows from the Lipschitz property of the KL and the fact we have conditioned on $\mathcal{E}_T$ (see Step 8 for further details). Hence,

$$|d(\mu_a, \lambda_a) - d(\hat{\mu}_{a,s}, \lambda_a)| \leqslant L\sqrt{2\sigma^2 \frac{f(s)}{N_{a,s}}}$$

which implies that

$$\sum_{s=1}^{t}\sum_{a=1}^{K} w_{a,s} d(\hat{\mu}_{a,t}, \lambda_a) \geqslant \sum_{s=1}^{t}\sum_{a=1}^{K} w_{a,s} d(\mu_a, \lambda_a) - L\sqrt{2\sigma^2 K t f(t)}.$$

Using the same result one more time yields

$$\sum_{s=1}^{t}\sum_{a=1}^{K} w_{a,s} d(\hat{\mu}_{a,t}, \lambda_a) \geqslant \sum_{s=1}^{t}\sum_{a=1}^{K} w_{a,s} d(\hat{\mu}_{a,s}, \lambda_a) - L\sqrt{2\sigma^2 K t f(t)} - 2L\sqrt{2\sigma^2 f(t)}\left(K^2 + 2\sqrt{2Kt}\right)$$

which gives the result in Equation (21).

Step 4: Defining Approximate Optimistic Saddle Point under Constraints. We now introduce concepts and properties that will help us to further lower bound the RHS in Equation (21). We extend the definition of an *approximate optimistic saddle point algorithm* from Degenne et al. (2019) to the constraint setting.

Definition 1. *An algorithm playing sequences of* $(\boldsymbol{w}_s, \boldsymbol{\lambda}_s)_{s \leqslant t} \in (\Pi \times \Lambda_{\mathcal{F}})^t$ *is said to be an* approximate optimistic saddle point algorithm *with slack* x_t *if*

$$\inf_{\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{s=1}^{t} \sum_{a=1}^{K} w_{s,a} d(\hat{\mu}_{a,s}, \lambda_a) \geqslant \max_{\boldsymbol{w} \in \Pi} \sum_{a=1}^{K} \sum_{s=1}^{t} w_a U_{a,s} - x_t, \tag{22}$$

where x_t *is defined in Eq.* (25) *and the confidence bound*

$$U_{a,s} = \max \left\{ \frac{f(t)}{N_{a,s}}, \max_{\xi \in [\alpha_{a,s}, \beta_{a,s}]} d(\xi, , \lambda_{a,s}) \right\}.$$

The difference in Definition 1 compared to the definition of an approximate optimistic saddle point algorithm in Degenne et al. (2019) is that we in Equation 22 take the maximum over Π and instead of arms as in Degenne et al. (2019). This is due to the fact that maximum over arms might not be in the set of feasible exploration policies Π.

Step 5: Definition of Regret of the Two Players. We define the regret of the allocation player, i.e. AdaGrad, as

$$R_t^{\boldsymbol{w}} = \max_{\boldsymbol{w} \in \Pi} \sum_{s=1}^{t} \sum_{a=1}^{K} w_a U_{a,s} - \sum_{s=1}^{t} \sum_{a=1}^{K} w_{a,t} U_{a,s} \tag{23}$$

and note that AdaGrad has an regret scaling of $R_{\boldsymbol{w}}^t \leqslant O(\sqrt{Qt})$ where Q is an upper bound on the losses such that $Q \geqslant \max_{x,y \in [\mu_{\min}, \mu_{\max}]} d(x, y)$. For the instance player we define the regret as

$$R_t^{\boldsymbol{\lambda}} = \sum_{s=1}^{t} \sum_{a=1}^{K} w_{a,s} d(\hat{\mu}_{a,s}, \lambda_{a,s}) - \inf_{\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{s=1}^{t} \sum_{a=1}^{K} w_{a,s} d(\hat{\mu}_{a,s}, \lambda_a) \tag{24}$$

and note that $R_{\boldsymbol{\lambda}}^t \leqslant 0$ since the instance player is performing a best-response against $\boldsymbol{w}_s$ at each s.

Step 6: CGE is an Approximate Optimistic Saddle Point Algorithm We now show that the CGE is an approximate optimistic saddle point algorithm. From the regret properties of $\boldsymbol{\lambda}$ player we have

$$\inf_{\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{s=1}^{t} \sum_{a=1}^{K} w_{s,a} d(\hat{\mu}_{a,s}, \lambda_a) \geqslant \inf_{\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{s=1}^{t} \sum_{a=1}^{K} w_{s,a} d(\hat{\mu}_{a,s}, \lambda_{a,s})$$

since $R_{\boldsymbol{\lambda}}^t \leqslant 0$.

Let $C_{a,s} = U_{a,s} - d(\hat{\mu}_{a,s}, \lambda_{a,s})$. We have

$$\inf_{\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{s=1}^{t} \sum_{a=1}^{K} w_{s,a} d(\hat{\mu}_{a,s}, \lambda_a) \geqslant \inf_{\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{s=1}^{t} \sum_{a=1}^{K} w_{s,a} U_{a,s}(\boldsymbol{\lambda}) - \sum_{s=1}^{t} \sum_{a=1}^{K} w_{s,a} C_{a,s}.$$

Now, we can combine Eq. (21) and (23) to get

$$c(t, \delta) \geqslant \inf_{\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}} \boldsymbol{\mu}} \sum_{s=1}^{t} \sum_{a=1}^{K} w_{s,a} U_{a,s}(\boldsymbol{\lambda}) - \sum_{s=1}^{t} \sum_{a=1}^{K} w_{s,a} C_{a,s} - (1 + \sqrt{t})K - O(\sqrt{t \log t})$$

Now we use the properties of $R_t^{\boldsymbol{w}}$ to get

$$\inf_{\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{s=1}^{t} \sum_{a=1}^{K} w_{s,a} d(\hat{\mu}_{a,s}, \lambda_s) \geqslant \max_{\boldsymbol{w} \in \Pi} \sum_{a=1}^{K} \sum_{s=1}^{t} w_a U_{a,s} - R_{\boldsymbol{w}}^t - \sum_{s=1}^{t} \sum_{a=1}^{K} w_{a,s} C_{a,s}$$

which shows that CGE is an approximate optimistic saddle point algorithm with slack

$$x_t = R_t^{\boldsymbol{w}} + \sum_{s=1}^{t}\sum_{a=1}^{K} w_{s,a} C_{a,t}. \tag{25}$$

Step 7: Plug slack x_t into Equation (21)**.** We now use the fact that CGE is an approximate optimistic saddle point algorithm in Equation (21)

$$c(t,\delta) \geqslant \max_{\boldsymbol{w}\in\Pi} \sum_{a=1}^{K}\sum_{s=1}^{t} w_a U_{a,s} - R_t^{\boldsymbol{w}} - \sum_{s=1}^{t}\sum_{a=1}^{K} w_{s,a} C_{a,t} - (1+\sqrt{t})K - O(\sqrt{t\log t}) \tag{26}$$

Step 8: Concentration of $\sum_{a=1}^{K} w_{s,a} C_{a,t}$

Assume the event $\mathcal{E}_T$. We have

$$|d(\mu_a,\lambda_a) - d(\hat{\mu}_{a,s},\lambda_a)| \leqslant L d(\hat{\mu}_{a,s},\mu_a)$$

due to the Lipschitz property of the KL-divergence and under the event $\mathcal{E}_T$ we have

$$|d(\mu_a,\lambda_a) - d(\hat{\mu}_{a,s},\lambda_a)| \leqslant L\sqrt{2\sigma^2\frac{f(s)}{N_{a,s}}}.$$

This implies that

$$\sup_{\xi\in[\alpha_{a,s},\beta_{a,s}]} U_{a,s} - d(\xi,\lambda_{a,s}) \leqslant \max\left\{2L\sqrt{2\sigma^2\frac{f(s)}{N_{a,s}}}, \frac{f(s)}{N_{a,s}}\right\}$$

since either $U_{a,s} = \max_{\xi\in[\alpha_{a,s},\beta_{a,s}]} d(\xi,,\lambda_{a,s})$ and the above is equal to the width of the confidence interval, or $U_{a,s} = \frac{f(s)}{N_{a,s}}$ and the above is trivially bounded $\frac{f(s)}{N_{a,s}}$ since the KL divergence is non-negative. Hence,

$$\begin{aligned}
\sum_{s=K+1}^{t}\sum_{a=1}^{K} w_{s,a} C_{a,s} &\leqslant \sum_{s=K+1}^{t}\sum_{a=1}^{K} w_{s,a}\left(\frac{f(s)}{N_{a,s}} + 2L\sqrt{2\sigma^2\frac{f(s)}{N_{a,s}}}\right) \\
&\leqslant f(t)\sum_{s=K+1}^{t}\sum_{a=1}^{K}\frac{w_{s,a}}{N_{a,s}} + 2L\sqrt{2\sigma^2 f(t)}\sum_{s=K+1}^{t}\sum_{a=1}^{K}\frac{w_{s,a}}{\sqrt{N_{a,s}}} \\
&\leqslant f(t)\left(K^2 + 2K\log\frac{t}{K}\right) + 2L\sqrt{2\sigma^2 f(t)}\left(K^2 + 2\sqrt{2Kt}\right) \\
&\leqslant O(\sqrt{t\log t}).
\end{aligned}$$

We have

$$c(t,\delta) \geqslant \max_{\boldsymbol{w}\in\Pi} \sum_{a=1}^{K}\sum_{s=1}^{t} w_a U_{a,s} - R_t^{\boldsymbol{w}} - O(\sqrt{t\log t}) - (1+\sqrt{t})K - O(\sqrt{t\log t}).$$

Step 9: Optimism

We now use the fact that $U_{a,s} \geqslant d(\mu_a,\lambda_a)$ under the event $\mathcal{E}_T$. Hence,

$$c(t,\delta) \geqslant \max_{\boldsymbol{w}\in\Pi} \sum_{a=1}^{K}\sum_{s=1}^{t} w_a d(\mu_a,\lambda_{a,s}) - R_t^{\boldsymbol{w}} - O(\sqrt{t\log t}) - (1+\sqrt{t})K - O(\sqrt{t\log t}).$$

Step 10: Get the Characteristic Time We note that

$$\max_{\boldsymbol{w}\in\Pi} \sum_{a=1}^{K}\sum_{s=1}^{t} w_a d(\mu_a,\lambda_{a,s}) \geqslant t \inf_{\lambda\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \max_{\boldsymbol{w}\in\Pi} \sum_{a=1}^{K} w_a d(\mu_a,\lambda_a)$$

$$\geqslant \max_{\boldsymbol{w}\in\Pi} \inf_{\lambda\in\Lambda_{\mathcal{F}}(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) = tT_{\mathcal{F}}^{-1}(\boldsymbol{\mu}).$$

Rearanging yields

$$t \leqslant T_{\mathcal{F}}(\boldsymbol{\mu})c(t,\delta) + R_t^{\boldsymbol{w}} + O(\sqrt{t\log t})$$

Step 11: Current Recommendation is the Wrong Policy. The above result is conditioned on the fact that our current recommendation is correct. We now bound the number of time steps where the current recommendation is wrong, using similar argument as in Degenne et al. (2019).

We define the Chernoff information as $\text{ch}(x,y) \triangleq \inf_{u\in\mathcal{D}} : d(u,x) + d(u,y)$. Assumption 1 gives that there $\exists\ \epsilon > 0$ such that $\forall \boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})$, $\exists a'$ such that $\text{ch}(\lambda_{a'}, \mu_{a'}) > \epsilon$.

Assume that $\boldsymbol{\pi}^* \neq \arg\max_{\boldsymbol{\pi}\in\mathcal{F}} \hat{\boldsymbol{\mu}}_t^\top \boldsymbol{\pi}$, i.e. if we stop we would recommend the wrong policy. This implies that $\hat{\boldsymbol{\mu}}_t \in \Lambda_{\mathcal{F}}(\boldsymbol{\mu})$ and $\text{ch}(\hat{\mu}_{a,t}, \mu_a) \geqslant \epsilon$ for some arm a. Under the good event $\mathcal{E}_T$ we have $N_{a,t} d(\hat{\mu}_{a,t}, \mu_a) \leqslant f(t)$ which implies that $\frac{f(t)}{N_{a,t}} \geqslant \epsilon$, since $\text{ch}(\hat{\mu}_{a,t}, \mu_a) \leqslant d(\hat{\mu}_{a,t}, \mu_a)$.

Let $\boldsymbol{\pi}_s \triangleq \arg\max_{\boldsymbol{\pi}\in\mathcal{F}} \hat{\boldsymbol{\mu}}_s^\top \boldsymbol{\pi}$, let $n_{\boldsymbol{\pi}'}(t)$ be the number of stages where $\boldsymbol{\pi}_s = \boldsymbol{\pi}'$. Our goal is to upper bound $n_{\boldsymbol{\pi}'}(t)$ for all extreme points $\boldsymbol{\pi}' \in \mathcal{F}$ such that $'\boldsymbol{\pi}' \neq \boldsymbol{\pi}^*$. For any $\boldsymbol{\lambda}$ such that $\boldsymbol{\pi}' = \arg\max_{\boldsymbol{\pi}\in\mathcal{F}} \boldsymbol{\lambda}^\top \boldsymbol{\pi}$ we have that $\boldsymbol{\mu} \in \Lambda_{\mathcal{F}}(\boldsymbol{\lambda})$ which gives

$$\epsilon_t = \sum_{s=1,\boldsymbol{\pi}_s\neq\boldsymbol{\pi}^*}^{t} \sum_{a=1}^{K} w_{a,s} d(\hat{\mu}_{a,s}, \mu_a) \geqslant \sum_{\boldsymbol{\pi}'\neq\boldsymbol{\pi}^*} \inf_{\boldsymbol{\lambda}:\boldsymbol{\pi}'\neq\arg\max_{\boldsymbol{\pi}} \boldsymbol{\lambda}^\top\boldsymbol{\pi}} \sum_{s=1,\boldsymbol{\pi}_s=\boldsymbol{\pi}'}^{t} \sum_{a=1}^{K} w_{a,s} d(\hat{\mu}_{a,s}, \lambda_a).$$

We use the fact that on the time steps where $\boldsymbol{\pi}_s = \boldsymbol{\pi}'$ CGE is a optimistic saddle point algorithm with slack $x = R^{\boldsymbol{w}}_{n_{\boldsymbol{\pi}'}(t)} + \sum_{s=1,\boldsymbol{\pi}_s=\boldsymbol{\pi}'}^{t} \sum_{a=1}^{K} w_{s,a} C_{a,t}$. Hence,

$$\inf_{\boldsymbol{\lambda}:\boldsymbol{\pi}'\neq\arg\max_{\boldsymbol{\pi}} \boldsymbol{\lambda}^\top\boldsymbol{\pi}} \sum_{s=1,\boldsymbol{\pi}_s=\boldsymbol{\pi}'}^{t} \sum_{a=1}^{K} w_{a,s} d(\hat{\mu}_{a,s}, \lambda_a) \geqslant$$

$$\max_{\boldsymbol{\pi}\in\Pi} \sum_{s=1,\boldsymbol{\pi}_s=\boldsymbol{\pi}'}^{t} \sum_{a=1}^{K} w_a U_{a,s} - R^{\boldsymbol{w}}_{n_{\boldsymbol{\pi}'}(t)} - \sum_{s=1,\boldsymbol{\pi}_s=\boldsymbol{\pi}'}^{t} \sum_{a=1}^{K} w_{a,s} C_{a,s}.$$

Under the event $\mathcal{E}_T$, and $s \leqslant t$ such that $\boldsymbol{\pi}_s = \boldsymbol{\pi}'$ there is an arm a_s such that $U_{a_s,s} \geqslant \epsilon$. This implies that the sum $\max_{\boldsymbol{\pi}\in\Pi} \sum_{s=1,\boldsymbol{\pi}_s=\boldsymbol{\pi}'}^{t} \sum_{a=1}^{K} w_a U_{a,s}$ is increasing linearly in $n_{\boldsymbol{\pi}'}(t)$ since it is at least $\epsilon n_{\boldsymbol{\pi}'}(t)$ under the concentration event $\mathcal{E}_T$. Thus,

$$\inf_{\boldsymbol{\lambda}:\boldsymbol{\pi}'\neq\arg\max_{\boldsymbol{\pi}} \boldsymbol{\lambda}^\top\boldsymbol{\pi}} \sum_{s=1,\boldsymbol{\pi}_s=\boldsymbol{\pi}'}^{t} \sum_{a=1}^{K} w_{a,s} d(\hat{\mu}_{a,s}, \lambda_a) \geqslant \epsilon n_{\boldsymbol{\pi}'}(t) - R^{\boldsymbol{w}}_{n_{\boldsymbol{\pi}'}(t)} - \sum_{s=1,\boldsymbol{\pi}_s=\boldsymbol{\pi}'}^{t} \sum_{a=1}^{K} w_{a,s} C_{a,s}$$

and we know that $R^{\boldsymbol{w}}_{n_{\boldsymbol{\pi}'}(t)} = O(\sqrt{Q n_{\boldsymbol{\pi}'}(t)})$ and $\sum_{s=1,\boldsymbol{\pi}_s=\boldsymbol{\pi}'}^{t} \sum_{a=1}^{K} w_{a,s} C_{a,s} = O(\sqrt{n_{\boldsymbol{\pi}'}(t)\log n_{\boldsymbol{\pi}'}(t)})$. This shows that ϵ_T increases at least linear in $n_{\boldsymbol{\pi}'}(t)$ and thus also linearly in the number of time steps for whitch $\boldsymbol{\pi}_s \neq \boldsymbol{\pi}^*$. However, we have

$$\epsilon_t = \sum_{s=1,\boldsymbol{\pi}_s\neq\boldsymbol{\pi}^*}^{t} \sum_{a=1}^{K} w_{a,s} d(\hat{\mu}_{a,s}, \mu_a) \leqslant \sum_{s=1}^{t} \sum_{a=1}^{K} w_{a,s} \frac{f(s)}{N_{a,s}}$$

$$\leqslant f(t)(K^2 + 2K\log\frac{t}{K}).$$

This implies that the current recommendation $\boldsymbol{\pi}_s = \arg\max \hat{\boldsymbol{\mu}}_t^\top \boldsymbol{\pi}$ differs from $\boldsymbol{\pi}^*$ at most $O(\sqrt{t\log t})$ number of times.

Step 12: Final Bound. We know from the concentration of $\mathcal{E}_T$ that the number of times the compliment happens is upper bounded by CK where C is some problem independent constant. Putting it all together, we get that $\mathbb{E}[\tau] \leqslant T_0(\delta) + CK$, where

$$T_0(\delta) := \max\left\{t \in \mathbb{N} : t \leqslant T_{\mathcal{F}}(\boldsymbol{\mu})c(t,\delta) + O(\sqrt{tQ}) + O(\sqrt{t\log t})\right\}.$$

□

D Finding ϵ-good policies under linear constraints

In some cases one might be more interested in finding a policy that is ϵ-close to the optimal one, i.e. finding $\boldsymbol{\pi}'$ such that $\boldsymbol{\mu}^\top(\boldsymbol{\pi}^*_{\boldsymbol{\mu}} - \boldsymbol{\pi}') \leqslant \epsilon$, since this might have a much smaller sample complexity compared to searching for the optimal policy, see for example (Garivier and Kaufmann, 2021) and (Kocák and Garivier, 2021). Both CTnS and CGE can in principle be extended to this case by changing the definition of the Alt-set. Given an instance $\boldsymbol{\mu}$ let $\Omega_{\mathcal{F},\epsilon}(\boldsymbol{\mu}) := \{\boldsymbol{\pi} \in \mathcal{N}_{\mathcal{F}} : \mu^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi} \leqslant \epsilon)\}$ be the set of ϵ-good policies where $\mathcal{N}_{\mathcal{F}}$ is the set of all extreme points in the polytope $\mathcal{F}$. For each $\boldsymbol{\pi} \in \Omega_{\mathcal{F},\epsilon}(\boldsymbol{\mu})$ we get the following Alt-set

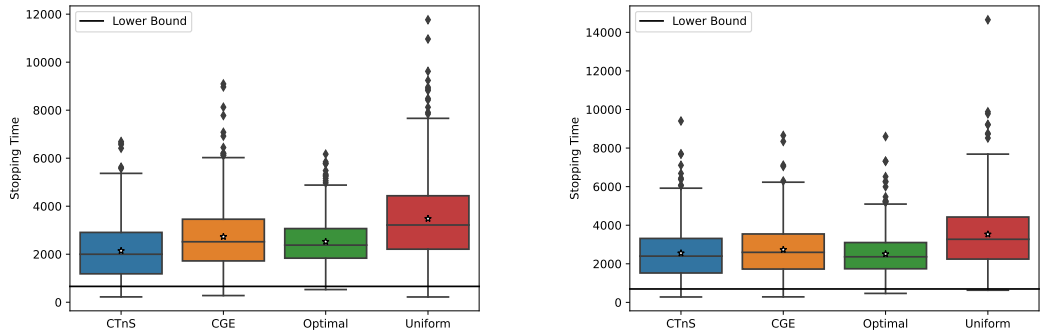
$$\Lambda_{\mathcal{F},\epsilon}(\boldsymbol{\mu}, \boldsymbol{\pi}) := \left\{\boldsymbol{\lambda} : \boldsymbol{\lambda}^\top \left(\boldsymbol{\pi}^*_{\boldsymbol{\lambda}} - \boldsymbol{\pi}\right) > \epsilon\right\}.$$

Hence, the sample complexity might be different depending on which near-optimal policy the learner is considering. To handle this we would have to augment CTnS and CGE with the "sticky" approach developed in (Degenne and Koolen, 2019), where the learner commits to a recommendation since otherwise the learner might oscillate between near-optimal policies and a mixture of their optimal allocations might not be optimal since $w^*(\boldsymbol{\mu})$ is no longer ensured to be convex. Furthermore, due to $\epsilon > 0$ it is no longer sufficient to project onto the normal cone and a naive implementation would have to optimize over $|\mathcal{N}_{\mathcal{F}}|$ convex sets which might only be tractable for a small set of constraints and/or arms.

E Additional experimental analysis

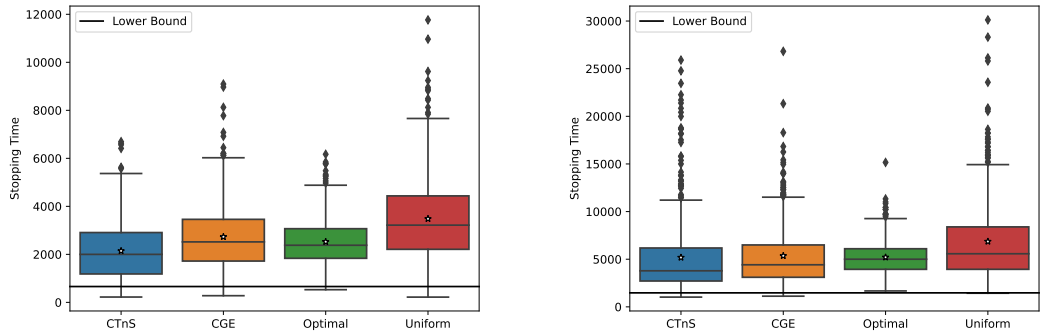
In Figure 5 and 6 we present results for arms with Bernoulli distributions and in Figure 7 and 8 we present additional results for arms with Gaussian distributions. CTnS and CGE outperforms the uniform baseline in all cases and are usually on par with or better than the learner that always sample according to the asymptotically optimal allocation. We also see that the algorithms tend to be close to the lower bound in all cases. An interesting observation, which we commented on already in the main text, is that there tend to be a larger difference between all sampling rules for end-of-time constraints compared to anytime constraints. This is due to the fact that anytime constraints can be very restrictive on which sampling allocations are allowed and there might not be less room for an adaptive learner.

In the case of arms with Bernoulli distributions we did not use a close-form projection, as for Gaussian distributions, and instead computed the projection numerically by minimizing the KL-divergence subject to $\boldsymbol{\lambda}^\top(\boldsymbol{\pi}^* - \boldsymbol{\pi}') = 0$, which is a convex problem. We discuss the effect of this in Section E.1



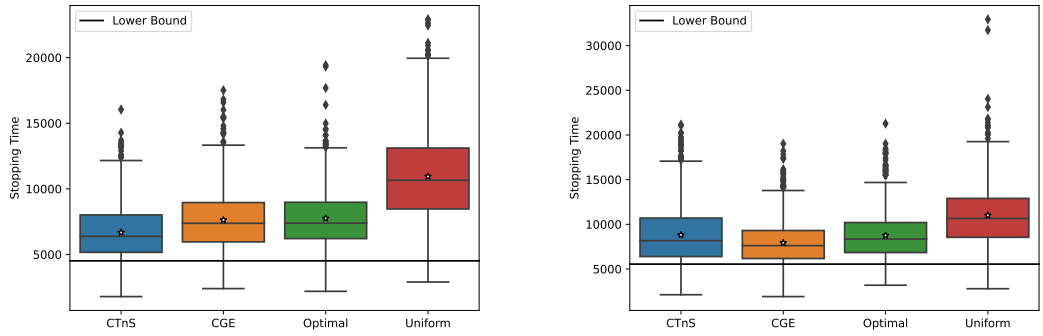
(a) End-of-time constraints (b) Anytime constraints

Figure 5: End-of-time and Anytime constraints with *Bernoulli* arms. The reward vector is $\boldsymbol{\mu} = (0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2)$ and the constraints are $\boldsymbol{\pi}_1 + \boldsymbol{\pi}_2 \leqslant 0.5$ and $\boldsymbol{\pi}_3 + \boldsymbol{\pi}_4 \leqslant 0.5$. Average over 500 seeds and $\delta = 0.1$. Optimal policy is $\boldsymbol{\pi}_1 = 0.5$ and $\boldsymbol{\pi}_3 = 0.5$.



(a) End-of-time constraints (b) Anytime constraints

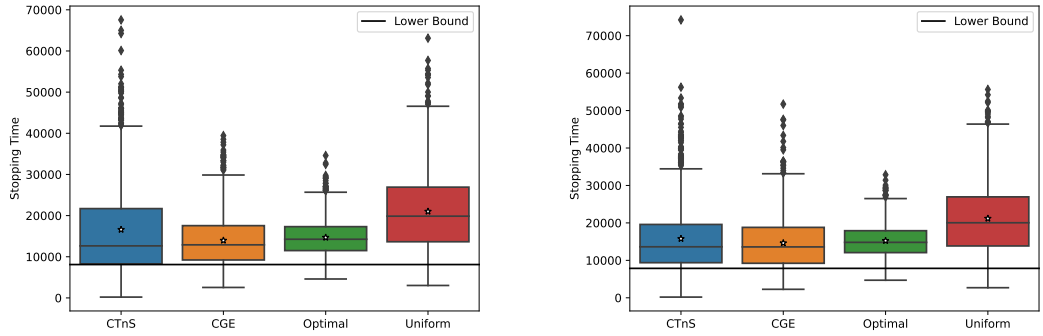
Figure 6: End-of-time and Anytime constraints with Bernoulli arms. The reward vector is $\boldsymbol{\mu} = (0.8, 0.7, 0.6, 0.5, 0.4)$ and the constraints are $4\boldsymbol{\pi}_1 - \boldsymbol{\pi}_5 \leqslant 1$ and $3\boldsymbol{\pi}_2 - \boldsymbol{\pi}_4 \leqslant 1$. Average over 500 seeds and $\delta = 0.1$. Optimal policy is $\boldsymbol{\pi}_1 = 0.25$, $\boldsymbol{\pi}_2 = 0.33$ and $\boldsymbol{\pi}_3 = 0.42$.



(a) End-of-time constraints

(b) Anytime constraints

Figure 7: End-of-time and Anytime constraints with *Gaussian* arms $\sigma^2 = 1$. The reward vector is $\boldsymbol{\mu} = (2.0, 1.5, 1.45, 0.5, 0.3, -1.0, -1.0)$ and the constraints are $4\boldsymbol{\pi}_1 + \boldsymbol{\pi}_2 \leqslant 0.7$ and $\boldsymbol{\pi}_2 + 2\boldsymbol{\pi}_3 \leqslant 0.5$. Average over 1000 seeds and $\delta = 10^{-4}$. Optimal policy is $\boldsymbol{\pi}_1 = 0.05$, $\boldsymbol{\pi}_2 = 0.5$ and $\boldsymbol{\pi}_4 = 0.45$.



(a) End-of-time constraints

(b) Anytime constraints

Figure 8: End-of-time and Anytime constraints with Gaussian arms $\sigma^2 = 1$. The reward vector is $\boldsymbol{\mu} = [1.0, 0.5, 0.4, 0.3, 0.2, 0.1]$ and the constraints are $\boldsymbol{\pi}_1 - \boldsymbol{\pi}_4 - \boldsymbol{\pi}_5 - \boldsymbol{\pi}_6 \leqslant 0.3$ and $\boldsymbol{\pi}_2 \leqslant 0.7$. Average over 1000 seeds and $\delta = 10^{-3}$. Optimal policy is $\boldsymbol{\pi}_1 = 0.65$ and $\boldsymbol{\pi}_4 = 0.35$.

E.1 Running Times

In Table 2 we present the average time it take for the algorithms to check the stopping criterion and select a new arm to play. The test was performed on 1 core of a Intel Xeon Gold 6130 CPU with 2.1 GHz. Gaussian indicates the experiments in Figure 7a, Bernoulli the experiments in Figure 5a and IMDB the experiments in Figure 4b. As expected CTnS is the algorithm requiring most computational time and the excessive running time it has on the experiment with Bernoulli distributions is due to the fact that we numerically solve the projection instead of relying on a close-form expression as in the case of Gaussian distributions. In contrast, we see that CGE has a relatively light computational footprint in all cases. Another advantage of CGE is that it performs a finite number of $\max$ calls at each iteration which can easily be parallelized for larger bandit instances with many constraints.

Algorithm	Bernoulli	Gaussian	IMDB
CTnS	1.00 ± 0.244	0.030 ± 0.006	0.033 ± 0.015
CGE	0.02 ± 0.001	0.005 ± 0.001	0.008 ± 0.001
Uniform	$0.009 \pm 3 \times 10^{-4}$	$0.001 \pm 1 \times 10^{-4}$	$0.002 \pm 2 \times 10^{-4}$

Table 2: Average time, in seconds, it takes to check the stopping criterion and select a new arm for the different algorithms. The$\pm$ indicates one standard deviation. We omitted the optimal sampler since this one has the same running time as the uniform sampler.

E.2 IMDB environment

For resproducibility, here we provide the specifics of the IMDB data in the Table 3 as used in the experiments (Figure 4).

Movie	Average Rating	σ	Action	Drama	Family
The Net	3.67	1.26	1	1	0
Happily N'Ever After	2.97	1.30	0	0	1
Tomorrowland	2.94	1.31	1	0	1
American Hero	3.52	1.33	1	1	0
Das Boot	3.18	1.30	0	1	0
Final Destination 3	2.02	0.93	0	0	0
Licence to Kill	2.79	1.22	1	0	0
The Hundred-Foot Journey	2.97	1.31	0	1	0
The Matrix	2.32	1.14	1	0	0
Creature	2.53	1.20	0	0	0
The Basket	2.55	1.19	0	1	0
Star Trek: The Motion Picture	2.54	1.16	0	0	0

Table 3: Movies used in the experiments presented in Figure 4. The optimal policy is $\boldsymbol{\pi}_1^* = 0.3$, $\boldsymbol{\pi}_2^* = 0.3$ and $\boldsymbol{\pi}_5^* = 0.4$. We used the maximum σ in the algorithms. This means that the algorithms didn't have access to the true σ of each arm and instead modelled them all as Gaussian distributions with $\sigma = 1.33$ but the rewards were sampled from the environment using the true σ.

F Useful definitions and results

Definition 2 (Upper hemicontinuity)**.** *We say that a set-valued function $C : \Theta \to \Omega$ is upper hemicontinuous at the point $\theta \in \Theta$ if for any open set $S \subset \Omega$ 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 5 (Berge's maximum theorem (Berge, 1963))**.** *Let X and Θ be topological spaces. Let $f : X \times \Theta \to \mathbb{R}$ be a continuous function and let $C : \Theta \to X$ 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.*

Below we restate the upper bound on the sample complexity of the Gamified Explorer (GE) of Degenne et al. (2019).

Theorem 6 (Theorem 2 in Degenne et al. (2019))**.** *The sample complexity of GE is*

$$\mathbb{E}[\tau] \leqslant T_0(\delta) + \frac{eK}{a}$$

where

$$T_0(\delta) = \max\{t \in \mathbb{N} : t \leqslant T(\boldsymbol{\mu})c(t, \delta) + C_{\boldsymbol{\mu}}(R_t^{\boldsymbol{\lambda}} + R_t^{\boldsymbol{w}} + O(\sqrt{t \log t}))\}$$

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.