Stochastic Bandits with Linear Constraints

Aldo Pacchiano
UC Berkeley
pacchiano@berkeley.edu

Mohammad Ghavamzadeh
Google Research
ghavamza@google.com

Peter Bartlett
UC Berkeley
peter@berkeley.edu

Heinrich Jiang
Google Research
heinrichj@google.com

Abstract

We study a constrained contextual linear bandit setting, where the goal of the agent is to produce a sequence of policies, whose expected cumulative reward over the course of T rounds is maximum, and each has an expected cost below a certain threshold τ. We propose an upper-confidence bound algorithm for this problem, called *optimistic pessimistic linear bandit* (OPLB), and prove an $\widetilde{\mathcal{O}}(\frac{d\sqrt{T}}{\tau-c_0})$ bound on its T-round regret, where the denominator is the difference between the constraint threshold and the cost of a known feasible action. We further specialize our results to multi-armed bandits and propose a computationally efficient algorithm for this setting. We prove a regret bound of $\widetilde{\mathcal{O}}(\frac{\sqrt{KT}}{\tau-c_0})$ for this algorithm in K-armed bandits, which is a $\sqrt{K}$ improvement over the regret bound we obtain by simply casting multi-armed bandits as an instance of contextual linear bandits and using the regret bound of OPLB. We also prove a lower-bound for the problem studied in the paper and provide simulations to validate our theoretical results.

1 Introduction

A *multi-armed bandit* (MAB) [Lai and Robbins, 1985, Auer et al., 2002, Lattimore and Szepesvári, 2019] is an online learning problem in which the agent acts by pulling arms. After an arm is pulled, the agent receives its *stochastic reward*. The goal of the agent is to maximize its expected cumulative reward without knowledge of the arms' distributions. To achieve this goal, the agent has to balance its *exploration* and *exploitation*: to decide when to *explore* and learn about the arms, and when to *exploit* and pull the arm with the highest estimated reward thus far. A *stochastic linear bandit* [Dani et al., 2008, Rusmevichientong and Tsitsiklis, 2010, Abbasi-Yadkori et al., 2011] is a generalization of MAB to the setting where each of (possibly) infinitely many arms is associated with a feature vector. The mean reward of an arm is the dot product of its feature vector and an unknown parameter vector, which is shared by all the arms. This formulation contains time-varying action (arm) sets and feature vectors, and thus, includes the *linear contextual bandit* setting. These models capture many practical applications spanning clinical trials [Villar et al., 2015], recommendation systems [Li et al., 2010, Balakrishnan et al., 2018], wireless networks [Maghsudi and Hossain, 2016], sensors [Washburn, 2008], and strategy games [Ontanón, 2013]. The most popular exploration strategies in stochastic bandits are *optimism in the face of uncertainty* (OFU) [Auer et al., 2002] and *Thompson sampling* (TS) [Thompson, 1933, Agrawal and Goyal, 2013a, Russo et al., 2018] that are relatively well understood in both multi-armed and linear bandits [Dani et al., 2008, Abbasi-Yadkori et al., 2011, Agrawal and Goyal, 2013b, Lattimore and Szepesvári, 2019].

In many practical problems, the agent requires to satisfy certain operational constraints while maximizing its cumulative reward. Depending on the form of the constraints, several *constrained*

stochastic bandit settings have been formulated and analyzed. One such setting is what is known as knapsack bandits. In this setting, pulling each arm, in addition to producing a reward signal, results in a random consumption of a global budget, and the goal is to maximize the cumulative reward before the budget is fully consumed (e.g., Badanidiyuru et al. 2013, 2014, Agrawal and Devanur 2014, Wu et al. 2015, Agrawal and Devanur 2016). Another such setting is referred to as *conservative bandits*. In this setting, there is a baseline arm or policy, and the agent, in addition to maximizing its cumulative reward, should ensure that at each round, the difference between its cumulative reward and that of the baseline remains below a predefined fraction of the baseline cumulative reward [Wu et al., 2016, Kazerouni et al., 2017, Garcelon et al., 2020]. In these two settings, the constraint applies to a cumulative quantity (budget consumption or reward) over the entire run of the algorithm. Thus, the set of feasible actions at each round is a function of the history of the algorithm.

Another constrained bandit setting is where each arm is associated with two (unknown) distributions, generating reward and cost signals. The goal is to maximize the cumulative reward, while making sure that with high probability, the expected cost of the arm pulled at each round is below a certain threshold. Here the constraint is stage-wise, and unlike the last two settings, is independent of the history. Amani et al. [2019] and Moradipari et al. [2019] have recently studied this setting for linear bandits and derived and analyzed explore-exploit [Amani et al., 2019] and Thompson sampling [Moradipari et al., 2019] algorithms for it.

This setting is the closest to the one we study in this paper. In our setting, we also assume two distributions for each arm, one for reward and for cost. At each round the agent constructs a policy according to which it takes its action. The goal of the agent is to produce a sequence of policies with maximum expected cumulative reward, while making sure that the expected cost of the constructed policy (not the pulled arm) at each round is below a certain threshold. This is a linear constraint and can be easily extended to more constraints by having more cost distributions associated to each arm, one per each constraint. Compared to the previous setting, our constraint is more relaxed (from high-probability to expectation), and as a result, it would be possible for us to obtain a solution with larger expected cumulative reward. We will have a detailed discussion on the relationship between these two settings and the similarities and differences of our results with those reported in Amani et al. [2019] and Moradipari et al. [2019] in Section 7.

In this paper, we study the above setting for contextual linear bandits. After defining the setting in Section 2, we propose an upper-confidence bound (UCB) algorithm for it, called *optimistic pessimistic linear bandit* (OPLB), in Section 3. We prove an $\widetilde{\mathcal{O}}(\frac{d\sqrt{T}}{\tau-c_0})$ bound on the T-round regret of OPLB in Section 4, where d is the action dimension and $\tau - c_0$ is the difference between the constraint threshold and the cost of a known feasible action. The action set considered in our contextual linear bandit setting is general enough to include MAB. However, in Section 5, we further specialize our results to MAB and propose a computationally efficient algorithm for this setting, called *optimistic pessimistic bandit* (OPB). We show that in the MAB case, there always exists a feasible optimal policy with probability mass on at most $m + 1$ arms, where m is the number of linear constraints. This property plays an important role in the computational efficiency of OPB. We prove a regret bound of $\widetilde{\mathcal{O}}(\frac{\sqrt{KT}}{\tau-c_0})$ for OPB in K-armed bandits, which is a $\sqrt{K}$ improvement over the regret bound we obtain by simply casting MAB as an instance of contextual linear bandit and using the regret bound of OPLB. We also prove a lower-bound for the problem studied in the paper and provide simulations to validate our theoretical results.

2 Problem Formulation

We adopt the following notation. The set $\{1, \ldots, T\}$ is denoted by $[T]$. We represent the set of distributions with support over a compact set $\mathcal{S}$ by $\Delta_{\mathcal{S}}$. We denote by $\langle x, y \rangle := x^\top y \in \mathbb{R}$, the inner product of two vectors $x, y \in \mathbb{R}^d$, and by $\|x\| := \sqrt{x^\top x}$, the ℓ_2-norm of vector x.

The setting we study in this paper is *contextual linear bandit* with *linear constraints*. In each round t, the agent is given an decision set $\mathcal{A}_t \subset \mathbb{R}^d$ from which it has to choose an action x_t. Upon taking action $x_t \in \mathcal{A}$, it observes a pair (r_t, c_t), where $r_t = \langle x_t, \theta_* \rangle + \xi_t^r$ and $c_t = \langle x_t, \mu_* \rangle + \xi_t^c$ are the reward and cost signals, respectively. In the reward and cost definitions, $\theta_* \in \mathbb{R}^d$ and $\mu_* \in \mathbb{R}^d$ are the unknown *reward* and *cost parameters*, and ξ_t^r and ξ_t^c are reward and cost noise, satisfying conditions that will be specified soon. The agent selects its action $x_t \in \mathcal{A}_t$ in each round t according to its policy $\pi_t \in \Delta_{\mathcal{A}_t}$ at that round, i.e., $x_t \sim \pi_t$.

The goal of the agent is to produce a sequence of policies $\{\pi_t\}_{t=1}^T$ with maximum *expected cumulative reward* over the course of T rounds, while satisfying the **linear constraint**

$$\mathbb{E}_{x\sim\pi_t}[\langle x,\mu_*\rangle] \leq \tau, \quad \forall t \in [T], \qquad (\tau \geq 0 \text{ is referred to as the } \textit{constraint threshold}). \quad (1)$$

Thus, the policy π_t selected by the agent in each round $t \in [T]$ should belong to the set of *feasible policies* over the action set $\mathcal{A}_t$, i.e., $\Pi_t = \{\pi \in \Delta_{\mathcal{A}_t} : \mathbb{E}_{x\sim\pi}[\langle x,\mu_*\rangle] \leq \tau\}$. Maximizing the expected cumulative reward in T rounds is equivalent to minimizing the T-round *constrained pseudo-regret*,[1]

$$\mathcal{R}_\Pi(T) = \sum_{t=1}^T \mathbb{E}_{x\sim\pi_t^*}[\langle x,\theta_*\rangle] - E_{x\sim\pi_t}[\langle x,\theta_*\rangle], \quad (2)$$

where $\pi_t, \pi_t^* \in \Pi_t \;\forall t \in [T]$ and π_t^* is the *optimal feasible* policy at round t, i.e., $\pi_t^* \in \max_{\pi\in\Pi_t}\mathbb{E}_{x\sim\pi}[\langle x,\theta_*\rangle]$. The terms $\mathbb{E}_{x\sim\pi}[\langle x,\theta_*\rangle]$ and $\mathbb{E}_{x\sim\pi}[\langle x,\mu_*\rangle]$ in (1) and (2) are the expected reward and cost of policy π, respectively. Thus, a feasible policy is the one whose expected cost is below the constraint threshold τ, and the optimal feasible policy is a feasible policy with maximum expected reward. We use the shorthand notations $x_\pi := \mathbb{E}_{x\sim\pi}[x]$, $r_\pi := \mathbb{E}_{x\sim\pi}[\langle x,\theta_*\rangle]$ and $c_\pi := \mathbb{E}_{x\sim\pi}[\langle x,\mu_*\rangle]$ for the expected action, reward and cost of a policy π. With these shorthand notations, we may write the T-round pseudo-regret as $\mathcal{R}_\Pi(T) = \sum_{t=1}^T r_{\pi_t^*} - r_{\pi_t}$.

We make the following assumptions for our setting. The first four assumptions are standard in linear bandits. The fifth one is necessary to guarantee constraint satisfaction (*safety*).

Assumption 1. *For all $t \in [T]$, the reward and cost noise random variables ξ_t^r and ξ_t^c are conditionally R-sub-Gaussian, i.e.,*

$$\mathbb{E}[\xi_t^r \mid \mathcal{F}_{t-1}] = 0, \qquad \mathbb{E}[\exp(\alpha\xi_t^r) \mid \mathcal{F}_{t-1}] \leq \exp(\alpha^2 R^2/2), \quad \forall\alpha\in\mathbb{R},$$
$$\mathbb{E}[\xi_t^c \mid \mathcal{F}_{t-1}] = 0, \qquad \mathbb{E}[\exp(\alpha\xi_t^c) \mid \mathcal{F}_{t-1}] \leq \exp(\alpha^2 R^2/2), \quad \forall\alpha\in\mathbb{R},$$

where $\mathcal{F}_t$ is the filtration that includes all the events $(x_{1:t+1}, \xi_{1:t}^r, \xi_{1:t}^c)$ until the end of round t.

Assumption 2. *There is a known constant $S > 0$, such that $\|\theta_*\| \leq S$ and $\|\mu_*\| \leq S$.[2]*

Assumption 3. *The ℓ_2-norm of all actions is bounded, i.e., $\max_{t\in[T]}\max_{x\in\mathcal{A}_t}\|x\| \leq L$.*

Assumption 4. *For all $t \in [T]$ and $x \in \mathcal{A}_t$, the mean rewards and costs are bounded, i.e., $\langle x,\theta_*\rangle \in [0,1]$ and $\langle x,\mu_*\rangle \in [0,1]$.*

Assumption 5. *There is a known safe action $x_0 \in \mathcal{A}_t$, $\forall t \in [T]$ with known cost c_0, i.e., $\langle x_0,\mu_*\rangle = c_0 < \tau$. We will show how the assumption of knowing c_0 can be relaxed later in the paper.*

Notation: We conclude this section with introducing another set of notations that will be used in the rest of the paper. We define the normalized safe action as $e_0 := x_0/\|x_0\|$ and the span of the safe action as $\mathcal{V}_o := \text{span}(x_0) = \{\eta x_0 : \eta \in \mathbb{R}\}$. We denote by $\mathcal{V}_o^\perp$, the orthogonal complement of $\mathcal{V}_o$, i.e., $\mathcal{V}_o^\perp = \{x \in \mathbb{R}^d : \langle x,y\rangle = 0, \forall y \in \mathcal{V}_o\}$.[3] We define the projection of a vector $x \in \mathbb{R}^d$ into the sub-space $\mathcal{V}_o$, as $x^o := \langle x,e_0\rangle e_0$, and into the sub-space $\mathcal{V}_o^\perp$, as $x^{o,\perp} := x - x^o$. We also define the projection of a policy π into $\mathcal{V}_o$ and $\mathcal{V}_o^\perp$, as $x_\pi^o := \mathbb{E}_{x\sim\pi}[x^o]$ and $x_\pi^{o,\perp} := \mathbb{E}_{x\sim\pi}[x^{o,\perp}]$.

3 Optimistic-Pessimistic Linear Bandit Algorithm

In this section, we propose an algorithm, called *optimistic-pessimistic linear bandit* (OPLB), whose pseudo-code is shown in Algorithm 1. Our OPLB algorithm balances a pessimistic assessment of the set of available policies, while acting optimistically within this set. Our principal innovation is the use of confidence intervals with asymmetric radii, proportional to α_r and α_c, for the reward and cost signals. This will prove crucial in the regret analysis of the algorithm.

---

[1]In the rest of the paper, we simply refer to the T-round constrained pseudo-regret $\mathcal{R}_\Pi(T)$ as T-round regret.
[2]The choice of the same upper-bounds for θ_* and μ_* is just for simplicity.
[3]In the case of $x_0 = \mathbf{0} \in \mathbb{R}^d$, we define $\mathcal{V}_o$ as the empty subspace and $\mathcal{V}_o^\perp$ as the whole $\mathbb{R}^d$.

Algorithm 1 Optimistic-Pessimistic Linear Bandit (OPLB)

Input: Horizon T, Confidence Parameter δ, Regularization Parameter λ, Constants $\alpha_r, \alpha_c \geq 1$

for $t = 1, \ldots, T$ **do**

  1. Compute RLS estimates $\widehat{\theta}_t$ and $\widehat{\mu}_t^{o,\perp}$ *(see Eqs. 3 to 5)*
  2. Construct sets $\mathcal{C}_t^r(\alpha_r)$ and $\mathcal{C}_t^c(\alpha_c)$ *(see Eq. 7)*
  3. Observe $\mathcal{A}_t$ and construct the (estimated) safe policy set Π_t *(see Eq. 12)*
  4. Compute policy $(\pi_t, \widetilde{\theta}_t) = \arg\max_{\pi \in \Pi_t, \, \theta \in \mathcal{C}_t^r(\alpha_r)} \mathbb{E}_{x \sim \pi}[\langle x, \theta \rangle]$
  5. Take action $x_t \sim \pi_t$ and observe reward and cost (r_t, c_t)

Line 1 of OPLB: At each round $t \in [T]$, given the actions $\{x_s\}_{s=1}^{t-1}$, rewards $\{r_s\}_{s=1}^{t-1}$, and costs $\{c_s\}_{s=1}^{t-1}$ observed until the end of round $t-1$, OPLB first computes the ℓ_2-regularized least-squares (RLS) estimates of θ_* and $\mu_*^{o,\perp}$ (projection of the cost parameter μ_* into the sub-space $\mathcal{V}_o^\perp$) as

$$\widehat{\theta}_t = \Sigma_t^{-1} \sum_{s=1}^{t-1} r_s x_s, \qquad \widehat{\mu}_t^{o,\perp} = (\Sigma_t^{o,\perp})^{-1} \sum_{s=1}^{t-1} c_s^{o,\perp} x_s^{o,\perp}, \tag{3}$$

where $\lambda > 0$ is the regularization parameter, and

$$\Sigma_t = \lambda I + \sum_{s=1}^{t-1} x_s x_s^\top, \qquad \Sigma_t^{o,\perp} = \lambda I_{\mathcal{V}_o^\perp} + \sum_{s=1}^{t-1} x_s^{o,\perp}(x_s^{o,\perp})^\top, \tag{4}$$

$$c_t^{o,\perp} = c_t - \frac{\langle x_t, e_0 \rangle}{\|x_0\|} c_0, \qquad I_{\mathcal{V}_o^\perp} = I_{d \times d} - \frac{1}{\|x_0\|^2} x_0 x_0^\top. \tag{5}$$

In (4), Σ_t and $\Sigma_t^{o,\perp}$ are the Gram matrices of actions and projection of actions into the sub-space $\mathcal{V}_o^\perp$. Note that $\Sigma_t^{o,\perp}$ is a rank deficient matrix, but with abuse of notation, we use $(\Sigma_t^{o,\perp})^{-1}$ to denote its pseudo-inverse throughout the paper. In (5), $I_{\mathcal{V}_o^\perp}$ is the projection of the identity matrix, I, into $\mathcal{V}_o^\perp$, and $c_t^{o,\perp}$ is the noisy projection of the cost c_t incurred by taking action x_t into $\mathcal{V}_o^\perp$, i.e.,[4]

$$c_t^{o,\perp} = \langle x_t^{o,\perp}, \mu_*^{o,\perp} \rangle + \xi_t^c = \langle x_t, \mu_* \rangle - \langle x_t^o, \mu_*^o \rangle + \xi_t^c = c_t - \langle x_t^o, \mu_*^o \rangle = c_t - \frac{\langle x_t, e_0 \rangle}{\|x_0\|} c_0. \tag{6}$$

Line 2: Using the RLS estimates $\widehat{\theta}_t$ and $\widehat{\mu}_t^{o,\perp}$ in (3), OPLB constructs the two *confidence sets*

$$\mathcal{C}_t^r(\alpha_r) = \left\{ \theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}_t\|_{\Sigma_t} \leq \alpha_r \beta_t(\delta, d) \right\}, \quad \mathcal{C}_t^c(\alpha_c) = \left\{ \mu \in \mathcal{V}_o^\perp : \|\mu - \widehat{\mu}_t^{o,\perp}\|_{\Sigma_t^{o,\perp}} \leq \alpha_c \beta_t(\delta, d-1) \right\}, \tag{7}$$

where $\alpha_r, \alpha_c \geq 1$ and $\beta_t(\delta, d)$ in the radii of these *confidence ellipsoids* is defined by the following theorem, originally proved in Abbasi-Yadkori et al. [2011].

Theorem 1. *[Thm. 2 in Abbasi-Yadkori et al. 2011] Let Assumptions 1 and 2 hold, $\widehat{\theta}_t$, $\widehat{\mu}_t^{o,\perp}$, Σ_t, and $\Sigma_t^{o,\perp}$ defined by (3) and (4), and $\mathcal{C}_t^r(\cdot)$ and $\mathcal{C}_t^c(\cdot)$ defined by (7). Then, for a fixed $\delta \in (0, 1)$ and*

$$\beta_t(\delta, d) = R\sqrt{d \log\left(\frac{1 + (t-1)L^2/\lambda}{\delta}\right)} + \sqrt{\lambda} \, S, \tag{8}$$

with probability at least $1 - \delta$ and for all $t \geq 1$, it holds that $\theta_* \in \mathcal{C}_t^r(1)$ and $\mu_*^{o,\perp} \in \mathcal{C}_t^c(1)$.

Since $\alpha_r, \alpha_c \geq 1$, for all rounds $t \in [T]$, the sets $\mathcal{C}_t^r(\alpha_r)$ and $\mathcal{C}_t^c(\alpha_c)$ also contain θ_*, the reward parameter, and $\mu_*^{o,\perp}$, the projection of the cost parameter into $\mathcal{V}_o^\perp$, respectively, with high probability.

Given these confidence sets, we define the *optimistic reward* and *pessimistic cost* of any policy π in round t as

$$\widetilde{r}_{\pi,t} := \max_{\theta \in \mathcal{C}_t^r(\alpha_r)} \mathbb{E}_{x \sim \pi}[\langle x, \theta \rangle], \qquad \widetilde{c}_{\pi,t} := \frac{\langle x_\pi^o, e_0 \rangle c_0}{\|x_0\|} + \max_{\mu \in \mathcal{C}_t^c(\alpha_c)} \mathbb{E}_{x \sim \pi}[\langle x, \mu \rangle]. \tag{9}$$

Proposition 1. *We may write (9) in closed-form as*   *(proof in Appendix A.1)*

$$\widetilde{r}_{\pi,t} = \langle x_\pi, \widehat{\theta}_t \rangle + \alpha_r \beta_t(\delta, d) \|x_\pi\|_{\Sigma_t^{-1}}, \tag{10}$$

$$\widetilde{c}_{\pi,t} = \frac{\langle x_\pi^o, e_0 \rangle c_0}{\|x_0\|} + \langle x_\pi^{o,\perp}, \widehat{\mu}_t^{o,\perp} \rangle + \alpha_c \beta_t(\delta, d-1) \|x_\pi^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}}. \tag{11}$$

---

[4]In the derivation of (6), we use the fact that $\langle x_t, \mu_* \rangle = \langle x_t^o + x_t^{o,\perp}, \mu_*^o + \mu_*^{o,\perp} \rangle = \langle x_t^o, \mu_*^o \rangle + \langle x_t^{o,\perp}, \mu_*^{o,\perp} \rangle$.

Line 3: After observing the action set $\mathcal{A}_t$, OPLB constructs its (estimated) feasible (safe) policy set

$$\Pi_t = \{\pi \in \Delta_{\mathcal{A}_t} : \widetilde{c}_{\pi,t} \leq \tau\}, \tag{12}$$

where $\widetilde{c}_{\pi,t}$ is the pessimistic cost of policy π in round t defined by (11). Note that Π_t is not empty since π_0, the policy that plays the safe action x_0 with probability (w.p.) 1, is always in Π_t. This is because $x_{\pi_0}^o = x_0$, $x_{\pi_0}^{o,\perp} = 0$, and $\frac{\langle x_{\pi_0}^o, e_0 \rangle c_0}{\|x_0\|} = c_0$. In the following proposition, whose proof is reported in Appendix A.2, we prove that all policies in Π_t are feasible with high probability.

Proposition 2. *With probability at least $1 - \delta$, for all rounds $t \in [T]$, all policies in Π_t are feasible.*

Line 4: The agent computes its policy, π_t, as the one that is safe (belongs to Π_t) and attains the maximum optimistic reward. We refer to $\widetilde{\theta}_t$ as the *optimistic reward parameter*. Thus, we write the optimistic reward of policy π_t as $\widetilde{r}_{\pi_t,t} = \langle x_{\pi_t}, \widetilde{\theta}_t \rangle$.

Line 5: Finally, the agent selects an action $x_t \sim \pi_t$ and observes the reward-cost pair (r_t, c_t).

Computational Complexity of OPLB. As shown in Line 4 of Algorithm 1 and in Proposition 1, in each round t, OPLB solves the following optimization problem:

$$\max_{\pi \in \Delta_{\mathcal{A}_t}} \langle x_\pi, \widehat{\theta}_t \rangle + \alpha_r \beta_t(\delta, d) \|x_\pi\|_{\Sigma_t^{-1}} \tag{13}$$

$$\text{s.t.} \quad \frac{\langle x_\pi^o, e_0 \rangle c_0}{\|x_0\|} + \langle x_\pi^{o,\perp}, \widehat{\mu}_t^{o,\perp} \rangle + \alpha_c \beta_t(\delta, d-1) \|x_\pi^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}} \leq \tau.$$

However, solving (13) can be challenging. The bottleneck is computing the safe policy set Π_t, which is the intersection between $\Delta_{\mathcal{A}_t}$ and the ellipsoidal constraint.

Remark 1. *The main challenge in obtaining a regret bound for OPLB is to ensure that optimism holds in each round t, i.e., the solution $(\pi_t, \widetilde{\theta}_t)$ of (13) satisfy $\widetilde{r}_{\pi_t,t} = \langle x_{\pi_t}, \widetilde{\theta}_t \rangle \geq r_{\pi_t^*}$. This is not obvious, since the (estimated) safe policy set Π_t may not contain the optimal policy π_t^*. Our main algorithmic innovation is the use of asymmetric confidence intervals $\mathcal{C}_t^r(\alpha_r)$ and $\mathcal{C}_t^c(\alpha_c)$ for θ_* and $\mu_*^{o,\perp}$, which allows us to guarantee optimism, by appropriately selecting the ratio $\gamma = \alpha_r/\alpha_c$. Of course, this comes at the cost of scaling the regret by a factor γ. As it will be shown in our analysis in Section 4, γ depends on the inverse gap $1/(\tau - c_0)$, which indicates when $\tau - c_0$ is small (the cost of the safe arm is close to the constraint threshold), the agent will have a difficult time to identify a safe arm and to compete against the optimal feasible policy π_t^*. We will formalize this in Lemma 4.*

Remark 2. *If the cost of the safe arm c_0 is unknown, we start by taking the safe action x_0 for T_0 rounds to produce a conservative estimate $\hat{\delta}_c$ of $\tau - c_0$ that satisfies $\hat{\delta}_c \geq \frac{\tau - c_0}{2}$. We warm start our estimators for θ_* and μ_* using the data collected by playing x_0. However, instead of estimating $\mu_*^{o,\perp}$, we build an estimator for μ_* over all its directions, including e_0, similar to what OPLB does for θ_*. We then set $\frac{\alpha_r}{\alpha_c} = 1/\hat{\delta}_c$ and run Algorithm 1 for rounds $t > T_0$ (see Appendix B.4 for more details).*

4 Regret Analysis

In this section, we prove the following regret bound for OPLB (Algorithm 1).

Theorem 2 (Regret of OPLB). *Let $\alpha_c = 1$ and $\alpha_r = \frac{2 + \tau - c_0}{\tau - c_0}$. Then, with probability at least $1 - 2\delta$, the regret of OPLB satisfies*

$$\mathcal{R}_\Pi(T) \leq \frac{2L(\alpha_r + 1)\beta_T(\delta, d)}{\sqrt{\lambda}} \sqrt{2T \log(1/\delta)} + (\alpha_r + 1)\beta_T(\delta, d) \sqrt{2Td \log(1 + \frac{TL^2}{\lambda})}. \tag{14}$$

We start the proof of Theorem 2, by defining the following event that holds w.p. at least $1 - \delta$:

$$\mathcal{E} := \left\{ \|\widehat{\theta}_t - \theta_*\|_{\Sigma_t} \leq \beta_t(\delta, d) \ \wedge \ \|\widehat{\mu}_t^{o,\perp} - \mu_*^{o,\perp}\|_{\Sigma_t^{o,\perp}} \leq \beta_t(\delta, d-1), \ \forall t \in [T] \right\}. \tag{15}$$

The regret $\mathcal{R}_\Pi(T)$ in (2) can be decomposed as ($\widetilde{r}_{\pi_t,t}$ is the optimistic reward defined by Eq. 9)

$$\mathcal{R}_\Pi(T) = \underbrace{\sum_{t=1}^{T} r_{\pi_t^*} - \widetilde{r}_{\pi_t,t}}_{\text{(I)}} + \underbrace{\sum_{t=1}^{T} \widetilde{r}_{\pi_t,t} - r_{\pi_t}}_{\text{(II)}}. \tag{16}$$

We first bound the term (II) in (16). To bound (II), we further decompose it as

$$(\text{II}) = \underbrace{\sum_{t=1}^{T} \langle x_{\pi_t}, \widetilde{\theta}_t \rangle - \langle x_t, \widetilde{\theta}_t \rangle}_{(\text{III})} + \underbrace{\sum_{t=1}^{T} \langle x_t, \widetilde{\theta}_t \rangle - \langle x_t, \theta_* \rangle}_{(\text{IV})} + \underbrace{\sum_{t=1}^{T} \langle x_t, \theta_* \rangle - \langle x_{\pi_t}, \theta_* \rangle}_{(\text{V})}. \tag{17}$$

In the following lemmas, we first bound the sum of (III) and (V) terms, and then bound (IV).

Lemma 1. *On the event $\mathcal{E}$ defined by* (15)*, for any $\gamma \in (0,1)$, w.p. at least $1 - \gamma$, we have*

$$(\text{III}) + (\text{V}) \leq \frac{2L(\alpha_r + 1)\beta_T(\delta, d)}{\sqrt{\lambda}} \cdot \sqrt{2T \log(1/\gamma)}.$$

Proof. We write $(\text{III}) + (\text{V}) = \sum_{t=1}^{T} \langle x_{\pi_t} - x_t, \widetilde{\theta}_t - \theta_* \rangle$. By Cauchy-Schwartz, we have $|\langle x_{\pi_t} - x_t, \widetilde{\theta}_t - \theta_* \rangle| \leq \|x_{\pi_t} - x_t\|_{\Sigma_t^{-1}} \|\widetilde{\theta}_t - \theta_*\|_{\Sigma_t}$. Since $\widetilde{\theta}_t \in \mathcal{C}_t^r(\alpha_r)$, on event $\mathcal{E}$, we have $\|\widetilde{\theta}_t - \theta_*\|_{\Sigma_t} \leq (\alpha_r + 1)\beta_t(\delta, d)$. Also from the definition of Σ_t, we have $\Sigma_t \succeq \lambda I$, and thus, $\|x_{\pi_t} - x_t\|_{\Sigma_t^{-1}} \leq \|x_{\pi_t} - x_t\|/\sqrt{\lambda} \leq 2L/\sqrt{\lambda}$. Therefore, $Y_t = \sum_{s=1}^{t} \langle x_{\pi_s} - x_s, \widetilde{\theta}_s - \theta_* \rangle$ is a martingale sequence with $|Y_t - Y_{t-1}| \leq 2L(\alpha_r + 1)\beta_t(\delta, d)/\sqrt{\lambda}$, for $t \in [T]$. By the Azuma–Hoeffding inequality and since β_t is an increasing function of t, i.e., $\beta_t(\delta, d) \leq \beta_T(\delta, d)$, $\forall t \in [T]$, w.p. at least $1 - \gamma$, we have $\mathbb{P}\big(Y_T \geq 2L(\alpha_r + 1)\beta_T(\delta, d)\sqrt{2T \log(1/\gamma)/\lambda}\big) \leq \gamma$, which concludes the proof. $\square$

Lemma 2. *On event $\mathcal{E}$, we have* $(\text{IV}) \leq (\alpha_r + 1)\beta_T(\delta, d)\sqrt{2Td \log\big(1 + \frac{TL^2}{\lambda}\big)}$.

We report the proof of Lemma 2 in Appendix B.1. After bounding all the terms in (II), we now process the term (I) in (16). Before stating the main result for this term in Lemma 4, we need to prove the following lemma (proof in Appendix B.2).

Lemma 3. *For any policy π, the following inequality holds:*

$$\|x_\pi^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}} \leq \|x_\pi\|_{\Sigma_t^{-1}}. \tag{18}$$

In the following lemma, we prove that by appropriately setting the parameters α_r and α_c, we can guarantee that at each round $t \in [T]$, OPLB selects an optimistic policy, i.e., a policy π_t, whose optimistic reward, $\widetilde{r}_{\pi_t,t}$, is larger than the reward of the optimal policy $r_{\pi_t^*}$, given the event $\mathcal{E}$. This means that with our choice of parameters α_r and α_c, the term (I) in (16) is always non-positive.

Lemma 4. *On the event $\mathcal{E}$, if we set α_r and α_c, such that $\alpha_r, \alpha_c \geq 1$ and $1 + \alpha_c \leq (\tau - c_0)(\alpha_r - 1)$, then for any $t \in [T]$, we have $\widetilde{r}_{\pi_t,t} \geq r_{\pi_t^*}$.*

Here we provide a proof sketch for Lemma 4. The detailed proof is reported in Appendix B.3.

Proof Sketch. We divide the proof into two cases, depending on whether in each round t, the optimal policy π_t^* belongs to the (estimated) set of feasible policies Π_t, or not.

Case yes 1. If $\pi_t^* \in \Pi_t$, then its optimistic reward is less than that of the policy π_t selected at round t (by the definition of π_t on Line 4 of Algorithm 1), i.e., $\widetilde{r}_{\pi_t^*,t} \leq \widetilde{r}_{\pi_t,t}$. This together with the fact that the optimistic reward of any policy π is larger than its expected reward, i.e., $\widetilde{r}_{\pi,t} \geq r_\pi$, gives us the desired result that $\widetilde{r}_{\pi_t,t} \geq r_{\pi_t^*}$.

Case 2. If $\pi_t^* \notin \Pi_t$, then we define a mixture policy $\widetilde{\pi}_t = \eta_t \pi_t^* + (1 - \eta_t)\pi_0$, where π_0 is the policy that always selects the safe action x_0 and $\eta_t \in [0, 1]$ is the maximum value of η for which the mixture policy belongs to the set of feasible actions, i.e., $\widetilde{\pi}_t \in \Pi_t$. Conceptually, we can think of η_t as a measure for safety of the optimal policy π_t^*. Mathematically, η_t is the value at which the pessimistic cost of the mixture policy equals to the constraint threshold, i.e., $\widetilde{c}_{\widetilde{\pi}_t,t} = \tau$. In the rest of the proof, we first write $\widetilde{c}_{\widetilde{\pi}_t,t}$ in terms of the pessimistic cost of the optimal policy as $\widetilde{c}_{\widetilde{\pi}_t,t} = (1 - \eta_t)c_0 + \eta_t \widetilde{c}_{\pi_t^*,t}$ (c_0 is the expected cost of the safe action x_0), and find a lower-bound for η_t (see Eq. 25 in Appendix B.3). We then use the fact that since $\widetilde{\pi}_t \in \Pi_t$, its optimistic reward is less than that of π_t, i.e., $\widetilde{r}_{\pi_t,t} \geq \widetilde{r}_{\widetilde{\pi}_t,t}$, and obtain a lower-bound for $\widetilde{r}_{\widetilde{\pi}_t,t}$ as a function of $r_{\pi_t^*}$ (see Eq. 26 in Appendix B.3). Finally, we conclude the proof by using this lower-bound and finding the relationship between the parameters α_r and α_c for which the desired result $\widetilde{r}_{\pi_t,t} \geq r_{\pi_t^*}$ is obtained, i.e., $1 + \alpha_c \leq (\tau - c_0)(\alpha_r - 1)$. $\square$

Proof of Theorem 2. The proof follows from the fact that the term (I) is negative (Lemma 4), and by combining the upper-bounds on the term (II) from Lemmas 1 and 2, and setting $\gamma = \delta$. $\square$

5 Constrained Multi-Armed Bandits

In this section, we specialize our results for contextual linear bandits to multi-armed bandits (MAB) and show that the structure of the MAB problem allows a computationally efficient implementation of the algorithm and an improvement in the regret bound.

In the MAB setting, the action set consists of K arms $\mathcal{A} = \{1, \ldots, K\}$. Each arm $a \in [K]$ has a reward and a cost distribution with means $\bar{r}_a, \bar{c}_a \in [0, 1]$. In each round $t \in [T]$, the agent constructs a policy π_t over $\mathcal{A}$, pulls an arm $a_t \sim \pi_t$, and observes a reward-cost pair (r_{a_t}, c_{a_t}) sampled i.i.d. from the reward and cost distributions of arm a_t. Similar to the constrained contextual linear case, the goal of the agent is to produce a sequence of policies $\{\pi_t\}_{t=1}^T$ with maximum expected cumulative reward over T rounds, i.e., $\sum_{t=1}^T \mathbb{E}_{a_t \sim \pi_t}[\bar{r}_{a_t}]$, while satisfying the ***linear constraint*** $\mathbb{E}_{a_t \sim \pi_t}[\bar{c}_{a_t}] \leq \tau, \ \forall t \in [T]$. Moreover, arm 1 is assumed to be the known safe arm, i.e., $\bar{c}_1 \leq \tau$.

Optimistic Pessimistic Bandit (OPB) Algorithm. Let $\{T_a(t)\}_{a=1}^K$ and $\{\hat{r}_a(t), \hat{c}_a(t)\}_{a=1}^K$ be the total number of times that arm a has been pulled and the estimated mean reward and cost of arm a up until round t. In each round $t \in [T]$, OPB relies on the high-probability upper-bounds on the mean reward and cost of the arms, i.e., $\{u_a^r(t), u_a^c(t)\}_{a=1}^K$, where $u_a^r(t) = \hat{r}_a(t) + \alpha_r \beta_a(t)$, $u_a^c(t) = \hat{c}_a(t) + \alpha_c \beta_a(t)$, $\beta_a(t) = \sqrt{2 \log(1/\delta')/T_a(t)}$, and constants $\alpha_r, \alpha_c \geq 1$. In order to produce a feasible policy, OPB solves the following linear program (LP) in each round $t \in [T]$:

$$\max_{\pi \in \Delta_K} \sum_{a \in \mathcal{A}} \pi_a \, u_a^r(t), \qquad \text{s.t.} \sum_{a \in \mathcal{A}} \pi_a \, u_a^c(t) \leq \tau. \qquad (19)$$

As shown in (19), OPB selects its policy by being optimistic about reward (using an upper-bound for r) and pessimistic about cost (using an upper-bound for c). We report the details of OPB and its pseudo-code (Algorithm 2) in Appendix C.1.

Computational Complexity of OPB. Unlike OPLB, whose optimization problem might be complex, OPB can be implemented extremely efficiently. Lemma 5, whose proof we report in Appendix C.2, show that (19) always has a solution (policy) with support of at most 2. This property allows us to solve (19) in closed form, without a LP solver, and implement OPB quite efficiently.

Lemma 5. *There exists a policy that solves* (19) *and has at most 2 non-zero entries.*

Regret Analysis of OPB. We prove the following regret-bound for OPB in Appendix C.3.

Theorem 3 (Regret of OPB). *Let $\delta = 4KT\delta'$, $\alpha_c = 1$, and $\alpha_r = 1 + 2/(\tau - \bar{c}_1)$. Then, with probability at least $1 - \delta$, the regret of OPB satisfies*

$$\mathcal{R}_\Pi(T) \leq \left(1 + \frac{2}{\tau - \bar{c}_1}\right) \times \left(2\sqrt{2KT \log(4KT/\delta)} + 4\sqrt{T \log(2/\delta) \log(4KT/\delta)}\right).$$

The main component in the proof of Theorem 3 is the following lemma, whose proof is reported in Appendix C.3. This lemma is the analogous to Lemma 4 in the contextual linear bandit case.

Lemma 6. *If we set the parameters α_r and α_c, such that $\alpha_r, \alpha_c \geq 1$ and $\alpha_c \leq (\tau - \bar{c}_1)(\alpha_r - 1)$, then with high probability, for any $t \in [T]$, we have $\mathbb{E}_{a \sim \pi_t}[u_a^r(t)] \geq \mathbb{E}_{a \sim \pi^*}[\bar{r}_a]$.*

Our contextual linear bandit formulation is general enough to include MAB. The regret analysis of OPLB (Theorem 2) yields a regret bound of order $\widetilde{\mathcal{O}}(\frac{K\sqrt{T}}{\tau - \bar{c}_1})$ for MAB. However, our OPB regret bound in Theorem 3 is of order $\widetilde{\mathcal{O}}(\frac{\sqrt{KT}}{\tau - \bar{c}_1})$, which shows a $\sqrt{K}$ improvement over simply casting MAB as an instance of contextual linear bandit and using the regret bound of OPLB.

Extension to m Constraints. In this case, the agent receives m cost signals after pulling each arm. The cost vector of the safe arm $\boldsymbol{c}_1$ satisfies $\boldsymbol{c}_1(i) < \tau_i, \forall i \in [m]$, where $\{\tau_i\}_{i=1}^m$ are the constraint thresholds. Similar to single-constraint OPB, multi-constraint OPB is also computationally efficient. The main reason is that the LP of m-constraint OPB has a solution with at most $m + 1$ non-zero entries. We obtain a regret bound of $\widetilde{\mathcal{O}}(\frac{\sqrt{KT}}{\min_i \tau_i - \boldsymbol{c}_1(i)})$ for m-constraint OPB in Appendix C.5.

Lower-bound. We also prove a mini-max lower-bound for this constrained MAB problem that shows no algorithm can attain a regret better than $\mathcal{O}(\max(\sqrt{KT}, \frac{1}{(\tau - \bar{c}_1)^2}))$. The formal statement of the lower-bound and the proof are reported in Appendix C.6.

6 Experiments

We run a set of experiments to show the behavior of OPB and validate our theoretical results. We consider a $K = 4$-armed bandits in which the reward and cost distributions of the arms are Bernoulli with means $\bar{r} = (.1, .2, .4, .7)$ and $\bar{c} = (0, .4, .5, .2)$. So, the cost of the safe arm is $\bar{c}_1 = 0$. In Figures 1 to 3, we gradually reduce the constraint threshold τ, and as a result the complexity of the problem $\tau - \bar{c}_1$, and show the regret *(left)* and the cost *(middle)* and reward *(right)* evolution of OPB. All the results are averaged over 10 runs and the shade is the $\pm.5$ standard deviation around the regret.

Our results show that the regret of OPB grows as we reduce τ *(left)*. They also indicate that the algorithm is successful in satisfying the constraint *(middle)* and reaching the optimal reward/performance *(right)*. In Figure 3, the reason that the cost evolution of OPB is the same as that of the optimal policy *(middle)* is that in this case, the cost of the best arm (arm 4) is equal to the constraint threshold $\tau = .2$.

7 Related Work

As described in Section 1, our setting is the closest to the one studied by Amani et al. [2019] and Moradipari et al. [2019]. They study a slightly different setting, in which the mean cost of the action that the agent takes should satisfy the constraint, i.e., $\langle x_t, \mu_* \rangle \leq \tau$, not the mean cost of the policy it computes, i.e., $\langle x_{\pi_t}, \mu_* \rangle \leq \tau$, as in our case. Clearly, the setting studied in our paper is more relaxed, and thus, is expected to obtain more rewards. Moradipari et al. [2019] propose a TS algorithm for their setting and prove an $\widetilde{O}(d^{3/2}\sqrt{T}/\tau)$ regret bound for it. They restrict themselves to linear bandits, i.e., $\mathcal{A}_t = \mathcal{A}, \forall t \in [T]$, and the safe action being the origin, i.e., $x_0 = \mathbf{0}$ and $c_0 = 0$. This is why c_0 does not appear in their bounds. They consider their action set to be any convex compact subset of $\mathbb{R}^d$ that contains the origin. Although later in their proofs, to guarantee that their algorithm does not violate the constraint
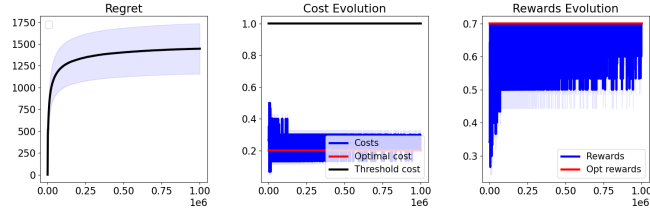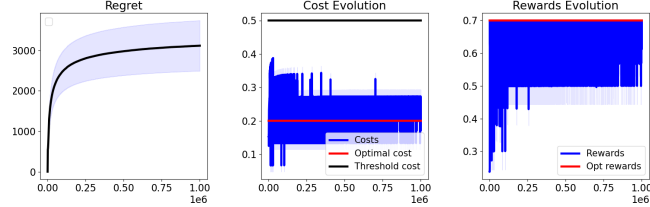

Figure 1: Constraint Threshold $\tau = 1$.
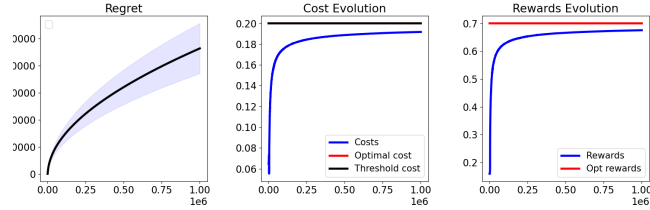

Figure 2: Constraint Threshold $\tau = 0.5$.

Figure 3: Constraint Threshold $\tau = 0.2$.

OPB. Bernoulli arms. $\bar{r} = (.1, .2, .4, .7), \bar{c} = (0, .4, .5, .2), \bar{c}_1 = 0$.

in the first round, they require the action set to also contain the ball with radius τ/S around the origin. Therefore, our action set is more general than theirs. Moreover, unlike us, their action set does not allows their results to be immediately applicable to MAB. Our regret bound also has a better dependence on d and $\log T$ than theirs, similar to the best regret results for UCB vs. TS. However, their algorithm is TS, and thus, is less complex than ours. Although it can be still intractable, even when $\mathcal{A}$ is convex. They needed to do several approximations in order to make their algorithm tractable in their experiments.

In Amani et al. [2019], reward and cost have the same unknown parameter θ_*, and the cost is defined as $c_t = x_t^\top B \theta_* \leq \tau$, where B is a known matrix. They derive and analyze an explore-exploit algorithm for this setting. Although our rate is better than theirs, i.e., $\widetilde{O}(T^{2/3})$, our algorithm cannot immediately give a $\widetilde{O}(\sqrt{T})$ regret for their setting, unless in special cases.

8 Conclusions

We derived a UCB-style algorithm for a new constrained contextual linear bandit setting, in which the goal is to produce a sequence of policies with maximum expected cumulative reward, while

each policy has an expected cost below a certain threshold τ. We proved a T-round regret bound of $\widetilde{\mathcal{O}}(\frac{d\sqrt{T}}{\tau-c_0})$ for our algorithm, which shows that the difficulty of the problem depends on the difference between the constraint threshold and the cost of a known feasible action c_0. We further specialized our results to MAB and proposed and analyzed a computationally efficient algorithm for this setting. We also proved a lower-bound for our constrained bandit problem and provided simulations to validate our theoretical results. A future direction is to use the optimism-pessimism idea behind our algorithm in other constrained bandit settings, including deriving a UCB-style algorithm for the setting studied in Amani et al. [2019] and Moradipari et al. [2019].

References

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pages 2312–2320, 2011.

S. Agrawal and N. Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the Fifteenth ACM conference on Economics and computation*, pages 989–1006, 2014.

S. Agrawal and N. Devanur. Linear contextual bandits with knapsacks. In *Advances in Neural Information Processing Systems 29*, pages 3450–3458, 2016.

S. Agrawal and N. Goyal. Further optimal regret bounds for Thompson sampling. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, pages 99–107, 2013a.

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, pages 127–135, 2013b.

S. Amani, M. Alizadeh, and C. Thrampoulidis. Linear stochastic bandits under safety constraints. In *Advances in Neural Information Processing Systems*, pages 9252–9262, 2019.

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

A. Badanidiyuru, R. Kleinberg, and A. Slivkins. Bandits with knapsacks. In *IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 207–216, 2013.

A. Badanidiyuru, J. Langford, and A. Slivkins. Resourceful contextual bandits. In *Proceedings of The 27th Conference on Learning Theory*, pages 1109–1134, 2014.

A. Balakrishnan, D. Bouneffouf, N. Mattei, and F. Rossi. Using contextual bandits with behavioral constraints for constrained online movie recommendation. In *IJCAI*, pages 5802–5804, 2018.

V. Dani, T. Hayes, and S. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 355–366, 2008.

E. Garcelon, M. Ghavamzadeh, A. Lazaric, and M. Pirotta. Improved algorithms for conservative exploration in bandits. In *AAAI*, 2020.

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

A. Kazerouni, M. Ghavamzadeh, Y. Abbasi Yadkori, and B. Van Roy. Conservative contextual linear bandits. In *Advances in Neural Information Processing Systems*, pages 3910–3919, 2017.

T. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2019.

L. Li, W. Chu, J. Langford, and R. Schapire. A contextual-bandit approach to personalized news article recommendation. In *WWW*, pages 661–670, 2010.

S. Maghsudi and E. Hossain. Multi-armed bandits with application to 5G small cells. *IEEE Wireless Communications*, 23(3):64–73, 2016.

A. Moradipari, S. Amani, M. Alizadeh, and C. Thrampoulidis. Safe linear thompson sampling with side information. *preprint arXiv:1911.02156*, 2019.

S. Ontanón. The combinatorial multi-armed bandit problem and its application to real-time strategy games. In *Ninth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2013.

P. Rusmevichientong and J. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

D. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen. A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018.

W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

S. Villar, J. Bowden, and J. Wason. Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statistical Science*, 30(2):199–215, 2015.

R. Washburn. Application of multi-armed bandits to sensor management. In *Foundations and Applications of Sensor Management*, pages 153–175. Springer, 2008.

H. Wu, R. Srikant, X. Liu, and C. Jiang. Algorithms with logarithmic or sub-linear regret for constrained contextual bandits. In *Advances in Neural Information Processing Systems 28*, pages 433–441, 2015.

Y. Wu, R. Shariff, T. Lattimore, and C. Szepesvári. Conservative bandits. In *International Conference on Machine Learning*, pages 1254–1262, 2016.

A   Proofs of Section 3

A.1   Proof of Proposition 1

Proof. We only prove the statement for the optimistic reward, $\widetilde{r}_{\pi,t}$. The proof for the pessimistic cost, $\widetilde{c}_{\pi,t}$, is analogous. From the definition of the confidence set $\mathcal{C}_t^r(\alpha_r)$ in (7), any vector $\theta \in \mathcal{C}_t^r(\alpha_r)$ can be written as $\widehat{\theta}_t + v$, where v satisfying $\|v\|_{\Sigma_t} \leq \alpha_r \beta_t(\delta, d)$. Thus, we may write

$$
\widetilde{r}_{\pi,t} = \max_{\theta \in \mathcal{C}_t^r(\alpha_r)} \mathbb{E}_{x \sim \pi}[\langle x, \theta \rangle] = \max_{\theta \in \mathcal{C}_t^r(\alpha_r)} \langle x_\pi, \theta \rangle = \langle x_\pi, \widehat{\theta}_t \rangle + \max_{v: \|v\|_{\Sigma_t} \leq \alpha_r \beta_t(\delta, d)} \langle x_\pi, v \rangle
$$

$$
\overset{(a)}{\leq} \langle x_\pi, \widehat{\theta}_t \rangle + \alpha_r \beta_t(\delta, d) \|x_\pi\|_{\Sigma_t^{-1}}.
$$

(a) By Cauchy-Schwartz, for all v, we have $\langle x_\pi, v \rangle \leq \|x_\pi\|_{\Sigma_t^{-1}} \|v\|_{\Sigma_t}$. The result follows from the condition on v in the maximum, i.e., $\|v\|_{\Sigma_t} \leq \alpha_r \beta_t(\delta, d)$.

Let us define $v^* := \frac{\alpha_r \beta_t(\delta, d) \Sigma_t^{-1} x_\pi}{\|x_\pi\|_{\Sigma_t^{-1}}}$. This value of v^* is feasible because

$$
\|v^*\|_{\Sigma_t} = \frac{\alpha_r \beta_t(\delta, d)}{\|x_\pi\|_{\Sigma_t^{-1}}} \sqrt{x_\pi^\top \Sigma_t^{-1} \Sigma_t \Sigma_t^{-1} x_\pi} = \frac{\alpha_r \beta_t(\delta, d)}{\|x_\pi\|_{\Sigma_t^{-1}}} \sqrt{x_\pi^\top \Sigma_t^{-1} x_\pi} = \alpha_r \beta_t(\delta, d).
$$

We now show that v^* also achieves the upper-bound in the above inequality resulted from Cauchy-Schwartz

$$
\langle x_\pi, v^* \rangle = \frac{\alpha_r \beta_t(\delta, d) x_\pi^\top \Sigma_t^{-1} x_\pi}{\|x_\pi\|_{\Sigma_t^{-1}}} = \alpha_r \beta_t(\delta, d) \|x_\pi\|_{\Sigma_t^{-1}}.
$$

Thus, v^* is the maximizer and we can write

$$
\widetilde{r}_{\pi,t} = \langle x_\pi, \widehat{\theta}_t \rangle + \langle x_\pi, v^* \rangle = \langle x_\pi, \widehat{\theta}_t \rangle + \alpha_r \beta_t(\delta, d) \|x_\pi\|_{\Sigma_t^{-1}},
$$

which concludes the proof. $\qquad\square$

A.2   Proof of Proposition 2

Proof. Recall that $\widetilde{c}_{\pi,t} = \frac{\langle x_\pi^o, e_0 \rangle c_0}{\|x_0\|} + \langle x_\pi^{o,\perp}, \widehat{t}_\pi^{o,\perp} \rangle + \alpha_c \beta_t(\delta, d-1) \|x_\pi^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}} \leq \tau$.

Conditioned on the event $\mathcal{E}$ as defined in equation 15, it follows that:

$$
|\langle x_\pi^{o,\perp}, \widehat{\mu}_t^{o,\perp} - \mu_*^{o,\perp} \rangle| \leq \|\mu_*^{o,\perp} - \widehat{\mu}_t^{o,\perp}\|_{\Sigma_t^{o,\perp}} \|x_\pi\|_{(\Sigma_t^{o,\perp})^{-1}}
$$

$$
\leq \langle x_\pi^{o,\perp}, \widehat{\mu}_t^{o,\perp} - \mu_*^{o,\perp} \rangle \beta_t(\delta, d-1) \|x_\pi\|_{(\Sigma_t^{o,\perp})^{-1}}
$$

And therefore:

$$
0 \leq \langle x_\pi^{o,\perp}, \widehat{\mu}_t^{o,\perp} - \mu_*^{o,\perp} \rangle + \beta_t(\delta, d-1) \|x_\pi\|_{(\Sigma_t^{o,\perp})^{-1}} \tag{20}
$$

Observe that:

$$
c_\pi = \frac{\langle x_\pi^o, e_0 \rangle c_0}{\|x_0\|} + \langle x_\pi^{o,\perp}, \mu_*^{o,\perp} \rangle
$$

$$
\leq \underbrace{\frac{\langle x_\pi^o, e_0 \rangle c_0}{\|x_0\|} + \langle x_\pi^{o,\perp}, \widehat{\mu}_t^{o,\perp} \rangle + \alpha_c \beta_t(\delta, d-1) \|x_\pi^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}}}_{\text{I}} \tag{21}
$$

The last inequality holds by adding Inequality 20 to Inequality 21. Since by assumption for all $\pi \in \Pi_t$ term $I \leq \tau$, we obtain that $c_\pi \leq \tau$. The result follows. $\qquad\square$

B Proofs of Section 4

B.1 Proof of Lemma 2

We first state the following proposition that is used in the proof of Lemma 2. This proposition is a direct consequence of Eq. 20.9 and Lemma 19.4 in Lattimore and Szepesvári [2019]. Similar result has also been reported in the appendix of Amani et al. [2019].

Proposition 3. *For any sequence of actions* $(x_1, \ldots, x_t)$*, let* Σ_t *be its corresponding Gram matrix defined by* (4) *with* $\lambda \geq 1$*. Then, for all* $t \in [T]$*, we have*

$$\sum_{s=1}^{T} \|x_s\|_{\Sigma_s^{-1}} \leq \sqrt{2Td \log\left(1 + \frac{TL^2}{\lambda}\right)}.$$

We now state the proof of Lemma 2.

Proof of Lemma 2. We prove this lemma through the following sequence of inequalities:

$$\sum_{t=1}^{T} \langle x_t, \widetilde{\theta}_t \rangle - \langle x_t, \theta_* \rangle \stackrel{(a)}{\leq} \sum_{t=1}^{T} \|x_t\|_{\Sigma_t^{-1}} \|\widetilde{\theta}_t - \theta_*\|_{\Sigma_t} \stackrel{(b)}{\leq} \sum_{t=1}^{T} (1 + \alpha_r)\beta_t(\delta, d) \|x_t\|_{\Sigma_t^{-1}}$$

$$\stackrel{(c)}{\leq} (1 + \alpha_r)\beta_T(\delta, d) \sum_{t=1}^{T} \|x_t\|_{\Sigma_t^{-1}} \stackrel{(d)}{\leq} (1 + \alpha_r)\beta_T(\delta, d) \sqrt{2Td \log\left(1 + \frac{TL^2}{\lambda}\right)}$$

(a) This is by Cauchy-Schwartz.

(b) This follows from the fact that $\widetilde{\theta}_t \in \mathcal{C}_t^r(\alpha_r)$ and we are on event $\mathcal{E}$.

(c) This is because $\beta_t(\delta, d)$ is an increasing function of t, i.e., $\beta_T(\delta, d) \geq \beta_t(\delta, d)$, $\forall t \in [T]$.

(d) This is a direct result of Proposition 3. $\qquad \square$

B.2 Proof of Lemma 3

Proof. In order to prove the desired result it is enough to show that:

$$\left(x_\pi^{o,\perp}\right)^\top \left(\Sigma_t^{o,\perp}\right)^\dagger x_\pi^{o,\perp} \leq x_\pi^\top \Sigma_t^{-1} x_\pi$$

w.l.o.g. we can assume $x_o = e_1$, the first basis vector. Notice that in this case $\Sigma_t^{o,\perp}$ can be thought of as a submatrix of Σ_t such that $\Sigma_t[2:, 2:] = \Sigma_t^{o,\perp}$, where $\Sigma_t[2:, 2:]$ denotes the submatrix with row and column indices from 2 onwards.

Using the following formula for the inverse of a psd symmetric matrix:

$$\begin{bmatrix} Z & \delta \\ \delta^\top & A \end{bmatrix} = \begin{bmatrix} \frac{1}{D} & -\frac{A^{-1}\delta}{D} \\ -\frac{\delta^\top A^{-1}}{D} & A^{-1} + \frac{A^{-1}\delta\delta^\top A^{-1}}{D} \end{bmatrix}$$

Where $D = z - \delta^\top A^{-1}\delta$. In our case $D = \Sigma_t[1,1] - \Sigma_t[2:d]^\top \left(\Sigma_t^{o,\perp}\right)^{-1} \Sigma_t[2:d] \in \mathbb{R}$. Observe that since Σ_t is PSD, $D \geq 0$. Therefore:

$$\Sigma_t^{-1} = \begin{bmatrix} 1/D & -\frac{\left(\Sigma_t^{o,\perp}\right)^{-1}\Sigma_t[2,:d]}{D} \\ -\frac{\Sigma_t^\top[2:d]\left(\Sigma_t^{o,\perp}\right)^{-1}}{D} & \left(\Sigma_t^{o,\perp}\right)^{-1} + \frac{\left(\Sigma_t^{o,\perp}\right)^{-1}\Sigma_t[2:d]\Sigma_t[2:d]\left(\Sigma_t^{o,\perp}\right)^{-1}}{D} \end{bmatrix}$$

Then:

$$x_\pi^\top \left(\Sigma_t^{-1}\right)^{-1} x_\pi = \frac{x_\pi(1)^2 - 2x_\pi(1)\Sigma_t[2:d]^\top \left(\Sigma_t^{o,\perp}\right)^{-1} x_\pi[2:d]}{D} +$$

$$\frac{x_\pi[2:d]^\top \left(\Sigma_t^{o,\perp}\right)^{-1} \Sigma_t[2:d]\Sigma_t[2:d]^\top \left(\Sigma_t^{o,\perp}\right)^{-1} x_\pi[2:d]}{D}$$

$$+ x_\pi[2:d]^\top \left(\Sigma_t^{o,\perp}\right)^{-1} x_\pi[2:d]$$

$$\geq x_\pi[2:d]^\top \left(\Sigma_t^{o,\perp}\right)^{-1} x_\pi[2:d]$$

The result follows by noting that $x_\pi[2:d] = x_\pi^{o,\perp}$. $\qquad\square$

B.3 Proof of Lemma 4

Proof. For any policy π, we have

$$\widetilde{r}_{\pi,t} = \max_{\theta \in \mathcal{C}_t^r(\alpha_r)} \langle x_\pi, \theta \rangle \geq \langle x_\pi, \theta_* \rangle = r_\pi. \tag{22}$$

If $\pi_t^* \in \Pi_t$, then by the definition of π_t (Line 4 of Algorithm 1), we have

$$\widetilde{r}_{\pi_t,t} \geq \widetilde{r}_{\pi_t^*,t}. \tag{23}$$

Combining (22) and (23), we may conclude that $\widetilde{r}_{\pi_t,t} \geq r_{\pi_t^*}$ as desired.

We now focus on the case that $\pi_t^* \notin \Pi_t$, i.e.,

$$\widetilde{c}_{\pi_t^*,t} = \frac{\langle x_{\pi_t^*}^o, e_0\rangle c_0}{\|x_0\|} + \langle x_{\pi_t^*}^{o,\perp}, \widehat{\mu}_t^{o,\perp}\rangle + \alpha_c\beta_t(\delta, d-1)\|x_{\pi_t^*}^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}} > \tau.$$

We define a mixture policy $\widetilde{\pi}_t = \eta_t\pi_t^* + (1-\eta_t)\pi_0$, where π_0 is the policy that always selects the safe action x_0 and $\eta_t \in [0,1]$ is the maximum value of η such that $\left(\eta\pi_t^* + (1-\eta)\pi_0\right) \in \Pi_t$. Conceptually, η_t shows how close is the optimal policy π_t^* to the set of safe policies Π_t.

By the definition of $\widetilde{\pi}_t$, we have

$$x_{\widetilde{\pi}_t}^o = \eta_t x_{\pi_t^*}^o + (1-\eta_t)x_0, \qquad x_{\widetilde{\pi}_t}^{o,\perp} = \eta_t x_{\pi_t^*}^{o,\perp}, \tag{24}$$

which allows us to write

$$\widetilde{c}_{\widetilde{\pi}_t,t} = \frac{\eta_t\langle x_{\pi_t^*}^o, e_0\rangle + (1-\eta_t)\langle x_0, e_0\rangle}{\|x_0\|} \cdot c_0 + \eta_t\langle x_{\pi_t^*}^{o,\perp}, \widehat{\mu}_t^{o,\perp}\rangle + \eta_t\alpha_c\beta_t(\delta, d-1)\|x_{\pi_t^*}^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}}$$

$$= \frac{(1-\eta_t)\langle x_0, e_0\rangle c_0}{\|x_0\|} + \eta_t\widetilde{c}_{\pi_t^*,t}.$$

From the definition of η_t, we have $\widetilde{c}_{\widetilde{\pi}_t,t} = \frac{(1-\eta_t)\langle x_0,e_0\rangle c_0}{\|x_0\|} + \eta_t\widetilde{c}_{\pi_t^*,t} = \tau$, and thus, we may write

$$\eta_t = \frac{\tau - \frac{\langle x_0,e_0\rangle c_0}{\|x_0\|}}{\widetilde{c}_{\pi_t^*,t} - \frac{\langle x_0,e_0\rangle c_0}{\|x_0\|}} = \frac{\tau - c_0}{\frac{\langle x_{\pi_t^*}^o,e_0\rangle c_0}{\|x_0\|} + \langle x_{\pi_t^*}^{o,\perp}, \widehat{\mu}_t^{o,\perp}\rangle + \alpha_c\beta_t(\delta, d-1)\|x_{\pi_t^*}^{o,\perp}\|_{(\Sigma_t^{0,\perp})^{-1}} - c_0}$$

$$= \frac{\tau - c_0}{\frac{\langle x_{\pi_t^*}^o,e_0\rangle c_0}{\|x_0\|} + \langle x_{\pi_t^*}^{o,\perp}, \mu_*\rangle + \langle x_{\pi_t^*}^{o,\perp}, \widehat{\mu}_t^{o,\perp} - \mu_*\rangle + \alpha_c\beta_t(\delta, d-1)\|x_{\pi_t^*}^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}} - c_0}$$

$$\overset{(a)}{\geq} \frac{\tau - c_0}{\frac{\langle x_{\pi_t^*}^o,e_0\rangle c_0}{\|x_0\|} + \langle x_{\pi_t^*}^{o,\perp}, \mu_*\rangle + (1+\alpha_c)\beta_t(\delta, d-1)\|x_{\pi_t^*}^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}} - c_0}$$

$$\overset{(b)}{\geq} \frac{\tau - c_0}{\tau + (\alpha_c + 1)\beta_t(\delta, d-1)\|x_{\pi_t^*}^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}} - c_0}. \tag{25}$$

13

(a) This holds because

$$\langle x^{o,\perp}_{\pi^*_t}, \widehat{\mu}^{o,\perp}_t - \mu_* \rangle = \langle x^{o,\perp}_{\pi^*_t}, \widehat{\mu}^{o,\perp}_t - \mu^{o,\perp}_* \rangle \leq \|\widehat{\mu}^{o,\perp}_t - \mu^{o,\perp}_*\|_{\Sigma^{o,\perp}_t} \|x^{o,\perp}_{\pi^*_t}\|_{(\Sigma^{o,\perp}_t)^{-1}} \leq \beta_t(\delta, d-1)\|x^{o,\perp}_{\pi^*_t}\|_{(\Sigma^{o,\perp}_t)^{-1}},$$

where the last inequality is because we are on the event $\mathcal{E}$.

(b) This passage is due to the fact that the optimal policy π^*_t is feasible, and thus, $\mathbb{E}_{x \sim \pi^*_t}[\langle x, \mu_* \rangle] \leq \tau$. Therefore, we may write

$$\mathbb{E}_{x \sim \pi^*_t}[\langle x, \mu_* \rangle] = \mathbb{E}_{x \sim \pi^*_t}[\langle x^o, \mu_* \rangle] + \langle x^{o,\perp}_{\pi^*_t}, \mu_* \rangle = \mathbb{E}_{x \sim \pi^*_t}[\langle \langle x, e_0 \rangle e_0, \mu_* \rangle] + \langle x^{o,\perp}_{\pi^*_t}, \mu_* \rangle$$

$$= \mathbb{E}_{x \sim \pi^*_t}[\langle \langle x, e_0 \rangle \frac{x_0}{\|x_0\|}, \mu_* \rangle] + \langle x^{o,\perp}_{\pi^*_t}, \mu_* \rangle = \frac{c_0}{\|x_0\|}\mathbb{E}_{x \sim \pi^*_t}[\langle x, e_0 \rangle] + \langle x^{o,\perp}_{\pi^*_t}, \mu_* \rangle$$

$$= \frac{\langle x^o_{\pi^*_t}, e_0 \rangle c_0}{\|x_0\|} + \langle x^{o,\perp}_{\pi^*_t}, \mu_* \rangle \leq \tau.$$

Since $\widetilde{\pi}_t \in \Pi_t$, we have

$$\widetilde{r}_{\pi_t, t} \geq \widetilde{r}_{\widetilde{\pi}_t, t} = \langle x_{\widetilde{\pi}_t}, \widehat{\theta}_t \rangle + \alpha_r \beta_t(\delta, d)\|x_{\widetilde{\pi}_t}\|_{\Sigma^{-1}_t} = \langle x_{\widetilde{\pi}_t}, \theta_* \rangle + \langle x_{\widetilde{\pi}_t}, \widehat{\theta}_t - \theta_* \rangle + \alpha_r \beta_t(\delta, d)\|x_{\widetilde{\pi}_t}\|_{\Sigma^{-1}_t}$$

$$\overset{(a)}{\geq} \langle x_{\widetilde{\pi}_t}, \theta_* \rangle + (\alpha_r - 1)\beta_t(\delta, d)\|x_{\widetilde{\pi}_t}\|_{\Sigma^{-1}_t} \overset{(b)}{\geq} \langle x_{\widetilde{\pi}_t}, \theta_* \rangle + (\alpha_r - 1)\beta_t(\delta, d-1)\|x^{o,\perp}_{\widetilde{\pi}_t}\|_{(\Sigma^{o,\perp}_t)^{-1}}$$

$$\overset{(c)}{=} \eta_t\langle x_{\pi^*}, \theta_* \rangle + (1 - \eta_t)\langle x_0, \theta_* \rangle + \eta_t(\alpha_r - 1)\beta_t(\delta, d-1)\|x^{o,\perp}_{\pi^*_t}\|_{(\Sigma^{o,\perp}_t)^{-1}}$$

$$\overset{(d)}{\geq} \eta_t\langle x_{\pi^*_t}, \theta_* \rangle + \eta_t(\alpha_r - 1)\beta_t(\delta, d-1)\|x^{o,\perp}_{\pi^*}\|_{(\Sigma^{o,\perp}_t)^{-1}}$$

$$\overset{(e)}{\geq} \underbrace{\left(\frac{\tau - c_0}{\tau - c_0 + (\alpha_c + 1)\beta_t(\delta, d-1)\|x^{o,\perp}_{\pi^*_t}\|_{(\Sigma^{o,\perp}_t)^{-1}}} \right)\left(\langle x_{\pi^*_t}, \theta_* \rangle + (\alpha_r - 1)\beta_t(\delta, d-1)\|x^{o,\perp}_{\pi^*_t}\|_{(\Sigma^{o,\perp}_t)^{-1}} \right)}_{C_0}.$$

$$\tag{26}$$

(a) This is because we may write

$$|\langle x_{\widetilde{\pi}_t}, \widehat{\theta}_t - \theta_* \rangle| \leq \|\widehat{\theta}_t - \theta_*\|_{\Sigma_t}\|x_{\widetilde{\pi}_t}\|_{\Sigma^{-1}_t} \leq \beta_t(\delta, d)\|x_{\widetilde{\pi}_t}\|_{\Sigma^{-1}_t},$$

where the last inequality is due to the fact that we are on the event $\mathcal{E}$. Thus, $\langle x_{\widetilde{\pi}_t}, \widehat{\theta}_t - \theta_* \rangle \geq -\beta_t(\delta, d)\|x_{\widetilde{\pi}_t}\|_{\Sigma^{-1}_t}$.

(b) This is a consequence of Lemma 3 stated in the paper and proved in Appendix B.2.

(c) This is from the definition of $\widetilde{\pi}$ and Eq. 24.

(d) This is because $\eta_t \in [0, 1]$ and from Assumption 4 we have that all expected rewards are positive (belong to $[0, 1]$), and thus, $\langle x_0, \theta_* \rangle \geq 0$.

(e) This is by lower-bounding η_t from (25).

Let us define the shorthand notation $C_1 := \beta_t(\delta, d-1)\|x^{o,\perp}_{\pi^*_t}\|_{(\Sigma^{o,\perp}_t)^{-1}}$. Thus, we may write C_0 as

$$C_0 = \frac{\tau - c_0}{\tau - c_0 + (1 + \alpha_c)C_1} \times \left(\langle x_{\pi^*_t}, \theta_* \rangle + (\alpha_r - 1)C_1 \right).$$

Note that $C_0 \geq \langle x_{\pi^*_t}, \theta_* \rangle = r_{\pi^*_t}$ (and as a results $\widetilde{r}_{\pi_t, t} \geq r_{\pi^*_t}$ as desired) iff:

$$(\tau - c_0)r_{\pi^*_t} + (\tau - c_0)(\alpha_r - 1)C_1 \geq (\tau - c_0)r_{\pi^*_t} + (1 + \alpha_c)C_1 r_{\pi^*_t},$$

which holds iff: $(\tau - c_0)(\alpha_r - 1)C_1 \geq (1 + \alpha_c)C_1 r_{\pi^*_t}$.

Since $r_{\pi^*_t} \leq 1$ from Assumption 4, this holds iff: $1 + \alpha_c \leq (\tau - c_0)(\alpha_r - 1)$. This concludes the proof as for both cases of $\pi^*_t \in \Pi_t$ and $\pi^*_t \notin \Pi_t$, we proved that $\widetilde{r}_{\pi_t, t} \geq r_{\pi^*_t}$. $\qquad\square$

B.4 Learning the safe policy's value

In this section we relax Assumption 5, and instead assume we only have the knowledge of a safe arm, but not any knowledge of its value c_0.

If the cost of the safe arm c_0 is unknown, we start by taking the safe action x_0 for T_0 rounds to produce first an empirical mean estimator for $\hat{c}_9$. Notice that for all $\delta \in (0,1)$, $\hat{c}_0$ satisfies:

$$\mathbb{P}\left(\hat{c}_0 \leq c_0 - \sqrt{\frac{2\log(1/\delta)}{T_0}}\right) \leq \delta \tag{27}$$

Let $\tilde{c}_0 = \hat{c}_0 + \sqrt{\frac{2\log(1/\delta)}{T_0}}$. By inequality 27, it follows that with probability at least $1 - \delta$:

$$\tilde{c}_0 \geq c_0$$

We select T_0 in an adaptive way. In other words, we do the following:

Let $\delta = \frac{1}{T^2}$. And let $\hat{c}_0(t)$ be the sample mean estimator of c_0, when using only t samples. Similarly define $\tilde{c}_0(t) = \hat{c}_0(t) + \sqrt{\frac{2\log(1/\delta)}{t}}$ Let's condition on the event $\mathcal{E}$ that for all $t \in [T]$:

$$|\hat{c}_0(t) - c_0| \leq \sqrt{\frac{2\log(1/\delta)}{t}}$$

By assumption $\mathbb{P}(\mathcal{E}) \geq 1 - T2\delta = 1 - \frac{2}{T}$. Let T_0 be the first time that $\tilde{c}_0(T_0) + 2\sqrt{\frac{2\log(1/\delta)}{T_0}} \leq \tau$.

Notice that in this case and conditioned on $\mathcal{E}$ and therefore on $\tilde{c}_0(T_0) \geq c_0$:

$$\sqrt{\frac{2\log(1/\delta)}{T_0}} \leq \frac{\tau - c_0}{2} \quad \text{i.e.} \quad T_0 \geq \frac{8\log(1/\delta)}{(\tau - c_0)^2}$$

In other words, this test does not stop until $T_0 \geq \frac{8\log(1/\delta)}{(\tau-c_0)^2}$. Now we see it won't take much longer than that to stop:

Conversely, let $T_0' \geq \frac{32\log(1/\delta)}{(\tau-c_0)^2}$. For any such T_0' we observe that by conditioning on $\mathcal{E}$:

$$\tilde{c}_0(T_0') + 2\sqrt{\frac{2\log(1/\delta)}{T_0'}} \leq c_0 + 4\sqrt{\frac{2\log(1/\delta)}{T_0'}} \leq \tau$$

Thus conditioned on $\mathcal{E}$, we conclude $\frac{8\log(1/\delta)}{(\tau-c_0)^2} \leq T_0 \leq \frac{32\log(1/\delta)}{(\tau-c_0)^2}$. Then,

Therefore $\hat{\delta}_c = \sqrt{\frac{8\log(1/\delta)}{T_0}}$ would serve as a conservative estimator for $\frac{\tau-c_0}{2}$ satisfying:

$$\frac{\tau - c_0}{2} \leq \hat{\delta}_c \leq \tau - c_0$$

We proceed by warm starting our estimators for θ_* and μ_* using the data collected by playing x_0. However, instead of estimating $\mu_*^{o,\perp}$, we build an estimator for μ_* over all its directions, including e_0, similar to what OPLB does for θ_*. We then set $\frac{\alpha_r}{\alpha_c} = 1/\hat{\delta}_c$ and run Algorithm 1 for rounds $t > T_0$. Since the scaling of α_r w.r.t. α_c is optimal up to constants, the same arguments hold.

C Constrained Multi-Armed Bandits

C.1 Optimism Pessimism

Here we reproduce the full pseudo-code for OPB:

Algorithm 2 Optimism-Pessimism

Input: Number of arms K, constants $\alpha_r, \alpha_c \geq 1$.
for $t = 1, \ldots, T$ **do**
> 1. Compute estimates $\{u_a^r(t)\}_{a \in \mathcal{A}}$, $\{u_a^c(t)\}_{a \in \mathcal{A}}$.
> 2. Form the approximate LP (19) using these estimates.
> 3. Find policy π_t by solving (19).
> 4. Play arm $a \sim \pi_t$

Similar to the case of OPLB, we define $\Pi_t = \{\pi \in \Delta_{\mathcal{A}} : \sum_{a \in \mathcal{A}} \pi_a u_a^c(t) \leq \tau\}$. We also define $\beta_a(0) = 0$ for all $a \in \mathcal{A}$.

C.2 The LP Structure

The main purpose of this section is to prove the optimal solutions of the linear program from (19) are supported on a set of size at most 2. This structural result will prove important to develop simple efficient algorithms to solve for solving it. Let's recall the form of the Linear program in 19 is:

$$\max_{\pi \in \Delta_K} \sum_{a \in \mathcal{A}} \pi_a u_a^r(t)$$

$$\text{s.t.} \sum_{a \in \mathcal{A}} \pi_a u_a^c(t) \leq \tau$$

Let's start by observing that in the case $K = 2$ with $\mathcal{A} = \{a_1, a_2\}$ and $u_{a_1}^c(t) < \tau < u_{a_2}^c(t)$, the optimal policy π^* is a mixture policy satisfying:

$$\pi_{a_1}^* = \frac{u_{a_2}^c(t) - \tau}{u_{a_2}^c(t) - u_{a_1}^c(t)}$$

$$\pi_{a_2}^* = \frac{\tau - u_{a_1}^c(t)}{u_{a_2}^c(t) - u_{a_1}^c(t)} \tag{28}$$

The main result in this section is the following Lemma:

Lemma 7 (π^* support). *If (19) is feasible, there exists an optimal solution with at most 2 non-zero entries.*

Proof. We start by inspecting the dual problem of (19):

$$\min_{\lambda \geq 0} \max_a \lambda(\tau - u_a^c(t)) + u_a^r(t) \tag{D}$$

This formulation is easily interpretable. The quantity $\tau - u_a^c(t)$ measures the feasibility gap of arm a, while $u_a^r(t)$ introduces a dependency on the reward signal. Let λ^* be the optimal value of the dual variable λ. Define $\mathcal{A}^* \subseteq \mathcal{A}$ as $\mathcal{A}^* = \arg\max_a \lambda^*(\tau - u_a^c(t)) + u_a^r(t)$. By complementary slackness the set of nonzero entries of π^* must be a subset of $\mathcal{A}^*$.

If $|\mathcal{A}^*| = 1$, complementary slackness immediately implies the desired result. If a_1, a_2 are two elements of $\mathcal{A}^*$, it is easy to see that:

$$u_{a_1}^r(t) - \lambda^* u_{a_1}^c(t) = u_{a_2}^r(t) - \lambda^* u_{a_2}^c(t),$$

and thus,

$$\lambda^* = \frac{u_{a_2}^r(t) - u_{a_1}^r(t)}{u_{a_2}^c(t) - u_{a_1}^c(t)} \tag{29}$$

If $\lambda^* = 0$, the optimal primal value is achieved by concentrating all mass on any of the arms in $\mathcal{A}^*$. Otherwise, plugging 29 back into the objective of (D) and rearranging the terms, we obtain

$$s$$

$$\begin{aligned}
\text{(D)} &= \lambda^*(\tau - u_{a_1}^c(t)) + u_{a_1}^r(t) \\
&= u_{a_1}^r(t)\left(\frac{\tau - u_{a_1}^c(t)}{u_{a_2}^c(t) - u_{a_1}^c(t)}\right) + u_{a_2}^r(t)\left(\frac{u_{a_2}^c(t) - \tau}{u_{a_2}^c(t) - u_{a_1}^c(t)}\right).
\end{aligned}$$

If $u_{a_2}^c(t) \geq \tau \geq u_{a_1}^c(t)$, we obtain a feasible value for the primal variable $\pi_{a_1}^* = \frac{\tau - u_{a_1}^c(t)}{u_{a_2}^c(t) - u_{a_1}^c(t)}$, $\pi_{a_2}^* = \frac{u_{a_2}^c(t) - \tau}{u_{a_2}^c(t) - u_{a_1}^c(t)}$ and zero for all other $a \in \mathcal{A}\backslash\{a_1, a_2\}$. Since we have assumed (19) to be feasible there must be either one arm $a^* \in \mathcal{A}^*$ satisfying $a^* = \arg\max_{a \in \mathcal{A}^*} u_a^r(t)$ and $u_{a^*}^c(t) \leq \tau$ or two such arms a_1 and a_2 in $\mathcal{A}^*$ that satisfy $u_{a_2}^c(t) \geq \tau \geq u_{a_1}^c(t)$, since otherwise it would be impossible to produce a feasible primal solution without having any of its supporting arms a satisfying $u_a^c(t) \leq \tau$, there must exist an arm $a \in \mathcal{A}^*$ with $u_a^c(t) < \tau$. This completes the proof. $\qquad\square$

From the proof of Lemma 5 we can conclude the optimal policy is either a delta mass centered at the arm with the largest reward - whenever this arm is feasible - or it is a strict mixture supported on two arms.

A further consequence of Lemma 7 is that it is possible to find the optimal solution π^* to problem 19 by simply enumerating all pairs of arms (a_i, a_j) and all singletons, compute their optimal policies (if feasible) using Equation 28 and their values and selecting the feasible pair (or singleton) achieving the largest value. More sophisticated methods can be developed by taking into account elimination strategies to prune out arms that can be determined in advance not to be optimal nor to belong to an optimal pair. Overall this method is more efficient than running a linear programming solver on (19).

If we had instead m constraints, a similar statement to Lemma 5 holds, namely it is possible to show the optimal policy will have support of size at most $m + 1$. The proof is left as an exercise for the reader.

C.3 Regret analysis

In order to show a regret bound for Algorithm 2, we start with the following regret decomposition:

$$\begin{aligned}
\mathcal{R}_\Pi(T) &= \sum_{t=1}^T \mathbb{E}_{a \sim \pi^*}[\bar{r}_a] - \mathbb{E}_{a \sim \pi_t}[\bar{r}_a] \\
&= \underbrace{\left(\sum_{t=1}^T \mathbb{E}_{a \sim \pi^*}[\bar{r}_a] - \mathbb{E}_{a \sim \pi_t}[u_a^r(t)]\right)}_{(i)} + \underbrace{\left(\sum_{t=1}^T \mathbb{E}_{a \sim \pi_t}[u_a^r(t)] - \mathbb{E}_{a \sim \pi_t}[\bar{r}_a]\right)}_{(ii)}.
\end{aligned}$$

In order to bound $\mathcal{R}_\Pi(T)$, we independently bound terms (i) and (ii).

We start by bounding term (i). We proceed by first proving an Lemma 6, the equivalent version of Lemma 4 for the multi armed bandit problem.

C.4 Proof of Lemma 6

Proof. Throughout this proof we denote as π_0 to the delta function over the safe arm 1. We start by noting that under $\mathcal{E}$, and because $\alpha_r, \alpha_c \geq 1$, then:

$$(\alpha_r - 1)\beta_a(t) \leq \xi_a^r(t) \leq (\alpha_r + 1)\beta_a(t) \; \forall a \quad \text{and} \quad (\alpha_c - 1)\beta_a(t) \leq \xi_a^c(t) \leq (\alpha_c + 1)\beta_a(t) \; \forall a \neq 0. \tag{30}$$

If $\pi^* \in \Pi_t$, it immediately follows that:

$$\mathbb{E}_{a \sim \pi^*}[\bar{r}_a] \leq \mathbb{E}_{a \sim \pi^*}[u_a^r(t)] \leq \mathbb{E}_{a \sim \pi_t}[u_a^r(t)]. \tag{31}$$

Let's now assume $\pi^* \notin \Pi_t$, i.e., $\mathbb{E}_{a\sim\pi^*}[u_a^c(t)] > \tau$. Let $\pi^* = \rho^*\bar{\pi}^* + (1-\rho)\pi_0$ with $\bar{\pi}^* \in \Delta_K[2 : K]^5$.

Consider a mixture policy $\widetilde{\pi}_t = \gamma_t\pi^* + (1-\gamma_t)\pi_0 = \gamma_t\rho^*\bar{\pi}^* + (1-\gamma_t\rho^*)\pi_0$, where γ_t is the maximum $\gamma_t \in [0,1]$ such that $\widetilde{\pi}_t \in \Pi_t$. It can be easily established that

$$
\gamma_t = \frac{\tau - \bar{c}_1}{\rho^*\mathbb{E}_{a\sim\bar{\pi}^*}[u_a^c(t)] - \rho^*\bar{c}_1} = \frac{\tau - \bar{c}_1}{\mathbb{E}_{a\sim\bar{\pi}^*}[\rho^*(\bar{c}_a + \xi_a^c(t))] - \rho^*\bar{c}_1}
$$
$$
\overset{(i)}{\geq} \frac{\tau - \bar{c}_1}{\tau - \bar{c}_1 + \rho^*(1+\alpha_c)\mathbb{E}_{a\sim\bar{\pi}^*}[\beta_a(t)]}.
$$

(i) is a consequence of (30) and of the observation that since π^* is feasible $\rho^*\mathbb{E}_{a\sim\bar{\pi}^*}[\bar{c}_a] + (1-\rho^*)\bar{c}_1 \leq \tau$. Since $\widetilde{\pi}_t \in \Pi_t$, we have

$$
\mathbb{E}_{a\sim\pi_t}[u_a^r(t)] \geq \underbrace{\gamma_t\mathbb{E}_{a\sim\pi^*}[u_a^r(t)] + (1-\gamma_t)u_0^r(t)}_{\mathbb{E}_{a\sim\widetilde{\pi}_t}[u_a^r(t)]}
$$
$$
\overset{(ii)}{\geq} \frac{\tau - \bar{c}_1}{\tau - \bar{c}_1 + \rho^*(1+\alpha_c)\mathbb{E}_{a\sim\bar{\pi}^*}[\beta_a(t)]} \times \mathbb{E}_{a\sim\pi^*}[u_a^r(t)]
$$
$$
= \frac{\tau - \bar{c}_1}{\tau - \bar{c}_1 + \rho^*(1+\alpha_c)\mathbb{E}_{a\sim\bar{\pi}^*}[\beta_a(t)]} \times \left(\mathbb{E}_{a\sim\pi^*}[\bar{r}_a] + \mathbb{E}_{a\sim\pi^*}[\xi_a^r(t)]\right)
$$
$$
\overset{(iii)}{\geq} \frac{\tau - \bar{c}_1}{\tau - \bar{c}_1 + \rho^*(1+\alpha_c)\mathbb{E}_{a\sim\bar{\pi}^*}[\beta_a(t)]} \times \left(\mathbb{E}_{a\sim\pi^*}[\bar{r}_a] + (\alpha_r - 1)\mathbb{E}_{a\sim\pi^*}[\beta_a(t)]\right)
$$
$$
\overset{(iv)}{\geq} \underbrace{\frac{\tau - \bar{c}_1}{\tau - \bar{c}_1 + (1+\alpha_c)\mathbb{E}_{a\sim\pi^*}[\beta_a(t)]} \times \left(\mathbb{E}_{a\sim\pi^*}[\bar{r}_a] + (\alpha_r - 1)\mathbb{E}_{a\sim\pi^*}[\beta_a(t)]\right)}_{C_0}.
$$

(ii) holds because $u_0^r(t) \geq 0$. **(iii)** is a consequence of (30) and **(iv)** follows because $\mathbb{E}_{a\sim\pi^*}[\beta_a(t)] = \rho^*\mathbb{E}_{a\sim\bar{\pi}^*}[\beta_a(t)] + (1-\rho^*)\beta_0(t) \geq \rho^*\mathbb{E}_{a\sim\bar{\pi}^*}[\beta_a(t)]$ since $\beta_a(t) \geq 0$ for all a and t.

Let $C_1 = \mathbb{E}_{a\sim\pi^*}[\beta_a(t)]$. The following holds:

$$
C_0 = \frac{\tau - \bar{c}_1}{\tau - \bar{c}_1 + (1+\alpha_c)C_1} \times \left(\mathbb{E}_{a\sim\pi^*}[\bar{r}_a] + (\alpha_r - 1)C_1\right).
$$

Note that $C_0 \geq \mathbb{E}_{a\sim\pi^*}[\bar{r}_a]$ iff:

$$
(\tau - \bar{c}_1)\mathbb{E}_{a\sim\pi^*}[\bar{r}_a] + (\tau - \bar{c}_1)(\alpha_r - 1)C_1 \geq (\tau - \bar{c}_1)\mathbb{E}_{a\sim\pi^*}[\bar{r}_a] + (1+\alpha_c)C_1\mathbb{E}_{a\sim\pi^*}[\bar{r}_a],
$$

which holds iff:

$$
(\tau - \bar{c}_1)(\alpha_r - 1)C_1 \geq (1+\alpha_c)C_1\mathbb{E}_{a\sim\pi^*}[\bar{r}_a].
$$

Since $\mathbb{E}_{a\sim\pi^*}[\bar{r}_a] \leq 1$, this holds if $1 + \alpha_c \leq (\tau - \bar{c}_1)(\alpha_r - 1)$. $\qquad\square$

Proposition 4. *If $\delta = \frac{\epsilon}{4KT}$ for $\epsilon \in (0,1)$, $\alpha_r, \alpha_c \geq 1$ with $\alpha_c \leq \tau(\alpha_r - 1)$, then with probability at least $1 - \frac{\epsilon}{2}$, we have*

$$
\sum_{t=1}^{T} \mathbb{E}_{a\sim\pi^*}[\bar{r}_a] - \mathbb{E}_{a\sim\pi_t}[u_a^r(t)] \leq 0
$$

Proof. A simple union bound implies that $\mathbb{P}(\mathcal{E}) \geq 1 - \frac{\epsilon}{2}$. Combining this observation with Lemma 6 yields the result. $\qquad\square$

Term (ii) can be bound using the confidence intervals radii:

Proposition 5. *If $\delta = \frac{\epsilon}{4KT}$ for an $\epsilon \in (0,1)$, then with probability at least $1 - \frac{\epsilon}{2}$, we have*

$$
\sum_{t=1}^{T} \mathbb{E}_{a\sim\pi_t}[u_a^r(t)] - \mathbb{E}_{a\sim\pi_t}[\bar{r}_a] \leq (\alpha_r + 1)\left(2\sqrt{2TK\log(1/\delta)} + 4\sqrt{T\log(2/\epsilon)\log(1/\delta)}\right)
$$

---

[5]In other words, the support of $\bar{\pi}^*$ does not contain the safe arm 1.

Proof. Under these conditions $\mathbb{P}(\mathcal{E}) \geq 1 - \frac{\epsilon}{2}$. Recall $u_a^r(t) = \hat{r}_a(t) + \alpha_r \beta_a(t)$ and that conditional on $\mathcal{E}$, $\bar{r}_a \in [\hat{r}_a(t) - \beta_a(t), \hat{r}_a(t) + \beta_a(t)]$ for all $t \in [T]$ and $a \in \mathcal{A}$. Thus, for all t, we have

$$\mathbb{E}_{a \sim \pi_t}[u_a^r(t)] - \mathbb{E}_{a \sim \pi_t}[\bar{r}_a] \leq (\alpha_r + 1)\mathbb{E}_{a \sim \pi_t}[\beta_a(t)].$$

Let $\mathcal{F}_{t-1}$ be the sigma algebra defined up to the choice of π_t and a'_t be a random variable distributed as $\pi_t \mid \mathcal{F}_{t-1}$ and conditionally independent from a_t, i.e., $a'_t \perp a_t \mid \mathcal{F}_{t-1}$. Note that by definition the following equality holds:

$$\mathbb{E}_{a \sim \pi_t}[\beta_a(t)] = \mathbb{E}_{a'_t \sim \pi_t}[\beta_a(t) \mid \mathcal{F}_{t-1}].$$

Consider the following random variables $A_t = \mathbb{E}_{a'_t \sim \pi_t}[\beta_{a'_t}(t) \mid \mathcal{F}_{t-1}] - \beta_{a_t}(t)$. Note that $M_t = \sum_{i=1}^{t} A_i$ is a martingale. Since $|A_t| \leq 2\sqrt{2\log(1/\delta)}$, a simple application of Azuma-Hoeffding[6] implies:

$$\mathbb{P}\left(\underbrace{\sum_{t=1}^{T} \mathbb{E}_{a \sim \pi_t}[\beta_a(t)] \geq \sum_{t=1}^{T} \beta_{a_t}(t) + 4\sqrt{T\log(2/\epsilon)\log(1/\delta)}}_{\mathcal{E}_A^c}\right) \leq \epsilon/2.$$

We can now upper-bound $\sum_{t=1}^{T} \beta_{a_t}(t)$. Note that $\sum_{t=1}^{T} \beta_{a_t}(t) = \sum_{a \in \mathcal{A}} \sum_{t=1}^{T} \mathbf{1}\{a_t = a\}\beta_a(t)$. We start by bounding for an action $a \in \mathcal{A}$:

$$\sum_{t=1}^{T} \mathbf{1}\{a_t = a\}\beta_a(t) = \sqrt{2\log(1/\delta)} \sum_{t=1}^{T_a(T)} \frac{1}{\sqrt{t}} \leq 2\sqrt{2T_a(T)\log(1/\delta)}.$$

Since $\sum_{a \in \mathcal{A}} T_a(T) = T$ and by concavity of $\sqrt{\cdot}$, we have

$$\sum_{a \in \mathcal{A}} 2\sqrt{2T_a(T)\log(1/\delta)} \leq 2\sqrt{2TK\log(1/\delta)}.$$

Conditioning on the event $\mathcal{E} \cap \mathcal{E}_A$ whose probability satisfies $\mathbb{P}(\mathcal{E} \cap \mathcal{E}_A) \geq 1 - \epsilon$ yields the result. $\square$

We can combine these two results into our main theorem:

Theorem 4 (Main Theorem). *If $\epsilon \in (0, 1)$, $\alpha_c = 1$ and $\alpha_r = \frac{2}{\tau - \bar{c}_1} + 1$, then with probability at least $1 - \epsilon$, Algorithm 2 satisfies the following regret guarantee:*

$$\mathcal{R}_\Pi(T) \leq \left(\frac{2}{\tau - \bar{c}_1} + 1\right)\left(2\sqrt{2TK\log(4KT/\epsilon)} + 4\sqrt{T\log(2/\epsilon)\log(4KT/\epsilon)}\right)$$

Proof. This result is a direct consequence of Propositions 4 and 5 by setting $\delta = 4KT\epsilon$. $\square$

C.5 Multiple constraints

We consider the problem where the learner must satisfy M constraints with threshold values $\tau_1, \cdots, \tau_M$. Borrowing from the notation in the previous sections, we denote by as $\{\bar{r}_a\}_{a \in \mathcal{A}}$ the mean reward signals and $\{\bar{c}_a^{(i)}\}$ the mean cost signals for $i = 1, \cdots, M$. The full information optimal policy can be obtained by solving the following linear program:

$$\max_{\pi \in \Delta_K} \sum_{a \in \mathcal{A}} \pi_a \bar{r}_a, \tag{P-M}$$

$$\text{s.t. } \sum_{a \in \mathcal{A}} \pi_a \bar{c}_a^{(i)} \leq \tau_i \text{ for } i = 1, \cdots, M.$$

In order to ensure the learner's ability to produce a feasible policy at all times, we make the following assumption:

Assumption 6. *The learner has knowledge of $\bar{c}_1^{(i)} < \tau_i$ for all $i = 1, \cdots, M$.*

---

[6]We use the following version of Azuma-Hoeffding: if $X_n, n \geq 1$ is a martingale such that $|X_i - X_{i-1}| \leq d_i$, for $1 \leq i \leq n$, then for every $n \geq 1$, we have $\mathbb{P}(X_n > r) \leq \exp\left(-\frac{r^2}{2\sum_{i=1}^{n} d_i^2}\right)$.

We denote by $\{\widehat{r}_a\}_{a\in\mathcal{A}}$ and $\{\widehat{c}_a^{(i)}\}_{a\in\mathcal{A}}$ for $i = 1, \cdots, M$ the empirical means of the reward and cost signals. We call $\{u_a^r(t)\}_{a\in\mathcal{A}}$ to the upper confidence bounds for our reward signal and $\{u_a^c(t,i)\}_{a\in\mathcal{A}}$ for $i = 1, \cdots, M$ the costs' upper confidence bounds:

$$u_a^r(t) = \widehat{r}_a(t) + \alpha_r \beta_a(t), \qquad u_a^c(t,i) = \widehat{c}_a^{(i)}(t) + \alpha_c \beta_a(t),$$

where $\beta_a(t) = \sqrt{2\log(1/\delta)/T_a(t)}$, $\delta \in (0,1)$ as before. A straightforward extension of Algorithm 2 considers instead the following $M-$constraints LP:

$$\max_{\pi\in\Delta_K} \quad \sum_{a\in\mathcal{A}} \pi_a\, u_a^r(t) \hspace{4cm} (\widehat{P-M})$$

$$\text{s.t.} \quad \sum_{a\in\mathcal{A}} \pi_a\, u_a^c(t,i) \leq \tau_i, \text{ for } i = 1, \cdots, M.$$

We now generalize Lemma 6:

Lemma 8. *Let $\alpha_r, \alpha_c \geq 1$ satisfying $\alpha_c \leq \min_i(\tau_i - \bar{c}_1^{(i)})(\alpha_r - 1)$. Conditioning on $\mathcal{E}_a(t)$ ensures that with probability $1 - \delta$:*

$$\mathbb{E}_{a\sim\pi_t}\left[u_a^r(t)\right] \geq \mathbb{E}_{a\sim\pi^*}\left[\bar{r}_a\right].$$

Proof. The same argument as in the proof of Lemma 6 follows through, the main ingredient is to realize that γ_t satisfies the sequence of inequalities in the lemma with $\tau - \bar{c}_1$ substituted by $\min \tau_i - \bar{c}_1^{(i)}$. $\qquad\square$

The following result follows:

Theorem 5 (Multiple Constraints Main Theorem). *If $\epsilon \in (0,1)$, $\alpha_c = 1$ and $\alpha_r = \frac{2}{\min_i \tau_i - \bar{c}_1^{(i)}} + 1$, then with probability at least $1 - \epsilon$, Algorithm 2 satisfies the following regret guarantee:*

$$\mathcal{R}_\Pi(T) \leq \left(\frac{2}{\min_i \tau_i - \bar{c}_1^{(i)}} + 1\right)\left(2\sqrt{2TK\log(4KT/\epsilon)} + 4\sqrt{T\log(2/\epsilon)\log(4KT/\epsilon)}\right)$$

Proof. The proof follows the exact same argument we used for the proof of Theorem 3 substituting $\tau - \bar{c}_1$ by $\min_i \tau_i - \bar{c}_1^{(i)}$. $\qquad\square$

C.6 Lower bound

We start by proving a generalized version of the divergence decomposition lemma for bandits.

Lemma 9. *[Divergence decomposition for constrained multi armed bandits] Let $\nu = ((P_1, Q_1), \cdots, (P_K, Q_K))$ be the reward and constraint distributions associated with one instance of the single constraint multi-armed bandit, and let $\nu' = ((P_1', Q_1'), \cdots, (P_K', Q_K'))$ be the reward and constraint distributions associated with another constrained bandit instance. Fix some algorithm $\mathcal{A}$ and let $\mathbb{P}_\nu = \mathbb{P}_{\nu\mathcal{A}}$ and $\mathbb{P}_{\nu'} = \mathbb{P}_{\nu'\mathcal{A}}$ be the probability measures on the cannonical bandit model (See section 4.6 of Lattimore and Szepesvári [2019]) induced by the T round interconnection of $\mathcal{A}$ and ν (respectively $\mathcal{A}$ and ν'). Then:*

$$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{a=1}^{K} \mathbb{E}_\nu[T_a(T)]\mathrm{KL}((P_a, Q_a), (P_a', Q_a'))$$

Where $T_a(T)$ denotes the number of times arm a was pulled until by $\mathcal{A}$ and up to time T.

Proof. The same proof as in Lemma 15.1 from Lattimore and Szepesvári [2019] applies in this case. $\qquad\square$

The following two lemmas will prove useful as well:

Lemma 10. *[Gaussian Divergence] The divergence between two multivariate normal distributions and means $\mu_1, \mu_2 \in \mathbb{R}^d$ with spherical identity covariance $\mathbb{I}_d$ equals:*

$$\mathrm{KL}(\mathcal{N}(\mu_1, \mathbb{I}_d), \mathcal{N}(\mu_2, \mathbb{I}_d)) = \frac{\|\mu_1 - \mu_2\|^2}{2}$$

Define the binary relative entropy to be:

$$d(x, y) = x \log(\frac{x}{y}) + (1 - x) \log(\frac{1 - x}{1 - y})$$

and satisfies:

$$d(x, y) \geq (1/2) \log(1/4y) \tag{32}$$

for $x \in [1/2, 1]$ and $y \in (0, 1)$. Adapted from Kaufmann et al. [2016], Lemma 1.

Lemma 11. *Let ν, ν' be two constrained bandit models with K arms. Borrow the setup, definitions and notations of Lemma 9, then for any measurable event $\mathcal{B} \in \mathcal{F}_T$:*

$$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{a=1}^{K} \mathbb{E}_\nu[T_a(T)] \mathrm{KL}((P_a, Q_a), (P'_a, Q'_a)) \geq d(\mathbb{P}_\nu(\mathcal{B}), \mathbb{P}_{\nu'}(\mathcal{B})) \tag{33}$$

We now present a worst-case lower bound for the constrained multi armed bandit problem. We restrict ourselves to Gaussian instances with mean reward and cost vectors $\bar{r}, \bar{c} \in [0, 1]^K$. Let $\mathcal{A}$ be an algorithm for policy selection in the constrained MAB problem. For the purpose of this section we denote as $\mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}, \bar{c})$ as the constrained regret of algorithm $\mathcal{A}$ in the Gaussian instance $\mathcal{N}(\bar{r}, \mathbb{I})$, $\mathcal{N}(\bar{c}, \mathbb{I})$. The following theorem holds:

Theorem 6. *Let $\tau, \bar{c}_1 \in (0, 1)$, $K \geq 4$, and $B := \max\left(\frac{1}{27}\sqrt{(k-1)T}, \frac{1}{6(\tau - \bar{c}_1)^2}\right)$ and assume[7] $T \geq \max(K - 1, 24eB)$ and let τ be the maximum allowed cost. Then for any algorithm $\mathcal{A}$ there is a pair of mean vectors $\bar{r}, \bar{c} \in [0, 1]^K$ such that:*

$$\mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}, \bar{c}) \geq B$$

Proof. If $\max\left(\frac{1}{27}\sqrt{(k-1)T}, \frac{1}{6(\tau - \bar{c}_1)^2}\right) = \sqrt{KT}$, then the argument in Theorem 15.2 of Lattimore and Szepesvári [2019] yields the desired result by noting that the framework of constrained bandits subsumes unconstrained multi armed bandits when all costs equal zero. In this case we conclude there is an instance $\bar{r}, \bar{c}$ with $\bar{c}_a = 0$ for all $a \in \mathcal{A}$ satisfying:

$$\mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}, \bar{c}) \geq \frac{1}{27}\sqrt{(k-1)T}$$

Let's instead focus on the case where $B = \max\left(\frac{1}{27}\sqrt{(k-1)T}, \frac{1}{6(\tau - \bar{c}_1)^2}\right) = \frac{1}{6(\tau - \bar{c}_1)^2}$.

Pick any algorithm. We want to show that the algorithm's regret on some environment is as large as B. If there was an instance $\bar{r}, \bar{c}$ such that $\mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}, \bar{c}) > B$ there would be nothing to be proven. Hence without loss of generality, we can assume that the algorithm satisfies $\mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}, \bar{c}) \leq B$ for all $\bar{r}, \bar{c} \in [0, 1]^K$ and having unit variance Gaussian rewards.

Let $c \in (0, 1)$ with $c = \tau - \bar{c}_1$. For the reader's convenience we will use the notation $\Delta = 1/2$. By treating the rewards in a symbolic way it is easier to understand the logic of the proof argument. Let's consider the following constrained bandit instance inducing measure ν:

$$\bar{c}^1 = (\tau - c, \quad \tau + 2c, \quad \tau - c, \quad \tau + 2c, \quad \cdots, \quad \tau + 2c)$$
$$\bar{r}^1 = (\Delta, \quad 8\Delta, \quad 0, \quad 4\Delta, \quad \cdots, \quad 4\Delta)$$

Notice that the optimal policy equals a mixture between arm 1 and 2, where arm 1 is chosen with probability $2/3$ and arm 2 with probability $1/3$. The value of this optimal policy equals $10/3\Delta$.

Recall we use the notation $\bar{T}_j(t)$ denote the total amount of probability mass that $\mathcal{A}$ allocated to arm j up to time t. Notice that the expected reward of all feasible policies that do not have arm 1 in their support have a gap (w.r.t the optimal feasible policy's expected reward) of at least $\frac{2\Delta}{3}$. Since by assumption, $\mathcal{A}$ satisfies $\mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}^1, \bar{c}^1) \leq B$:

$$B \geq \mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}^1, \bar{c}^1) \quad \geq \frac{2\Delta}{3}\left(\frac{2}{3}T - \frac{1}{2}T\right)\mathbb{P}\left(\bar{T}_1(T) < \frac{T}{2}\right) = \frac{\Delta}{9}T\mathbb{P}\left(\bar{T}_1(T) < \frac{T}{2}\right)$$

---

[7]This constraint on T translates to $T \geq C$ for some constant C.

And therefore:

$$\mathbb{P}\left(\bar{T}_1(T) \geq \frac{T}{2}\right) = 1 - \mathbb{P}\left(\bar{T}_1(T) < \frac{T}{2}\right) \geq 1 - \frac{9B}{\Delta T} \geq 1/2$$

The last inequality follows from the assumption $T \geq \max(K - 1, 24eB)$.

Let's now consider the following constrained bandit instance inducing measure ν':

$$\begin{aligned}
\bar{c}_2 &= (\tau - c, & \tau + 2c, & \quad 0, & \tau - c, & \cdots, & \tau + 2c) \\
\bar{r}_2 &= (\Delta, & 8\Delta, & \quad 0, & 4\Delta, & \cdots, & 4\Delta)
\end{aligned}$$

In this instance the optimal policy is to play arm 4 deterministically, which gets a reward of 4Δ. Notice that the expected reward of any feasible policy that does not contain arm 4 in its support has a gap (w.r.t. the optimal feasible policy's expected reward) of at least $\frac{2\Delta}{3}$. Since by assumption, $\mathcal{A}$ satisfies $\mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}^2, \bar{c}^2) \leq B$:

$$B \geq \mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}^2, \bar{c}^2) \qquad \geq \frac{2\Delta}{3}\left(\frac{1}{2}T\right)\mathbb{P}\left(\bar{T}_1(T) \geq \frac{T}{2}\right) = \frac{\Delta}{3}T\mathbb{P}\left(\bar{T}_1(T) \geq \frac{T}{2}\right)$$

And therefore:

$$\mathbb{P}\left(\bar{T}_1(T) \geq \frac{T}{2}\right) \leq \frac{3B}{\Delta T} \leq \frac{1}{4e}$$

The last inequality follows from the assumption $T \geq \max(K - 1, 24eB)$. As a consequence of inequality 32, Lemma 11 and 10:

$$\mathbb{E}_\nu[T_4(T)]\mathrm{KL}(\binom{\tau + 2c}{4\Delta}, \mathbb{I}_d), \mathcal{N}(\binom{\tau - c}{4\Delta}, \mathbb{I}_d)) = \mathbb{E}_\nu[T_4(T)]2c^2 \geq \frac{1}{2}$$

And therefore we can conclude:

$$\mathbb{E}[\bar{T}_4(T)] = \mathbb{E}[T_4(T)] \geq \frac{1}{4c^2} \tag{34}$$

Since in ν, any feasible policy with support in arm 4 and no support in arm 2 has a suboptimality gap of $4/3\Delta$, we conclude the regret $\mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}^2, \bar{c}^2)$ must satisfy:

$$\mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}^2, \bar{c}^2) \geq \frac{\Delta}{3c^2}$$

Since $\Delta = \frac{1}{2}$ and noting that in this case $\frac{\Delta}{3c^2} = B$. The result follows. $\qquad \square$