Stochastic Bandits with Linear Constraints

Aldo Pacchiano UC Berkeley
Mohammad Ghavamzadeh Google Research
Peter Bartlett UC Berkeley
Heinrich Jiang Google Research

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 multiple rounds is maximum, and each one of them has an expected cost below a certain threshold. We propose an upper-confidence bound algorithm for this problem, called optimistic pessimistic linear bandit (OPLB), and prove a sublinear bound on its regret that is inversely proportional to the difference between the constraint threshold and the cost of a known feasible action. Our algorithm balances exploration and constraint satisfaction using a novel idea that scales the radii of the reward and cost confidence sets with different scaling factors. We further specialize our results to multi-armed bandits and propose a computationally efficient algorithm for this setting and prove a a regret bound that is better than simply casting multi-armed bandits as an instance of 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. Finally, we show how our algorithm and analysis can be extended to multiple constraints and to the case when the cost of the feasible action is unknown.

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) or *upper confidence bound* (UCB) (Auer et al., 2002) and *Thompson sampling* (TS) (Thompson, 1933; Agrawal and Goyal, 2013a; Abeille and Lazaric, 2017; Russo et al., 2018) that are relatively well understood in both multi-armed and linear bandits (Abbasi-Yadkori et al., 2011; Agrawal and Goyal, 2013b).

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, its

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

cumulative reward remains above a predefined fraction of the cumulative reward of the baseline (Wu et al., 2016; Kazerouni et al., 2017; Garcelon et al., 2020). In these two settings, the constraint is *history-dependent*, i.e., it applies to a cumulative quantity, such as 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.

In this setting, each action has a context-dependent (unknown) cost and only actions should be taken, whose cost is below a certain threshold. This setting has many applications, for example, a recommendation system should not suggest an item to a customer that despite high probability of click (high reward) reduces her watch-time or her chance of coming back to the website (bounded cost), or a drug that may help with a certain symptom (high reward) should not have too many side-effects (bounded cost). It is important to note that the reward and cost in this setting can be viewed as different objectives according to which a recommendation or a medical diagnosis system are evaluated.

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.

As discussed above, the setting considered in this paper is a relaxation of the high probability stage-wise constrained setting described earlier. In many constrained or multi-objective problems, such as recommendation and medical diagnosis systems, making sure that the constraints are always satisfied or certain objectives are always within certain thresholds would result in a very conservative performance. A common solution to balance performance and constraint satisfaction is to replace conservative high probability constraints with more relaxed expectation ones.

In this paper, we study the above setting for contextual linear bandits. After defining the setting in Section 2, we propose an OFU-style 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 *safety gap*, i.e., the difference between the constraint threshold and the cost of a known feasible (safe) 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 constraints. This property plays an important role in the computational efficiency of OPB. We prove a regret bound of order $\widetilde{\mathcal{O}}(\frac{\sqrt{KT}}{\tau - c_0})$ 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 constrained bandit problem studied in the paper.

In our setting the learner interacts with arms whose costs are unknown while required to satisfy an upper bound on its policy's expected cost. Since the learner does not know the cost function in advance, she has to balance three competing objectives: 1) collect reward, 2) satisfy the cost constraint and 3) learn about the cost and reward functions. At any point in time and given the learner's knowledge of the reward and cost function, objective 2) may prevent her from even considering to execute the true optimal policy. This precludes the use of algorithms based solely on the principle of optimism. One of our main technical contributions is the introduction of a general and simple technique based on asymmetric confidence intervals that can be used to easily develop algorithms for bandits or reinforcement learning problems with unknown constraints.

2 Problem Formulation

Notation. We adopt the following notation through-

out the paper. We denote by $\langle x, y\rangle = x^\top y$ and $\langle x, y\rangle_A = x^\top A y$, for a positive definite matrix $A \in \mathbb{R}^{d\times d}$, the inner-product and weighted inner-product of the vectors $x, y \in \mathbb{R}^d$. Similarly, we denote by $\|x\| = \sqrt{x^\top x}$ and $\|x\|_A = \sqrt{x^\top A x}$, the ℓ_2 and weighted ℓ_2 norms of vector x. For any square matrix A, we denote by $A^\dagger$, its Moore-Penrose pseudo-inverse. We represent the set of distributions with support over a compact set $\mathcal{S}$ by $\Delta_{\mathcal{S}}$. The set $\{1, \dots, T\}$ is denoted by $[T]$. Finally, we use $\widetilde{\mathcal{O}}$ for the big-$\mathcal{O}$ notation up to logarithmic factors.

We study the following *constrained contextual linear bandit* setting in this paper. In each round t, the agent is given a 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 in Assumption 1. 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 *stage-wise linear constraint*

$$\mathbb{E}_{x\sim\pi_t}[\langle x, \mu_*\rangle] \le \tau, \quad \forall t \in [T], \tag{1}$$

where $\tau \ge 0$ is referred to as the *constraint threshold*. Thus, the policy π_t that the agent selects in each round $t \in [T]$ should belong to the set of *feasible policies* over the action set $\mathcal{A}_t$, i.e., $\Pi_t^* = \{\pi \in \Delta_{\mathcal{A}_t} : \mathbb{E}_{x\sim\pi}[\langle x, \mu_*\rangle] \le \tau\}$. Maximizing the expected cumulative reward in T rounds is equivalent to minimizing the T-round *constrained pseudo-regret*,[1]

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

where $\pi_t, \pi_t^* \in \Pi_t$, for all $t \in [T]$, and $\pi_t^* \in \max_{\pi\in\Pi_t^*} \mathbb{E}_{x\sim\pi}[\langle x, \theta_*\rangle]$ is the *optimal feasible* policy in round t. 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 notations, we may write the T-round regret as $\mathcal{R}_\Pi(T) = \sum_{t=1}^T r_{\pi_t^*} - r_{\pi_t}$.

[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.

We make the following assumptions for our setting. The first four assumptions are standard in linear bandits and the fifth one is necessary for constraint satisfaction.

Assumption 1 (sub-Gaussian noise). *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., for all $\alpha \in \mathbb{R}$,*

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

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 (bounded parameters). *There is a known constant $S > 0$, such that $\|\theta_*\| \le S$ and $\|\mu_*\| \le S$.*[2]

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

Assumption 4 (bounded rewards and costs). *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 (safe action). *There is a known safe action $x_0 \in \mathcal{A}_t$, $\forall t \in [T]$ with known cost c_0, i.e., $\langle x_0, \mu_*\rangle = c_0 < \tau$.*

Remark 1. *Knowing a safe action x_0 is absolutely necessary for solving the constrained contextual linear bandit problem studied in this paper, because it requires the constraint to be satisfied from the very first round. However, the assumption of knowing the expected cost of the safe action c_0 can be relaxed. We can think of the safe action as a baseline policy, the current strategy (e.g., resource allocation of a company), whose cost is known and reasonable, but its reward may still be improved. We will discuss how our proposed algorithm will change if c_0 is unknown in Section 3 and Appendix B.4.*

Notation. We conclude this section with introducing another set of notations that will be used in describing our algorithm and its analysis. 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 $x_\pi^{o,\perp} := \mathbb{E}_{x\sim\pi}[x^{o,\perp}]$, respectively.

[2] The choice of the same upper-bound S for both θ_* and μ_* is just for simplicity and convenience.

[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

Input: Horizon T, Confidence Parameter δ, Regularization Parameter λ, Constants $\alpha_r, \alpha_c \geq 1$
for $t = 1, \dots, T$ **do**
  Compute regularized least-squares estimates $\widehat{\theta}_t$ and $\widehat{\mu}_t^{o,\perp}$ *(Eqs. 3 to 5)*
  Construct sets $\mathcal{C}_t^r(\alpha_r)$ and $\mathcal{C}_t^c(\alpha_c)$ *(Eq. 7)*
  Observe the action set $\mathcal{A}_t$ and construct the feasible (safe) policy set Π_t *(Eq. 13)*
  Compute policy $(\pi_t, \widetilde{\theta}_t) = \arg\max_{\pi \in \Pi_t,\, \theta \in \mathcal{C}_t^r(\alpha_r)} \mathbb{E}_{x\sim\pi}[\langle x, \theta \rangle]$
  Take action $x_t \sim \pi_t$ and observe reward and cost (r_t, c_t)
end for

3 Algorithm

In this section, we propose a UCB-style algorithm for the setting described in Section 2. We call our algorithm *optimistic-pessimistic linear bandit* (OPLB) because it maintains a pessimistic assessment of the set of available policies, while acting optimistically within this set. Algorithm 1 contains the pseudo-code of OPLB. The novel idea in OPLB is to balance exploration and constraint satisfaction by *asymmetrically* scaling the radii of the reward and cost confidence sets with different scaling factors α_r and α_c. This will prove crucial in the regret analysis of OPLB. We now describe OPLB in details.

Line 3 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

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

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

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

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

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 $c_s^{o,\perp}$ $(\forall s \in [t-1])$ is the noisy projection of the cost c_s into $\mathcal{V}_o^\perp$, i.e.,[4]

$$c_s^{o,\perp} = \langle x_s^{o,\perp}, \mu_*^{o,\perp} \rangle + \xi_s^c = \langle x_s, \mu_* \rangle - \langle x_s^o, \mu_*^o \rangle + \xi_s^c$$
$$= c_s - \langle x_s^o, \mu_*^o \rangle = c_s - \frac{\langle x_s, e_0 \rangle}{\|x_0\|} c_0. \quad (6)$$

Line 4: Using the RLS estimates $\widehat{\theta}_t$ and $\widehat{\mu}_t^{o,\perp}$ in (3), OPLB constructs the reward and cost *confidence sets*

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

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*

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

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 $\mathcal{V}_o^\perp$, with high probability.

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

$$\widetilde{r}_{\pi,t} := \max_{\theta \in \mathcal{C}_t^r(\alpha_r)} \mathbb{E}_{x\sim\pi}[\langle x, \theta \rangle], \quad (9)$$

$$\widetilde{c}_{\pi,t} := \frac{\langle x_\pi^o, e_0 \rangle c_0}{\|x_0\|} + \max_{\mu \in \mathcal{C}_t^c(\alpha_c)} \mathbb{E}_{x\sim\pi}[\langle x, \mu \rangle]. \quad (10)$$

We provide closed-form expressions for $\widetilde{r}_{\pi,t}$ and $\widetilde{c}_{\pi,t}$ in the following proposition that we report its proof in Appendix A.1.

Proposition 1. *We may write (9) and (10) in closed-form as*

$$\widetilde{r}_{\pi,t} = \langle x_\pi, \widehat{\theta}_t \rangle + \alpha_r \beta_t(\delta, d) \|x_\pi\|_{\Sigma_t^{-1}}, \quad (11)$$

$$\widetilde{c}_{\pi,t} = \frac{\langle x_\pi^o, e_0 \rangle c_0}{\|x_0\|} + \langle x_\pi^{o,\perp}, \widehat{\mu}_t^{o,\perp} \rangle + \alpha_c \beta_t(\delta, d-1) \|x_\pi^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}}. \quad (12)$$

Line 5: After observing the action set $\mathcal{A}_t$, OPLB constructs its feasible (safe) policy set as

$$\Pi_t = \{\pi \in \Delta_{\mathcal{A}_t} : \widetilde{c}_{\pi,t} \leq \tau\}, \quad (13)$$

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

where $\widetilde{c}_{\pi,t}$ is the pessimistic cost of policy π in round t defined by (12). Note that Π_t is an approximation to Π_t^* and 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 $x^o_{\pi_0} = x_0$, $x^{o,\perp}_{\pi_0} = 0$, and $\frac{\langle x^o_{\pi_0}, e_0\rangle c_0}{\|x_0\|} = c_0$. 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 6: 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 7: 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 6 of Algorithm 1, in each round t, OPLB solves the following optimization problem:

$$\max_{\pi\in\Delta_{\mathcal{A}_t}} \langle x_\pi, \widehat{\theta}_t\rangle + \alpha_r\beta_t(\delta,d)\|x_\pi\|_{\Sigma_t^{-1}} \tag{14}$$
$$\text{s.t.} \quad \frac{\langle x^o_\pi, e_0\rangle c_0}{\|x_0\|} + \langle x^{o,\perp}_\pi, \widehat{\mu}^{o,\perp}_t\rangle + \alpha_c\beta_t(\delta,d-1)\|x^{o,\perp}_\pi\|_{(\Sigma^{o,\perp}_t)^{-1}} \le \tau.$$

However, solving (14) 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.

Main Challenge in Regret Analysis. The main challenge in obtaining a regret bound for OPLB is to ensure that optimism holds in each round $t \in [T]$, i.e., the solution $(\pi_t, \widetilde{\theta}_t)$ of (14) satisfies $\widetilde{r}_{\pi_t,t} = \langle x_{\pi_t}, \widetilde{\theta}_t\rangle \ge r_{\pi_t^*}$. This is not obvious, since the safe policy set Π_t might have been constructed such that it does 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 α_r/α_c. Of course, this comes at the cost of scaling the regret by the same ratio. As we will show in our analysis in Section 4, α_r/α_c depends on the inverse safety gap $1/(\tau - c_0)$, which indicates that 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.

Unknown 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 a conservative estimate $\hat{\delta}_c$ of the safety gap $\tau - c_0$ that satisfies $\hat{\delta}_c \ge \frac{\tau - c_0}{2}$. 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 our OPLB algorithm.

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*

$$\mathcal{R}_\Pi(T) \le \frac{2L(\alpha_r+1)\beta_T(\delta,d)}{\sqrt{\lambda}}\sqrt{2T\log(1/\delta)} + (\alpha_r+1)\beta_T(\delta,d)\sqrt{2Td\log(1+\frac{TL^2}{\lambda})}. \tag{15}$$

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

$$\mathcal{E} = \big\{\|\widehat{\theta}_t - \theta_*\|_{\Sigma_t} \le \beta_t(\delta,d) \wedge \|\widehat{\mu}_t^{o,\perp} - \mu_*^{o,\perp}\|_{\Sigma_t^{o,\perp}} \le \beta_t(\delta,d-1),\ \forall t\in[T]\big\}. \tag{16}$$

The regret $\mathcal{R}_\Pi(T)$ in (2) can be decomposed as

$$\mathcal{R}_\Pi(T) = \underbrace{\sum_{t=1}^T r_{\pi_t^*} - \widetilde{r}_{\pi_t,t}}_{\text{(I)}} + \underbrace{\sum_{t=1}^T \widetilde{r}_{\pi_t,t} - r_{\pi_t}}_{\text{(II)}}. \tag{17}$$

where $\widetilde{r}_{\pi_t,t}$ is the optimistic reward defined by (9) and (11). We first bound (II) in (17). To bound (II), we further decompose it as

$$\text{(II)} = \underbrace{\sum_{t=1}^T \langle x_{\pi_t}, \widetilde{\theta}_t\rangle - \langle x_t, \widetilde{\theta}_t\rangle}_{\text{(III)}} + \underbrace{\sum_{t=1}^T \langle x_t, \widetilde{\theta}_t\rangle - \langle x_t, \theta_*\rangle}_{\text{(IV)}} + \underbrace{\sum_{t=1}^T \langle x_t, \theta_*\rangle - \langle x_{\pi_t}, \theta_*\rangle}_{\text{(V)}}. \tag{18}$$

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

Lemma 1. *On event $\mathcal{E}$ defined by (16), for any $\gamma \in (0,1)$, with probability at least $1-\gamma$, we have*

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

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

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

Proof. We report the proof in Appendix B.1. □

After bounding all the terms in (II), we now process the term (I). Before stating the main result for this term in Lemma 4, we need to prove the following lemma.

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

$$\|x_\pi^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}} \le \|x_\pi\|_{\Sigma_t^{-1}}. \tag{19}$$

Proof. We report the proof in Appendix B.2. □

In the following lemma, we prove that by appropriately setting the scaling 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 (17) is always non-positive.

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

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 set of feasible policies Π_t, or not.

Case 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 6 of Algorithm 1), i.e., $\widetilde{r}_{\pi_t^*,t} \le \widetilde{r}_{\pi_t,t}$. This together with the fact that the optimistic reward of any policy π is larger than its expected reward, i.e., $\widetilde{r}_{\pi,t} \ge r_\pi$, gives us the desired result that $\widetilde{r}_{\pi_t,t} \ge r_{\pi_t^*}$.

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 policies, 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. 26 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., $\widetilde{r}_{\pi_t,t} \ge \widetilde{r}_{\widetilde{\pi}_t,t}$, and obtain a lower-bound for $\widetilde{r}_{\widetilde{\pi}_t,t}$ as a function of $r_{\pi_t^*}$ (see Eq. 27 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 $\widetilde{r}_{\pi_t,t} \ge r_{\pi_t^*}$ is obtained, i.e., $1 + \alpha_c \le (\tau - c_0)(\alpha_r - 1)$. □

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 $\gamma = \delta$. □

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 our OPLB algorithm and an improvement in its regret bound.

In the MAB setting, the action set consists of K arms $\mathcal{A} = \{1, \dots, K\}$. 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 *stage-wise linear constraint* $\mathbb{E}_{a_t \sim \pi_t}[\bar{c}_{a_t}] \le \tau$, $\forall t \in [T]$. Moreover, arm 1 is assumed to be the known safe arm, i.e., $\bar{c}_1 \le \tau$.

Optimistic Pessimistic Bandit (OPB) Algorithm. Let $\{T_a(t)\}_{a=1}^K$ and $\{\widehat{r}_a(t), \widehat{c}_a(t)\}_{a=1}^K$ 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 $u_a^r(t) = \widehat{r}_a(t) + \alpha_r \beta_a(t)$, $u_a^c(t) = \widehat{c}_a(t) + \alpha_c \beta_a(t)$,

$\beta_a(t) = \sqrt{2\log(1/\delta')/T_a(t)}$, 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 $t \in [T]$:

$$\max_{\pi \in \Delta_K} \sum_{a \in \mathcal{A}} \pi_a \, u_a^r(t), \qquad \text{s.t.} \quad \sum_{a \in \mathcal{A}} \pi_a \, u_a^c(t) \leq \tau. \quad (20)$$

As shown in (20), 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 to solve, OPB can be implemented extremely efficiently. Lemma 5, whose proof we report in Appendix C.2, show that (20) always has a solution (policy) with support of at most 2. This property allows us to solve (20) in closed-form, without a LP solver, and implement OPB very efficiently.

Lemma 5. *There exists a policy that solves (20) 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*

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

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

Remark 2. *Our contextual linear bandit formulation is general enough to include MAB. The regret analysis of OPLB (Thm. 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 Thm. 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.*

Lower-bound. We also prove a mini-max lower-bound for our constrained MAB problem. Our lower-bound shows that no algorithm can attain a regret better than $\mathcal{O}\big(\max(\sqrt{KT}, \frac{1}{(\tau - \bar{c}_1)^2})\big)$ for this problem. The formal statement of the lower-bound and its proof are reported in Appendix C.5.

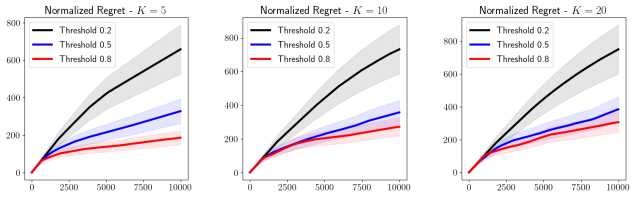

Figure 1: Regret of OPB for three instances of the randomly generated constrained multi-armed bandit problems with the number of arms equal to 5 *(left)*, 10 *(middle)*, and 20 *(right)*.

Extension to Multiple Constraints. In the case of m constraints, 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 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 also obtain a regret bound of $\widetilde{\mathcal{O}}(\frac{\sqrt{KT}}{\min_{i\in[K]}(\tau_i - \boldsymbol{c}_1(i))})$ for m-constraint OPB. The proofs and details of this case are reported in Appendix C.6.

6 Experimental Results

We run a set of experiments to show the behavior of the OPB algorithm and validate our theoretical results. We produce random instances of our constrained multi-armed bandit problem. We select one arm with mean reward and cost 0 to be the safe arm. We sample the mean rewards and costs of the rest of the arms uniformly at random from the interval $[0, 1]$. In Figure 1, we report the regret of OPB for each of the number of arms K equal to 5 *(left)*, 10 *(middle)*, and 20 *(right)*, and for three constraint threshold τ values, 0.8 *(red)*, 0.5 *(blue)*, and 0.2 *(black)*. For each parameter setting we sample 10 random problem instances and report the average regret curves with a shaded region corresponding to the ± 0.5 standard deviation around the regret. Figure 1 also shows that the regret of OPB grows inversely with the safety gap.

In the next experiment, we consider a $K = 4$-armed bandit problem in which the reward and cost distributions of the arms are Bernoulli with means $\bar{r} = (0.1, 0.2, 0.4, 0.7)$ and $\bar{c} = (0, 0.4, 0.5, 0.2)$. Arm 1 is the safe arm with the expected cost $\bar{c}_1 = 0$. In Figures 2 to 5, we gradually reduce the constraint threshold τ, and as a result, the *safety gap* $\tau - \bar{c}_1$, and show the regret *(left)*, cost *(middle)*, and reward *(right)* evolution of OPB. The cost and reward of OPB are in blue and the optimal cost and reward are in red. All results are averaged over 10 runs and the shade is the ± 0.5 standard deviation around the regret.

In figures figs. 2 to 5 show that the regret of OPB grows as we reduce τ, and as a result the safety gap *(left)*.

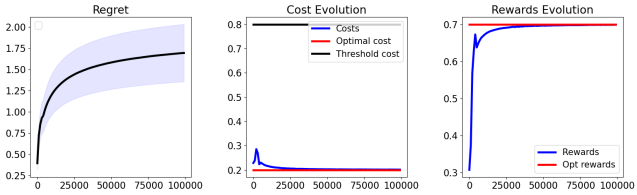

Figure 2: Constraint Threshold $\tau = 0.8$.

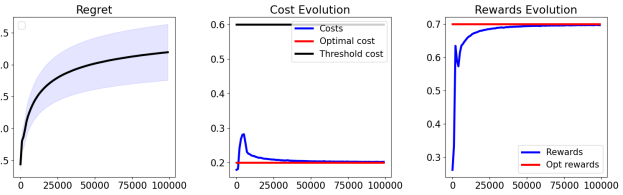

Figure 3: Constraint Threshold $\tau = 0.6$.

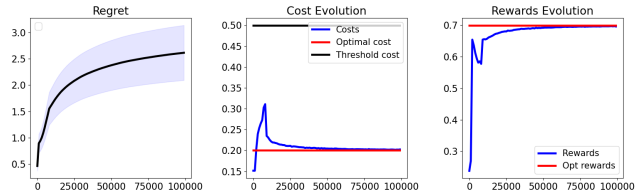

Figure 4: Constraint Threshold $\tau = 0.5$.

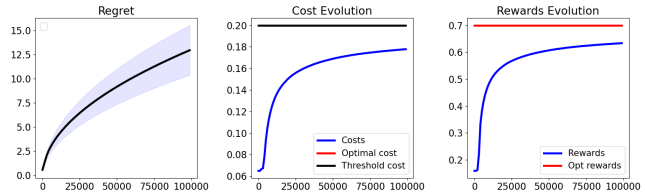

Figure 5: Constraint Threshold $\tau = 0.2$.

Regret *(left)*, cost *(middle)*, and reward *(right)* evolution of OPB in a 4-armed bandit problem with Bernoulli reward and cost distributions with means $\bar{r} = (.1, .2, .4, .7)$ and $\bar{c} = (0, .4, .5, .2)$. The cost of the safe arm (Arm 1) is $\bar{c}_1 = 0$.

This is in support of our theories that identified the safety gap as the complexity of this constrained bandit problem. The results also indicate that the algorithm is successful in satisfying the constraint *(middle)* and in reaching the optimal reward/performance *(right)*. In Figure 5, the cost of the best arm (Arm 4) is equal to the constraint threshold $\tau = 0.2$. Thus, the cost of the optimal policy *(red)* and the constraint threshold *(black)* overlap in the cost evolution *(middle)* sub-figure.

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. As also discussed in Section 1, 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 define their action set to be any convex compact subset of $\mathbb{R}^d$ that contains the origin. Therefore, they restrict their *"known"* safe action to be the origin, $x_0 = \mathbf{0}$, with the *"known"* cost $c_0 = 0$. This is why c_0 does not appear in their bounds. Although later in their proofs, to guarantee that their algorithm does not violate the constraint 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. Similarly a TS version of OPLB suitable for our setting can also be derived whose regret bound will also suffer from a suboptimal $d^{3/2}$ scaling.

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 novel 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 order $\widetilde{\mathcal{O}}(\frac{d\sqrt{T}}{\tau - c_0})$ for our algorithm, which shows that the difficulty of the problem depends on the *safety gap* $\tau - c_0$, i.e., the difference between the constraint threshold and the cost of a known feasible action. 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, showed how our algorithm and analysis can be extended to multiple constraints and to the case when the cost of the safe action, c_0, is unknown, and provided simulations to validate our theoretical results. A future direction is to use the optimism-pessimism idea 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).

Bibliography

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.

M. Abeille and A. Lazaric. Linear Thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.

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.