Journal of Machine Learning Research 26 (2025) 1-57 Submitted 2/24; Revised 2/25; Published 8/25

Contextual Bandits with Stage-wise Constraints

Aldo Pacchiano PACCHIAN@BU.EDU
Faculty of Computing & Data Sciences
Boston University
Boston, MA, USA

Mohammad Ghavamzadeh GHAVAMZA@AMAZON.COM
Amazon AGI
Sunnyvale, CA, USA

Peter Bartlett PETER@BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences
University of California
Berkeley, CA, USA

Editor: Quanquan Gu

Abstract

We study contextual bandits in the presence of a stage-wise constraint when the constraint must be satisfied both with high probability and in expectation. We start with the linear case where both the reward function and the stage-wise constraint (cost function) are linear. In each of the high probability and in expectation settings, we propose an upper-confidence bound algorithm for the problem and prove a T-round regret bound for it. We also prove a lower-bound for this constrained problem, show how our algorithms and analyses can be extended to multiple constraints, and provide simulations to validate our theoretical results. In the high probability setting, we describe the minimum requirements for the action set for our algorithm to be tractable. In the setting that the constraint is in expectation, we specialize our results to multi-armed bandits and propose a computationally efficient algorithm for this setting with regret analysis. Finally, we extend our results to the case where the reward and cost functions are both non-linear. We propose an algorithm for this case and prove a regret bound for it that characterize the function class complexity by the eluder dimension.

Keywords: multi-armed bandits, contextual bandits, constraints, safety, eluder dimension

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* sampled from the distribution of the arm. 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 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. 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 their watch-time or their 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). Another example is a company whose goal is to optimize its app's strategy for sending notification to its customers. Here the reward signal is often related to the customer's engagement with the app, and the cost signal depends on the probability that the customer gets tired of the notifications and opt out. Thus, the goal is to derive a strategy that while maximizes customer's engagement with the app, keeps the churn below a certain threshold. 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 or an app's notification strategy are evaluated.

Amani et al. (2019) and Moradipari et al. (2019) 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. We start the paper by studying the same setting for contextual linear bandits. After defining the setting in Section 2, we propose a *UCB-style* algorithm for it, called Linear Constraint Linear UCB (LC-LUCB), in Section 4.1. We prove a T-round regret bound for LC-LUCB in Sections 4.2, which clearly identifies the main components that control the hardness of this problem. We also prove a lower-bound for this setting in Section 4.3, show how this setting can be extended to multiple constraints (multiple cost distributions for each arm) in Section 4.4, and

report experimental results as a proof of concept for LC-LUCB in Section 4.5. We provide a detailed comparison between our results and those in Amani et al. (2019) and Moradipari et al. (2019) in Section 3.

We then switch to a slightly different setting in Section 5 in which we relax the high probability constraint and replace it with a constraint in expectation. High probability constraints are often quite restrictive and result in overly conservative strategies. This is why in many applications we may want to relax them to obtain strategies with higher expected cumulative reward. We describe this relaxed setting in Section 5.1 and propose an algorithm for it, called Optimistic-Pessimistic Linear Bandit (OPLB) (Section 5.1.1), with regret analysis (Section 5.1.2). We then specialize our results to multi-armed bandits (Section 5.2) and report experimental results as a proof of concept for OPLB (Section 5.4). Finally in Section 5.3, we extend our results to the case where the reward and cost functions are non-linear. We propose an algorithm, called Optimistic Pessimistic Nonlinear Bandit (OPNLB), and prove a regret bound for it. We use a characterization of function class complexity based on the eluder dimension (Russo and Van Roy, 2013) in our regret bound. This part of the paper is an extension of our earlier work (Pacchiano et al., 2021). Here we improve the regret bounds reported in Pacchiano et al. (2021) for both contextual linear and multi-armed bandit settings to better show their dependence on the components that contribute to the hardness of the problem. We also show how our results can be extended to non-linear reward and cost(s).

2. Problem Formulation

Notation. We adopt the following notation throughout the paper. We denote by $\langle x, y\rangle = x^\top y$ and $\langle x, y\rangle_{\mathbf{A}} = x^\top \mathbf{A} y$, for a positive definite matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, the inner-product and weighted inner-product of vectors $x, y \in \mathbb{R}^d$. Similarly, we denote by $\|x\| = \sqrt{x^\top x}$ and $\|x\|_{\mathbf{A}} = \sqrt{x^\top \mathbf{A} x}$, the ℓ_2 and weighted ℓ_2 norms of vector $x \in \mathbb{R}^d$. For any square matrix $\mathbf{A}$, we denote by $\mathbf{A}^\dagger$, its Moore-Penrose pseudo-inverse. We represent the set of distributions with support over a compact set $\mathcal{S}$ by $\Delta_{\mathcal{S}}$. We use upper-case letters for random variables (e.g., X), and their corresponding lower-case letters for a particular instantiation of that random variable (e.g., $X = x$). 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 $t \in [T]$, the agent (also referred to as learner) is given a decision set $\mathcal{A}_t \subset \mathbb{R}^d$ from which it has to choose an action x_t. Upon taking an action $X_t \in \mathcal{A}_t$, the agent 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 aims to maximize its *expected T-round reward*, i.e., $\sum_{t=1}^T \langle X_t, \theta_*\rangle$, while is required to satisfy a ***stage-wise linear constraint***, i.e., $\langle X_t, \mu_*\rangle \le \tau,\ \forall t \in [T]$, with high probability. The *constraint threshold* $\tau \ge 0$ is a positive constant that is known to the agent.

Because of the constraint, in each round t, the agent should pull an arm from the set of *feasible actions* in that round, i.e., $\mathcal{A}_t^f = \{x \in \mathcal{A}_t : \langle x, \mu_*\rangle \le \tau\}$. Of course this set is unknown, because the agent does not know the cost parameter μ_*. Maximizing the expected T-round reward is equivalent to minimizing the *expected T-round (constrained) (pseudo)-regret*, i.e.,

$$\mathcal{R}_{\mathcal{C}}(T) = \sum_{t=1}^{T} \langle x_t^*, \theta_*\rangle - \langle X_t, \theta_*\rangle = \sum_{t=1}^{T} \langle x_t^* - X_t, \theta_*\rangle, \tag{1}$$

where x_t^* is the *optimal feasible action* in round t, i.e., $x_t^* \in \arg\max_{x \in \mathcal{A}_t^f} \langle x, \theta_* \rangle$, and X_t is the action taken by the agent in round t, which belongs to the set of feasible actions in that round, i.e., $X_t \in \mathcal{A}_t^f$, with high probability.

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 $\alpha \in \mathbb{R}$,*

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

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

Assumption 2 (bounded parameters) *There is a known constant $S > 0$, such that $\|\theta_*\| \leq S$ and $\|\mu_*\| \leq S$.*[1]

Assumption 3 (bounded actions) *The ℓ_2-norm of all actions are bounded by $L > 0$, i.e.,*

$$\max_{t \in [T]} \max_{x \in \mathcal{A}_t} \|x\| \leq L.$$

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., $\langle x_0, \mu_* \rangle = c_0 < \tau$, and known reward r_0.*

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 its expected reward r_0 and cost 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 that is safe (i.e., its cost $c_0 < \tau$) and has a reasonable performance (i.e., its reward r_0 is not low). In this case, it makes sense to assume that the reward r_0 and cost c_0 of this action (policy) are both known. We will discuss how our proposed algorithms will change if r_0 and c_0 are unknown in Sections 4.1 and 5.1.1, and Appendix F.1.

We will show later that the difficulty of solving the above constrained bandit problem is directly related to the quality of the safe action x_0, more specifically to its *safety gap* and *sub-optimality*.

Definition 6 (safety gap & sub-optimality) *The safety gap and sub-optimality of a safe action x_0 quantify how close its cost c_0 and reward r_0 are to the constraint threshold τ and the maximum achievable reward 1, and are defined as $(\tau - c_0)$ and $(1 - r_0)$, respectively.*

Notation. We conclude this section with introducing another set of notations that will be used in describing our algorithms and their analyses. 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., $\mathcal{V}_o^\perp = \{x \in \mathbb{R}^d : \langle x, y \rangle = 0, \ \forall y \in \mathcal{V}_o\}$.[2] We define the projection of a vector $x \in \mathbb{R}^d$ into the subspace $\mathcal{V}_o$, as $x^o := \langle x, e_0 \rangle e_0$, and into the subspace $\mathcal{V}_o^\perp$, as $x^{o,\perp} := x - x^o$.

1. The choice of the same upper-bound S for both θ_* and μ_* is just for simplicity and convenience.
2. 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 entire $\mathbb{R}^d$.

3. Related Work

As described in Section 1, the high probability constraint satisfaction setting that we study in Section 4 is similar to the one in Moradipari et al. (2019) and Amani et al. (2019). Moradipari et al. (2019) propose a Thompson sampling (TS) algorithm for this setting and prove an $\widetilde{O}(d^{3/2}\sqrt{T}/\tau)$ regret bound for it. Our algorithm is UCB-style and our regret bound is $\widetilde{\mathcal{O}}((1+\frac{1-r_0}{\tau-c_0})d\sqrt{T})$, which not only has a better dependence on d, but also clearly identifies the ratio between the sub-optimality $(1-r_0)$ of the safe action x_0 and its safety gap $(\tau - c_0)$ as the measure of hardness for the problem. The $\widetilde{\mathcal{O}}(\sqrt{d})$ advantage to their bound is similar to the best regret results for UCB vs. TS. Moreover, 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 contain the ball with radius τ/S around the origin. Hence, our setting and action set are more general than theirs. We also prove a lower-bound for the problem and show how our algorithm and analysis can be extended to multiple constraints and to the case when the reward and cost of the safe action are unknown. Finally, unlike us, their action set does not allow their results to be immediately applicable to MAB. However, their algorithm is TS, and thus, is less complex than ours. Although it can still be intractable, even when the action set $\mathcal{A}$ is convex, as we can see they require several approximations in their experiments. Unlike them, we describe the minimum requirements on the action set in order for our algorithm to be tractable.

Amani et al. (2019) propose an explore-exploit algorithm for a slightly different setting than ours, in which reward and cost have the same unknown parameter θ_*, and the constraint is defined as $c_t = x_t^\top B\theta_* \leq \tau$, for a *"known"* matrix B. They prove a regret bound of $\widetilde{\mathcal{O}}(T^{2/3})$ for their algorithm. Although our algorithm has a better regret rate $\widetilde{\mathcal{O}}(\sqrt{T})$, it cannot immediately give the same rate for the setting studied in Amani et al. (2019), except in special cases, such as when all $\mathcal{A}_t$ are convex and $B = I$.

Several authors have extended the constrained problem studied in this paper to other constrained bandit settings. Chen et al. (2022) modified the constraint to cumulative and obtained an $o(T)$ bound for cumulative constraint violation while obtaining an $\mathcal{O}(\log(T)^2)$ instance-dependent bound for the cumulative regret. Other extensions include to anytime cumulative constraints (Liu et al., 2021b), kernel setting (Zhou and Ji, 2022), best arm identification (Wang et al., 2022), and online convex optimization (Chaudhary and Kalathil, 2022).

Our stage-wise constrained bandit problem has also been extended to reinforcement learning (RL) where the goal is to find a policy with maximum expected cumulative reward while the learner is required to keep the expected cumulative cost below a threshold at every single round. Here the learner has access to a safe policy that can be deployed while the learner does not have sufficient knowledge of the safety constraint. This RL setting has been studied in tabular (Efroni et al., 2020; Liu et al., 2021a; Wei et al., 2021; Bura et al., 2022) and linear (Ding et al., 2021; Ghosh et al., 2022) MDPs. It is notable that Liu et al. (2021a) make use of the optimism-pessimism principle that we developed in our earlier work (Pacchiano et al., 2021) and used in the analysis of this paper. Their result is a direct extension of ours to constrained RL.

Amani et al. (2021) extended this constrained RL setting to per-step (from per-round) constraints, i.e., the expected cost of the action taken at every visited state should be below a threshold. The key idea is that some actions are unsafe and need to be avoided at every step. Here the assumption is the

access to a safe action whose expected cost is below the threshold. They use the same geometric conditions that we impose in the analysis of our LC-LUCB algorithm (see Definition 8) to ensure that knowledge of a safe action is sufficient for safe exploration. Moreover, their algorithmic techniques rely heavily on our optimism-pessimism principle. Shi et al. (2023) later extended the per-step constrained RL work of Amani et al. (2021) to the case where some state/action combinations are unsafe.

3.1 A Summary of our Results

In this paper, we introduce several algorithms for constrained linear bandits in high-probability and in-expectation settings. The learner's objective is to achieve low regret while playing actions that satisfy a cost constraint. An action (or policy) is safe if its expected cost is upper-bounded by a known cost threshold τ. In order to achieve this, the learner has access to a safe arm x_0 that belongs to all contexts, and has an expected reward $\langle x_0, \theta_* \rangle = r_0$ and an expected cost $\langle x_0, \mu_* \rangle = c_0$ satisfying the constraint $c_0 < \tau$ (see Assumption 5). All our algorithms satisfy a regret bound of order $\mathcal{O}\left(\frac{1-r_0}{\tau - c_0} d\sqrt{T}\right)$ (ignoring logarithmic factors). In contrast, previous approaches such as Safe-TS satisfy a regret bound of order $\mathcal{O}\left(\frac{1}{\tau} d^{3/2}\sqrt{T}\right)$ for problems where $c_0 = 0$. In the following table, we compare and contrast our algorithms with Safe-TS. We also highlight the requirements that our approaches require for computational tractability.

Algorithm	**Contextual**	**Action Space**	$x_0 = \mathbf{0}$
Safe-LTS (Moradipari et al., 2019)	X	Convex and Compact	✓
LC-LUCB (Algorithm 1)	✓	Star Convex	X
OPLB (Algorithm 2)	✓	Arbitrary	X
OPB (Algorithm 3)	✓	Multi-Armed Bandits	X

Safe-LTS is not adapted to contextual scenarios and requires the action space to be convex and compact, and to contain the safe action $x_0 = \mathbf{0}$. In contrast, our algorithms LC-LUCB and OPLB are adapted to the contextual scenario. LC-LUCB achieves high probability guarantees and is tractable when the contexts are finite star-convex centered around the safe action x_0. In contrast, the OPLB algorithm achieves in-expectation guarantees and is not tractable for general context spaces. Finally, the OPB algorithm attains in-expectation guarantees in multi-armed bandit problems. Note that the OPB policies can be computed by solving a linear program.

Algorithm	**Regret Bound**	**Tractability**
Safe-LTS (Moradipari et al., 2019)	$\mathcal{O}\left(\frac{1}{\tau} d^{3/2}\sqrt{T}\right)$	✓
LC-LUCB (Algorithm 1)	$\mathcal{O}\left(\frac{1-r_0}{\tau - c_0} d\sqrt{T}\right)$	Finite Star-Convex
OPLB (Algorithm 2)	$\mathcal{O}\left(\frac{1-r_0}{\tau - c_0} d\sqrt{T}\right)$	X
OPB (Algorithm 3)	$\mathcal{O}\left(\frac{1-r_0}{\tau - c_0} \sqrt{KT}\right)$	Linear Program

4. High Probability Constraint Satisfaction

As described in Section 2, we study a *contextual linear bandit* setting in which each action (arm) is associated with two distributions, generating reward $R_t = \langle X_t, \theta_* \rangle + \xi_t^r$ and cost $C_t = \langle X_t, \mu_* \rangle + \xi_t^c$ signals. The agent aims to maximize its *expected cumulative reward* in T rounds, i.e., $\sum_{t=1}^{T} \langle X_t, \theta_* \rangle$, while is required to satisfy the ***stage-wise linear constraint***

$$\langle X_t, \mu_* \rangle \leq \tau, \quad \forall t \in [T], \tag{2}$$

with probability at least $1 - \delta$. The agent knows the constraint threshold $\tau \geq 0$ and has access to a safe action $x_0 \in \mathcal{A}_t$ with known cost $c_0 = \langle x_0, \mu_* \rangle < \tau$ and reward r_0 (Assumption 5).

Remark 7 *It is important to note that the high probability constrained setting described above cannot be solved for multi-armed bandits (MABs). This is because there is no generalization among the arms/actions in MABs, and thus, we cannot have an estimate of the cost of an arm without pulling it, which may itself violate the constraint* (2)*. In other words, pulling the safe action/arm, x_0, does not give us any information about the cost of the other arms in MABs. Thus, only interaction with decision sets $\mathcal{A}_t$ that allow for the safe exploration of progressively better actions may yield provable guarantees. We capture this intuition via a geometric condition on the decision sets $\mathcal{A}_t$ known as star-convexity. This is in contrast with the in-expectation constrained setting that we study in Section 5, where it is possible to guarantee safety by playing a distribution over the arms. Extensions to reinforcement learning, such as in Amani et al. (2021), follow the same in-expectation structure that we study in Section 5 and cannot be achieved in the high probability setting studied in this section.*

Definition 8 (star-convex set) *We call a set $\mathcal{A}$ star-convex around a point $x \in \mathcal{A}$ if for all other points $a \in \mathcal{A}$, the ray $[x, a]$ (the line between x and a) is in $\mathcal{A}$. When all action sets are star convex centered around x_0 the family of star-convex sets is rich enough to contain all convex sets (i.e., any convex set is star-convex).*

Definition 8 subsumes the case where the action sets $\mathcal{A}_t$ are convex, and thus, assuming $\mathcal{A}_t$'s are star-convex is weaker than assuming that they are convex. In this section, we make the following assumption:

Assumption 9 *All action sets $\mathcal{A}_t$ are star convex centered around the safe action x_0.*

Here we first propose an algorithm for the high probability *contextual linear bandit* setting described above. We provide its regret analysis under Assumption 9, prove a lower-bound for it, discuss how this setting can be extended to multiple constraints, and finally conclude with a set of experimental results as a proof of concept.

4.1 Algorithm

Let $\{X_s\}_{s=1}^t$ be the sequence of actions played by the agent up to time t, and $\{R_s = \langle X_s, \theta_* \rangle + \xi_s^r\}_{s=1}^t$ and $\{C_s = \langle X_s, \mu_* \rangle + \xi_s^c\}_{s=1}^t$ be the rewards and costs it observes in the same duration. Since the agent knows the cost of the safe action, i.e., $c_0 = \langle x_0, \mu_* \rangle$, it can compute the (random) cost incurred by X_t along the subspace $\mathcal{V}_o^{\perp}$, i.e., $C_t^{\perp} = C_t - \frac{\langle X_t, e_0 \rangle c_0}{\|x_0\|}$. The knowledge of c_0 allows us to build a (regularized) least-squares estimator for μ_* without estimating it along the e_0 direction

Algorithm 1 Linear Constraint Linear UCB (LC-LUCB)

Input: Safe action x_0 with reward r_0 and cost c_0; Constraint threshold $\tau \geq 0$; Scaling parameters $\alpha_r, \alpha_c \geq 1$
for $t = 1, \ldots, T$ **do**
  Observe star-convex $\mathcal{A}_t$ and build the estimated feasible action set $\widetilde{\mathcal{A}}_t^f$ using (6) and (7)
  Compute action $X_t = \arg\max_{x \in \widetilde{\mathcal{A}}_t^f} \widetilde{V}_t^r(x)$ *(see* (8) *and* (9) *for the definition of* $\widetilde{V}_t^r$*)*
  Take action X_t and observe reward and cost signals (R_t, C_t)
end for

(recall $x_0 = \|x_0\| e_0$). For any regularization parameter $\lambda > 0$, we define the regularized covariance matrix in round t as

$$\Sigma_t = \lambda \mathrm{I} + \sum_{s=1}^{t-1} X_s X_s^\top, \qquad \Sigma_t^{o,\perp} = \lambda \mathrm{I}^{o,\perp} + \sum_{s=1}^{t-1} X_s^{o,\perp} (X_s^{o,\perp})^\top, \tag{3}$$

where $\mathrm{I}^{o,\perp} = \mathrm{I} - e_0 e_0^\top$, and Σ_t and $\Sigma_t^{o,\perp}$ are the Gram matrices of the actions and projection of actions into the sub-space $\mathcal{V}_o^\perp$, respectively. Using (3), we define the regularized least-squares estimates $\widehat{\theta}_t$ and $\widehat{\mu}_t^{o,\perp}$ of the reward θ_* and cost $\mu_*^{o,\perp}$ parameters as

$$\widehat{\theta}_t = \Sigma_t^{-1} \sum_{s=1}^{t-1} R_s X_s, \qquad \widehat{\mu}_t^{o,\perp} = (\Sigma_t^{o,\perp})^\dagger \sum_{s=1}^{t-1} C_s^\perp X_s^{o,\perp}. \tag{4}$$

To define high probability confidence sets around estimators $\widehat{\theta}_t$ and $\widehat{\mu}_t^{o,\perp}$, and to capture how far they are from θ_* and $\mu_*^{o,\perp}$, we make use of Theorem 2 in Abbasi-Yadkori et al. (2011). These confidence sets, and in particular their radii, will play an important role in our algorithm.

Theorem 10 (Thm. 2 in Abbasi-Yadkori et al., 2011) *For a fixed* $\delta \in (0,1)$ *and*

$$\beta_t(\delta, d) = R\sqrt{d \log\left(\frac{1 + (t-1)L^2/\lambda}{\delta}\right)} + \sqrt{\lambda} S, \qquad \forall t \in [T],$$

it holds with probability (w.p.) at least $1 - \delta$ *that*

$$\|\widehat{\theta}_t - \theta_*\|_{\Sigma_t} \leq \beta_t(\delta, d), \qquad \|\widehat{\mu}_t^{o,\perp} - \mu_*^{o,\perp}\|_{\Sigma_t^{o,\perp}} \leq \beta_t(\delta, d-1).$$

Using Theorem 10, we now define the following confidence sets (ellipsoids):

$$\begin{aligned} \mathcal{C}_t^r(\alpha_r) &= \{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}_t\|_{\Sigma_t} \leq \alpha_r \beta_t(\delta, d)\}, \\ \mathcal{C}_t^c(\alpha_c) &= \{\mu \in \mathcal{V}_0^\perp : \|\mu - \widehat{\mu}_t^{o,\perp}\|_{\Sigma_t^{o,\perp}} \leq \alpha_c \beta_t(\delta, d-1)\}, \end{aligned} \tag{5}$$

around the estimates $\widehat{\theta}_t$ and $\widehat{\mu}_t^{o,\perp}$ with *scaling parameters* $\alpha_r, \alpha_c \geq 1$. It is important to note that these confidence sets are *asymmetrically scaled*, i.e., their radii have been scaled with different scaling parameters. Theorem 10 suggests that $\theta_* \in \mathcal{C}_t^r(\alpha_r)$ and $\mu_*^{o,\perp} \in \mathcal{C}_t^c(\alpha_c)$, each with probability at least $1 - \delta$.

Algorithm 1 contains the pseudo-code of our upper confidence bound (UCB) algorithm, which we call Linear Constraint Linear UCB (LC-LUCB). Our algorithm leverages the asymmetrically scaled confidence sets in (5) to appropriately balance its optimism about rewards and pessimism about costs. LC-LUCB starts by constructing a feasible (safe) action set $\widetilde{\mathcal{A}}_t^f$ from the original action set $\mathcal{A}_t$. In each round t, this is done by first computing a ***pessimistic cost value*** for an action x as

$$\widetilde{V}_t^c(x) = \underbrace{\frac{\langle x^o, e_0\rangle c_0}{\|x_0\|}}_{\text{known cost along } e_0} + \underbrace{\max_{\mu^{o,\perp} \in \mathcal{C}_t^c(\alpha_c)} \langle x^{o,\perp}, \mu^{o,\perp}\rangle}_{\text{max possible cost in } \mathcal{V}_o^\perp}. \tag{6}$$

Note that the known cost of x along e_0 equals $\frac{\langle x^o, e_0\rangle c_0}{\|x_0\|}$, since $\frac{c_0}{\|x_0\|}$ is the unit cost in direction e_0. Whenever the confidence interval $\mathcal{C}_t^c(\alpha_c)$ holds, $\widetilde{V}_t^c(x)$ overestimates the cost of action x (pessimistic). The feasible action set constructed by LC-LUCB in round t, i.e., $\widetilde{\mathcal{A}}_t^f$, contains all actions whose pessimistic cost value $\widetilde{V}_t^c(\cdot)$ is at most τ, i.e.,

$$\widetilde{\mathcal{A}}_t^f = \big\{x \in \mathcal{A}_t : \widetilde{V}_t^c(x) \le \tau\big\}. \tag{7}$$

We construct $\widetilde{\mathcal{A}}_t^f$ pessimistically in order to ensure that all its actions are indeed feasible. It is important to note that $\widetilde{\mathcal{A}}_t^f$ is always non-empty, since as a consequence of Assumption 5, the safe action x_0 is always in $\widetilde{\mathcal{A}}_t^f$.

LC-LUCB then proceeds by playing optimistically w.r.t. the reward signal, but only makes use of the feasible actions $x \in \widetilde{\mathcal{A}}_t^f$. In each round t, this is done by first computing an ***optimistic reward value*** for every action $x \in \mathcal{A}_t$ as

$$\widetilde{V}_t^r(x) = \max_{\theta \in \mathcal{C}_t^r(\alpha_r)} \langle x, \theta\rangle, \tag{8}$$

and then playing the arm X_t that maximizes it over the feasible action set $\widetilde{\mathcal{A}}_t^f$ (see Lines 2 and 3 of Algorithm 1). The following proposition contains the closed-form expressions for the *pessimistic* cost and *optimistic* reward values defined by (6) and (8).

Proposition 11 *The optimistic reward and pessimistic cost values in* (8) *and* (6) *can be written in closed-form as*

$$\widetilde{V}_t^r(x) = \langle x, \widehat{\theta}_t\rangle + \alpha_r \beta_t(\delta, d)\|x\|_{\Sigma_t^{-1}}, \tag{9}$$

$$\widetilde{V}_t^c(x) = \frac{\langle x^o, e_0\rangle c_0}{\|x_0\|} + \langle x^{o,\perp}, \widehat{\mu}_t^{o,\perp}\rangle + \alpha_c \beta_t(\delta, d-1)\|x^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}}. \tag{10}$$

Proof See Appendix A. ■

The leading term $\frac{\langle x^o, e_0\rangle c_0}{\|x_0\|}$ in (10) accounts for the knowledge of $\mu_*^{o,\perp}$ derived from the information we possess about the safe action x_0 and its cost c_0. Later we use (9) and (10) to derive a computationally efficient implementation of Algorithm 1 for a specific form of the action sets $\{\mathcal{A}_t\}_{t=1}^T$.

Remark 12 (unknown r_0 and c_0) *As discussed in Section 2, knowing a safe action x_0 is absolutely necessary for solving the constrained contextual linear bandit setting studied in this paper, otherwise,*

it would be impossible to satisfy the constraint from the very first round. However, we can relax the assumption of knowing the reward r_0 and cost c_0 of the safe arm. In this case, we start by playing x_0 for T_0 rounds in order to construct conservative estimates $\widehat{\Delta}_r$ and $\widehat{\Delta}_c$ of the quantities $1 - r_0$ and $\tau - c_0$ that satisfy $\widehat{\Delta}_r \geq \frac{1-r_0}{2}$ and $\widehat{\Delta}_c \geq \frac{\tau - c_0}{2}$. We then warm-start our estimators for θ_* and μ_* using the data collected by playing x_0 and instead of only estimating $\mu_*^{o,\perp}$, we build an estimator for μ_* over all its directions, including e_0, just as LC-LUCB already does for θ_*. Finally, we set $\frac{\alpha_r}{\alpha_c} = \frac{\widehat{\Delta}_r}{\widehat{\Delta}_c}$ and run Algorithm 1 for rounds $t > T_0$. The regret incurred during these first T_0 rounds can be upper bounded by $\mathcal{O}\left(\log(T/\delta)\max\left(\frac{1-r_0}{(\tau-c_0)^2}, \frac{1}{1-r_0}\right)\right)$. We report the details of this modification of LC-LUCB in Appendix F.1.

4.1.1 COMPUTATIONAL TRACTABILITY OF LC-LUCB

As described above, each round of LC-LUCB involves computing a feasible action set followed by selecting an action that maximizes a linear function over this set. Unfortunately, even if the action set $\mathcal{A}_t$ is convex, the feasible set $\widetilde{\mathcal{A}}_t^f$ can have a form for which maximizing the linear function is intractable.[3] Here we show (see Lemma 14) that whenever the action set $\mathcal{A}_t$ is *star-convex* and *finite*, (see Definition 8), the optimization in Line 2 of LC-LUCB can be solved efficiently.

Definition 13 (finite star-convex set) *We say a star-convex set (see Definition 13) is **finite**, if there exist finitely many points $\{x_i\}_{i=1}^M$ such that $\mathcal{A} = \cup_{i=1}^M \{[x, x_i]\}$.*

It is important to emphasize that according to Definition 13, a finite star-convex set is not necessarily a finite set and can have infinitely many members. We now report the main result of this section that shows when the action sets $\{\mathcal{A}_t\}_{t=1}^T$ are all star-convex and finite, the LC-LUCB algorithm is tractable. We also empirically evaluate LC-LUCB in Section 4.5.

Lemma 14 *If all action sets $\{\mathcal{A}_t\}_{t=1}^T$ are star-convex around the safe action x_0 and finite, then LC-LUCB can be implemented in polynomial time.*

Proof We may write each action set $\mathcal{A}_t$ as $\mathcal{A}_t = \cup_{i=1}^M \{[x_0, x_i]\}$, because it is star-convex around x_0 and finite. Since $x_0 \in \widetilde{\mathcal{A}}_t^f$, the feasible action set constructed by LC-LUCB, $\widetilde{\mathcal{A}}_t^f = \mathcal{A}_t \cap \{x : \widetilde{V}_t^c(x) \leq \tau\}$, is also a finite star-convex set around x_0 and can be written as $\widetilde{\mathcal{A}}_t^f = \cup_{i=1}^M \{[x_0, \widetilde{x}_i]\}$, where $\widetilde{x}_i = \alpha_i^* x_i$ and $\alpha_i^* = \arg\max_{\alpha \in [0,1],\, \alpha x_i \in \widetilde{\mathcal{A}}_t^f} \alpha$. Solving for α_i^* can be done by a simple line search, hence, Line 2 in Algorithm 1 can be executed by optimizing over each ray $[x_0, \widetilde{x}_i],\ \forall i \in [M]$. This optimization is easy because $\widetilde{V}_t^r(x)$ is a convex function of x (see Eq. 9), and thus, its maximum over the one dimensional set $[x_0, \widetilde{x}_i]$ is achieved at either x_0 or $\widetilde{x}_i$. ■

4.2 Regret Analysis

In this section, we prove a regret bound for Algorithm 1. Although LC-LUCB can be used in the presence of arbitrary action sets $\mathcal{A}_t$, we require $\mathcal{A}_t$ to be star convex around x_0 for our regret

3. Note that even in unconstrained linear bandits, the optimization problem that needs to be solved in each round of OFU-style algorithms (e.g., Abbasi-Yadkori et al. 2011) can be intractable even when the set is convex. This is because the problem of maximizing a quadratic form over a convex set can be hard in general.

analysis. Let $\{X_t\}_{t=1}^T$ be the sequence of actions selected by Algorithm 1 and $\{\widetilde{V}_t^r(X_t)\}_{t=1}^T$ be their corresponding *optimistic reward values* defined by (8) and (9). We start by adding $\{\widetilde{V}_t^r(X_t)\}_{t=1}^T$ to and subtracting them from the regret defined by (1), and rewriting it as

$$\mathcal{R}_{\mathcal{C}}(T) = \underbrace{\sum_{t=1}^{T} V_t^r(x_t^*) - \widetilde{V}_t^r(X_t)}_{\text{(I)}} + \underbrace{\sum_{t=1}^{T} \widetilde{V}_t^r(X_t) - V_t^r(X_t)}_{\text{(II)}}, \tag{11}$$

where for any action $x \in \mathcal{A}_t$, we denote its ***true reward value*** by $V_t^r(x) = \langle x, \theta_* \rangle$.

Optimism via Asymmetric Scaling. In the unconstrained bandit algorithms that are based on the OFU principle (e.g., Abbasi-Yadkori et al. 2011), term (I) in (11) is upper-bounded by 0. This is because most of such algorithms select action X_t that maximizes an optimistic reward value $\widetilde{V}_t^r : \mathcal{A}_t \to \mathbb{R}$, and thus, satisfies $\widetilde{V}_t^r(x_t) \geq V_t^r(x),\ \forall x \in \mathcal{A}_t$. Unfortunately, this property does not hold for LC-LUCB, because it selects X_t as the maximizer of $\widetilde{V}_t^r(x)$ over the pessimistic set $\widetilde{A}_t^f$ (see Eq. 8 and Line 2 in Algorithm 1), hence it is possible that $x_t^* \notin \widetilde{\mathcal{A}}_t^f$. Therefore, it does not immediately follow that $\widetilde{V}_t^r(X_t) \geq V_t^r(x_t^*)$ in LC-LUCB. We get around this limitation using the asymmetrically scaled confidence sets $\mathcal{C}_t^r(\alpha_r)$ and $\mathcal{C}_t^c(\alpha_c)$ defined in (5). By selecting α_r to be much larger than α_c, we ensure that the scaling of $\widetilde{V}_t^r(x)$ is enough to overcome the potential absence of x_t^* in $\widetilde{\mathcal{A}}_t^f$. This imbalanced scaling allows us to enjoy the benefits of optimism without requiring the optimal action x_t^* to be in the estimated set of feasible actions $\widetilde{\mathcal{A}}_t^f$. Although stretching the optimistic reward value $\widetilde{V}_t^r(x)$ allows us to control (I), the extra scaling causes challenges in bounding (II). As we will show in Lemma 18, the amount of stretching needed for the argument to work for (II) depends on the ratio between the *sub-optimality*, $1 - r_0$, and *safety gap*, $\tau - c_0$, of the safe action x_0. Our results indicate that the smaller the value of $\frac{1-r_0}{\tau - c_0}$, the harder learning becomes.

Before bounding the two terms in (11), we define the following event that according to Theorem 10 holds with probability at least $1 - \delta$:

$$\mathcal{E} := \left\{ \|\widehat{\theta}_t - \theta_*\|_{\Sigma_t} \leq \beta_t(\delta, d) \ \wedge\ \|\widehat{\mu}_t^{o,\perp} - \mu_*^{o,\perp}\|_{\Sigma_t^{o,\perp}} \leq \beta_t(\delta, d-1),\ \forall t \in [T] \right\}. \tag{12}$$

Bounding (II)**:** Let $\widetilde{\theta}_t = \arg\max_{\theta \in \mathcal{C}_t^r(\alpha_r)} \max_{x \in \widetilde{\mathcal{A}}_t^f} \langle x, \theta \rangle$ be the parameter attaining the optimistic maximum. Since $\widetilde{V}_t^r(X_t) = \langle X_t, \widetilde{\theta}_t \rangle$, we may write (II) $= \sum_{t=1}^T \langle X_t, \widetilde{\theta}_t - \theta_* \rangle$. We now state the following proposition that is used in bounding (II). 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 15 *For any given (possibly random) sequence of actions $\{x_s\}_{s=1}^t$, let Σ_t be its corresponding Gram matrix defined by* (3) *with $\lambda \geq 1$. Then, for all $t \in [T]$, we have*

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

Armed with Proposition 15, we now prove an upper-bound for (II) in the following lemma.

Lemma 16 *On event $\mathcal{E}$ defined by* (12) *(that holds with probability at least $1-\delta$), we have*

$$\text{(II)} \leq \alpha_r \beta_T(\delta, d) \sqrt{2Td \log\left(1 + \frac{TL^2}{\lambda}\right)}.$$

Proof The following inequalities hold on event $\mathcal{E}$:

$$\begin{aligned}
\sum_{t=1}^{T} \langle X_t, \widetilde{\theta}_t \rangle - \langle X_t, \theta_* \rangle &\overset{\text{(a)}}{\leq} \sum_{t=1}^{T} \|x_t\|_{\Sigma_t^{-1}} \|\widetilde{\theta}_t - \theta_*\|_{\Sigma_t} \\
&\overset{\text{(b)}}{\leq} \sum_{t=1}^{T} \alpha_r \beta_t(\delta, d) \|X_t\|_{\Sigma_t^{-1}} \overset{\text{(c)}}{\leq} \alpha_r \beta_T(\delta, d) \sum_{t=1}^{T} \|X_t\|_{\Sigma_t^{-1}} \\
&\overset{\text{(d)}}{\leq} \alpha_r \beta_T(\delta, d) \sqrt{2Td \log\left(1 + \frac{TL^2}{\lambda}\right)}.
\end{aligned}$$

(a) follows from Cauchy Schwartz. **(b)** is a direct consequence of conditioning on $\mathcal{E}$ that implies $\|\widetilde{\theta}_t - \theta_*\| \leq \alpha_r \beta_t(\delta, d)$. **(c)** holds because $\beta_t(\delta, d)$ is an increasing function of t. **(d)** follows from Proposition 15. ■

Bounding (I)**:** Here we show that by appropriately selecting the reward and cost scaling parameters α_r and α_c, we can guarantee optimism for our constrained linear bandit formulation, i.e., in each round $t \in [T]$, the optimistic reward value of the action selected by Algorithm 1, $\widetilde{V}_t^r(X_t)$, overestimates the true reward value of the optimal action, $V_t^r(x_t^*)$. This result implies that (I) can be upper-bounded by 0. Before proving the main result of this section (Lemma 18), we state the following supporting lemma, whose proof is reported in Appendix A.

Lemma 17 *For any $x \in \mathbb{R}^d$, the following inequality holds:*

$$\|x^{o,\perp}\|_{(\Sigma_t^{o,\perp})^\dagger} \leq \|x\|_{\Sigma_t^{-1}}. \tag{13}$$

We now find the appropriate conditions on α_r and α_c in order to ensure optimism for Algorithm 1.

Lemma 18 *If the scaling parameters α_r and α_c are set such that $\alpha_r, \alpha_c \geq 1$ and $(1+\alpha_c)(1-r_0) \leq (\tau - c_0)(\alpha_r - 1)$, then for all $t \in [T]$, with probability at least $1-\delta$, we have $\widetilde{V}_t^r(X_t) \geq V_t^r(x_t^*)$.*

Proof On event $\mathcal{E}$, for any action $x \in \mathcal{A}_t$, we have

$$\widetilde{V}_t^r(x) = \max_{\theta \in \mathcal{C}_t^r(\alpha_r)} \langle x, \theta \rangle \geq \langle x, \theta_* \rangle = V_t^r(x). \tag{14}$$

We divide the proof into two cases depending on whether in each round $t \in [T]$, the optimal action x_t^* belongs to the set of feasible actions $\widetilde{\mathcal{A}}_t^f$, or not.

Case 1. When $x_t^* \in \widetilde{\mathcal{A}}_t^f$, the result follows immediately, since by definition X_t is a maximizer of $\widetilde{V}_t^r(x)$ over $\widetilde{\mathcal{A}}_t^f$, and thus, we have

$$\widetilde{V}_t^r(X_t) \geq \widetilde{V}_t^r(x_t^*). \tag{15}$$

Combining (14) and (15), we can conclude that $\widetilde{V}_t(X_t) \geq V_t^r(x_t^*)$ as desired.

Case 2. When $x_t^* \notin \widetilde{\mathcal{A}}_t^f$, we know that the *pessimistic cost value* of the optimal action violates the constraint, i.e., $\widetilde{V}_t^c(x_t^*) > \tau$, while its ***true cost value*** satisfies the constraint, i.e., $V_t^c(x_t^*) := \langle x_t, \mu_* \rangle \leq \tau$. Since $\mathcal{A}_t$ is assumed to be star-convex around x_0, action $\gamma x_t^* + (1-\gamma)x_0 \in \mathcal{A}_t$ for all $\gamma \in [0,1]$. Now consider the following mixture action $\widetilde{x}_t = \gamma_t x_t^* + (1-\gamma_t)x_0$, where $\gamma_t \in [0,1]$ is the maximum value of γ for which the mixture action belongs to the estimated set of feasible actions, i.e., $\widetilde{x}_t \in \widetilde{\mathcal{A}}_t^f$. Since $\mathcal{A}_t$ is star-convex, all actions $\gamma x_t^* + (1-\gamma)x_0$ for $\gamma \leq \gamma_t$ are in $\widetilde{\mathcal{A}}_t^f$. From the definition of $\widetilde{x}_t$, we have

$$\widetilde{x}_t^{o,\perp} = \gamma_t x_t^{*,o,\perp}, \tag{16}$$

which allows us to write

$$\widetilde{V}_t^c(x_t^*) \overset{\text{(a)}}{=} \frac{\langle x_t^{*,o}, e_0 \rangle c_0}{\|x_0\|} + \langle x_t^{*,o,\perp}, \widehat{\mu}_t^{o,\perp} \rangle + \alpha_c \beta_t(\delta, d-1) \|x_t^{*,o,\perp}\|_{(\Sigma_t^{o,\perp})^\dagger}, \tag{17}$$

$$\begin{aligned}\widetilde{V}_t^c(\widetilde{x}_t) &\overset{\text{(b)}}{=} \frac{\left(\gamma_t \langle x_t^{*,o}, e_0 \rangle + (1-\gamma_t)\langle x_0, e_0 \rangle\right) c_0}{\|x_0\|} + \gamma_t \langle x_t^{*,o,\perp}, \widehat{\mu}_t^{o,\perp} \rangle + \gamma_t \alpha_c \beta_t(\delta, d-1) \|x_t^{*,o,\perp}\|_{(\Sigma_t^{o,\perp})^\dagger} \\ &\overset{\text{(c)}}{=} (1-\gamma_t)c_0 + \gamma_t \widetilde{V}_t^c(x_t^*).\end{aligned} \tag{18}$$

(a) is from the definition of pessimistic cost value in (10), **(b)** is obtained from the definition of $\widetilde{x}_t$, together with (10) and (16), and finally, **(c)** comes directly from (17).

Since $x_t^* \notin \widetilde{\mathcal{A}}_t^f$, from the definition of γ_t, it is easy to see that $\widetilde{V}_t^c(\widetilde{x}_t) = \tau$. Using this fact and (18), we first write γ_t in terms of $\widetilde{V}_t^c(x_t^*)$ and then with the following chain of inequalities obtain a lower-bound on γ_t as

$$\begin{aligned}\gamma_t &= \frac{\tau - c_0}{\widetilde{V}_t^c(x_t^*) - c_0} \\ &= \frac{\tau - c_0}{\frac{\langle x_t^{*,o}, e_0 \rangle c_0}{\|x_0\|} + \langle x_t^{*,o,\perp}, \widehat{\mu}_t^{o,\perp} \rangle + \alpha_c \beta_t(\delta, d-1) \|x_t^{*,o,\perp}\|_{(\Sigma_t^{0,\perp})^\dagger} - c_0} \\ &= \frac{\tau - c_0}{\frac{\langle x_t^{*,o}, e_0 \rangle c_0}{\|x_0\|} + \langle x_t^{*,o,\perp}, \mu_*^{o,\perp} \rangle + \langle x_t^{*,o,\perp}, \widehat{\mu}_t^{o,\perp} - \mu_*^{o,\perp} \rangle + \alpha_c \beta_t(\delta, d-1) \|x_t^{*,o,\perp}\|_{(\Sigma_t^{o,\perp})^\dagger} - c_0} \\ &\overset{\text{(a)}}{\geq} \frac{\tau - c_0}{\frac{\langle x_t^{*,o}, e_0 \rangle c_0}{\|x_0\|} + \langle x_t^{*,o,\perp}, \mu_*^{o,\perp} \rangle + (1+\alpha_c) \beta_t(\delta, d-1) \|x_t^{*,o,\perp}\|_{(\Sigma_t^{o,\perp})^\dagger} - c_0} \\ &\overset{\text{(b)}}{\geq} \frac{\tau - c_0}{\tau - c_0 + (1+\alpha_c) \beta_t(\delta, d-1) \|x_t^{*,o,\perp}\|_{(\Sigma_t^{o,\perp})^\dagger}}.\end{aligned} \tag{19}$$

(a) holds because

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

(b) holds because x_t^* is the optimal action in round t, and thus, $\frac{\langle x_t^{*,o}, e_0 \rangle c_0}{\|x_0\|} + \langle x_t^{*,o,\perp}, \mu_*^{o,\perp} \rangle \leq \tau$.

Now let's assume that $\langle x_0, \theta_* \rangle = \langle x_t^*, \theta_* \rangle - \Delta_t$ for all $t \in [T]$. Since both X_t and $\widetilde{x}_t$ are in the feasible set $\widetilde{\mathcal{A}}_t^f$, and given the definition of X_t, we may write

$$\begin{aligned}
\widetilde{V}_t^r(X_t) \geq \widetilde{V}_t^r(\widetilde{x}_t) &= \langle \widetilde{x}_t, \widehat{\theta}_t \rangle + \alpha_r \beta_t(\delta, d) \|\widetilde{x}_t\|_{\Sigma_t^{-1}} \\
&= \langle \widetilde{x}_t, \theta_* \rangle + \langle \widetilde{x}_t, \widehat{\theta}_t - \theta_* \rangle + \alpha_r \beta_t(\delta, d) \|\widetilde{x}_t\|_{\Sigma_t^{-1}} \\
&\overset{\text{(a)}}{\geq} \langle \widetilde{x}_t, \theta_* \rangle + (\alpha_r - 1) \beta_t(\delta, d) \|\widetilde{x}_t\|_{\Sigma_t^{-1}} \\
&\overset{\text{(b)}}{\geq} \langle \widetilde{x}_t, \theta_* \rangle + (\alpha_r - 1) \beta_t(\delta, d-1) \|\widetilde{x}_t^{o,\perp}\|_{(\Sigma_t^{o,\perp})^\dagger} \\
&\overset{\text{(c)}}{=} \gamma_t \langle x_t^*, \theta_* \rangle + (1 - \gamma_t) \langle x_0, \theta_* \rangle + \gamma_t (\alpha_r - 1) \beta_t(\delta, d-1) \|x_t^{*,o,\perp}\|_{(\Sigma_t^{o,\perp})^\dagger} \\
&\overset{\text{(d)}}{=} \langle x_t^*, \theta_* \rangle - (1 - \gamma_t) \Delta_t + \gamma_t (\alpha_r - 1) \beta_t(\delta, d-1) \|x_t^{*,o,\perp}\|_{(\Sigma_t^{o,\perp})^\dagger} \\
&= \langle x_t^*, \theta_* \rangle + \underbrace{\gamma_t \left((\alpha_r - 1) \beta_t(\delta, d-1) \|x_t^{*,o,\perp}\|_{(\Sigma_t^{o,\perp})^\dagger} + \Delta_t \right) - \Delta_t}_{\text{(V)}}. \qquad (20)
\end{aligned}$$

(a) follows from the definition of event $\mathcal{E}$ in (12) and Cauchy Schwartz, i.e.,

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

(b) is a consequence of Lemma 17. **(c)** is from the definition of $\widetilde{x}_t$ and (16). **(d)** follows from the assumption that $\langle x_0, \theta_* \rangle = \langle x_t^*, \theta_* \rangle - \Delta_t$.

Now we derive conditions under which term (V) in (20) is non-negative. To reduce notation clutter let $C_1 := \beta_t(\delta, d-1) \|x_t^{*,o,\perp}\|_{(\Sigma_t^{o,\perp})^\dagger}$. Then, the following inequality holds for (V):

$$\text{V} \geq \frac{\tau - c_0}{\tau - c_0 + (1 + \alpha_c) C_1} \big((\alpha_r - 1) C_1 + \Delta_t \big) - \Delta_t,$$

where the inequality follows by lower-bounding γ_t using (19).

Consequently if $\frac{\tau - c_0}{\tau - c_0 + (1+\alpha_c) C_1} \left((\alpha_r - 1) C_1 + \Delta_t \right) - \Delta_t \geq 0$, then (V) will be non-negative, which holds whenever

$$(\tau - c_0)(\alpha_r - 1) \geq (1 + \alpha_c) \Delta_t. \qquad (21)$$

By the definition of Δ_t and the fact that rewards are bounded in $[0, 1]$ (Assumption 4), we have $\Delta_t \leq 1 - r_0$. Thus, inequality (21) holds if $(\tau - c_0)(\alpha_r - 1) \geq (1 + \alpha_c)(1 - r_0)$. This concludes the proof, since we proved that $\widetilde{V}_t^r(X_t) \geq V_t^r(x_t^*)$ in both cases where $x_t^* \in \widetilde{\mathcal{A}}_t^f$ and $x_t^* \notin \widetilde{\mathcal{A}}_t^f$. ■

After bounding the two terms in (11) using Lemmas 16 to 18, we are now ready to state the main theorem of this section, which is a regret bound for Algorithm 1.

Theorem 19 (regret bound for LC-LUCB) *Let $\alpha_c = 1$ and $\alpha_r = 1 + \frac{2(1-r_0)}{\tau - c_0}$. Then, with probability at least $1 - \delta$, the regret of Algorithm 1 can be upper-bounded as*

$$\mathcal{R}_{\mathcal{C}}(T) \leq \alpha_r \beta_T(\delta, d) \sqrt{2Td \log \left(1 + \frac{TL^2}{\lambda} \right)}. \qquad (22)$$

Proof The proof follows directly from bounding (I) and (II) in the regret decomposition (11) using Lemmas 16 to 18. ■

Remark 20 *When $\lambda = 1$, ignoring logarithmic dependencies on T and $1/\delta$, the term $\beta_T(\delta, d)$ in* (22) *is of order $\sqrt{d}$. Thus, this parameter setting yields a regret bound of order $\mathcal{R}_\mathcal{C}(T) = \widetilde{\mathcal{O}}\big((1 + \frac{1-r_0}{\tau-c_0})d\sqrt{T}\big)$, which shows that LC-LUCB recovers the same dependence on d and T as unconstrained OFU-style linear bandit algorithms (e.g., Abbasi-Yadkori et al. 2011). The extra term of $\widetilde{\mathcal{O}}(\frac{1-r_0}{\tau-c_0}d\sqrt{T})$ is the cost of satisfying the constraint and the multiplier $\frac{1-r_0}{\tau-c_0}$ represents the hardness of the constrained problem.*

4.3 Lower Bound

We also prove a min-max lower-bound for the constrained contextual linear bandit setting described in Section 2. We prove in Theorem 21 that no algorithm can obtain a regret better than $\mathcal{O}\big(\max(d\sqrt{T}, \frac{1-r_0}{(\tau-c_0)^2})\big)$ on all such constrained contextual linear bandit instances. This result substantiates our intuition that learning while satisfying linear constraints is statistically harder than the unconstrained case, particularly when the safety gap $\tau - c_0$ is small w.r.t. the horizon T and the reward suboptimality $1 - r_0$.

Theorem 21 *Let $\tau, c_0, r_0 \in (0, 1)$, $B = \max\big(\frac{d\sqrt{T}}{8e^2}, \frac{1-r_0}{21(\tau-c_0)^2}\big)$, and assume $T \geq \max(d - 1, \frac{168eB}{1-r_0})$. Then, for any algorithm $\mathfrak{A}$, there is a pair of reward and cost parameters (θ_*, μ_*), such that $\mathcal{R}_\mathcal{C}(T) \geq B$.*

Proof See Appendix E. ■

Theorem 21 shows that if $\frac{1}{\tau-c_0} = \Omega(\sqrt{T})$ and $r_0 \leq 1/2$, learning while satisfying linear constraints is impossible in a min-max sense since in this case our lower bound indicates the regret must grow at least linearly. As an additional example, if $\frac{1}{\tau-c_0} = \sqrt{d}\, T^{3/8}$ and $r_0 \leq 1/2$, Theorem 21 implies that a constrained learner must incur $\Omega(d\, T^{3/4})$ regret, while unconstrained learning can achieve a regret rate of order $d\sqrt{T}$. This shows the existence of a fundamental statistical separation between constrained and unconstrained learning as a function of the ratio between the safety gap $\tau - c_0$ and the reward suboptimality $1 - r_0$. The question of whether the quadratic dependence on $\tau - c_0$ is optimal in this lower-bound remains open.

4.4 Extension to Multiple Constraints

The formulation, algorithm, and analysis of the previous sections can be extended to multiple constraints. In this setting, when the agent takes an action X_t in each round $t \in [T]$, in addition to the reward signal R_t, it observes a vector of m cost signals $C_t^{(i)} = \langle X_t, \mu_*^{(i)} \rangle + \xi_t^{c(i)},\ \forall i \in [m]$, where the reward and costs satisfy the assumptions listed in Section 2. The agent is required to satisfy m ***stage-wise linear constraints*** $\langle X_t, \mu_*^{(i)} \rangle \leq \tau_i,\ \forall i \in [m]$. Here we also need the following assumption for the safe action, which is a generalization of Assumption 5 to multiple constraints.

Assumption 22 (safe action) *There is a known* safe action $x_0 \in \mathcal{A}_t,\ \forall t \in [T]$, *with known reward r_0 and costs $c_0^{(i)} = \langle x_0, \mu_*^{(i)} \rangle \leq \tau_i,\ \forall i \in [m]$.*

In extending LC-LUCB to multiple constraints, we maintain estimators $\{\widehat{\mu}_t^{o,\perp(i)}\}_{i=1}^m$ for all cost parameters $\{\mu_*^{(i)}\}_{i=1}^m$, and construct the feasible action set as

$$\widetilde{\mathcal{A}}_t^f = \{x \in \mathcal{A}_t : \widetilde{V}_t^{c(i)}(x) \leq \tau_i,\ \forall i \in [m]\}, \tag{23}$$

where $\widetilde{V}_t^{c(i)}(\cdot)$ is the pessimistic cost value for the i'th cost signal. The rest of the algorithm remains unchanged.

To derive a regret bound for the extension of LC-LUCB to multiple constraints, it suffices to prove the following extension of Lemma 18.

Lemma 23 *If we set the scaling parameters α_r and α_c such that $\alpha_r, \alpha_c \geq 1$ and $(1+\alpha_c)(1-r_0) \leq \Delta_c^*(\alpha_r - 1)$, where $\Delta_c^* = \min_{i\in[m]}(\tau - c_0^{(i)})$, then for all $t \in [T]$, with probability at least $1-\delta$, we have $\widetilde{V}_t^r(X_t) \geq V^r(x_t^*)$.*

Proof A simple modification to the proof of Lemma 18 yields the desired result. Note that substituting $\tau - c_0$ with Δ_c^* and following the same argument as in the derivation of inequality (19) yields

$$\gamma_t \geq \frac{\Delta_c^*}{\Delta_c^* + (1+\alpha_c)\beta_t(\delta, d-1)\|x_t^{*,o,\perp}\|_{(\Sigma_t^{o,\perp})^\dagger}}. \tag{24}$$

Plugging this result into (20) and continuing with the proof logic of Lemma 18 concludes the proof. ■

Lemma 23 allows us to derive a regret bound for the extension of LC-LUCB to multiple constraints, identical to the one we proved for the single-constraint case in Theorem 19.

Theorem 24 *Let $\alpha_c = 1$ and $\alpha_r = 1 + 2(1-r_0)/\Delta_c^*$. Then, with probability at least $1-\delta$, the regret of the extension of LC-LUCB to multiple constraints can be upper-bounded as*

$$\mathcal{R}_{\mathcal{C}}(T) \leq \alpha_r \beta_T(\delta, d)\sqrt{2Td\log\left(1 + \frac{TL^2}{\lambda}\right)}. \tag{25}$$

We show in Appendix F.1 how the multi-constraint algorithm (similar to the single-constraint case) can be changed to handle the scenario where the reward r_0 and costs $\{c_0^{(i)}\}_{i=1}^m$ of the safe action are unknown.

4.5 Experiments

In this section we compare the performance of LC-LUCB and the Safe-LTS algorithm of Moradipari et al. (2019) in two simulation-based experiments. In each of these scenarios we show that LC-LUCB performs better than Safe-LTS. In all our experiments, we run a regularized least-squares regression by setting $\lambda = 1$.

In our first experiment, presented in Figure 1, we consider a linear bandit problem in which the safe action is the zero-vector $x_0 = 0$ and the arm sets, $\mathcal{A}_t$, are 10 dimensional star-convex sets generated by the 10 cyclic shifted versions of the vector $v/\|v\|$, where $v = (0, 1, \ldots, 9)$. For all t, the action set $\mathcal{A}_t$ is the star-convex set defined by this set of actions and the lines emanating from

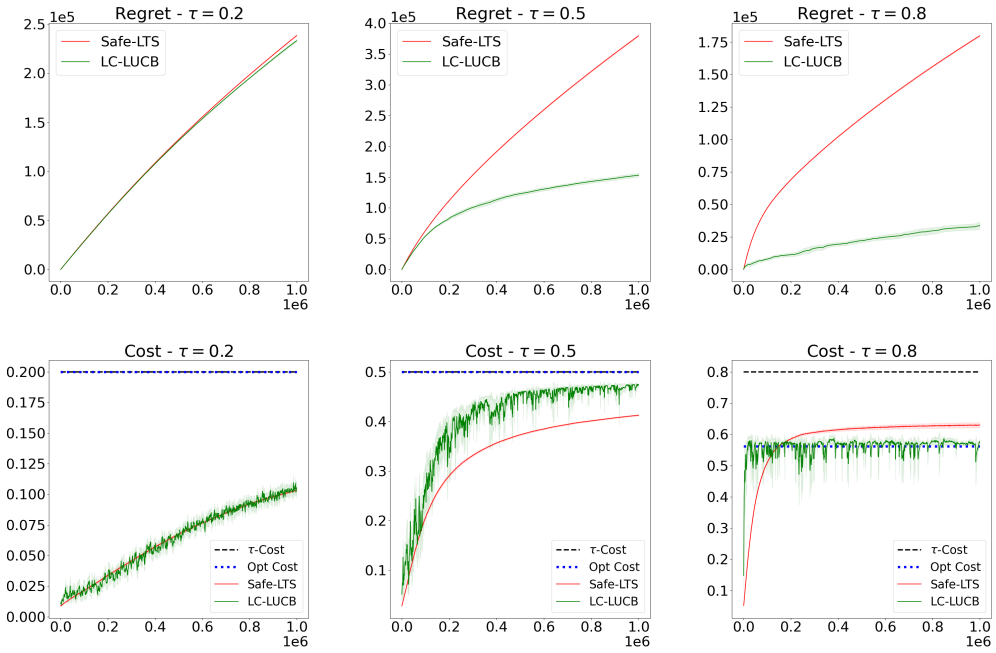

Figure 1: Regret and cost evolution of **LC-LUCB** and **Safe-LTS**. **Left:** Constraint Threshold $\tau = 0.2$. **Center:** Constraint Threshold $\tau = 0.5$. **Right:** Constraint Threshold $\tau = 0.8$. Learning is harder for smaller thresholds τ. The arm sets $\mathcal{A}_t$ are 10 dimensional star-convex sets generated by the 10 cyclic shifted versions of the vector $v/\|v\|$, where $v = (0, 1, \ldots, 9)$. There exist optimal unconstrained solutions with cost less than 0.561. This is less than the cost threshold $\tau = 0.8$. This is why the lower right cost evolution plot shows convergence to a level below the 0.8 threshold.

the zero vector. We set $\theta_* = v/\|v\|$ where $v = (0, 1, \ldots, 9)$ and[4] $\mu_* = (9, 8, \ldots, 0)/\|v\|$ to be the ℓ_2 normalized version of v and $(9, \ldots, 0)$. In Figure 1, we plot the regret and cost evolution of LC-LUCB for different threshold values τ, and compare them with those for the Safe-LTS algorithm of Moradipari et al. (2019). The safe action is the zero vector and each plot is an average over 10 runs. We show that as the threshold τ is driven to 0, the problem gets progressively harder. The results show that for all threshold values and dimensions, LC-LUCB has a better regret profile than Safe-LTS, while satisfying the constraint. We report the results for dimensions $d = 3$ and $d = 5$ of this problem, and also show the reward evolution (in addition to regret and cost) for LC-LUCB and Safe-LTS in Appendix 5.4.

In our second experiment, presented in Figure 2, we consider the setting where the action sets $\mathcal{A}_t$ are the unit ball (infinite) and the safe action is the zero vector. Our plots compare the regret across time of Safe-LTS and LC-LUCB when averaged across problem instances. We generate different problem instances by sampling θ_* and μ_* vectors uniformly from the unit sphere and also generate thresholds τ by sampling uniformly from the interval $[0, 1]$. Each sample run of this experiment corresponds to a sample problem instance. In order to make the optimization problem at each round of LC-LUCB tractable for this infinite size action set, we approximate $\mathcal{A}_t$ by sampling 100 vectors $\{v_i\}_{i=1}^{1000}$ uniformly from the unit sphere and defining an approximate (still infinite) action set $\tilde{\mathcal{A}}_t$,

4. The vector $(9, 8, \ldots, 0)$ is the flipped version of v.

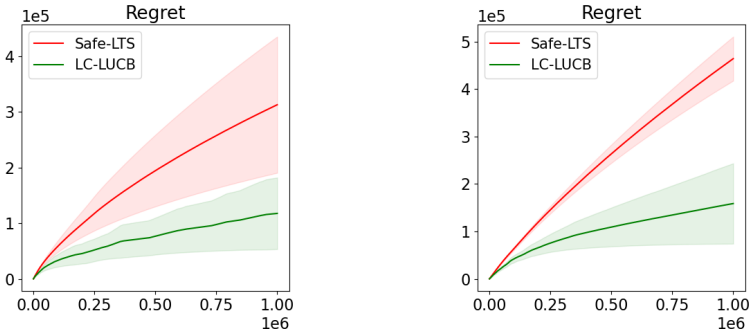

Figure 2: Unit sphere, **left** dimension $d = 5$, **right** dimension $d = 10$.

consisting of the rays from zero to each of the v_i. Figure 2 compares the regret of LC-LUCB with Safe-LTS for dimensions $d = 5$ and $d = 10$ of this problem. Each plot is averaged over 10 sample runs and the shaded regions around the curves correspond to 1 standard deviation. Similar to the previous experiment, LC-LUCB shows better performance than Safe-LTS.

5. Constraint Relaxation: From High Probability to Expectation

In Section 4, we studied a constrained contextual linear bandit setting in which the agent maximizes its expected T-round cumulative reward (minimizes its expected T-round constrained pseudo-regret) while satisfying a stage-wise linear high probability constraint defined in (2). In many constrained or multi-objective problems, making sure that the constraints are not violated or certain objectives are within certain thresholds with high probability would result in overly conservative strategies. A common solution to balance performance and constraint satisfaction is to replace conservative *high probability* constraints with more relaxed *in-expectation* constraints. In this section, we study such a relaxation in which the high probability constraint (2) is replaced with a constraint in expectation. We describe this relaxed setting in Section 5.1. We then propose an algorithm for this setting, provide its regret analysis, specialize our results to multi-armed bandits (MABs),[5] and report experimental results as a proof of concept. Additionally, in Section 5.3 we extend these results to the scenario where the reward and cost functions are non-linear. We propose an algorithm for this setting with regret analysis. We use a characterization of function class complexity based on the eluder dimension (Russo and Van Roy, 2013) in the derivation of this regret bound.

5.1 Linear Contextual Bandits with In-expectation Stage-Wise Linear Constraints

In the relaxed setting we study in this section, in each round $t \in [T]$, the agent selects its action $X_t \in \mathcal{A}_t$ according to its policy $\pi_t \in \Pi_t = \Delta_{\mathcal{A}_t}$, 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***

$$\mathbb{E}_{X\sim\pi_t}[\langle X, \mu_* \rangle] \leq \tau, \quad \forall t \in [T], \tag{26}$$

5. Unlike the high probability constrained setting of Section 4 that is impossible to solve it in MABs (see Remark 7), the relaxed setting we study in this section can be solved for MABs.

Algorithm 2 Optimistic-Pessimistic Linear Bandit (OPLB)

Input: Safe action x_0 with reward r_0 and cost c_0; Constraint threshold $\tau \geq 0$; Scaling parameters $\alpha_r, \alpha_c \geq 1$
for $t = 1, \ldots, T$ **do**
  Compute regularized least-squares estimates $\widehat{\theta}_t$ and $\widehat{\mu}_t^{o,\perp}$ using (4)
  Construct confidence sets $\mathcal{C}_t^r(\alpha_r)$ and $\mathcal{C}_t^c(\alpha_c)$ using (5)
  Observe $\mathcal{A}_t$ and construct the estimated feasible policy set $\widetilde{\Pi}_t^f$ using (33)
  Compute policy $(\pi_t, \widetilde{\theta}_t) = \arg\max_{\pi \in \widetilde{\Pi}_t^f,\, \theta \in \mathcal{C}_t^r(\alpha_r)} \mathbb{E}_{x\sim\pi}[\langle x, \theta\rangle]$
  Take action $X_t \sim \pi_t$ and observe reward and cost signals (R_t, C_t)
end for

where $\tau \geq 0$ is the *constraint threshold*. Note that unlike the constraint (2) studied in Section 4 which is in high probability, this constraint is in expectation.

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., $\Pi_t^f = \{\pi \in \Pi_t : \mathbb{E}_{X\sim\pi}[\langle X, \mu_*\rangle] \leq \tau\}$. Maximizing the expected cumulative reward in T rounds is equivalent to minimizing the T-round *constrained (pseudo)-regret*

$$\mathcal{R}_\Pi(T) = \sum_{t=1}^{T} \mathbb{E}_{X\sim\pi_t^*}[\langle X, \theta_*\rangle] - \mathbb{E}_{X\sim\pi_t}[\langle X, \theta_*\rangle], \tag{27}$$

where $\pi_t, \pi_t^* \in \Pi_t$ for all $t \in [T]$, and $\pi_t^* \in \max_{\pi\in\Pi_t^f} \mathbb{E}_{X\sim\pi}[\langle X, \theta_*\rangle]$ is the *optimal feasible* policy in round t. The terms $\mathbb{E}_{X\sim\pi}[\langle X, \theta_*\rangle]$ in (27) and $\mathbb{E}_{X\sim\pi}[\langle X, \mu_*\rangle]$ in (26) 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, respectively. With these notations, we may write the T-round regret in (27) as

$$\mathcal{R}_\Pi(T) = \sum_{t=1}^{T} R_{\pi_t^*} - R_{\pi_t}. \tag{28}$$

Notation. Here we use some extra notations in addition to those defined in Section 2. We 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.

5.1.1 ALGORITHM

We propose a UCB-style algorithm for this setting, called *optimistic-pessimistic linear bandit* (OPLB), because it maintains a pessimistic assessment of the set of available policies, while acting optimistically within this set. Algorithm 2 contains the pseudo-code of OPLB. Similar to LC-LUCB (Algorithm 1), the main idea behind Algorithm 2 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 in detail the steps of OPLB that differ from the LC-LUCB algorithm.

Just as in the analysis of LC-LUCB, the choice of $\alpha_r, \alpha_c \geq 1$ and Theorem 10 suggest that $\theta_* \in \mathcal{C}_t^r(\alpha_r)$ and $\mu_*^{o,\perp} \in \mathcal{C}_t^c(\alpha_c)$, each w.p. at least $1 - \delta$. Replacing actions with policies in (6)

and (8), we define the *optimistic reward* and *pessimistic cost* values for any policy π in round t as

$$\widetilde{V}_t^r(\pi) := \max_{\theta \in \mathcal{C}_t^r(\alpha_r)} \mathbb{E}_{X\sim\pi}[\langle X, \theta\rangle], \tag{29}$$

$$\widetilde{V}_t^c(\pi) := \frac{\langle x_\pi^o, e_0\rangle c_0}{\|x_0\|} + \max_{\mu \in \mathcal{C}_t^c(\alpha_c)} \mathbb{E}_{X\sim\pi}[\langle X, \mu\rangle]. \tag{30}$$

Similar to Proposition 11, we can derive the following closed-form expressions for $\widetilde{V}_t^r(\pi)$ and $\widetilde{V}_t^c(\pi)$.

Proposition 25 *We may write $\widetilde{V}_t^r(\pi)$ and $\widetilde{V}_t^c(\pi)$, defined in* (29) *and* (30)*, in closed-form as*

$$\widetilde{V}_t^r(\pi) = \langle x_\pi, \widehat{\theta}_t\rangle + \alpha_r\beta_t(\delta, d)\|x_\pi\|_{\Sigma_t^{-1}}, \tag{31}$$

$$\widetilde{V}_t^c(\pi) = \frac{\langle x_\pi^o, e_0\rangle c_0}{\|x_0\|} + \langle x_\pi^{o,\perp}, \widehat{\mu}_t^{o,\perp}\rangle + \alpha_c\beta_t(\delta, d-1)\|x_\pi^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}}. \tag{32}$$

Proof See Appendix C. ■

Following the same logic as in the high probability formulation, we adapt the pessimistic estimation of the feasible action set $\widetilde{\mathcal{A}}_t^f$ in (7) to the policy setting. After observing the action set $\mathcal{A}_t$, OPLB constructs its feasible (safe) policy set as

$$\widetilde{\Pi}_t^f = \big\{\pi \in \Delta_{\mathcal{A}_t} : \widetilde{V}_t^c(\pi) \le \tau\big\}. \tag{33}$$

Note that $\widetilde{\Pi}_t^f$ is an approximation of Π_t^f, i.e., the set of feasible policies over the action set $\mathcal{A}_t$, and is not an empty set because π_0 is always in $\widetilde{\Pi}_t^f$. We can think of the safe action x_0 as a policy π_0 whose probability mass is all on x_0, and thus, we have $x_{\pi_0}^o = x_0$ and $x_{\pi_0}^{o,\perp} = 0$. Plugging π_0 into (32) yields $\widetilde{V}_t^c(\pi_0) = \frac{\langle x_{\pi_0}^o, e_0\rangle c_0}{\|x_0\|} = c_0 \le \tau$. We prove in the following proposition that all policies in $\widetilde{\Pi}_t^f$ are feasible with high probability.

Proposition 26 *With probability at least $1-\delta$, for all rounds $t \in [T]$, all policies in $\widetilde{\Pi}_t^f$ are feasible.*

Proof See Appendix C. ■

In Line 6 of Algorithm 2, the agent computes its policy π_t as the one that is safe (belongs to $\widetilde{\Pi}_t^f$) and attains the maximum optimistic reward value, i.e., $\widetilde{V}_t^r(\pi) = \max_{\theta\in\mathcal{C}_t^r(\alpha_r)}\langle x_\pi, \theta\rangle = \langle x_{\pi_t}, \widetilde{\theta}_t\rangle$. We refer to $\widetilde{\theta}_t$ as the *optimistic reward parameter*.

Computational Complexity of OPLB. As shown in Line 6 of Algorithm 2, in each round t, OPLB solves the following optimization problem:

$$\begin{aligned} \max_{\pi\in\Delta_{\mathcal{A}_t}} \ & \langle x_\pi, \widehat{\theta}_t\rangle + \alpha_r\beta_t(\delta, d)\|x_\pi\|_{\Sigma_t^{-1}}, \\ \text{s.t.} \quad & \frac{\langle x_\pi^o, e_0\rangle c_0}{\|x_0\|} + \langle x_\pi^{o,\perp}, \widehat{\mu}_t^{o,\perp}\rangle + \alpha_c\beta_t(\delta, d-1)\|x_\pi^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}} \le \tau. \end{aligned} \tag{34}$$

However, solving (34) could be challenging. The bottleneck is in computing the safe policy set $\widetilde{\Pi}_t^f$, which is the intersection of $\Delta_{\mathcal{A}_t}$ and the ellipsoidal constraint. This is a consequence of the intractability of the optimization problem that needs to be solved in each round of OFU-style algorithms (e.g., Abbasi-Yadkori et al. 2011). In contrast to LC-LUCB (see Section 4.1.1), solving (34) could be intractable even when the action set $\mathcal{A}_t$ is finite star-convex around the safe action x_0.

Unknown r_0 and c_0. In case that the reward r_0 and cost c_0 of the safe action are unknown, OPLB may use the same warm-starting sub-routine as in LC-LUCB to estimate them with sufficient accuracy. The regret incurred during these first T_0 rounds can be upper bounded by $\mathcal{O}\left(\log(T/\delta)\max\left(\frac{1-r_0}{(\tau-c_0)^2}, \frac{1}{1-r_0}\right)\right)$. We discuss this in more details in Appendix F.1.

5.1.2 REGRET ANALYSIS

In this section, we prove a regret bound for the OPLB algorithm. The main challenge in obtaining this regret bound is to ensure that optimism holds in each round $t \in [T]$, i.e., the solution $(\pi_t, \widetilde{\theta}_t)$ of (34) satisfies $\widetilde{V}_t^r(\pi_t) = \langle x_{\pi_t}, \widetilde{\theta}_t \rangle \geq V_t^r(\pi_t^*)$. This is not obvious, since the approximate feasible policy set $\widetilde{\Pi}_t^f$ might have been constructed such that it does not contain the optimal policy π_t^*. Similar to the analysis of LC-LUCB in Section 4.2, 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 $\mu_*^{o,\perp}$, respectively, that allows us to guarantee optimism by appropriately selecting the ratio $\frac{\alpha_r}{\alpha_c}$. We will show in our analysis that similar to the case of LC-LUCB, $\frac{\alpha_r}{\alpha_c}$ depends on the ratio $\frac{1-r_0}{\tau-c_0}$.

Theorem 27 (Regret Bound for OPLB) *Let $\alpha_c = 1$ and $\alpha_r = 1 + \frac{2(1-r_0)}{\tau-c_0}$. Then, with probability at least $1-2\delta$, the regret of OPLB satisfies*

$$\mathcal{R}_\Pi(T) \leq \frac{2L(\alpha_r+1)\beta_T(\delta,d)}{\sqrt{\lambda}}\sqrt{2T\log(1/\delta)} + (\alpha_r+1)\beta_T(\delta,d)\sqrt{2Td\log\left(1+\frac{TL^2}{\lambda}\right)}\,. \quad (35)$$

We provide a proof sketch for Theorem 27 here. Most of the results are obtained in a similar fashion as those in the analysis of the LC-LUCB algorithm in Section 4.2. We use the high probability event $\mathcal{E}$ defined by (12). We first decompose the regret $\mathcal{R}_\Pi(T)$ in (27) as

$$\mathcal{R}_\Pi(T) = \underbrace{\sum_{t=1}^T V_t^r(\pi_t^*) - \widetilde{V}_t^r(\pi_t)}_{\text{(I)}} + \underbrace{\sum_{t=1}^T \widetilde{V}_t^r(\pi_t) - V_t^r(\pi_t)}_{\text{(II)}}\,, \quad (36)$$

where for any policy $\pi \in \Pi_t^f$, we denote by $V_t^r(\pi) = \langle x_\pi, \theta_* \rangle$, its ***true reward value*** and by $\widetilde{V}_t^r(\pi_t)$ its optimistic reward value defined by (29) and (31). We first bound term (II) in (36) by further decomposing it as

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

When the event $\mathcal{E}$ defined by (12) holds, (IV) can be bounded by Lemma 16 as

$$(\mathrm{IV}) \leq \alpha_r\beta_T(\delta,d)\sqrt{2Td\log\left(1+\frac{TL^2}{\lambda}\right)}\,.$$

We bound the sum of (III) and (V) in the following lemma.

Lemma 28 *On the event $\mathcal{E}$ in* (12)*, for any $\delta' \in (0,1)$, with probability at least $1-\delta'$, we have*

$$\text{(III)} + \text{(V)} \leq \frac{2L(\alpha_r+1)\beta_T(\delta,d)}{\sqrt{\lambda}} \cdot \sqrt{2T\log(1/\delta')}\,.$$

Proof Applying Cauchy-Schwartz to $\text{(III)} + \text{(V)} = \sum_{t=1}^{T} \langle x_{\pi_t} - x_t, \widetilde{\theta}_t - \theta_* \rangle$, we may write $|\langle x_{\pi_t} - X_t, \widetilde{\theta}_t - \theta_* \rangle| \leq \|x_{\pi_t} - X_t\|_{\Sigma_t^{-1}} \|\widetilde{\theta}_t - \theta_*\|_{\Sigma_t}$. Since $\widetilde{\theta}_t \in \mathcal{C}_t^r(\alpha_r)$ on event $\mathcal{E}$, we have $\|\widetilde{\theta}_t - \theta_*\|_{\Sigma_t} \leq \alpha_r \beta_t(\delta,d)$. From the definition of Σ_t, we have $\Sigma_t \succeq \lambda I$, and thus, $\|x_{\pi_t} - X_t\|_{\Sigma_t^{-1}} \leq \|x_{\pi_t} - X_t\|/\sqrt{\lambda} \leq 2L/\sqrt{\lambda}$. Hence, $Y_t = \sum_{s=1}^{t} \langle x_{\pi_s} - X_s, \widetilde{\theta}_s - \theta_* \rangle$ is a martingale sequence with $|Y_t - Y_{t-1}| \leq 2L(\alpha_r+1)\beta_t(\delta,d)/\sqrt{\lambda}$, for all $t \in [T]$. Using the Azuma–Hoeffding inequality and since β_t is an increasing function of t, i.e., $\beta_t(\delta,d) \leq \beta_T(\delta,d),\ \forall t \in [T]$, with probability at least $1-\delta'$, we may write $\mathbb{P}\big(Y_T \geq 2L\alpha_r\beta_T(\delta,d)\sqrt{2T\log(1/\delta')/\lambda}\big) \leq \delta'$, which concludes the proof. ■

We now bound term (I) in (36). Similar to the regret proof for LC-LUCB, setting the values of α_r and α_c to 1 and then solving for π_t is not enough to ensure $\widetilde{V}_t^r(\pi_t) \geq V_t^r(\pi_t^\star)$. However, an appropriate choice of radii α_r and α_c for the confidence intervals can help us to get around this issue. Lemma 29 contains the main result in which we prove that by appropriately setting α_r and α_c, we can guarantee that in each round $t \in [T]$, OPLB selects an optimistic policy, i.e., a policy π_t whose optimistic reward, $\widetilde{V}_t^r(\pi_t)$, is larger than the reward of the optimal policy, $V_t^r(\pi_t^*)$, on event $\mathcal{E}$. This means that with this choice of α_r and α_c, (I) is always non-positive. This result is the in-expectation version of the one proved in Lemma 18.

Lemma 29 *On the event $\mathcal{E}$ defined by* (12)*, if we set α_r and α_c such that $\alpha_r, \alpha_c \geq 1$ and $(1+\alpha_c)(1-r_0) \leq (\tau - c_0)(\alpha_r - 1)$, then for any $t \in [T]$, we have $\widetilde{V}_t^r(\pi_t) \geq V_t^r(\pi_t^*)$.*

Lemma 29 is a corollary of Lemma 18. The exact same proof argument holds with a few notational substitutions. The optimal action $x_t^* \in \widetilde{\mathcal{A}}_t^f$ is substituted with $x_{\pi_t^*}$, which is the average action vector of $\pi_t^* \in \widetilde{\Pi}_t^f$. Consequently, the mixture action $\widetilde{X}_t$ is substituted with $x_{\widetilde{\pi}_t}$, where $\widetilde{\pi}_t = \eta_t \pi_t^* + (1-\eta_t)\pi_0$, π_0 is the safe 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., $\widetilde{\pi}_t \in \widetilde{\Pi}_t^f$. The rest of the proof remains unchanged.

5.2 Specializing to Multi-Armed Bandits

We now specialize the results of this section to multi-armed bandits (MABs) and show that the structure of the MAB problem allows a computationally efficient implementation of our OPLB algorithm as well as an improvement in its regret bound.

In the MAB setting, the action set consists of K arms $\mathcal{A} = \{1, \ldots, K\}$, where 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 bandit case studied above, 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* $\mathbb{E}_{A_t \sim \pi_t}[\bar{c}_{A_t}] \leq \tau,\ \forall t \in [T]$. Moreover, Arm 1 is assumed to be the known safe arm, i.e., its mean reward $\bar{r}_1$ and cost $\bar{c}_1$ are known, and $\bar{c}_1 \leq \tau$.

Algorithm 3 Optimism-Pessimism Bandit (OPB)

1: **Input:** Number of arms K; Mean reward $\bar{r}_1$ and cost $\bar{c}_1$ of the safe arm; Constraint threshold $\tau \geq 0$; Scaling parameters $\alpha_r, \alpha_c \geq 1$
2: **for** $t = 1, \ldots, T$ **do**
3: Compute estimates $\{u_a^r(t)\}_{a \in \mathcal{A}}$ and $\{u_a^c(t)\}_{a \in \mathcal{A}}$
4: Form the approximate LP (37) using the estimates in Line 3
5: Compute policy π_t by solving (37)
6: Play arm $a \sim \pi_t$
7: **end for**

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 $t \in [T]$:

$$\max_{\pi \in \Delta_K} \sum_{a \in \mathcal{A}} \pi_a \; u_a^r(t), \qquad \text{s.t.} \; \sum_{a \in \mathcal{A}} \pi_a \; u_a^c(t) \leq \tau. \tag{37}$$

As (37) indicates, OPB selects its policy by being optimistic about reward and pessimistic about cost (using an upper-bound for both r and c). Algorithm 3 contains the pseudo-code of OPB. Note that similar to OPLB, OPB constructs an (approximate) feasible (safe) policy set of the form $\widetilde{\Pi}_t^f = \left\{\pi \in \Delta_K : \sum_{a \in \mathcal{A}} \pi_a u_a^c(t) \leq \tau\right\}$ (see Eq. 37) and sets $\beta_a(0) = 0, \; \forall a \in \mathcal{A}$.

Computational Complexity of OPB. Unlike OPLB whose optimization problem might be complex to solve, OPB can be implemented extremely efficiently. The following lemma shows that (37) always has a solution (policy) with support of at most 2. This property allows us to solve (37) in closed-form without a LP solver and to implement OPB very efficiently.

Lemma 30 *There exists a policy that solves* (37) *and has at most* 2 *non-zero entries.*

Proof See Appendix D.1. ■

Regret Analysis of OPB. We also prove the following regret-bound for OPB.

Theorem 31 *Let* $\delta = 4KT\delta'$, $\alpha_c = 1$, *and* $\alpha_r = 1 + \frac{2(1-\bar{r}_1)}{\tau - \bar{c}_1}$. *Then, with probability at least* $1 - \delta$, *the regret of OPB satisfies*

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

Proof See Appendix D.2. ■

The main component of this proof is the following lemma, which is the analogous to Lemma 29 (and therefore also Lemma 18) in the contextual linear bandit case.

Lemma 32 *If we set α_r and α_c such that $\alpha_r, \alpha_c \geq 1$ and $(1+\alpha_c)(1-\bar{r}_1) \leq (\tau - \bar{c}_1)(\alpha_r - 1)$, then with high probability, for any $t \in [T]$, we have $\mathbb{E}_{a\sim\pi_t}[u_a^r(t)] \geq \mathbb{E}_{a\sim\pi^*}[\bar{r}_a]$.*

Proof See Appendix D.2. ■

Remark 33 *The constrained contextual linear bandit formulation of Section 5.1 is general enough to include the constrained MAB one described here. As a result the regret bound of OPLB in Theorem 27 can be instantiated for the constrained MAB setting, in which case it yields a regret bound of order $\widetilde{\mathcal{O}}\big((1+\frac{1-\bar{r}_1}{\tau-\bar{c}_1})K\sqrt{T}\big)$. However, our OPB regret bound in Theorem 31 is of order $\widetilde{\mathcal{O}}((1+\frac{1-\bar{r}_1}{\tau-\bar{c}_1})\sqrt{KT})$, which shows $\sqrt{K}$ 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 min-max lower-bound for our constrained MAB problem. Our lower-bound shows that no algorithm can attain a regret better than $\mathcal{O}\big(\max(\sqrt{KT}, \frac{1-\bar{r}_1}{(\tau-\bar{c}_1)^2})\big)$ for this problem. The formal statement of the lower-bound can be found below.

Theorem 34 *Let $\tau, \bar{c}_1, \bar{r}_1 \in (0,1)$, $B = \max\big(\frac{1}{27}\sqrt{(K-1)T}, \frac{1-\bar{r}_1}{21(\tau-\bar{c}_1)^2}\big)$, and assume $T \geq \max(K-1, \frac{168eB}{1-\bar{r}_1})$. Then, for any algorithm there is a pair of reward and cost parameters (θ_*, μ_*), such that $\mathcal{R}_{\mathcal{C}}(T) \geq B$.*

Proof See Appendix E. ■

Extension to Multiple Constraints. In the case of m constraints, the learner 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 D.4.

Unknown $\bar{c}_1$ and $\bar{r}_1$. The same warm starting sub-routine as in LC-LUCB and OPLB can be used for computing sufficiently accurate estimators of $\bar{r}_1$ and $\bar{c}_1$ for OPB. A detailed explanation can be found in Appendix F.1.

5.3 Extension to Nonlinear Rewards and Costs

Building on a rich line of research that extends bandit problems to function approximation—such as Russo and Van Roy (2013), Zhou et al. (2020), Foster and Rakhlin (2020), and Liu and Wang (2023)—we investigate how to adapt constrained bandit problems to this setting. In this work, we leverage the eluder dimension framework to develop our bounds and algorithms.

In this section, we explore constrained learning beyond linearity and extend the OPLB algorithm to the setting where the reward and cost functions are possibly nonlinear functions of the actions. We call the resulting algorithm Optimistic-Pessimistic Nonlinear Bandit algorithm (OPNLB). In each round $t \in [T]$, the agent is given an action set $\mathcal{A}_t \subseteq \mathcal{A}$, where $\mathcal{A}$ is a formal action set. Upon taking action $X_t \in \mathcal{A}_t$ the learner observes a reward-cost pair (R_t, C_t) such that $R_t = \theta_*(X_t) + \xi_t^r$

and $C_t = \mu_*(X_t) + \xi_t^c$, where $\theta_*(\cdot) \in \mathcal{G}_r$ and $\mu_*(\cdot) \in \mathcal{G}_c$ are the mean reward and mean cost functions that belong to known function classes $\mathcal{G}_r$ and $\mathcal{G}_c$, and ξ_t^r and ξ_t^c are conditionally zero-mean sub-Gaussian random variables. The rewards and costs satisfy the following assumption.

Assumption 35 (Bounded Responses) *For all $t \in [T]$ and $x \in \mathcal{A}_t$, the mean rewards and costs are bounded, i.e., $\theta_*(x), \mu_*(x) \in [0,1]$. Moreover, the rewards and costs observed in all rounds of the algorithm are also bounded, i.e., $R_t, C_t \in [0,1],\ \forall t \in [T]$.*

Moreover, the action sets $\mathcal{A}_t$ satisfy the safe action Assumption 5, i.e., there is an action $x_0 \in \mathcal{A}_t$, $\forall t \in [T]$, with known average reward $r_0 = \theta_*(x_0)$ and known average cost $c_0 = \mu_*(x_0)$, such that $c_0 < \tau$. The policy π_t according to which an action is taken in round t is an element of $\Delta_{\mathcal{A}_t}$. We denote by $\Pi_t^f = \{\pi \in \Delta_{\mathcal{A}_t} : \mathbb{E}_{X\sim\pi}[\mu_*(X)] \leq \tau\}$ and $\pi_t^* = \arg\max_{\pi\in\Pi_t^f} \mathbb{E}_{X\sim\pi}[\theta_*(X)]$ the set of feasible polices and the optimal policy in round $t \in [T]$. Finally, we define the T-round regret as

$$\mathcal{R}_\Pi(T) = \sum_{t=1}^{T} \mathbb{E}_{X\sim\pi_t^*}\big[\theta_*(X)\big] - \mathbb{E}_{X\sim\pi_t}\big[\theta_*(X)\big]. \tag{38}$$

The nonlinear reward and cost model that we discuss in this section does not allow for a high probability constraint satisfaction scenario without making strong assumptions on the action set. The star-convexity requirement of Definition 8 does not extend to non-linear action spaces. This is the reason why we study an expected constraint scenario instead. Before introducing our algorithm for this setting, we formally define the eluder dimension, i.e., a notion of complexity relevant to adaptive selection procedures introduced in Russo and Van Roy (2013).

Definition 36 (Action Independence and Eluder Dimension) *Let $\varepsilon > 0$ and $\{x_i\}_{i=1}^n$ be a set of actions. Then, we have the following definitions:*

- *An action x is ε-dependent on $\{x_i\}_{i=1}^n$ w.r.t. the function space $\mathcal{G}$, if any $f, f' \in \mathcal{G}$ satisfying $\sqrt{\sum_{i=1}^n (f(x_i) - f'(x_i))^2} \leq \varepsilon$ also satisfy $|f(x) - f'(x)| \leq \varepsilon$. An action x is ε-independent of $\{x_i\}_{i=1}^n$ w.r.t. $\mathcal{G}$ if it is not ε-dependent on $\{x_i\}_{i=1}^n$.*
- *The ε-eluder dimension $d_{\text{eluder}}(\mathcal{G}, \varepsilon)$ is the length of the longest sequence of elements in $\{x_i\}_{i=1}^n$ such that for some $\varepsilon' \geq \varepsilon$, every element is ε'-independent of its predecessors.*

Throughout this section, we will use the notation $d_{\text{eluder}}^r = d_{\text{eluder}}(\mathcal{G}_r, 1/T)$ and $d_{\text{eluder}}^c = d_{\text{eluder}}(\mathcal{G}_c, 1/T)$ to denote the eluder dimensions of the function spaces $\mathcal{G}_r$ and $\mathcal{G}_c$, respectively.

Optimistic-Pessimistic Non-Linear Bandit (OPNLB). In each round $t \in [T]$, we define two confidence sets

$$C_t^r(\delta) = \big\{\theta \in \mathcal{G}_r : \|\theta - \widehat{\theta}_t\|_{\mathcal{D}_t} \leq \gamma_r(t, \delta/2)\big\},$$
$$C_t^c(\delta) = \big\{\mu \in \mathcal{G}_c : \|\mu - \widehat{\mu}_t\|_{\mathcal{D}_t} \leq \gamma_c(t, \delta/2) \ \text{ and } \ \mu(x_0) = c_0\big\},$$

where $\mathcal{D}_t = \{(X_s, R_s, C_s)\}_{s=1}^{t-1}$ is the dataset of actions, rewards, and costs observed up until the beginning of round t, $\|f\|_{\mathcal{D}_t} = \sqrt{\sum_{x\in\mathcal{D}_t} f^2(x)}$ is the norm defined by the dataset $\mathcal{D}_t$ for any function $f : \mathcal{A}_t \to \mathbb{R}$, and $\gamma_r(t,\delta)$ and $\gamma_c(t,\delta)$ are the reward and cost confidence set radii defined as

$$\gamma_r(t,\delta) = 512 \log\left(\frac{24|\mathcal{G}_r|\log(2t)}{\delta}\right), \qquad \gamma_c(t,\delta) = 512 \log\left(\frac{24|\mathcal{G}_c|\log(2t)}{\delta}\right).$$

Algorithm 4 Optimistic-Pessimistic Nonlinear Bandit (OPNLB)

Input: Safe action x_0 with reward r_0 and cost c_0; Constraint threshold $\tau \geq 0$; Scaling parameters $\alpha_r, \alpha_c \geq 1$
for $t = 1, \ldots, T$ **do**
  Compute $\widehat{\theta}_t, \widehat{\mu}_t$ using least squares
  Observe action set $\mathcal{A}_t$ and construct the estimated feasible policy set $\widetilde{\Pi}_t^f$ using (40).
  Compute policy $\pi_t = \arg\max_{\pi \in \widetilde{\Pi}_t^f} \widetilde{V}_t^r(\pi)$.
  Take action $X_t \sim \pi_t$ and observe reward and cost signals (R_t, C_t)
end for

Let $\mathcal{E}$ be the event defined as

$$\mathcal{E} := \{\theta_* \in C_t^r(\delta) \ \wedge \ \mu_* \in C_t^r(\delta), \ \forall t \in [T]\}. \tag{39}$$

This is the same event as in (12) for the linear case where the confidence sets $C_t^r(\delta)$ and $C_t^c(\delta)$ satisfy $\theta_* \in C_t^r(\delta)$ and $\mu_* \in C_t^c(\delta)$ for all $t \in [T]$. Corollary 55 in Appendix G.1 implies that $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$.

We define pessimistic cost $\widetilde{V}_t^c(\pi)$ and optimistic reward $\widetilde{V}_t^r(\pi)$ values for a policy π in each round t as

$$\widetilde{V}_t^c(\pi) = \max_{\mu \in C_t^c(\delta)} \mathbb{E}_{x\sim\pi}\big[\mu(x)\big],$$
$$\widetilde{V}_t^r(\pi) = \max_{\theta \in C_t^r(\delta)} \mathbb{E}_{x\sim\pi}\big[\theta(x)\big] + \alpha_r \max_{\mu', \mu'' \in C_t^c(\delta)} \mathbb{E}_{x\sim\pi}\big[\mu'(x)\big] - \mathbb{E}_{x\sim\pi}\big[\mu''(x)\big].$$

Note that $\widetilde{V}_t^r$ is the combination of an optimistic reward estimator plus an artificially inflated confidence interval that depends on the cost function class. We define the set of feasible policies in round t as

$$\widetilde{\Pi}_t^f = \big\{\pi \in \Delta_{\mathcal{A}_t} : \widetilde{V}_t^c(\pi) \leq \tau\big\}. \tag{40}$$

We now show that by appropriately setting the scaling parameters α_r and α_c, the policy π_t selected by Algorithm 4 is feasible and a basic optimistic relationship holds between π_t and π_t^*, i.e., $\widetilde{V}_t^r(\pi_t) \geq V_t^r(\pi_t^*)$. The following lemma is the equivalent of Lemma 29 from the linear case.

Lemma 37 *If the event $\mathcal{E}$ defined by* (39) *holds and the scaling parameter satisfies $\alpha_r = \frac{1-r_0}{\tau - c_0}$, then for all $t \in [T]$, we have $\widetilde{V}_t^r(\pi_t) \geq V_t^r(\pi_t^*) = \theta_*(\pi_t^*, \mathcal{A}_t)$.*

Proof See Appendix G.2. ■

The proof of Lemma 37 follows a similar logic as that of Lemmas 18 and 29. Given this result, we now prove a regret bound for OPNLB in terms of the eluder dimensions d_{eluder}^r and d_{eluder}^c. When

$\mathcal{E}$ holds for all $t \in [T]$, the following inequalities are satisfied

$$\begin{aligned}
\mathcal{R}_\Pi(T) &= \sum_{t=1}^{T} \mathbb{E}_{X\sim\pi_t^*}\big[\theta_*(X)\big] - \mathbb{E}_{X\sim\pi_t}\big[\theta_*(X)\big] \\
&\overset{\text{(a)}}{\leq} \sum_{t=1}^{T} \widetilde{V}_t^r(\pi_t) - \mathbb{E}_{X\sim\pi_t}\big[\theta_*(X)\big] \\
&= \sum_{t=1}^{T} \max_{\theta\in C_t^r(\delta)} \mathbb{E}_{X\sim\pi_t}\big[\theta(X)\big] - \mathbb{E}_{X\sim\pi_t}\big[\theta_*(X)\big] + \alpha_r \max_{\mu,\mu'\in C_t^c(\delta)} \mathbb{E}_{X\sim\pi_t}\big[\mu(X)\big] - \mathbb{E}_{X\sim\pi_t}\big[\mu'(X)\big] \\
&\overset{\text{(b)}}{\leq} \sum_{t=1}^{T} \max_{\theta,\theta'\in C_t^r(\delta)} \mathbb{E}_{X\sim\pi_t}\big[\theta(X)\big] - \mathbb{E}_{X\sim\pi_t}\big[\theta'(X)\big] + \alpha_r \max_{\mu,\mu'\in C_t^c(\delta)} \mathbb{E}_{X\sim\pi_t}\big[\mu(X)\big] - \mathbb{E}_{X\sim\pi_t}\big[\mu'(X)\big].
\end{aligned}$$

(a) holds because of Lemma 37. **(b)** holds because $\theta_* \in C_t^r(\delta)$ for all $t \in [T]$ with probability at least $1-\delta$. The inequality sequence above implies that the regret can be upper-bounded by a weighted sum of uncertainty widths. We bound the sum of the uncertainty widths using Lemma 3 in Chan et al. (2023) (by setting the parallelism parameter $P = 1$) as

$$\sum_{t=1}^{T} \max_{\theta,\theta'\in C_t^r(\delta)} \mathbb{E}_{X\sim\pi_t}\big[\theta(X)\big] - \mathbb{E}_{X\sim\pi_t}\big[\theta'(X)\big] = \mathcal{O}\left(d^r_{\text{eluder}} + \sqrt{T d^r_{\text{eluder}} \gamma_r(T, \delta/2)}\right)$$

and

$$\sum_{t=1}^{T} \max_{\mu,\mu'\in C_t^c(\delta)} \mathbb{E}_{X\sim\pi_t}\big[\mu(X)\big] - \mathbb{E}_{X\sim\pi_t}\big[\mu'(X)\big] = \mathcal{O}\left(d^c_{\text{eluder}} + \sqrt{T d^c_{\text{eluder}} \gamma_c(T, \delta/2)}\right).$$

Combining these results and using $\mathbb{P}(\mathcal{E}) \geq 1-\delta$, we obtain the main result of this section, which is a regret bound for the OPNLB algorithm (Algorithm 4).

Theorem 38 (OPNLB regret-bound) *With probability at least $1-\delta$, the regret of Algorithm 4 satisfies*

$$\mathcal{R}_\Pi(T) = \mathcal{O}\left(\sqrt{T d^r_{\text{eluder}} \gamma_r(T, \delta/2)} + \frac{1-r_0}{\tau - c_0}\sqrt{T d^c_{\text{eluder}} \gamma_c(T, \delta/2)} + d^r_{\text{eluder}} + \frac{1-r_0}{\tau-c_0} d^c_{\text{eluder}}\right).$$

Remark 39 *It is not possible to extend the results of this section (the non-linear case) to the high probability setting studied in Section 4. In the linear high probability scenario, star-convexity around the safe action x_0 allows the learner to form a model of $\mu_\star(x)$ by playing a convex combination of the safe action and any other action x. In the non-linear setting, since $\mathcal{A}_t \subset \mathcal{A}$ is a formal action set, closure of $\mathcal{A}_t$ under convexity is not defined. Thus, it is possible to have actions x that are safe, i.e., $\mu_\star(x) < \tau$, but can never b explored safely.*

Computational Tractability of ONPLB. Step 5 of Algorithm 4 involves solving a constrained optimization problem that, in general, can be intractable. It remains an open question how to design tractable algorithms for constrained non-linear bandit problems.

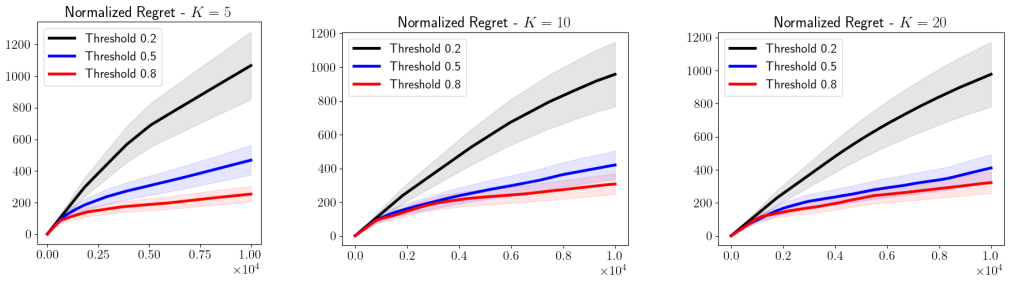

Figure 3: Regret of OPB for three instances of the randomly generated constrained multi-armed bandit problem with the number of arms equal to 5 *(a)*, 10 *(b)*, and 20 *(c)*. The cost and reward of the safe arm are set to $\bar{c}_1 = \bar{r}_1 = 0$.

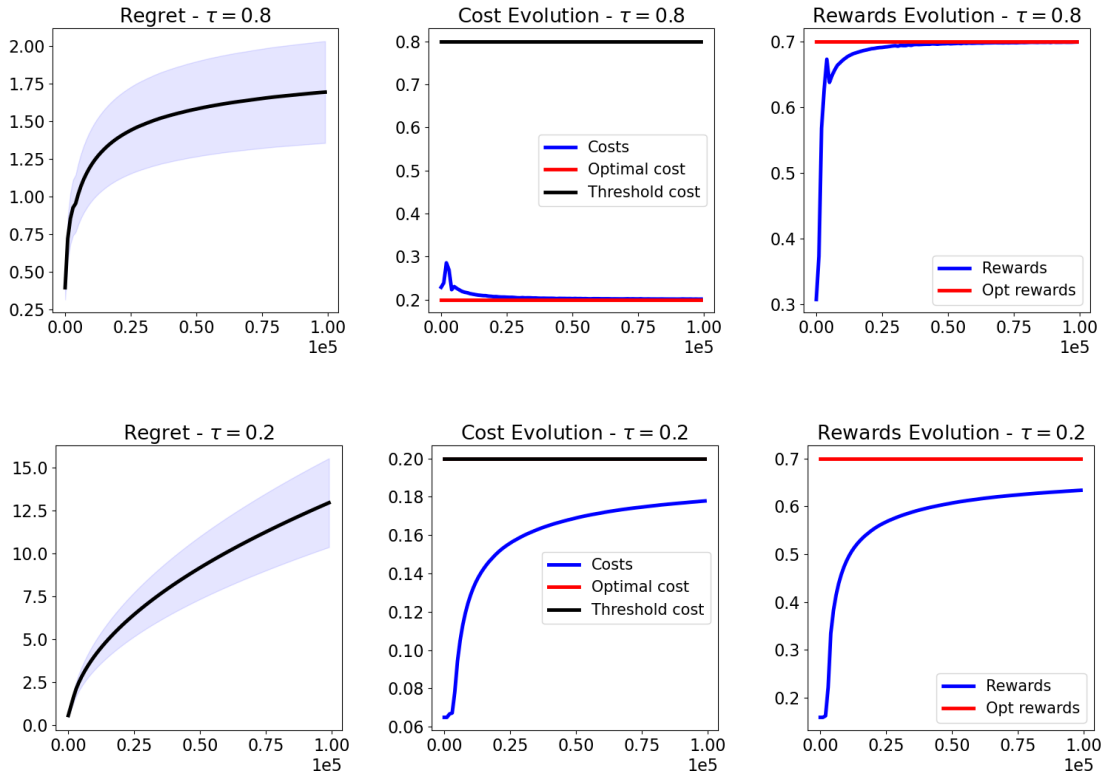

Figure 4: Regret *(left)*, cost *(middle)*, and reward *(right)* evolution of OPB in a 4-armed bandit problem with Bernoulli reward and cost distributions with means $\bar{r} = (0.1, 0.2, 0.4, 0.7)$ and $\bar{c} = (0, 0.4, 0.5, 0.2)$. The cost of the safe arm (Arm 1) is $\bar{c}_1 = 0$. The constraint threshold is set to $\tau = 0.8$ *(top)* and $\tau = 0.2$ *(bottom)*.

5.4 Experimental Results

We run a set of experiments to show the behavior of the OPB algorithm and validate our theoretical results. In our first experiment, presented in Figure 3, we produce random instances of our constrained multi-armed bandit problem. In all the instances, we set the safe arm to have reward and cost 0. We generate different problem instances by sampling the Bernoulli mean rewards and costs of the rest of the arms uniformly at random from the interval $[0, 1]$. Each sample run in this experiment corresponds to a sample problem instance. In Figure 3, 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 3 also shows that the regret of OPB grows inversely with the safety gap.

In the next experiment, presented in Figure 4, 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 $\bar{c}_1 = 0$. 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.

Figure 4 shows that the regret of OPB grows as we reduce τ, and as a result the safety gap *(left)*. 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 the bottom three plots of Figure 4, 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. In Figure 8 in Appendix 5, we report more experiments with the same 4-armed bandit problem instance with constraint threshold values $\tau = 0.5$ and 0.6. Using these intermediate threshold values we provide further support to our results showing the the safety gap governs the complexity in this constrained bandit problem.

6. Conclusions

In this work, we expand the frontier of the study of constrained bandit problems with anytime cost constraints. We extend the results of Pacchiano et al. (2021) in a variety of ways. First, we introduce the high probability constraint satisfaction regime for linear bandit problems with stage-wise constraints along with the LC-LUCB algorithm (Section 4). This formulation captures problems where an in-expectation constraint is not sufficient to ensure safety. We show that in contrast with OPLB, when the action set is finite and star-convex, the LC-LUCB algorithm is computationally tractable (Section 4.1.1). This stands in marked contrasts with the case of OPLB, that only has a tractable form in the multi-armed bandit setting (see the OPB algorithm in Section 5.2). Second, we improve the regret-bound of OPLB reported in Pacchiano et al. (2021) to better identify the quantity representing the hardness of the constrained problem $\frac{1-r_0}{\tau - c_0}$.

Finally, we go beyond the scenario of linear rewards and cost functions and explore the nonlinear regime where the reward and cost functions come from arbitrary function classes of bounded eluder dimension (Section 5.3). When the reward and cost function classes are arbitrary and the requirement

is to satisfy an anytime expected cost constraint, we introduce the OPNLB algorithm and prove it satisfies a regret-bound equivalent to the regret-bound for OPLB where the eluder dimension of the reward and cost function classes plays the role of the linear dimension in the linear case. Since the eluder dimension of linear classes equals the dimension of the ambient space, these results subsume the regret-bounds for OPLB in Pacchiano et al. (2021).

The design of all of our algorithms (LC-LUCB, OPLB, OPB and OPNLB) relies on the principle of optimism-pessimism and the technique of asymmetric confidence intervals that enables the provable analysis of optimistic-pessimistic algorithms. We hope the results of this work can serve as inspiration to extend the study of stage-wise constrained problems to richer scenarios such as reinforcement learning and beyond.

Acknowledgments

A.P. would like to thank the support of the Eric and Wendy Schmidt Center at the Broad Institute of MIT and Harvard. This work was supported in part by funding from the Eric and Wendy Schmidt Center at the Broad Institute of MIT and Harvard.

Contents

1 Introduction **1**

2 Problem Formulation **3**

3 Related Work **5**
3.1 A Summary of our Results 6

4 High Probability Constraint Satisfaction **7**
4.1 Algorithm 7
4.1.1 Computational Tractability of LC-LUCB 10
4.2 Regret Analysis 10
4.3 Lower Bound 15
4.4 Extension to Multiple Constraints 15
4.5 Experiments 16

5 Constraint Relaxation: From High Probability to Expectation **18**
5.1 Linear Contextual Bandits with In-expectation Stage-Wise Linear Constraints . . . 18
5.1.1 Algorithm 19
5.1.2 Regret Analysis 21
5.2 Specializing to Multi-Armed Bandits 22
5.3 Extension to Nonlinear Rewards and Costs 24
5.4 Experimental Results 29

6 Conclusions **29**

A Proofs of Section 4 **33**

B Additional Experiments for Section 4 **34**

C Proofs of Section 5 **34**

D Constrained Multi-Armed Bandits **37**
D.1 The LP Structure 37
D.2 Regret Analysis 39
D.3 Proof of Lemma 32 39
D.4 Multiple Constraints 42

E Lower Bounds **43**

F Extensions **47**
F.1 Unknown c_0 and r_0 47

G Nonlinear Rewards **49**
G.1 Properties of Least Squares Estimators 49
G.2 Proof of Lemma 37 52

H Additional Experiments of Section 5 **54**

Appendix A. Proofs of Section 4

Proof of Proposition 11: We only prove the statement (9) for the optimistic reward $\widetilde{V}_t^r(x)$. The proof of statement (10) for the pessimistic cost $\widetilde{V}_t^c(x)$ is analogous. From the definition of the confidence set $\mathcal{C}_t^r(\alpha_r)$, 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

$$\begin{aligned}\widetilde{V}_t^r(x) &= \max_{\beta \in \mathcal{C}_t^r(\alpha_r)} \langle x, \theta \rangle = \langle x, \widehat{\theta}_t \rangle + \max_{v: \|v\|_{\Sigma_t} \leq \alpha_r \beta_t(\delta,d)} \langle x, v \rangle \\ &\overset{\text{(a)}}{\leq} \langle x, \widehat{\theta}_t \rangle + \alpha_r \beta_t(\delta, d) \|x\|_{\Sigma_t^{-1}}. \end{aligned} \tag{41}$$

(a) By Cauchy-Schwartz, for all v, we have $\langle x, v \rangle \leq \|x\|_{\Sigma_t^{-1}} \|v\|_{\Sigma_t}$. 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 $v^* := \frac{\alpha_r \beta_t(\delta,d) \Sigma_t^{-1} x}{\|x\|_{\Sigma_t^{-1}}}$. This value of v^* is feasible because

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

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

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

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

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

which concludes the proof. ■

Proof of Lemma 18: In order to prove the desired result, it is enough to show that

$$(x^{o,\perp})^\top (\Sigma_t^{o,\perp})^\dagger x^{o,\perp} \leq x^\top \Sigma_t^{-1} x.$$

Without loss of generality, we can assume $x_o = e_1$, where e_1 is the first basis vector. Note that in this case $\Sigma_t^{o,\perp}$ can be thought of as a sub-matrix of Σ_t such that $\Sigma_t[2:, 2:] = \Sigma_t^{o,\perp}$, where $\Sigma_t[2:, 2:]$ denotes the sub-matrix with row and column indices from 2 onward. Using the following formula for the inverse of a positive semi-definite (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 $D = z - \delta^\top A^{-1} \delta$, we may write Σ_t^{-1} as

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

where $D = \Sigma_t[1,1] - \Sigma_t[2:d]^\top (\Sigma_t^{o,\perp})^\dagger \Sigma_t[2:d] \in \mathbb{R}$. This allows us to write

$$\begin{aligned} x^\top(\Sigma_t^{-1})x &= \frac{x(1)^2 - 2x(1)\Sigma_t[2:d]^\top(\Sigma_t^{o,\perp})^\dagger x[2:d]}{D} \\ &+ \frac{x[2:d]^\top(\Sigma_t^{o,\perp})^\dagger \Sigma_t[2:d]\Sigma_t[2:d]^\top(\Sigma_t^{o,\perp})^\dagger x[2:d]}{D} + x[2:d]^\top(\Sigma_t^{o,\perp})^\dagger x[2:d] \\ &\geq x[2:d]^\top(\Sigma_t^{o,\perp})^\dagger x[2:d]. \end{aligned}$$

The result follows by noting that $x[2:d] = x^{o,\perp}$. ■

Appendix B. Additional Experiments for Section 4

In this section, we present a comprehensive set of results extending the experiments presented in Figure 1. We consider a linear bandit problem in which the safe action equals the zero vector $x_0 = 0$ and the arm sets $\mathcal{A}_t$ are d dimensional star convex sets generated by the d cyclic shifted versions of the vector $v/\|v\|$ where $v = (0, 1, \cdots, d-1)$. Just like in Figure 1, the action set $\mathcal{A}_t$ is the star convex set defined by this set of actions and the lines emanating from the zero vector. We let We let $\theta_* = v/\|v\|$ and $\mu_\star = (d-1, d-2, \cdots, 0)/\|v\|$, where $(d-1, d-2, \cdots, 0)$ is the flipped version of $(0, 1, \cdots, d-1)$.

In Figures 5, 6, and 7, we plot the regret and cost evolution of LC-LUCB for dimensions $d = 3, 5, 10$, and threshold values $\tau = 0.2, 0.5, 0.8$, and compare them with those for the Safe-LTS algorithm of Moradipari et al. (2019). The results for dimensions $d = 3, 5$ and 10 are presented in Figures 5, 6, and 7 respectively. We show that as the threshold τ is driven to 0, the problem gets progressively harder. The results show that LC-LUCB has a better regret profile than Safe-LTS, while satisfying the constraint, for all threshold values and dimensions.

Appendix C. Proofs of Section 5

Proof of Proposition 25: The proof follows the exact same structure as Proposition 11. Instead of using Equation 41 we utilize the following identity,

$$\begin{aligned} \widetilde{V}_t^r(\pi) &= \max_{\theta \in \mathcal{C}_t^r(\alpha_r)} \mathbb{E}_{X\sim\pi}[\langle X, \theta\rangle] = \max_{\theta \in \mathcal{C}_t^r(\alpha_r)} \langle x_\pi, \theta\rangle = \langle x_\pi, \widehat{\theta}_t\rangle + \max_{v:\|v\|_{\Sigma_t} \leq \alpha_r \beta_t(\delta,d)} \langle x_\pi, v\rangle \\ &\overset{(a)}{\leq} \langle x_\pi, \widehat{\theta}_t\rangle + \alpha_r \beta_t(\delta, d)\|x_\pi\|_{\Sigma_t^{-1}}. \end{aligned}$$

The rest of the argument remains the same, substituting x by x_π.

■

Proof of Proposition 26: Recall that

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

Dimension $d = 3$.

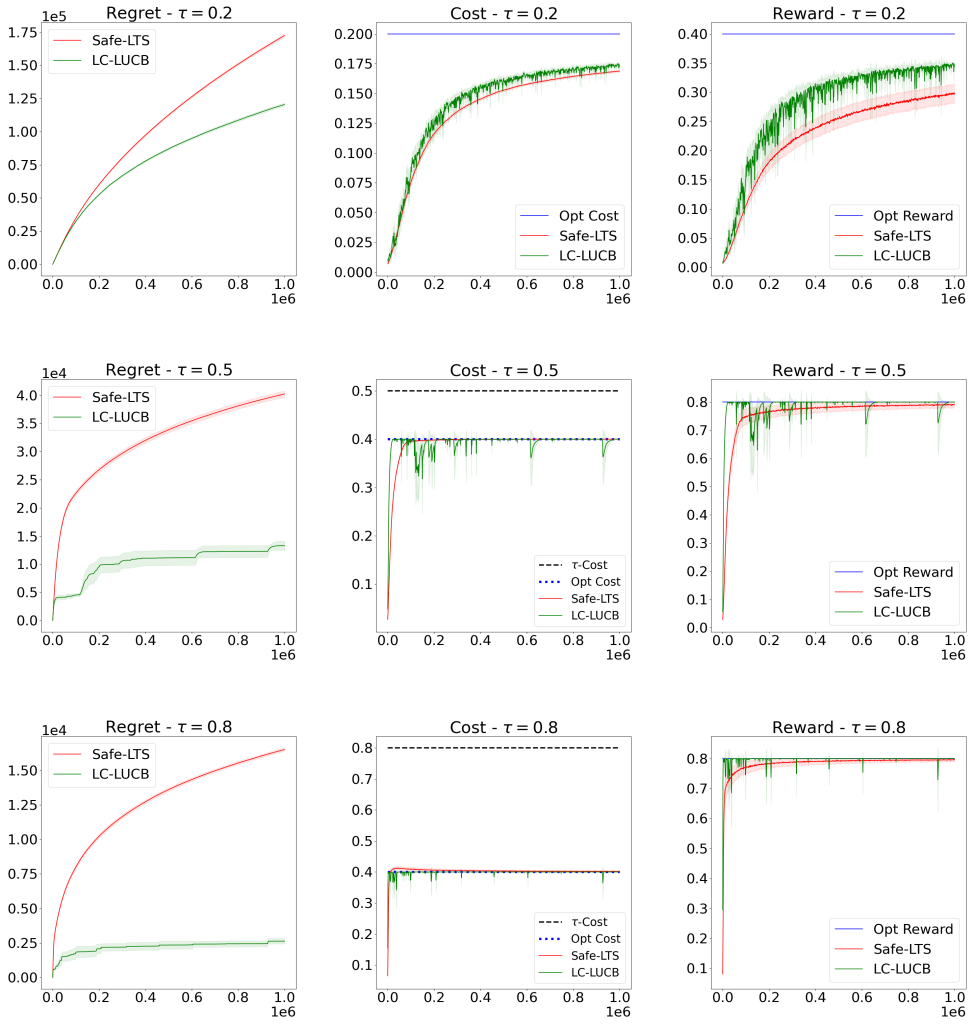

Figure 5: **LC-LUCB:** Dimension $d = 3$. **Top:** Constraint Threshold $\tau = 0.2$. **Center:** Constraint Threshold $\tau = 0.5$. **Bottom:** Constraint Threshold $\tau = 0.8$. The shaded regions around the curves correspond to one standard deviation.

Conditioned on the event $\mathcal{E}$ defined in Eq. 12, it follows that

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

Thus, we have

$$0 \leq \langle x_\pi^{o,\perp}, \widehat{\mu}_t^{o,\perp} - \mu_*^{o,\perp}\rangle + \beta_t(\delta, d-1) \|x_\pi\|_{(\Sigma_t^{o,\perp})^{-1}}. \tag{42}$$

Dimension $d = 5$.

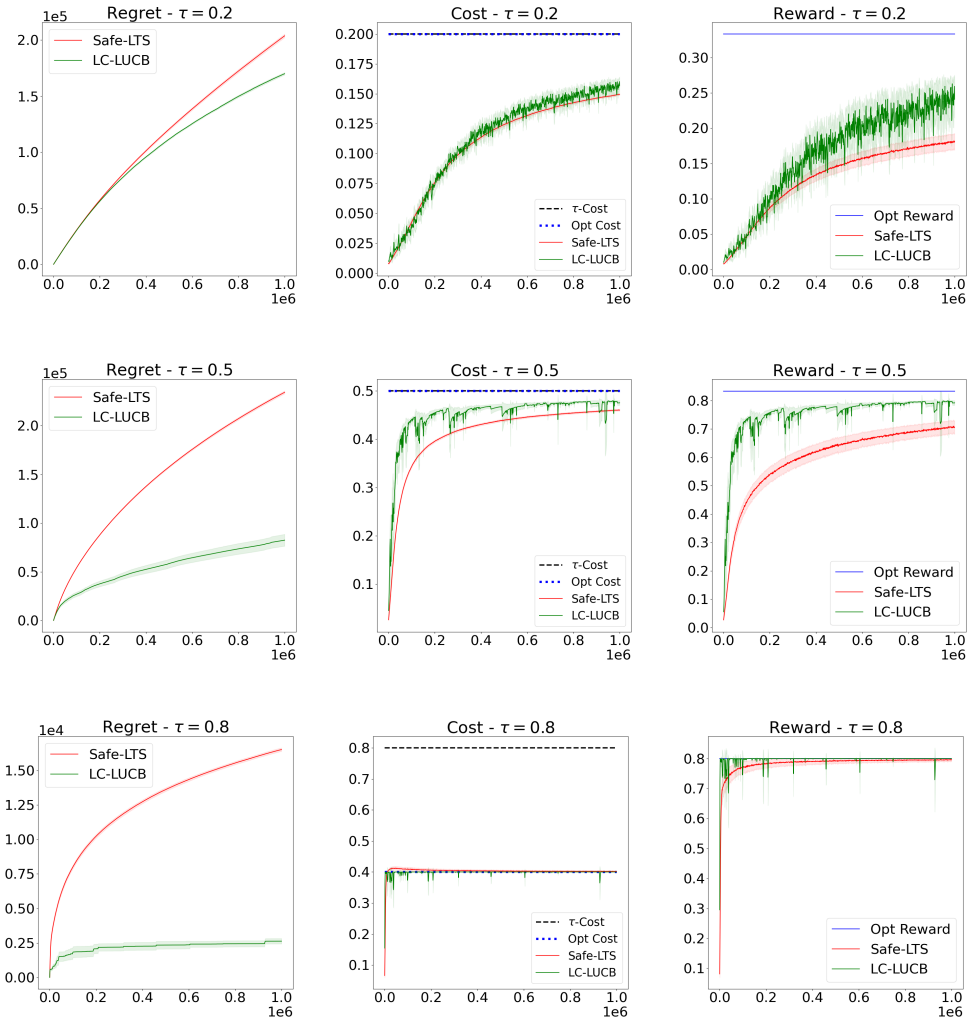

Figure 6: **LC-LUCB:** Dimension $d = 5$. **Top:** Constraint Threshold $\tau = 0.2$. **Center:** Constraint Threshold $\tau = 0.5$. **Bottom:** Constraint Threshold $\tau = 0.8$. The shaded regions around the curves correspond to one standard deviation.

Note that

$$c_\pi = \frac{\langle x_\pi^o, e_0\rangle c_0}{\|x_0\|} + \langle x_\pi^{o,\perp}, \mu_*^{o,\perp}\rangle \tag{43}$$
$$\leq \underbrace{\frac{\langle x_\pi^o, e_0\rangle c_0}{\|x_0\|} + \langle x_\pi^{o,\perp}, \widehat{\mu}_t^{o,\perp}\rangle + \alpha_c \beta_t(\delta, d-1)\|x_\pi^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}}}_{\text{(V)}}.$$

The above inequality holds by adding the inequality in Eq. 42 to Eq. 43. Since by assumption we have (V) $\leq \tau$ for all $\pi \in \Pi_t$, we obtain that $c_\pi \leq \tau$ which concludes the proof. ■

Dimension $d = 10$.

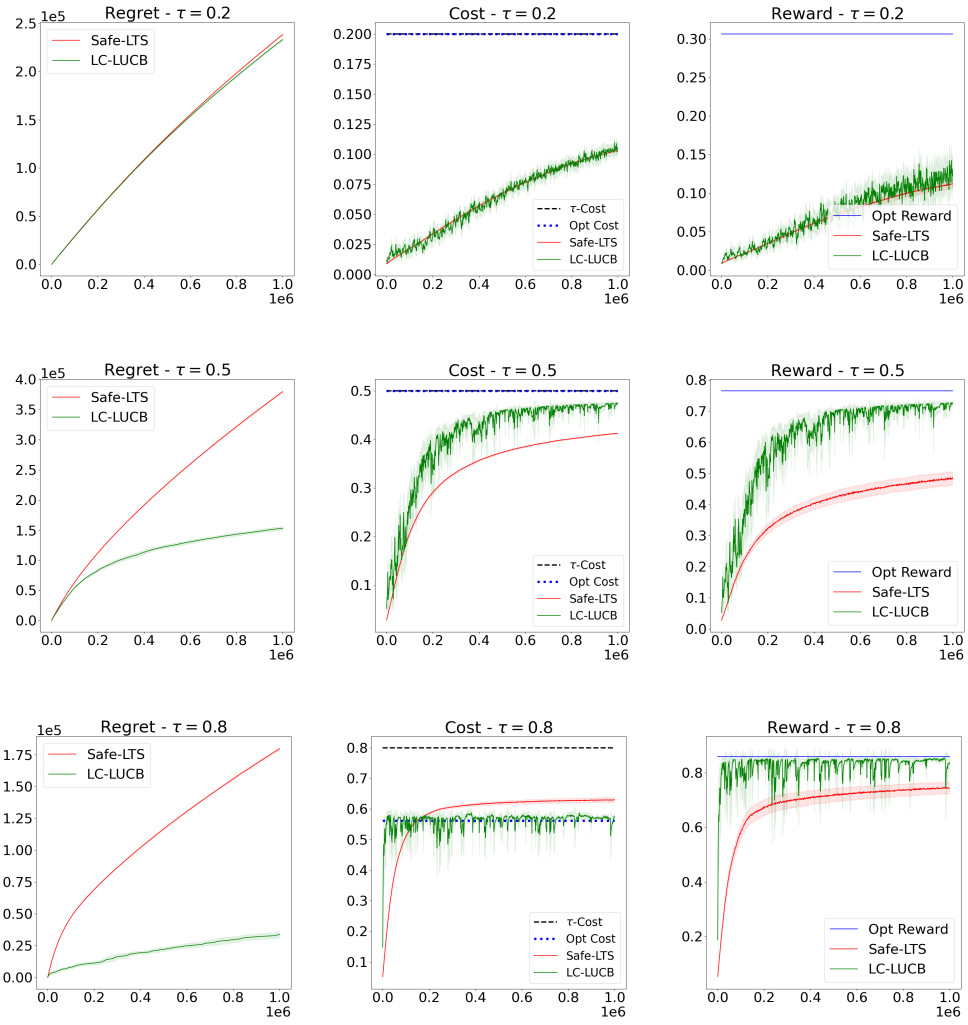

Figure 7: **LC-LUCB:** Dimension $d = 10$. **Top:** Constraint Threshold $\tau = 0.2$. **Center:** Constraint Threshold $\tau = 0.5$. **Bottom:** Constraint Threshold $\tau = 0.8$. The shaded regions around the curves correspond to one standard deviation.

Appendix D. Constrained Multi-Armed Bandits

D.1 The LP Structure

The main purpose of this section is to prove the optimal solutions of the linear program from (37) 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 (37), i.e.,

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

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:

$$\pi_{a_1}^* = \frac{u_{a_2}^c(t) - \tau}{u_{a_2}^c(t) - u_{a_1}^c(t)}, \qquad \pi_{a_2}^* = \frac{\tau - u_{a_1}^c(t)}{u_{a_2}^c(t) - u_{a_1}^c(t)}. \tag{44}$$

The main result in this section is the following Lemma:

Lemma 40 (support of π^*) *If* (37) *is feasible, there exists an optimal solution with at most* 2 *non-zero entries.*

Proof We start by inspecting the dual problem of (37):

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

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,

$$\lambda^* = \frac{u_{a_2}^r(t) - u_{a_1}^r(t)}{u_{a_2}^c(t) - u_{a_1}^c(t)}. \tag{45}$$

If $\lambda^* = 0$, the optimal primal value is achieved by concentrating all mass on any of the arms in $\mathcal{A}^*$. Otherwise, plugging (45) back into the objective of (D) and rearranging the terms, we obtain

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

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 (37) 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. ■

From the proof of Lemma 30 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 40 is that it is possible to find the optimal solution π^* to problem 37 by simply enumerating all pairs of arms (a_i, a_j) and all singletons, compute their optimal policies (if feasible) using Equation 44 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 (37).

If we had instead m constraints, a similar statement to Lemma 30 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.

D.2 Regret Analysis

In order to show a regret bound for Algorithm 3, we start with the following regret decomposition:

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

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 32, the equivalent version of Lemma 29 for the multi armed bandit problem.

D.3 Proof of Lemma 32

Lemma 41 *If we set α_r and α_c such that $\alpha_r, \alpha_c \geq 1$ and $(1 + \alpha_c)(1 - \bar{r}_1) \leq (\tau - \bar{c}_1)(\alpha_r - 1)$, then with high probability, for any $t \in [T]$, we have $\mathbb{E}_{a\sim\pi_t}[u_a^r(t)] \geq \mathbb{E}_{a\sim\pi^*}[\bar{r}_a]$.*

Proof Throughout this proof we denote as π_0 to the delta function over the safe arm 1. We'll use the notation $u_a^r(t) = \bar{r}_a + \xi_a^r(t)$ and $u_a^c(t) = \bar{c}_a + \xi_a^c(t)$. We start by noting that under $\mathcal{E}$, and because $\alpha_r, \alpha_c \geq 1$, then:

$$(\alpha_r - 1)\beta_a(t) \leq \xi_a^r(t) \leq (\alpha_r + 1)\beta_a(t)\ \forall a \quad \text{and} \quad (\alpha_c - 1)\beta_a(t) \leq \xi_a^c(t) \leq (\alpha_c + 1)\beta_a(t)\ \forall a \neq 0. \tag{46}$$

If $\pi^* \in \tilde{\Pi}_t$, it immediately follows that:

$$\mathbb{E}_{a\sim\pi^*}[\bar{r}_a] \leq \mathbb{E}_{a\sim\pi^*}[u_a^r(t)] \leq \mathbb{E}_{a\sim\pi_t}[u_a^r(t)]. \tag{47}$$

Let's now assume $\pi^* \notin \tilde{\Pi}_t$, i.e., $\mathbb{E}_{a\sim\pi^*}[u_a^c(t)] > \tau$. Let $\pi^* = \rho^*\bar{\pi}^* + (1 - \rho^*)\pi_0$ with $\bar{\pi}^* \in \Delta_K[2 : K]$[6].

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

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 $\widetilde{\pi}_t \in \widetilde{\Pi}_t$. It can be easily established that

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

(i) is a consequence of (46) and of the observation that since π^* is feasible $\rho^* \mathbb{E}_{a\sim\bar{\pi}^*}[\bar{c}_a] + (1-\rho^*)\bar{c}_1 \leq \tau$. Let $\Delta = \mathbb{E}_{a\sim\pi^*}[\bar{r}_a] - \mathbb{E}_{a\sim\pi_0}[\bar{r}_a]$. Since $\widetilde{\pi}_t \in \Pi_t$, we have

$$\begin{aligned}
\mathbb{E}_{a\sim\pi_t}[u_a^r(t)] &\geq \mathbb{E}_{a\sim\tilde{\pi}_t}[u_a^r(t)] = \mathbb{E}_{a\sim\tilde{\pi}_t}[\bar{r}_a] + \mathbb{E}_{a\sim\tilde{\pi}_t}[\xi_a(t)] \\
&\overset{(a)}{\geq} \mathbb{E}_{a\sim\tilde{\pi}_t}[\bar{r}_a] + (\alpha_r - 1)\mathbb{E}_{a\sim\tilde{\pi}_t}[\beta_a(t)] \\
&\overset{(b)}{\geq} \gamma_t \mathbb{E}_{a\sim\pi^*}[\bar{r}_a] + (1-\gamma_t)\mathbb{E}_{a\sim\pi_0}[\bar{r}_a] + (\alpha_r - 1)\gamma_t \rho^* \mathbb{E}_{a\sim\bar{\pi}^*}[\beta_a(t)] \\
&= \mathbb{E}_{a\sim\pi^*}[\bar{r}_a] + (1-\gamma_t)\mathbb{E}_{a\sim\pi_0}[\bar{r}_a] - (1-\gamma_t)\mathbb{E}_{a\sim\pi^*}[\bar{r}_a] + (\alpha_r - 1)\gamma_t \rho^* \mathbb{E}_{a\sim\bar{\pi}^*}[\beta_a(t)] \\
&= \mathbb{E}_{a\sim\pi^*}[\bar{r}_a] - (1-\gamma_t)\Delta + (\alpha_r - 1)\gamma_t \rho^* \mathbb{E}_{a\sim\bar{\pi}^*}[\beta_a(t)] \\
&= \mathbb{E}_{a\sim\pi^*}[\bar{r}_a] + \underbrace{\gamma_t\left(\Delta + (\alpha_r - 1)\rho^* \mathbb{E}_{a\sim\bar{\pi}^*}[\beta_a(t)]\right) - \Delta}_{\text{A}}.
\end{aligned}$$

Where (a) holds by Equation 46, (b) holds by definition of $\tilde{\pi}_t$ and because $\mathbb{E}_{a\sim\pi^*}[\beta_a(t)] \geq \mathbb{E}_{a\sim\bar{\pi}^*}[\beta_a(t)]$. Let $C_0 = \rho^* \mathbb{E}_{a\sim\bar{\pi}^*}[\beta_a(t)]$. Let's show conditions under which $\mathbb{I} \geq$. The following chain of inequalities holds,

$$\text{A} \overset{(a)}{\geq} \frac{\tau - \bar{c}_1}{\tau - \bar{c}_1 + (1+\alpha_c)C_0}\left(\Delta + (\alpha_r - 1)C_0 - \Delta\right) - \Delta.$$

Where (a) follows by substituting $\gamma \geq \frac{\tau - \bar{c}_1}{\tau - \bar{c}_1 + (1+\alpha_c)C_0}$. Following the same logic as in the analysis of Equation 21 in the proof of Lemma 29 we conclude that I is non-negative whenever,

$$(\tau - \bar{c}_1)(\alpha_r - 1) \geq (1+\alpha_c)\Delta.$$

Since $\Delta \leq 1 - \bar{r}_1$ this concludes the proof. ■

Proposition 42 *If $\delta = \frac{\delta'}{4KT}$ for $\delta' \in (0,1)$, $\alpha_r, \alpha_c \geq 1$ with $(1+\alpha_c)(1-\bar{r}_1) \leq (\tau - \bar{c}_1)(\alpha_r - 1)$, then with probability at least $1 - \frac{\delta'}{2}$, 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 $\mathbb{P}(\mathcal{E}) \geq 1 - \frac{\delta'}{2}$. Combining this observation with Lemma 32 yields the result. ■

Term (II) can be bounded using the confidence intervals radii:

Proposition 43 *If $\delta = \frac{\delta'}{4KT}$ for an $\delta' \in (0,1)$, then with probability at least $1 - \frac{\delta'}{2}$, we have*

$$\sum_{t=1}^{T} \mathbb{E}_{a\sim\pi_t}[u_a^r(t)] - \mathbb{E}_{a\sim\pi_t}[\bar{r}_a] \leq (\alpha_r + 1)\left(2\sqrt{2TK\log(1/\delta)} + 4\sqrt{T\log(2/\delta')\log(1/\delta)}\right).$$

Proof Under these conditions $\mathbb{P}(\mathcal{E}) \geq 1 - \frac{\varepsilon}{2}$. Recall $u_a^r(t) = \widehat{r}_a(t) + \alpha_r\beta_a(t)$ and that conditional on $\mathcal{E}$, $\bar{r}_a \in [\widehat{r}_a(t) - \beta_a(t), \widehat{r}_a(t) + \beta_a(t)]$ 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[7] implies:

$$\mathbb{P}\left(\underbrace{\sum_{t=1}^{T}\mathbb{E}_{a\sim\pi_t}[\beta_a(t)] \geq \sum_{t=1}^{T}\beta_{a_t}(t) + 4\sqrt{T\log(2/\delta')\log(1/\delta)}}_{\mathcal{E}_A^c}\right) \leq \varepsilon/2.$$

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 $\mathbb{P}(\mathcal{E} \cap \mathcal{E}_A) \geq 1 - \delta'$ yields the result. ■

We can combine these two results into our main theorem:

Theorem 31 *Let $\delta = 4KT\delta'$, $\alpha_c = 1$, and $\alpha_r = 1 + \frac{2(1-\bar{r}_1)}{\tau - \bar{c}_1}$. Then, with probability at least $1 - \delta$, the regret of OPB satisfies*

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

Proof This result is a direct consequence of Propositions 42 and 43 by setting $\delta = 4KT\delta'$. ■

7. 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 $\mathbb{P}(X_n > r) \leq \exp\left(-\frac{r^2}{2\sum_{i=1}^{n} d_i^2}\right)$.

D.4 Multiple Constraints

We consider the problem where the learner must satisfy M constraints with threshold values $\tau_1, \dots, \tau_M$. 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 $i = 1, \dots, M$. The full information optimal policy can be obtained by solving the following linear program:

$$\max_{\pi\in\Delta_K} \sum_{a\in\mathcal{A}} \pi_a \bar{r}_a, \qquad \text{s.t.} \ \sum_{a\in\mathcal{A}} \pi_a \bar{c}_a^{(i)} \leq \tau_i, \quad \text{for } \ i = 1, \dots, M. \tag{P-M}$$

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

Assumption 44 *The learner has knowledge of $\bar{c}_1^{(i)} < \tau_i$ for all $i = 1, \dots, M$.*

We denote by $\{\widehat{r}_a\}_{a\in\mathcal{A}}$ and $\{\widehat{c}_a^{(i)}\}_{a\in\mathcal{A}}$, for $i = 1, \dots, M$ 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 $i = 1, \dots, M$ 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 3 considers instead the following $M-$constraints LP:

$$\max_{\pi\in\Delta_K} \sum_{a\in\mathcal{A}} \pi_a \, u_a^r(t), \qquad \text{s.t.} \quad \sum_{a\in\mathcal{A}} \pi_a \, u_a^c(t,i) \leq \tau_i, \quad \text{for } \ i = 1, \dots, M. \tag{$\widehat{P-M}$}$$

We now generalize Lemma 32:

Lemma 45 *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 32 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 $\min \tau_i - \bar{c}_1^{(i)}$. ■

The following result follows:

Theorem 46 (Multiple Constraints Main Theorem) *If $\varepsilon \in (0,1)$, $\alpha_c = 1$ and $\alpha_r = \frac{2}{\min_i \tau_i - \bar{c}_1^{(i)}} + 1$, then with probability at least $1-\varepsilon$, Algorithm 3 satisfies the following regret guarantee:*

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

Proof The proof follows the exact same argument we used for the proof of Theorem 31 substituting $\tau - \bar{c}_1$ by $\min_i \tau_i - \bar{c}_1^{(i)}$. ■

Appendix E. Lower Bounds

In this section we prove the two lower bounds from the main text. We will do so by exhibiting a lower bound for the

We start by stating a generalized version of the divergence decomposition lemma for bandits. The proof is a direct application of Lemma 15.1 in Lattimore and Szepesvári (2019) to this case.

Lemma 47 (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 $\mathbb{P}_{\nu'} = \mathbb{P}_{\nu' A}$ 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,*

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

where $T_a(T)$ denotes the number of times that arm "a" has been pulled by $\mathcal{A}$ up to time T.

The following two lemmas will also be useful in our lower-bound proof, so we state them here.

Lemma 48 (Gaussian Divergence) *The divergence between two multi-variate normal distributions with means $\mu_1, \mu_2 \in \mathbb{R}^d$ and spherical identity covariance $\mathbb{I}_d$ is equal to*

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

Lemma 49 *The binary relative entropy to be*

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

and satisfies

$$d(x, y) \geq (1/2) \log(1/4y), \tag{48}$$

for $x \in [1/2, 1]$ and $y \in (0, 1)$.

Lemma 50 (Adapted from Kaufmann et al. 2016, Lemma 1.) *Let ν, ν' be two constrained bandit models with K arms. Borrow the setup, definitions and notations of Lemma 47, then for any measurable event $\mathcal{B} \in \mathcal{F}_T$:*

$$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{a=1}^{K} \mathbb{E}_\nu\big[T_a(T)\big]\mathrm{KL}\big((P_a, Q_a), (P'_a, Q'_a)\big) \geq d\big(\mathbb{P}_\nu(\mathcal{B}), \mathbb{P}_{\nu'}(\mathcal{B})\big). \tag{49}$$

We start by showing that under an appropriate noise assumption, it is possible to reduce the constrained (in expectation) Multi Armed Bandit (CE-MAB) problem studied in Pacchiano et al. (2021) to our setting. The argument behind the proof of the main result in this section, the lower bound Theorem 21 relies on the problem structure behind the LC-LUCB version of the CE-MAB problem given by this reduction.

Setup: Let's first describe the CE-MAB setup. In the constrained K-armed bandit setting, the action sets satisfy $\mathcal{A}_t = \Delta_K$, where Δ_K is the K-dimensional simplex. The reward and cost parameters are reduced to the K-dimensional vectors containing the mean reward and cost values of the K arms, i.e., $\theta_* = (\bar{r}_0, \dots, \bar{r}_{K-1})$ and $\mu_* = (\bar{c}_0, \dots, \bar{c}_{K-1})$.

In this case $X_t \in \Delta_K$ and we assume that abusing notation $a_t \sim X_t$, an index in $[K]$ is sampled from the distribution X_t, after which the reward value and the cost satisfy:

$$R_t = \bar{r}_{a_t} + \nu_t^r \ , \qquad\qquad C_t = \bar{c}_{a_t} + \nu_t^c \ ,$$

where ν_t^r and ν_t^c are conditionally zero mean sub-Gaussian random variables. The learner's objective is to play policies X_t such that for all t, $\langle X_t, \mu_* \rangle \leq \tau$ while at the same time maximizing $\langle X_t, \theta_* \rangle$. We work under the assumption that $\bar{c}_0$ is known to the learner and satisfies $\bar{c}_0 \leq \tau$.

Reduction: We now show that it is possible to reduce the CE-MAB problem to the LC-LUCB setup. Using the notation in Assumption 1 we define $\xi_t^r = R_t - \sum_{a \in [K]} X_t(a) \bar{r}_a$ and $\xi_t^r = C_t - \sum_{a \in [K]} X_t(a) \bar{c}_a$. Where $X_t(a)$ corresponds to the $a-$th coordinate of X_t. Notice that:

$$R_T = \langle X_t, \theta_* \rangle + \xi_t^r \ , \qquad\qquad C_T = \langle X_t, \mu_* \rangle + \xi_t^c \ ,$$

with ξ_t^r and ξ_t^c both conditionally zero mean subgaussian random variables:

Indeed since $\{\bar{r}_a, \bar{c}_a\}_{a \in [K]}$ are all assumed to be bounded, the conditional subgaussianity assumption of ξ_t^r and ξ_t^c is satisfied for an appropriate choice of subgaussianity parameter R, dependent on the subgaussianity parmeters of ν_t^r, ν_t^c and the boundedness of $\{\bar{r}_a, \bar{c}_a\}_{a \in [K]}$. This finalizes the reduction.

We now proceed to prove Theorem 21 the main result of this section:

Theorem 21 *Let $\tau, c_0, r_0 \in (0,1)$, $B = \max\left(\frac{d\sqrt{T}}{8e^2}, \frac{1-r_0}{21(\tau - c_0)^2}\right)$, and assume $T \geq \max(d-1, \frac{168eB}{1-r_0})$. Then, for any algorithm $\mathfrak{A}$, there is a pair of reward and cost parameters (θ_*, μ_*), such that $\mathcal{R}_{\mathcal{C}}(T) \geq B$.*

Proof If $\max\left(d\sqrt{T}, \frac{1-r_0}{21(\tau-c_0)^2}\right) = d\sqrt{T}$, then the argument in Theorem 24.1 of Lattimore and Szepesvári (2019) yields the desired result by noting that the framework of constrained bandits subsumes linear bandits. In this case we conclude there is a constrained linear bandit instance with $\theta_* \in \{-\frac{1}{\sqrt{T}}, \frac{1}{\sqrt{T}}\}^d$ and $\mathcal{A}_t = [-1,1]^d$ satisfying:

$$\mathcal{R}_{\mathcal{C}}(T) \geq \frac{d\sqrt{T}}{8e^2}.$$

Let's instead focus on the case where $B = \max\left(\frac{d\sqrt{T}}{8e^2}, \frac{1-r_0}{21(\tau-c_0)^2}\right) = \frac{1-r_0}{21(\tau-c_0)^2}$.

Pick any algorithm $\mathfrak{A}$. We want to show that the algorithm's regret on some environment is as large as B. For the remainder of the proof we restrict ourselves to instances where $\mathcal{A}_t = \Delta_d$ and $\theta_* = \bar{r}$, $\mu_* = \bar{c}$ with $\bar{r}, \bar{c} \in [0,1]^d$ parametrize a constrained Multi Armed Bandit problem such that arm 0, the (known) safe arm satisfies $\bar{r}_0 = r_0$ and $\bar{c}_0 = c_0$.

If there was any such instance $\bar{r}, \bar{c}$ such that $\mathcal{R}_{\mathcal{C}}(T) > B$ there would be nothing to prove. Hence without loss of generality, and by the way of contradiction we assume algorithm $\mathfrak{A}$ satisfies $\mathcal{R}_{\mathcal{C}}(T) \leq B$ for all $\bar{r}, \bar{c} \in [0,1]^d$ with $\bar{r}_0 = r_0$ and $\bar{c}_0 = c_0$ and where all arms have unit variance Gaussian rewards and costs.

In what follows we denote $\mathcal{R}_{\mathcal{C}}(T, \bar{r}, \bar{c})$ as the regret incurred by algorithm $\mathfrak{A}$ on instance $\theta_* = \bar{r}, \mu_* = \bar{c}$.

For simplicity we will introduce a couple of shorthand notational choices. Let $c = \tau - c_0$, $\Delta = \frac{1-r_0}{7}$ and $D = \frac{8r_0-1}{7}$ when $r_0 \geq \frac{1}{8}$. This will make it easier to explain the logic of the proof argument. Let's consider the following constrained bandit instance inducing measure ν:

$$\begin{aligned}
\bar{c}^1 &= (\tau - c, \quad \tau + 2c, \quad \tau - c, \quad \tau + 2c, \quad \ldots, \quad \tau + 2c) \\
\bar{r}^1 &= (D + \Delta, \quad D + 8\Delta, \quad D, \quad D + 4\Delta, \quad \ldots, \quad D + 4\Delta\ ).
\end{aligned}$$

Notice that by definition $\bar{c}_0^1 = c_0$ and $\bar{r}_0^1 = r_0$. For the $(\bar{r}^1, \bar{c}^1)$ problem instance, the optimal policy is a mixture between arms 0 and 1, where arm 0 is chosen with probability $2/3$ and arm 1 with probability $1/3$. The value of this optimal policy equals $D + \frac{10}{3}\Delta$.

Let $\bar{T}_j(T) \in [0, T]$ be the total amount of probability mass that $\mathfrak{A}$ allocated to arm j up to time T.

Let's lower bound the regret in the event that $\bar{T}_0(T) < \frac{T}{2}$. By the feasibility assumption it follows the average visitation policy π_T defined as $\pi_T(i) = \frac{\bar{T}_i(T)}{T}$ is feasible.

When the event $\{\bar{T}_0(T) < \frac{T}{2}\}$ holds policy π_T satisfies $\pi_T(0) \leq \frac{1}{2}$. A simple computation shows that to maximize its cumulative reward while maintaining feasibility constrained to $\pi_T(0) \leq \frac{1}{2}$, π_T's optimal mass allocation is

$$\pi_T(i) = \begin{cases} \frac{1}{2} & \text{if } i = 0 \\ \frac{1}{3} & \text{if } i = 1 \\ \frac{1}{6} & \text{if } i = 2 \end{cases}$$

This policy has a reward of $D + \frac{19\Delta}{6}$ and therefore the regret of π_T is lower bounded by $\frac{\Delta}{6}$. Thus, the regret $\mathcal{R}_{\mathcal{C}}(T, \bar{r}^1, \bar{c}^1)$ can be lower bounded bounded as

$$\mathcal{R}_{\mathcal{C}}(T, \bar{r}^1, \bar{c}^1) \geq \frac{\Delta T}{6} \mathbb{P}_\nu \left(\bar{T}_0(T) < \frac{T}{2} \right)$$

Since by assumption, $\mathfrak{A}$ satisfies $\mathcal{R}_{\mathcal{C}}(T, \bar{r}^1, \bar{c}^1) \leq B$:

$$B \geq \mathcal{R}_{\mathcal{C}}(T, \bar{r}^1, \bar{c}^1) \geq \frac{\Delta T}{6} \mathbb{P}_\nu \left(\bar{T}_0(T) < \frac{T}{2} \right)$$

And therefore:

$$\mathbb{P}_\nu \left(\bar{T}_0(T) \geq \frac{T}{2} \right) = 1 - \mathbb{P}_\nu \left(\bar{T}_0(T) < \frac{T}{2} \right) \geq 1 - \frac{6B}{\Delta T} \geq 1/2 \tag{50}$$

The last inequality follows from the assumption $T \geq \max(d - 1, \frac{168eB}{1-r_0})$ (recall that $\Delta = \frac{1-r_0}{7}$.

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

$$\begin{aligned}
\bar{c}^2 &= (\tau - c, \quad \tau + 2c, \quad \tau - c, \quad \tau - c, \quad \ldots, \quad \tau + 2c) \\
\bar{r}^2 &= (D + \Delta, \quad D + 8\Delta, \quad D, \quad D + 4\Delta, \quad \ldots, \quad D + 4\Delta\ ).
\end{aligned}$$

In this instance the optimal policy is to play arm 3 deterministically. This policy achieves a reward of $D + 4\Delta$.

We will now lower bound the regret $\mathcal{R}_{\mathcal{C}}(T, \bar{r}^2, \bar{c}^2)$ under measure ν'. We'll do so by lower bounding the regret in the event that $\{\bar{T}_0 \geq \frac{T}{2}\}$ holds.

Similar to the argument we expanded on above, when $\{\bar{T}_0 \geq \frac{T}{2}\}$ feasibility guarantees the average policy $\pi'_T(i) = \frac{\bar{T}_i(T)}{T}$ is feasible. When $\{\bar{T}_0 \geq \frac{T}{2}\}$ holds policy π'_T satisfies $\pi'_T(0) \geq \frac{1}{2}$. Simple computations show that to maximize the cumulative reward of policy π'_T while maintaining feasibility constrained to $\pi'_T(0) \geq \frac{1}{2}$ the optimal allocation satisfies,

$$\pi'_T(i) = \begin{cases} \frac{1}{2} & \text{if } i = 0 \\ \frac{1}{2} & \text{if } i = 3 \end{cases}$$

This policy achieves an expected reward of $D + \frac{5\Delta}{2}$ and therefore the regret of π'_T is lower bounded by $D + 4\Delta - D - \frac{5}{2}\Delta = \frac{3}{2}\Delta$. Therefore,

$$\mathcal{R}_{\mathcal{C}}(T, \bar{r}^2, \bar{c}^2) \geq \frac{3}{2}\Delta \mathbb{P}_{\nu'}\left(\bar{T}_0(T) \geq \frac{T}{2}\right)$$

Since by assumption $\mathfrak{A}$ satisfies $\mathcal{R}_{\mathcal{C}}(T, \bar{r}^2, \bar{c}^2) \leq B$, we have

$$\mathbb{P}_{\nu'}\left(\bar{T}_0(T) \geq \frac{T}{2}\right) \leq \frac{2B}{3\Delta T} \leq \frac{1}{4e}.$$

The last inequality follows from the assumption that $T \geq \max(d-1, \frac{168eB}{1-r_0})$. We now apply the results of Lemma 49 to this upper bound and the lower bound from Equation 50 and obtain,

$$d\left(\mathbb{P}_{\nu}\left(\bar{T}_0(T) \geq \frac{T}{2}\right), \mathbb{P}_{\nu'}\left(\bar{T}_0(T) \geq \frac{T}{2}\right)\right) \geq 1/2.$$

As a consequence of (48), Lemma 48 and 50, we have

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{\nu}, \mathbb{P}_{\nu'}) &= \mathbb{E}_{\nu}\big[T_3(T)\big] \times \mathrm{KL}\left(\mathcal{N}\left(\begin{pmatrix} \tau + 2c \\ 4\Delta \end{pmatrix}, \mathbb{I}_d\right), \mathcal{N}\left(\begin{pmatrix} \tau - c \\ 4\Delta \end{pmatrix}, \mathbb{I}_d\right)\right) \\
&= 2c^2 \mathbb{E}_{\nu}\big[T_3(T)\big] \\
&\geq d\left(\mathbb{P}_{\nu}\left(\bar{T}_0(T) \geq \frac{T}{2}\right), \mathbb{P}_{\nu'}\left(\bar{T}_0(T) \geq \frac{T}{2}\right)\right) \\
&\geq \frac{1}{2}.
\end{aligned}$$

Therefore, we can conclude

$$\mathbb{E}_{\nu}[\bar{T}_3(T)] = \mathbb{E}_{\nu}[T_3(T)] \geq \frac{1}{4c^2}. \tag{51}$$

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

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

Since $\Delta = \frac{1-r_0}{7}$ and $D = \frac{8r_0-1}{7}$ and noting that in this case $\frac{\Delta}{3c^2} = B$. The result follows. ■

Theorem 34 *Let* $\tau, \bar{c}_1, \bar{r}_1 \in (0,1)$, $B = \max\left(\frac{1}{27}\sqrt{(K-1)T}, \frac{1-\bar{r}_1}{21(\tau-\bar{c}_1)^2}\right)$, *and assume* $T \geq \max(K-1, \frac{168eB}{1-\bar{r}_1})$. *Then, for any algorithm there is a pair of reward and cost parameters* (θ_*, μ_*), *such that* $\mathcal{R}_{\mathcal{C}}(T) \geq B$.

Proof

If $\max\left(\frac{1}{27}\sqrt{(K-1)T}, \frac{1-r_0}{21(\tau-c_0)^2}\right) = \frac{1}{27}\sqrt{(K-1)T}$, 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 multi armed bandits. In this case we conclude there is a constrained multi armed bandit instance satisfying:

$$\mathcal{R}_{\mathcal{C}}(T) \geq \frac{1}{27}\sqrt{(K-1)T}.$$

When $B = \max\left(\frac{d\sqrt{T}}{8e^2}, \frac{1-r_0}{21(\tau-c_0)^2}\right) = \frac{1-r_0}{21(\tau-c_0)^2}$, the same argument as in the proof of Theorem 21 finalizes the result. ■

Appendix F. Extensions

F.1 Unknown c_0 and r_0

In this section, we relax Assumption 5, and instead assume that we only have the knowledge of a safe action x_0, and no knowledge about its cost c_0 and reward r_0. The same discussion applies to $\bar{c}_1$ and $\bar{r}_1$ in OPB. The objective will be to design an algorithm capable of estimating c_0 and r_0 up to an accuracy of $\tau - c_0$ and $1 - r_0$ for c_0 and r_0 respectively. We summarize the algorithm in the box below,

Algorithm 5 Unknown c_0, r_0 estimation.

Input: Safe arm x_0.
for $t = 1, \ldots, T$ **do**
- 1. Pull arm x_0.
- 2. Compute average cost estimator $\widehat{c}_0(t)$.
- 3. Compute average reward estimator $\widehat{r}_0(t)$.
- 4. **if** $\widehat{c}_0(t) + 3\sqrt{2\log(1/\delta)/t} \leq \tau$*:* **then**
  - Stop estimating c_0.
- **end**
- 5. **if** $\widehat{r}_0(t) + 3\sqrt{2\log(1/\delta)/t} \leq 1$*:* **then**
  - Stop estimating r_0.
- **end**
- 6. **if** $\widehat{c}_0(t) + 3\sqrt{2\log(1/\delta)/t} \leq \tau$ *and* $\widehat{r}_0(t) + 3\sqrt{2\log(1/\delta)/t} \leq 1$*:* **then**
  - Return $\widehat{c}_0(T_0^c)$ and $\widehat{r}_0(T_0^r)$.
- **end**

end

For all t rounds to produce empirical mean estimators $\widehat{c}_0$ and $\widehat{r}_0$. Note that for all $\delta \in (0,1)$ and all $t \leq T$, $\widehat{c}_0(t)$ and $\widehat{r}_0(t)$ satisfy,

$$\mathbb{P}\big(|\widehat{c}_0(t) - c_0| \leq \sqrt{2\log(2T/\delta)/t}\big) \geq 1 - \delta/2T \tag{52}$$

$$\mathbb{P}\big(|\widehat{r}_0(t) - r_0| \leq \sqrt{2\log(2T/\delta)/t}\big) \geq 1 - \delta/2T. \tag{53}$$

Let's define $\tilde{\mathcal{E}}$ as the event where Equations 52 and 53 hold for all $t \leq T$. The reasoning above implies $\mathbb{P}(\tilde{\mathcal{E}}) \geq 1 - \delta$.

Denote T_0^c, T_0^r as the times when conditions 4. and 5. of Algorithm 5 trigger. Let's analyze T_0^c. Since T_0^c is the first time when conditions 4. triggers thus,

$$c_0 + 2\sqrt{2\log(2T/\delta)/T_0^c} \overset{(i)}{\leq} \widehat{c}_0(T_0^c) + 3\sqrt{2\log(2T/\delta)/T_0^c} \leq \tau.$$

Where (i) holds because of equation 52. Thus

$$\sqrt{2\log(2T/\delta)/T_0^c} \leq \frac{\tau - c_0}{2}. \tag{54}$$

Since $\widetilde{\mathcal{E}}$ implies $\widehat{c}_0(T_0^c) \in \left[c_0 - \sqrt{2\log(2T/\delta)/t}, c_0 + \sqrt{2\log(2T/\delta)/t}\right]$ we have,

$$\tau - \widehat{c}_0(T_0^c) \in \left[\tau - c_0 - \sqrt{2\log(2T/\delta)/T_0^c}, \tau - c_0 + \sqrt{2\log(2T/\delta)/T_0^c}\right] \subseteq \left[\frac{\tau - c_0}{2}, \frac{3(\tau - c_0)}{2}\right]$$

Similarly we conclude that whenever $\widetilde{\mathcal{E}}$ holds,

$$1 - \widehat{r}_0(T_0^r) \in \left[1 - r_0 - \sqrt{2\log(2T/\delta)/T_0^r}, 1 - r_0 + \sqrt{2\log(2T/\delta)/T_0^r}\right] \subseteq \left[\frac{1 - r_0}{2}, \frac{3(1 - r_0)}{2}\right]$$

We define $\widehat{\Delta}_c = \tau - \widehat{c}_0(T_0^c)$ and $\widehat{\Delta}_r = 1 - \widehat{r}_0(T_0^r)$. The above discussion implies Δ_c and Δ_r are upper and lower bounded by constant multiples of $\tau - c_0$ and $1 - r_0$ respectively.

When $\widetilde{\mathcal{E}}$ holds, Equation 54 implies $T_0^c \geq \frac{8\log(2T/\delta)}{(\tau - c_0)^2}$ and $T_0^r \geq \frac{8\log(2T/\delta)}{(1 - r_0)^2}$. We now see that we can also upper bound these quantities, let's work through the argument for c_0. For all t such that $\sqrt{2\log(2T/\delta)/t} \leq \frac{\tau - c_0}{4}$, when $\widetilde{\mathcal{E}}$ holds,

$$\widehat{c}_0(t) + 3\sqrt{2\log(2T/\delta)/t} \leq c_0 + 4\sqrt{2\log(2T/\delta)/t} \leq \tau.$$

Thus, condition 4. of Algorithm 5 holds. Similarly for all t such that $\sqrt{2\log(2T/\delta)/t} \leq \frac{1 - r_0}{4}$, when $\widetilde{\mathcal{E}}$ holds,

$$\widehat{r}_0(t) + 3\sqrt{2\log(2T/\delta)/t} \leq r_0 + 4\sqrt{2\log(2T/\delta)/t} \leq 1.$$

Thus condition 5. of Algorithm 5 holds. This implies $T_0^c \leq \frac{32\log(2T/\delta)}{(\tau - c_0)^2}$ and $T_0^r \leq \frac{32\log(2T/\delta)}{(1 - r_0)^2}$. If we define as T_0 to the time-step when condition 6. of Algorithm 5 triggers, it follows that

$$T_0 \leq 32\log(2T/\delta)\max\left(\frac{1}{(\tau - c_0)^2}, \frac{1}{(1 - r_0)^2}\right).$$

We then set $\frac{\alpha_r}{\alpha_c} = \widehat{\Delta}_r / \widehat{\Delta}_c$ and run LC-LUCB for rounds $t > T_0$. Since the scaling of α_r w.r.t. α_c is optimal up to constants, the same regret bounds (plus the regret incurred up to T_0) would hold. We can upper bound the regret incurred during T_0,

$$32 \log(2T/\delta) \max\left(\frac{1-r_0}{(\tau - c_0)^2}, \frac{1}{1-r_0}\right)$$

Therefore, in case c_0 is unknown, the algorithm proceeds by warm-starting our estimates of θ_* and μ_* using the data collected by playing the safe arm 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 Algorithm 1 (LC-LUCB) and Algorithm 1 (OPLB) do for the reward parameter θ_*. For the multi-constrained setting the estimation procedure of Algorithm 5 can be used to estimate each of the cost signals simultaneously. An equivalent stopping condition yields a scheme to estimate the minimal cost gap up to constant accuracy. The same analysis as in the single constraint case holds.

Appendix G. Nonlinear Rewards

G.1 Properties of Least Squares Estimators

In this section we derive convergence properties of least squares estimators. These results will be crucial to analyze the NLC-LUCB algorithm in the following section. Let $\{X_t, Y_t\}_{t=1}^{\infty}$ be a martingale sequence such that $X_t \in \mathcal{X}$ and $Y_t \in \mathbb{R}$ with $Y_t = f_\star(X_t) + \xi_t$ where ξ_t satisfies Assumption 1. Throughout this section we will use the notation $\mathcal{F}_{t-1} = \sigma(X_1, Y_1 \cdots, X_{t-1}, Y_{t-1})$ to denote the sigma algebra generated by all previous outcomes.

Let $\mathcal{G}$ be a finite[8] class of functions such that $f_\star \in \mathcal{G}$ and for all $t \in \mathbb{N}$ consider the least squares regression estimator $\widehat{f}_t$ defined as,

$$\widehat{f}_t = \min_{f \in \mathcal{F}} \sum_{\ell=1}^{t} (f(X_\ell) - Y_\ell)^2$$

We assume that

Assumption 51 (Bounded responses) *There exists a $B > 0$ such that for all $X \in \mathcal{X}$, and all $f \in \mathcal{F}$,*

$$|f(X)| \leq B, \text{ and } |Y_i| \leq B.$$

Our results rely on the following Uniform Empirical Bernstein bound from Howard et al. (2021).

Lemma 52 (Uniform empirical Bernstein bound) *In the terminology of Howard et al. (2021), let $Z_t = \sum_{i=1}^{t} Y_i$ be a sub-ψ_P process with parameter $c > 0$ and variance process W_t. Then with probability at least $1 - \widetilde{\delta}$ for all $t \in \mathbb{N}$*

$$Z_t \leq 1.44\sqrt{\max(W_t, m)\left(1.4 \log\log\left(2\left(\max\left(\frac{W_t}{m}, 1\right)\right)\right) + \log\frac{5.2}{\widetilde{\delta}}\right)} + 0.41c\left(1.4 \log\log\left(2\left(\max\left(\frac{W_t}{m}, 1\right)\right)\right) + \log\frac{5.2}{\widetilde{\delta}}\right)$$

where $m > 0$ is arbitrary but fixed.

8. Our results can be easily extended to the case of infinite function classes with bounded metric entropy.

As a corollary of Lemma 52 we can show the following,

Lemma 53 (Freedman) *Suppose $\{X_t\}_{t=1}^{\infty}$ is a martingale difference sequence with $|X_t| \leq b$. Let $S_t = \sum_{\ell=1}^{t} X_\ell^2$ For any $\widetilde{\delta} \in (0,1)$, with probability at least $1-\widetilde{\delta}$,*

$$\sum_{\ell=1}^{t} X_\ell \leq 4\sqrt{S_t \log \frac{12\log 2t}{\widetilde{\delta}}} + 6b \log \frac{12 \log 2t}{\widetilde{\delta}}.$$

for all $t \in \mathbb{N}$ simultaneously.

Proof We are ready to use Lemma 52 (with $c = b$). Let $S_t = \sum_{\ell=1}^{t} X_t$ and $W_t = \sum_{\ell=1}^{t} \mathrm{Var}_\ell(X_\ell)$. Let's set $m = b^2$. It follows that with probability $1-\widetilde{\delta}$ for all $t \in \mathbb{N}$

$$\begin{aligned}
S_t &\leq 1.44\sqrt{\max(W_t, b^2)\left(1.4\log\log\left(2\left(\max\left(\frac{W_t}{b^2},1\right)\right)\right) + \log\frac{5.2}{\widetilde{\delta}}\right)} \\
&\quad + 0.41b\left(1.4\log\log\left(2\left(\max\left(\frac{W_t}{b},1\right)\right)\right) + \log\frac{5.2}{\widetilde{\delta}}\right) \\
&\leq 2\sqrt{\max(W_t, b^2)\left(2\log\log\left(2\left(\max\left(\frac{W_t}{b^2},1\right)\right)\right) + \log\frac{6}{\widetilde{\delta}}\right)} \\
&\quad + b\left(2\log\log\left(2\left(\max\left(\frac{W_t}{b^2},1\right)\right)\right) + \log\frac{6}{\widetilde{\delta}}\right) \\
&= 2\max(\sqrt{W_t}, b)A_t + bA_t^2 \\
&\leq 2\sqrt{W_t}A_t + 2bA_t + bA_t^2 \\
&\overset{(i)}{\leq} 2\sqrt{W_t}A_t + 3bA_t^2\,,
\end{aligned}$$

where $A_t = \sqrt{2\log\log\left(2\left(\max\left(\frac{W_t}{b^2},1\right)\right)\right) + \log\frac{6}{\widetilde{\delta}}}$. Inequality (i) follows because $A_t \geq 1$. By identifying $V_t = W_t$ we conclude that for any $\widetilde{\delta} \in (0,1)$ and $t \in \mathbb{N}$

$$\mathbb{P}\left(\sum_{\ell=1}^{t} X_\ell > 2\sqrt{V_t}A_t + 3bA_t^2\right) \leq \widetilde{\delta}$$

Where $A_t = \sqrt{2\log\log\left(2\left(\max\left(\frac{V_t}{b^2},1\right)\right)\right) + \log\frac{6}{\widetilde{\delta}}}$. Since $V_t \leq tb^2$ with probability 1,

$$\frac{V_t}{b^2} \leq t,$$

And therefore $2\log\log\left(2\max(\frac{V_t}{b^2},1)\right) \leq 2\log\log 2t$ implying,

$$A_t \leq \sqrt{2\log\frac{12\log t}{\widetilde{\delta}}}$$

Thus

$$\mathbb{P}\left(\sum_{\ell=1}^{t} X_\ell > 4\sqrt{V_t \log\frac{12\log 2t}{\widetilde{\delta}}}A + 6b\log\frac{12\log 2t}{\widetilde{\delta}}\right) \le \widetilde{\delta}$$

Since $V_t \le S_t$ the result follows. ■

Lemma 54 *Let $\tilde{\delta} \in (0,1)$. The estimator $\widehat{f}_t$ satisfies,*

$$\sum_{\ell=1}^{t}\left(\widehat{f}_t(X_\ell) - f_\star(X_\ell)\right)^2 \le \gamma(t,\tilde{\delta})$$

for all $t \in \mathbb{N}$ with probability at least $1-\tilde{\delta}$, where $\gamma(t,\tilde{\delta}) = 256B(B+1)\log\left(\frac{12|\mathcal{G}|\log 2t}{\tilde{\delta}}\right)$.

Proof Since $\widehat{f}_t$ is the minimizer of the square loss over the data up to time t,

$$\sum_{\ell=1}^{t}(\widehat{f}_t(X_\ell) - Y_\ell)^2 \le \sum_{\ell=1}^{t}(f_\star(X_\ell) - Y_\ell)^2$$

Plugging in the definition $Y_\ell = f_\star(X_\ell) + \xi_\ell$ and expanding both sides of the inequality yields,

$$\sum_{\ell=1}^{t}(\widehat{f}_t(X_\ell) - f_\star(X_\ell) - \xi_\ell)^2 \le \sum_{\ell=1}^{t}\xi_\ell^2$$

and therefore,

$$\sum_{\ell=1}^{t}(\widehat{f}_t(X_\ell) - f_\star(X_\ell))^2 \le 2\sum_{\ell=1}^{t}\xi_\ell\left(\widehat{f}_t(X_\ell) - f_\star(X_\ell)\right) \tag{55}$$

For any <u>fixed</u> $f \in \mathcal{F}$ consider the martingale difference process $\{Z_\ell\}_{\ell=1}^{\infty}$,

$$Z_\ell^f = \xi_\ell\left(f(X_\ell) - f_\star(X_\ell)\right).$$

Since $|Z_\ell| \le B$ it is easy to see that by the boundedness assumption, $\mathbb{E}\left[\left(Z_\ell^f\right)^2\middle|\mathcal{O}_{\ell-1}\right] \le B^2\left(f(X_\ell) - f_\star(X_\ell)\right)^2$. Thus, Freedman's inequality (Lemma 53) implies,

$$\begin{aligned}
\sum_{\ell=1}^{t} Z_\ell^f &\le 4\sqrt{\sum_{\ell=1}^{t}\mathbb{E}\left[\left(Z_\ell^f\right)^2\middle|\mathcal{O}_{\ell-1}\right]\log\left(\frac{12|\mathcal{G}|\log 2t}{\tilde{\delta}}\right)} + 6B\log\frac{12|\mathcal{G}|\log 2t}{\tilde{\delta}} \\
&\le 4B\sqrt{\left[\sum_{\ell=1}^{t}\left(f(X_\ell) - f_\star(X_\ell)\right)^2\right]\log\left(\frac{12|\mathcal{G}|\log 2t}{\tilde{\delta}}\right)} + 6B\log\frac{12|\mathcal{G}|\log 2t}{\tilde{\delta}} \\
&\overset{(i)}{\le} \frac{\sum_{\ell=1}^{t}\left(f(X_\ell) - f_\star(X_\ell)\right)^2}{4} + 64B^2\log\left(\frac{12|\mathcal{G}|\log 2t}{\tilde{\delta}}\right) + 6B\log\left(\frac{12|\mathcal{G}|\log 2t}{\tilde{\delta}}\right) \\
&\le \frac{\sum_{\ell=1}^{t}\left(f(X_\ell) - f_\star(X_\ell)\right)^2}{4} + 64B(B+1)\log\left(\frac{12|\mathcal{G}|\log 2t}{\tilde{\delta}}\right)
\end{aligned}$$

with probability at least $1 - \frac{\tilde{\delta}}{|\mathcal{G}|}$ for all $t \in \mathbb{N}$. Where (i) holds because of the inequality $2\sqrt{ab} \leq \alpha a + \frac{b}{\alpha}$ for any $\alpha > 0$. Plugging this back into equation 55 we obtain,

$$\sum_{\ell=1}^{t}(\widehat{f}_t(X_\ell) - f_\star(X_\ell))^2 \leq 128B(B+1)\log\left(\frac{12|\mathcal{G}|\log 2t}{\tilde{\delta}}\right) + \frac{1}{2}\sum_{\ell=1}^{t}\left(\widehat{f}_t(X_\ell) - f_\star(X_\ell)\right)^2$$

Canceling terms on both sides yields (and multiplying by two) yields,

$$\sum_{\ell=1}^{t}(\widehat{f}_t(X_\ell) - f_\star(X_\ell))^2 \leq 256B(B+1)\log\left(\frac{12|\mathcal{G}|\log 2t}{\tilde{\delta}}\right)$$

The result follows. ■

Corollary 55 *If* $\gamma_r(t,\delta) = 512\log\left(\frac{24|\mathcal{G}_r|\log 2t}{\delta}\right), \gamma_c(t,\delta) = 512\log\left(\frac{24|\mathcal{G}_c|\log 2t}{\delta}\right)$ *then* $\theta_\star \in C_t^r(\delta)$ *and* $\mu_\star \in C_t^c(\delta)$ *for all* $t \in \mathbb{N}$ *with probability at least* $1-\delta$.

Proof This result is an immediate consequence of setting $B = 1$ and $\tilde{\delta} = \delta/2$ in Lemma 54. ■

G.2 Proof of Lemma 37

Notice that for any policy π

$$\widetilde{V}_t^c(\pi) \leq \mu_\star(\pi) + \max_{\mu,\mu' \in C_t^c(\delta)} \mu(\pi) - \mu'(\pi). \tag{56}$$

with probability at least $1-\delta$ for all $t \in \mathbb{N}$. This is because $\mu_\star$ belongs to $C_t^c(\delta)$ w.h.p and therefore

$$\begin{aligned}\widetilde{V}_t^c(\pi) &= \max_{\mu \in C_t^c(\delta)} \mu(\pi) = \mu_\star(\pi) + \max_{\mu \in C_t^c(\delta)} \mu(\pi) - \mu_\star(\pi) \\ &\leq \mu_\star(\pi) + \max_{\mu,\mu' \in C_t^c(\delta)} \mu(\pi) - \mu'(\pi).\end{aligned}$$

Similarly since $\theta_\star \in C_t^r(\delta)$ with high probability, $\max_{\theta \in C_t^r(\delta)} \theta(\pi) \geq \theta_\star(\pi)$ and therefore

$$\widetilde{V}_t^r(\pi) \geq \underbrace{\theta_\star(\pi)}_{V_t^r(\pi)} + \alpha_r \max_{\mu',\mu'' \in C_t^c(\delta)} \mu'(\pi) - \mu''(\pi) \tag{57}$$

Lemma 56 *If the event* $\mathcal{E}$ *defined by* (39) *holds and the scaling parameter satisfies* $\alpha_r = \frac{1-r_0}{\tau - c_0}$, *then for all* $t \in [T]$, *we have* $\widetilde{V}_t^r(\pi_t) \geq V_t^r(\pi_t^*) = \theta_*(\pi_t^*, \mathcal{A}_t)$.

Proof We are going to prove this result by splitting it into two cases determined by whether $\pi_t^\star$ belongs to $\widetilde{\Pi}_t$ or not.

Case 1. $\pi_t^* \in \widetilde{\Pi}_t$**.** Recall that $\pi_t = \arg\max_{\pi \in \widetilde{\Pi}_t} \widetilde{V}_t^r(\pi)$. It follows that $\widetilde{V}_t^r(\pi_t) \geq \widetilde{V}_t^r(\pi_t^*) \geq V_t^r(\pi_t^*)$ where the last inequality is true because $\widetilde{V}_t^r(\pi)$ is an optimistic estimator of the value of all policies (see Equation 57).

Case 2. $\pi_t^* \notin \widetilde{\Pi}_t$**.** Let $\pi_0 = \delta(x_0)$. By definition for all $\mu \in C_t^c(\delta)$ it follows that $\mu(x_0) = c_0$. Consider a mixture policy $\widetilde{\pi}_t = \gamma_t \pi_t^* + (1-\gamma_t)\pi_0$ where γ_t is the smallest value in $[0,1]$ such that $\widetilde{\pi}_t \in \widetilde{\Pi}_t$. Let's see this value exists:

Let $\widetilde{\mu}_t = \arg\max_{\mu \in C_t^c(\delta) \text{ s.t } \mu(x_0)=c_0} \mu(\pi_t^\star, \mathcal{A}_t)$. Observe that $\widetilde{\mu}_t$ by definition also satisfies $\widetilde{\mu}_t = \arg\max_{\mu \in C_t^c(\delta) \text{ s.t } \mu(x_0)=c_0} \mu(\gamma\pi_t^* + (1-\gamma)\pi_0, \mathcal{A}_t)$ and that

$$\widetilde{V}_t^c(\gamma\pi_t^* + (1-\gamma)\pi_0) = \widetilde{\mu}_t(\gamma\pi_t^* + (1-\gamma)\pi_0, \mathcal{A}_t) = \gamma\widetilde{\mu}_t(\pi_t^*, \mathcal{A}_t) + (1-\gamma)c_0.$$

This shows there exists a value $\gamma_t \in [0,1]$ such that $\widetilde{V}_t^c(\widetilde{\pi}_t) = \widetilde{\mu}_t(\gamma_t\pi_t^* + (1-\gamma_t)\pi_0, \mathcal{A}_t) = \tau$. Let's start by proving a lower bound for γ_t. By definition

$$\widetilde{V}_t^c(\widetilde{\pi}_t) = \gamma_t \widetilde{V}_t^c(\pi_t^*) + (1-\gamma_t)c_0 = \tau.$$

And therefore,

$$\begin{aligned}\gamma_t = \frac{\tau - c_0}{\widetilde{V}_t^c(\pi_t^*) - c_0} &\overset{(i)}{\geq} \frac{\tau - c_0}{\mu_\star(\pi_t^*) - c_0 + \max_{\mu,\mu' \in C_t^c(\delta_c)} \mu(\pi_t^*, \mathcal{A}_t) - \mu'(\pi_t^*, \mathcal{A}_t)} \\ &\overset{(ii)}{\geq} \frac{\tau - c_0}{\tau - c_0 + \max_{\mu,\mu' \in C_t^c(\delta_c)} \mu(\pi_t^*, \mathcal{A}_t) - \mu'(\pi_t^*, \mathcal{A}_t)} \end{aligned} \tag{58}$$

Where (i) follows from 56 and (ii) holds because it satisfies $\mu_\star(\pi_t^*, \mathcal{A}_t) \leq \tau$. Let $r_0 = \theta_\star(x_0)$. Since π_t and $\widetilde{\pi}_t$ are both feasible, it follows that $\widetilde{V}_t^r(\pi_t) \geq \widetilde{V}_t^r(\widetilde{\pi}_t)$ and therefore,

$$\begin{aligned}\widetilde{V}_t^r(\pi_t) \geq \widetilde{V}_t^r(\widetilde{\pi}_t) &= \gamma_t \widetilde{V}_t^r(\pi_t^*) + (1-\gamma_t)r_0 \\ &\overset{(i)}{\geq} \gamma_t \left(\theta_\star(\pi_t^\star, \mathcal{A}_t) + \alpha_r \max_{\mu', \mu'' \in C_t^c(\delta)} \mu^{'}(\pi_t^\star, \mathcal{A}_t) - \mu^{''}(\pi_t^\star, \mathcal{A}_t) \right) + (1-\gamma_t)r_0.\end{aligned}$$

Where (i) is a result of inequality 57. Let $C_1 = \max_{\mu', \mu'' \in C_t^c(\delta)} \mu^{'}(\pi_t^\star, \mathcal{A}_t) - \mu^{''}(\pi_t^\star, \mathcal{A}_t)$. Substituting the γ_t lower bound from Equation 58 in the RHS of the equation above,

$$\gamma_t(\theta_\star(\pi_t^\star, \mathcal{A}_t) + \alpha_r C_1) + (1-\gamma_t)r_0 = \frac{\tau - c_0}{\tau - c_0 + C_1}(\theta_\star(\pi_t^\star, \mathcal{A}_t) + \alpha_r C_1) + \frac{C_1}{\tau - c_0 + C_1} r_0$$

Since $\theta_\star(\pi_t^*, \mathcal{A}_t) \leq 1$ (Assumption 35), setting $\alpha_r = \frac{1-r_0}{\tau - c_0}$ is enough to guarantee the inequality $\widetilde{V}_t^r(\pi_t) \geq \theta_\star(\pi_t^\star, \mathcal{A}_t)$ holds. ■

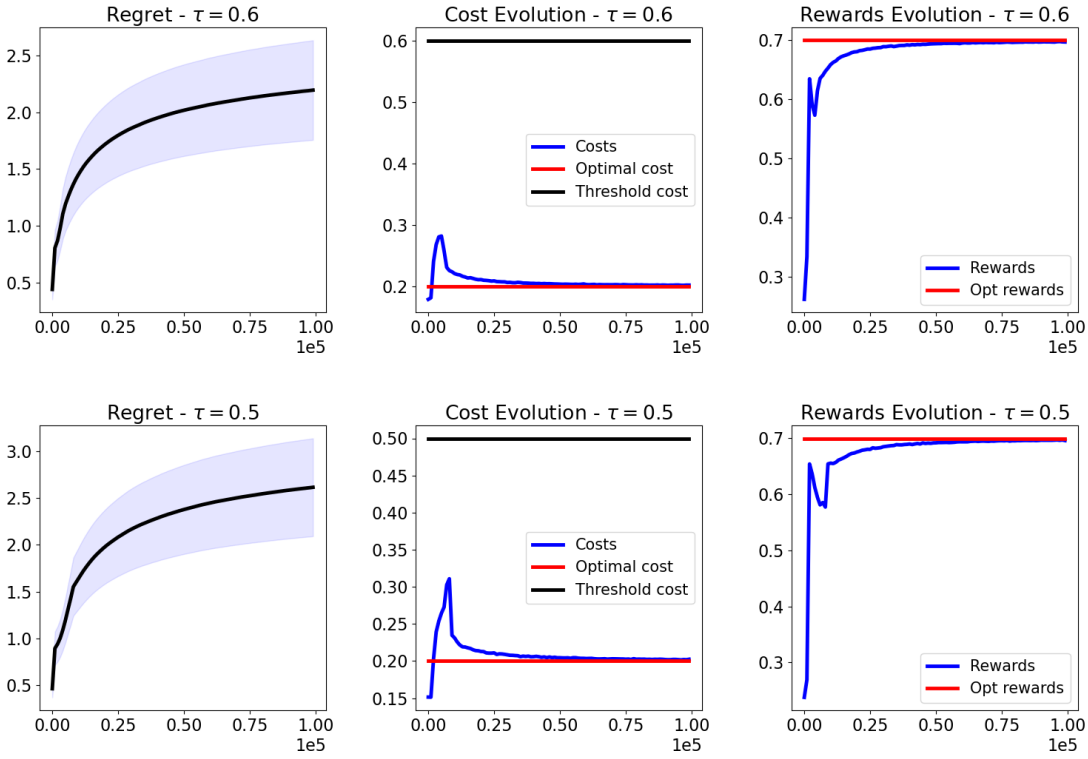

Figure 8: Regret *(left)*, cost *(middle)*, and reward *(right)* evolution of OPB in a 4-armed bandit problem with Bernoulli reward and cost distributions with means $\bar{r} = (0.1, 0.2, 0.4, 0.7)$ and $\bar{c} = (0, 0.4, 0.5, 0.2)$. The cost of the safe arm (Arm 1) is $\bar{c}_1 = 0$.

Appendix H. Additional Experiments of Section 5

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.

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.

S. Amani, C. Thrampoulidis, and L. Yang. Safe reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 243–253, 2021.

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.

A. Bura, A. Hasanzade Zonuzy, D. Kalathil, S. Shakkottai, and J.-F. Chamberland. Dope: Doubly optimistic and pessimistic exploration for safe reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

J. Chan, A. Pacchiano, N. Tripuraneni, Y. Song, P. Bartlett, and M. Jordan. Parallelizing contextual bandits. *arXiv:2105.10590*, 2023.

S. Chaudhary and D. Kalathil. Safe online convex optimization with unknown linear safety constraints. In *AAAI Conference on Artificial Intelligence*, pages 6175–6182, 2022.

T. Chen, A. Gangrade, and V. Saligrama. A doubly optimistic strategy for safe linear bandits. *arXiv preprint arXiv:2209.13694*, 2022.

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.

D. Ding, X. Wei, Z. Yang, Z. Wang, and M. Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3304–3312, 2021.

Y. Efroni, S. Mannor, and M. Pirotta. Exploration-exploitation in constrained mdps. *arXiv:2003.02189*, 2020.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

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

A. Ghosh, X. Zhou, and N. Shroff. Provably efficient model-free constrained RL with linear function approximation. In *Advances in Neural Information Processing Systems*, volume 35, pages 13303–13315, 2022.

S. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 2021.

E. Kaufmann, O. Cappé, and A. 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.

Chong Liu and Yu-Xiang Wang. Global optimization with parametric function approximation. In *International Conference on Machine Learning*, pages 22113–22136. PMLR, 2023.

T. Liu, R. Zhou, D. Kalathil, P. Kumar, and C. Tian. Learning policies with zero or bounded constraint violation for constrained MDPs. *Advances in Neural Information Processing Systems*, 34:17183–17193, 2021a.

X. Liu, B. Li, P. Shi, and L. Ying. An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. *Advances in Neural Information Processing Systems*, 34: 24075–24086, 2021b.

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. *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.

A. Pacchiano, M. Ghavamzadeh, P. Bartlett, and H. Jiang. Stochastic bandits with linear constraints. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, 2021.

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.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

M. Shi, Y. Liang, and N. Shroff. A near-optimal algorithm for safe reinforcement learning under instantaneous hard constraints. *arXiv preprint arXiv:2302.04375*, 2023.

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.

Z. Wang, A. Wagenmaker, and K. Jamieson. Best arm identification with safety constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 9114–9146, 2022.

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

H. Wei, X. Liu, and L. Ying. A provably-efficient model-free algorithm for constrained Markov decision processes. *arXiv preprint arXiv:2106.01577*, 2021.

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.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with ucb-based exploration. In *International Conference on Machine Learning*, pages 11492–11502. PMLR, 2020.

X. Zhou and B. Ji. On kernelized multi-armed bandits with constraints. *Advances in Neural Information Processing Systems*, 35:14–26, 2022.