Linear Stochastic Bandits Under Safety Constraints

Sanae Amani
University of California, Santa Barbara
samanigeshnigani@ucsb.edu

Mahnoosh Alizadeh
University of California, Santa Barbara
alizadeh@ucsb.edu

Christos Thrampoulidis
University of California, Santa Barbara
cthrampo@ucsb.edu

Abstract

Bandit algorithms have various application in safety-critical systems, where it is important to respect the system constraints that rely on the bandit's unknown parameters at every round. In this paper, we formulate a linear stochastic multi-armed bandit problem with safety constraints that depend (linearly) on an unknown parameter vector. As such, the learner is unable to identify all safe actions and must act conservatively in ensuring that her actions satisfy the safety constraint at all rounds (at least with high probability). For these bandits, we propose a new UCB-based algorithm called Safe-LUCB, which includes necessary modifications to respect safety constraints. The algorithm has two phases. During the pure exploration phase the learner chooses her actions at random from a restricted set of safe actions with the goal of learning a good approximation of the entire unknown safe set. Once this goal is achieved, the algorithm begins a safe exploration-exploitation phase where the learner gradually expands their estimate of the set of safe actions while controlling the growth of regret. We provide a general regret bound for the algorithm, as well as a problem dependent bound that is connected to the location of the optimal action within the safe set. We then propose a modified heuristic that exploits our problem dependent analysis to improve the regret.

1 Introduction

The stochastic multi-armed bandit (MAB) problem is a sequential decision-making problem where, at each step of a T-period run, a learner plays one of k arms and observes a corresponding loss that is sampled independently from an underlying distribution with unknown parameters. The learner's goal is to minimize the pseudo-regret, i.e., the difference between the expected T-period loss incurred by the decision making algorithm and the optimal loss if the unknown parameters were given. The linear stochastic bandit problem generalizes MAB to the setting where each arm is associated with a feature vector x and the expected loss of each arm is equal to the inner product of its feature vector x and an unknown parameter vector μ. There are several variants of linear stochastic bandits that consider finite or infinite number of arms, as well as the case where the set of feature vectors changes over time. A detailed account of previous work in this area will be provided in Section 1.2.

Bandit algorithms have found many applications in systems that repeatedly deal with unknown stochastic environments (such as humans) and seek to optimize a long-term reward by simultaneously learning and exploiting the unknown environment (e.g., ad display optimization algorithms with unknown user preferences, path routing, ranking in search engines). They are also naturally relevant for many cyber-physical systems with humans in the loop (e.g., pricing end-use demand in societal-scale infrastructure systems such as power grids or transportation networks to minimize system costs given the limited number of user interactions possible). However, existing bandit heuristics might not

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

be directly applicable in these latter cases. One critical reason is the existence of safety guarantees that have to be met at every single round. For example, when managing demand to minimize costs in a power system, it is required that the operational constraints of the power grid are not violated in response to our actions (these can be formulated as linear constraints that depend on the demand). Thus, for such systems, it becomes important to develop new bandit algorithms that account for critical safety requirements.

Given the high level of uncertainty about the system parameters in the initial rounds, any such bandit algorithm will be initially highly constrained in terms of safe actions that can be chosen. However, as further samples are obtained and the algorithm becomes more confident about the value of the unknown parameters, it is intuitive that safe actions become easier to distinguish and it seems plausible that the effect of the system safety requirements on the growth of regret can be diminished.

In this paper, we formulate a variant of linear stochastic bandits where at each round t, the learner's choice of arm should also satisfy a *safety constraint* that is dependent on the unknown parameter vector μ. While the formulation presented is certainly an abstraction of the complications that might arise in the systems discussed above, we believe that it is a natural first step towards understanding and evaluating the effect of safety constraints on the performance of bandit heuristics.

Specifically, we assume that the learner's goal is twofold: 1) Minimize the T-period cumulative pseudo-regret; 2) Ensure that a linear side constraint of the form $\mu^\dagger Bx \leq c$ is respected at every round during the T-period run of the algorithm, where B and c are known. See Section 1.1 for details. Given the learner's uncertainty about μ, the existence of this safety constraint effectively restricts the learner's choice of actions to what we will refer to as the *safe decision set* at each round t. To tackle this constraint, in Section 2, we present Safe-LUCB as a safe version of the standard linear UCB (LUCB) algorithm Dani et al. (2008); Abbasi-Yadkori et al. (2011); Rusmevichientong and Tsitsiklis (2010). In Section 3 we provide general regret bounds that characterize the effect of safety constraints on regret. We show that the regret of the modified algorithm is dependent on the parameter $\Delta = c - \mu^\dagger Bx^*$, where x^* denotes the optimal safe action given μ. When $\Delta > 0$ and is known to the learner, we show that the regret of Safe-LUCB is $\widetilde{\mathcal{O}}(\sqrt{T})$; thus, the effect of the system safety requirements on the growth of regret can be diminished (for large enough T). In Section 4, we also present a heuristic modification of Safe-LUCB that empirically approaches the same regret without a-priori knowledge of the value of Δ. On the other hand, when $\Delta = 0$, the regret of Safe-LUCB is $\widetilde{\mathcal{O}}(T^{2/3})$. Technical proofs and some further discussions are deferred to the appendix provided in the supplementary material.

Notation. The Euclidean norm of a vector x is denoted by $\|x\|_2$ and the spectral norm of a matrix M is denoted by $\|M\|$. We denote the transpose of any column vector x by $x^\dagger$. Let A be a positive definite $d \times d$ matrix and $v \in \mathbb{R}^d$. The weighted 2-norm of v with respect to A is defined by $\|v\|_A = \sqrt{v^\dagger Av}$. We denote the minimum and maximum eigenvalue of A by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$. The maximum of two numbers α, β is denoted $\alpha \vee \beta$. For a positive integer n, $[n]$ denotes the set $\{1, 2, \ldots, n\}$. Finally, we use standard $\widetilde{\mathcal{O}}$ notation for big-O notation that ignores logarithmic factors.

1.1 Safe linear stochastic bandit problem

Cost model. The learner is given a convex compact decision set $\mathcal{D}_0 \subset \mathbb{R}^d$. At each round t, the learner chooses an action $x_t \in \mathcal{D}_0$ which results in an observed loss ℓ_t that is linear on the unknown parameter μ with additive random noise η_t, i.e., $\ell_t := c_t(x_t) := \mu^\dagger x_t + \eta_t$.

Safety Constraint. The learning environment is subject to a side constraint that restricts the choice of actions by dividing $\mathcal{D}_0$ into a safe and an unsafe set. The learner is restricted to actions x_t from the *safe set* $\mathcal{D}_0^{\mathrm{S}}(\mu)$. As notation suggests, the safe set depends on the unknown parameter. Since μ is unknown, the learner is unable to identify the safe set and must act conservatively in ensuring that actions x_t are feasible for all t. In this paper, we assume that $\mathcal{D}_0^{\mathrm{S}}(\mu)$ is defined via a linear constraint

$$\mu^\dagger Bx_t \leq c, \tag{1}$$

which needs to be satisfied by x_t at all rounds t with high probability. Thus, $\mathcal{D}_0^{\mathrm{S}}(\mu)$ is defined as,

$$\mathcal{D}_0^{\mathrm{S}}(\mu) := \{x \in \mathcal{D}_0 \,:\, \mu^\dagger Bx \leq c\}. \tag{2}$$

The matrix $B \in \mathbb{R}^{d \times d}$ and the positive constant $c > 0$ are known to the learner. However, after playing any action x_t, the value $\mu^\dagger Bx_t$ is *not* observed by the learner. When clear from context, we drop the argument μ in the definition of the safe set and simply refer to it as $\mathcal{D}_0^{\mathrm{S}}$.

Regret. Let T be the total number of rounds. If $x_t,\ t \in [T]$ are the actions chosen, then the *cumulative pseudo-regret* (Audibert et al. (2009)) of the learner's algorithm for choosing the actions x_t is defined by $R_T = \sum_{t=1}^{T} \mu^\dagger x_t - \mu^\dagger x^*$, where x^* is the optimal *safe* action that minimizes the loss ℓ_t in expectation, i.e., $x^* \in \arg\min_{x \in \mathcal{D}_0^{\mathrm{S}}(\mu)} \mu^\dagger x$.

Goal. The goal of the learner is to keep R_T as small as possible. At the bare minimum, we require that the algorithm leads to $R_T/T \to 0$ (as T grows large). In contrast to existing linear stochastic bandit formulations, we require that the chosen actions $x_t, t \in [T]$ are safe (i.e., belong in $\mathcal{D}_0^{\mathrm{S}}$ (2)) with high probability. For the rest of this paper, we simply use regret to refer to the pseudo-regret R_T.

In Section 2.1 we place some further technical assumptions on $\mathcal{D}_0$ (bounded), on $\mathcal{D}_0^{\mathrm{S}}$ (non-empty), on μ (bounded) and on the distribution of η_t (subgaussian).

1.2 Related Works

Our algorithm relies on a modified version of the famous UCB algorithm known as UCB1, which was first developed by Auer et al. (2002). For linear stochastic bandits, the regret of the LUCB algorithm was analyzed by, e.g., Dani et al. (2008); Abbasi-Yadkori et al. (2011); Rusmevichientong and Tsitsiklis (2010); Russo and Van Roy (2014); Chu et al. (2011) and it was shown that the regret grows at the rate of $\sqrt{T}\log(T)$. Extensions to generalized linear bandit models have also been considered by, e.g., Filippi et al. (2010); Li et al. (2017). There are two different contexts where constraints have been applied to the stochastic MAB problem. The first line of work considers the MAB problem with global budget (a.k.a. knapsack) constraints where each arm is associated with a random resource consumption and the objective is to maximize the total reward before the learner runs out of resources, see, e.g., Badanidiyuru et al. (2013); Agrawal and Devanur (2016); Wu et al. (2015); Badanidiyuru et al. (2014). The second line of work considers stage-wise safety for bandit problems in the context of ensuring that the algorithm's regret performance stays above a fixed percentage of the performance of a baseline strategy at every round during its run Kazerouni et al. (2017); Wu et al. (2016). In Kazerouni et al. (2017), which is most closely related to our setting, the authors study a variant of LUCB in which the chosen actions are constrained such that the *cumulative* reward remains *strictly* greater than $(1-\alpha)$ times a given baseline reward for all t. In both of the above mentioned lines of work, the constraint applies to the cumulative resource consumption (or reward) across the entire run of the algorithm. As such, the set of permitted actions at each round vary depending on the round and on the history of the algorithm. This is unlike our constraint, which is applied at each individual round, is deterministic, and does *not* depend on the history of past actions.

In a more general context, the concept of safe learning has received significant attention in recent years from different communities. Most existing work that consider mechanisms for *safe exploration* in unknown and stochastic environments are in reinforcement learning or control. However, the notion of safety has many diverse definitions in this literature. For example, Moldovan and Abbeel (2012) proposes an algorithm that allows safe exploration in Markov Decision Processes (MDP) in order to avoid fatal absorbing states that must never be visited during the exploration process. By considering constrained MDPs that are augmented with a set of auxiliary cost functions and replacing them with surrogates that are easy to estimate, Achiam et al. (2017) purposes a policy search algorithm for constrained reinforcement learning with guarantees for near constraint satisfaction at each iteration. In the framework of global optimization or active data selection, Schreiter et al. (2015); Berkenkamp et al. (2016) assume that the underlying system is safety-critical and present active learning frameworks that use Gaussian Processes (GP) as non-parametric models to learn the safe decision set. More closely related to our setting, Sui et al. (2015, 2018) extend the application of UCB to *nonlinear* bandits with nonlinear constraints modeled through Gaussian processes (GPs). The algorithms in Sui et al. (2015, 2018) come with convergence guarantees, but *no* regret bounds as provided in our paper. Regret guarantees imply convergence guarantees from an optimization perspective (see Srinivas et al. (2010)), *but not the other way around*. Such approaches for safety-constrained optimization using GPs have shown great promise in robotics applications with safety constraints Ostafew et al. (2016); Akametalu et al. (2014).With a control theoretic point of view, Gillulay and Tomlin (2011) combines reachability analysis and machine learning for autonomously learning the dynamics of a target vehicle and Aswani et al. (2013) designs a learning-based MPC scheme that provides deterministic guarantees on robustness when the underlying system model is linear and has a known level of uncertainty. In a very recent related work Usmanova et al. (2019), the authors propose and analyze a (safe) variant of the Frank-Wolfe algorithm to solve a smooth

optimization problem with unknown linear constraints that are accessed by the learner via stochastic zeroth-order feedback. The main goal in Usmanova et al. (2019) is to provide a convergence rate for more general convex objective, whereas we aim to provide *regret bounds* for a linear but otherwise unknown objective.

2 A Safe-LUCB Algorithm

Our proposed algorithm is a safe version of LUCB. As such, it relies on the well-known heuristic principle of *optimism in the face of uncertainty* (OFU). The algorithm constructs a confidence set $\mathcal{C}_t$ at each round t, within which the unknown parameter μ lies with high probability. In the absence of any constraints, the learner chooses the most "favorable" environment μ from the set $\mathcal{C}_t$ and plays the action x_t that minimizes the expected loss in that environment. However, the presence of the constraint (1) complicates the choice of the learner. To address this, we propose an algorithm called *safe linear upper confidence bound* (Safe-LUCB), which attempts to minimize regret while making sure that the safety constraints (1) are satisfied. Safe-LUCB is summarized in Algorithm 1 and a detailed presentation follows in Sections 2.2 and 2.3, where we discuss the *pure-exploration* and *safe exploration-exploitation* phases of the algorithm, respectively. Before these, in Section 2.1 we introduce the necessary conditions under which our proposed algorithm operates and achieves good regret bounds as will be shown in Section 3.

2.1 Model Assumptions

Let $\mathcal{F}_t = \sigma(x_1, x_2, \dots, x_{t+1}, \eta_1, \eta_2, \dots, \eta_t)$ be the σ-algebra (or, history) at round t. We make the following standard assumptions on the noise distribution, on the parameter μ and on the actions.

Assumption 1 (Subgaussian Noise)**.** *For all t, η_t is conditionally zero-mean R-sub-Gaussian for fixed constant $R \geq 0$, i.e., $\mathbb{E}[\eta_t \,|\, x_{1:t}, \eta_{1:t-1}] = 0$ and $\mathbb{E}[e^{\lambda \eta_t} \,|\, \mathcal{F}_{t-1}] \leq \exp(\lambda^2 R^2/2), \quad \forall \lambda \in \mathbb{R}$.*

Assumption 2 (Boundedness)**.** *There exist positive constants S, L such that $\|\mu\|_2 \leq S$ and $\|x\|_2 \leq L, \forall x \in \mathcal{D}_0$. Also, $\mu^\dagger x \in [-1, 1], \forall x \in \mathcal{D}_0$.*

In order to avoid trivialities, we also make the following assumption. This, together with the assumption that $C > 0$ in (1), guarantee that the safe set $\mathcal{D}_0^{\mathrm{S}}(\mu)$ is non-empty (for every μ).

Assumption 3 (Non-empty safe set)**.** *The decision set $\mathcal{D}_0$ is a convex body in $\mathbb{R}^d$ that contains the origin in its interior.*

Algorithm 1 Safe-LUCB

Pure exploration phase:
for $t = 1, 2, \dots, T'$
  Randomly choose $x_t \in \mathcal{D}^w$ (defined in (3)) and observe loss $\ell_t = c_t(x_t)$.
end for
Safe exploration-exploitation phase:
for $t = T' + 1, 2, \dots, T$
  Set $A_t = \lambda I + \sum_{\tau=1}^{t-1} x_\tau x_\tau^\dagger$ and compute $\hat{\mu}_t = A_t^{-1} \sum_{\tau=1}^{t-1} \ell_\tau x_\tau$
  $\mathcal{C}_t = \{v \in \mathbb{R}^d : \|v - \hat{\mu}_t\|_{A_t} \leq \beta_t\}$ and β_t chosen as in (7)
  $\mathcal{D}_t^{\mathrm{S}} = \{x \in \mathcal{D}_0 : v^\dagger B x \leq c, \forall v \in \mathcal{C}_t\}$
  $x_t = \arg\min_{x \in \mathcal{D}_t^{\mathrm{S}}} \min_{v \in \mathcal{C}_t} v^\dagger x$
  Choose x_t and observe loss $\ell_t = c_t(x_t)$.
end for

2.2 Pure exploration phase

The pure exploration phase of the algorithm runs for rounds $t \in [T']$, where T' is passed as input to the algorithm. In Section 3, we will show how to appropriately choose its value to guarantee that the cumulative regret is controlled. During this phase, the algorithm selects random actions from a safe subset $\mathcal{D}^w \subset \mathcal{D}_0$ that we define next. For every chosen action x_t, we observe a loss ℓ_t. The collected action-loss pairs (x_t, ℓ_t) over the T' rounds are used in the second phase to obtain a good estimate of μ. We will see in Section 2.3 that this is important since the quality of the estimate of μ determines our belief of which actions are safe. Now, let us define the safe subset $\mathcal{D}^w$.

The safe set $\mathcal{D}_0^{\mathrm{S}}$ is unknown to the learner (since μ is unknown). However, it can be deduced from the constraint (1) and the boundedness Assumption 2 on μ, that the following subset $\mathcal{D}^w \subset \mathcal{D}_0$ is safe:

$$\mathcal{D}^w := \{x \in \mathcal{D}_0 : \max_{\|v\|_2 \leq S} v^\dagger B x \leq c\} = \{x \in \mathcal{D}_0 : \|Bx\|_2 \leq c/S\}. \tag{3}$$

Note that the set $\mathcal{D}^w$ is only a conservative (inner) approximation of $\mathcal{D}_0^{\mathrm{S}}$, but this is inevitable, since the learner has not yet collected enough information on the unknown parameter μ.

In order to make the choice of random actions $x_t, t \in [T']$ concrete, let $X \sim \mathrm{Unif}(\mathcal{D}^w)$ be a d-dimensional random vector uniformly distributed in $\mathcal{D}^w$ according to the probability measure given by the normalized volume in $\mathcal{D}^w$ (recall that $\mathcal{D}^w$ is a convex body by Assumption 3). During rounds $t \in [T']$, Safe-LUCB chooses safe IID actions $x_t \overset{\mathrm{iid}}{\sim} X$. For future reference, we denote the covariance matrix of X by $\Sigma = \mathbb{E}[XX^\dagger]$ and its minimum eigenvalue by

$$\lambda_- := \lambda_{\min}(\Sigma) > 0. \tag{4}$$

Remark 1. *Since $\mathcal{D}_0$ is compact with zero in its interior, we can always find $0 < \epsilon \leq C/S$ such that*

$$\widetilde{\mathcal{D}^w} := \{x \in \mathbb{R}^d \,|\, \|Bx\|_2 = \epsilon\} \subset \mathcal{D}^w. \tag{5}$$

Thus, an effective way to choose (random) actions x_t during the safe-exploration phase for which an explicit expression for λ_- is easily derived, is as follows. For simplicity, we assume B is invertible. Let ϵ be the largest value $0 < \epsilon \leq c/S$ such that (5) holds. Then, generate samples $x_t \sim \mathrm{Unif}(\widetilde{\mathcal{D}^{\mathrm{w}}}), t = 1, \ldots, T'$, by choosing $x_t = \epsilon B^{-1} z_t$, where z_t are iid samples on the unit sphere $\mathcal{S}^{d-1}$. Clearly, $\mathbb{E}[z_t z_t^\dagger] = \frac{1}{d} I$. Thus, $\Sigma := \mathbb{E}[x_t x_t^\dagger] = \frac{\epsilon^2}{d} \left(B^\dagger B\right)^{-1}$, from which it follows that $\lambda_- := \lambda_{\min}(\Sigma) = \frac{\epsilon}{d\, \lambda_{\max}\left(B^\dagger B\right)} = \frac{\epsilon^2}{d\|B\|^2}$.

2.3 Safe exploration-exploitation phase

We implement the OFU principle *while respecting the safety constraints*. First, at each $t = T' + 1, T' + 2 \ldots, T$, the algorithm uses the previous action-observation pairs and obtains a λ-regularized least-squares estimate $\hat{\mu}_t$ of μ with regularization parameter $\lambda > 0$ as follows:

$$\hat{\mu}_t = A_t^{-1} \sum_{\tau=1}^{t-1} \ell_\tau x_\tau, \text{ where } A_t = \lambda I + \sum_{\tau=1}^{t-1} x_\tau x_\tau^\dagger.$$

Then, based on $\hat{\mu}_t$ the algorithm builds a *confidence set*

$$\mathcal{C}_t := \{v \in \mathbb{R}^d : \|v - \hat{\mu}_t\|_{A_t} \leq \beta_t\}, \tag{6}$$

where, β_t is chosen according to Theorem 1 below (Abbasi-Yadkori et al. (2011)) to guarantee that $\mu \in \mathcal{C}_t$ with high probability.

Theorem 1 (Confidence Region, Abbasi-Yadkori et al. (2011))**.** *Let Assumptions 1 and 2 hold. Fix any $\delta \in (0, 1)$ and let β_t in (6) be chosen as follows,*

$$\beta_t = R\sqrt{d \log\left(\frac{1 + (t-1)L^2/\lambda}{\delta}\right)} + \lambda^{1/2} S, \qquad \textit{for all} \quad t > 0. \tag{7}$$

Then, with probability at least $1 - \delta$, for all $t > 0$, it holds that $\mu \in \mathcal{C}_t$.

The remaining steps of the algorithm also build on existing principles of UCB algorithms. However, here we introduce necessary modifications to account for the safety constraint (1). Specifically, we choose the actions with the following two principles.

Caution in the face of constraint violation. At each round t, the algorithm performs conservatively, to ensure that the constraint (1) is satisfied for the chosen action x_t. As such, at the beginning of each round $t = T' + 1, \ldots, T$, Safe-LUCB forms the so-called *safe decision set* denoted as $\mathcal{D}_t^{\mathrm{S}}$:

$$\mathcal{D}_t^{\mathrm{S}} = \{x \in \mathcal{D}_0 : v^\dagger B x \leq c, \forall v \in \mathcal{C}_t\}. \tag{8}$$

Recall from Theorem 1 that $\mu \in \mathcal{C}_t$ with high probability. Thus, $\mathcal{D}_t^{\mathrm{S}}$ is guaranteed to be a set of safe actions that satisfy (1) with the same probability. On the other hand, note that $\mathcal{D}_t^{\mathrm{S}}$ is still a

conservative inner approximation of $\mathcal{D}_0^{\mathrm{S}}(\mu)$ (actions in it are safe for *all* parameter vectors in $\mathcal{C}_t$, not only for the true μ). This (unavoidable) conservative definition of safe decision sets could contribute to the growth of the regret. This is further studied in Section 3.

Optimism in the face of uncertainty in cost. After choosing safe actions randomly at rounds $1, \ldots, T'$, the algorithm creates the safe decision set $\mathcal{D}_t^{\mathrm{S}}$ at all rounds $t \geq T' + 1$, and chooses an action x_t based on the OFU principle. Specifically, a pair $(x_t, \tilde{\mu}_t)$ is chosen such that

$$\tilde{\mu}_t^\dagger x_t = \min_{x \in \mathcal{D}_t^{\mathrm{S}}, v \in \mathcal{C}_t} v^\dagger x. \tag{9}$$

3 Regret Analysis of Safe-LUCB

3.1 The regret of safety

In the safe linear bandit problem, the safe set $\mathcal{D}_0^{\mathrm{S}}$ is not known, since μ is unknown. Therefore, at each round, the learner chooses actions from a conservative inner approximation of $\mathcal{D}_0^{\mathrm{S}}$. Intuitively, the better this approximation, the more likely that the optimistic actions of Safe-LUCB lead to good cumulant regret, ideally of the same order as that of LUCB in the original linear bandit setting.

A key difference in the analysis of Safe-LUCB compared to the classical LUCB is that x^* may not lie within the estimated safe set $\mathcal{D}_t^{\mathrm{S}}$ at each round. To see what changes, consider the standard decomposition of the instantaneous regret $r_t,\ t = T' + 1, \ldots, T$ in two terms as follows (e.g., Dani et al. (2008); Abbasi-Yadkori et al. (2011)):

$$r_t := \mu^\dagger x_t - \mu^\dagger x^* = \underbrace{\mu^\dagger x_t - \tilde{\mu}_t^\dagger x_t}_{\text{Term I}} + \underbrace{\tilde{\mu}_t^\dagger x_t - \mu^\dagger x^*}_{\text{Term II}}, \tag{10}$$

where, $(\tilde{\mu}_t, x_t)$ is the optimistic pair, i.e. the solution to the minimization in Step 10 of Algorithm 1. On the one hand, controlling Term I, is more or less standard and closely follows previous such bounds on UCB-type algorithms (e.g., Abbasi-Yadkori et al. (2011)); see Appendix B.2 for details. On the other hand, controlling Term II, which we call *the regret of safety* is more delicate. This complication lies at the heart of the new formulation with additional safety constraints. When safety constraints are absent, classical LUCB guarantees that Term II is non-positive. Unfortunately, this is *not* the case here: x^* does *not* necessarily belong to $\mathcal{D}_t^{\mathrm{S}}$ in (8), thus Term II can be positive. This extra regret of safety is the price paid by Safe-LUCB for choosing safe actions at each round. Our main contribution towards establishing regret guarantees is upper bounding Term II. We show in Section 3.2 that the pure-exploration phase is critical in this direction.

3.2 Learning the safe set

The challenge in controlling the regret of safety is that, in general, $\mathcal{D}_t^{\mathrm{S}} \neq \mathcal{D}_0^{\mathrm{S}}$. At a high level, we proceed as follows (see Appendix B.3 for details). First, we relate Term II with a certain notion of "distance" in the direction of x^* between the estimated set $\mathcal{D}_t^{\mathrm{S}}$ at rounds $t = T' + 1, \ldots, T$ and the true safe set $\mathcal{D}_0^{\mathrm{S}}$. Next, we show that this "distance" term can be controlled by appropriately lower bounding the minimum eigenvalue $\lambda_{\min}(A_t)$ of the Gram matrix A_t. Due to the interdependency of the actions x_t, it is difficult to directly establish such a lower bound for each round t. Instead, we use that $\lambda_{\min}(A_t) \geq \lambda_{\min}(A_{T'+1}),\ t \geq T' + 1$ and we are able to bound $\lambda_{\min}(A_{T'+1})$ thanks to the pure exploration phase of Safe-LUCB . Hence, the pure exploration phase guarantees that $\mathcal{D}_t^{\mathrm{S}}$ is a sufficiently good approximation to the true $\mathcal{D}_0^{\mathrm{S}}$ once the exploration-exploitation phase begins.

Lemma 1. *Let $A_{T'+1} = \lambda I + \sum_{t=1}^{T'} x_t x_t^\dagger$ be the Gram matrix corresponding to the first T' actions of Safe-LUCB (pure-exploration phase). Recall the definition of λ_- in (4). Then, for any $\delta \in (0, 1)$, it holds with probability at least $1 - \delta$,*

$$\lambda_{\min}(A_{T'+1}) \geq \lambda + \frac{\lambda_- T'}{2}, \tag{11}$$

provided that $T' \geq t_\delta := \frac{8L^2}{\lambda_-} \log(\frac{d}{\delta})$.

The proof of the lemma and technical details relating the result to a desired bound on Term II are deferred to Appendixes A and B.3, respectively.

3.3 Problem dependent upper bound

In this section, we present a problem-dependent upper bound on the regret of Safe-LUCB in terms of the following critical parameter, which we call the *safety gap*:

$$\Delta := c - \mu^\dagger B x^*. \tag{12}$$

Note that $\Delta \geq 0$. In this section, we assume that Δ is known to the learner. The next lemma shows that if $\Delta > 0$ [1] , then choosing $T' = \mathcal{O}(\log T)$ guarantees that $x^* \in \mathcal{D}_t^{\mathrm{S}}$ for all $t = T'+1, \ldots, T$.

Lemma 2 ($x^* \in \mathcal{D}_t^{\mathrm{S}}$)**.** *Let Assumptions 1, 2 and 3 hold. Fix any $\delta \in (0,1)$ and assume a positive safety gap $\Delta > 0$. Initialize Safe-LUCB with (recall the definition of t_δ in Lemma 1)*

$$T' \geq T_\Delta := \left(\frac{8L^2 \|B\|^2 \beta_T^2}{\lambda_- \Delta^2} - \frac{2\lambda}{\lambda_-} \right) \vee t_\delta. \tag{13}$$

Then, with probability at least $1-\delta$, for all $t = T'+1, \ldots, T$ it holds that $x^* \in \mathcal{D}_t^{\mathrm{S}}$.

In light of our discussion in Sections 3.1 and 3.2, once we have established that $x^* \in \mathcal{D}_t^{\mathrm{S}}$ for $t = T'+1, \ldots, T$, the regret of safety becomes nonpositive and we can show that the algorithm performs just like classical LUCB during the exploration-exploitation phase [2]. This is formalized in Theorem 2 showing that when $\Delta > 0$ (and is known), then the regret of Safe-LUCB is $\widetilde{\mathcal{O}}(\sqrt{T})$.

Theorem 2 (Problem-dependent bound; $\Delta > 0$)**.** *Let the same assumptions as in Lemma 2 hold. Initialize Safe-LUCB with $T' \geq T_\Delta$ specified in (13). Then, for $T \geq T'$, with probability at least $1-2\delta$, the cumulative regret of Safe-LUCB satisfies*

$$R_T \leq 2T' + 2\beta_T \sqrt{2d\,(T-T')\,\log\left(\frac{2TL^2}{d(\lambda_- T' + 2\lambda)} \right)}\,. \tag{14}$$

Specifically, choosing $T' = T_\Delta$ guarantees cumulant regret $\mathcal{O}(T^{1/2}\log T)$.

The bound in (14) is a contribution of two terms. The first one is a trivial bound on the regret of the exploration-only phase of Safe-LUCB and is proportional to its duration T'. Thanks to Lemma 2 the duration of the exploration phase is limited to T_Δ rounds and T_Δ is (at most) logarithmic in the total number of rounds T. Thus, the first summand in (14) contributes only $\mathcal{O}(\log T)$ in the total regret. Note, however, that T_Δ grows larger as the normalized safety gap $\Delta/\|B\|$ becomes smaller. The second summand in (14) contributes $\mathcal{O}(T^{1/2}\log T)$ and bounds the cumulant regret of the exploration-exploitation phase, which takes the bulk of the algorithm. More specifically, it bounds the contribution of Term I in (10) since the Term II is zeroed out once $x^* \in \mathcal{D}_t^{\mathrm{S}}$ thanks to Lemma 2. Finally, note that Theorem 2 requires the total number of rounds T to be large enough for the desired regret performance. This is the price paid for the extra safety constraints compared to the performance of the classical LUCB in the original linear bandit setting. We remark that existing lower bounds for the simpler problem without safety constraints (e.g. Rusmevichientong and Tsitsiklis (2010); Dani et al. (2008)), show that the regret $\widetilde{\mathcal{O}}(\sqrt{Td})$ of Theorem 2 cannot be improved modulo logarithmic factors. The proofs of Lemma 2 and Theorem 2 are in Appendix B.

3.4 General upper bound

We now extend the results of Section 3.3 to instances where the safety gap is zero, i.e. $\Delta = 0$. In this case, we cannot guarantee an exploration phase that results in $x^* \in \mathcal{D}_t^{\mathrm{S}}, t > T'$ in a reasonable time length T'. Thus, the regret of safety is not necessarily non-positive and it is unclear whether a sub-linear cumulant regret is possible.

Theorem 3 shows that Safe-LUCB achieves regret $\widetilde{\mathcal{O}}(T^{2/3})$ when $\Delta = 0$. Note that this (worst-case) bound is also applicable when the safety gap is unknown to the learner. While it is significantly worse than the performance guaranteed by Theorem 2, it proves that Safe-LUCB always leads to $R_T/T \to 0$ as T grows large. The proof is deferred to Appendix B.

[1] We remark that the case $\Delta > 0$ studied here is somewhat reminiscent of the assumption $\alpha r_\ell > 0$ in Kazerouni et al. (2017).

[2] Our simulation results in Appendix F emphasize the critical role of a sufficiently long pure exploration phase by Safe-LUCB as suggested by Lemma 2. Specifically, Figure 1b depicts an instance where *no* exploration leads to significantly worse order of regret.

Theorem 3 (General bound: worst-case)**.** *Suppose Assumptions 1, 2 and 3 hold. Fix any $\delta \in (0, 0.5)$. Initialize Safe-LUCB with $T' \geq t_\delta$ specified in Lemma 1. Then, with probability at least $1 - 2\delta$ the cumulative regret R_T of Safe-LUCB for $T \geq T'$ satisfies*

$$R_T \leq 2T' + 2\beta_T \sqrt{2d(T - T') \log\left(\frac{2TL^2}{d(\lambda_- T' + 2\lambda)}\right)} + \frac{2\sqrt{2}\|B\|L\beta_T(T - T')}{c\sqrt{\lambda_- T' + 2\lambda}}. \tag{15}$$

Specifically, choosing $T' = T_0 := \left(\frac{\|B\|L\beta_T T}{c\sqrt{2\lambda_-}}\right)^{\frac{2}{3}} \vee t_\delta$, guarantees regret $\mathcal{O}(T^{2/3} \log T)$.

Compared to Theorem 2, the bound in (15) is now comprised of three terms. The first one captures again the exploration-only phase and is linear in its duration T'. However, note that T' is now $\mathcal{O}(T^{2/3} \log T)$, i.e., of the same order as the total bound. The second term bounds the total contribution of Term I of the exploration-exploitation phase. As usual, its order is $\widetilde{\mathcal{O}}(T^{1/2})$. Finally, the additional third term bounds the regret of safety and is of the same order as that of the first term.

4 Unknown Safety Gap

In Section 3.3 we showed that when the safety gap $\Delta > 0$, then Safe-LUCB achieves good regret performance $\widetilde{\mathcal{O}}(\sqrt{T})$. However, this requires that the value of Δ, or at least a (non-trivial) lower bound on it, be known to the learner so that T' is initialized appropriately according to Lemma 2. This requirement might be restrictive in certain applications. When that is the case, one option is to run Safe-LUCB with a choice of T' as suggested by Theorem 3, but this could result in an unnecessarily long pure exploration period (during which regret grows linearly). Here, we present an alternative. Specifically, we propose a variation of Safe-LUCB refered to as *generalized safe linear upper confidence bound* (GSLUCB). The key idea behind GSLUCB is to build a lower confidence bound Δ_t for the safety gap Δ and calculate the length of the pure exploration phase associated with Δ_t, denoted as T'_t. This allows the learner to stop the pure exploration phase at round t such that condition $t \leq T'_{t-1}$ has been met. While we do not provide a separate regret analysis for GSLUCB, it is clear that the worst case regret performance would match that of Safe-LUCB with $\Delta = 0$. However, our numerical experiment highlights the improvements that GSLUCB can provide for the cases where $\Delta \neq 0$. We give a full explanation of GSLUCB, including how we calculate the lower confidence bound Δ_t, in Appendix E.

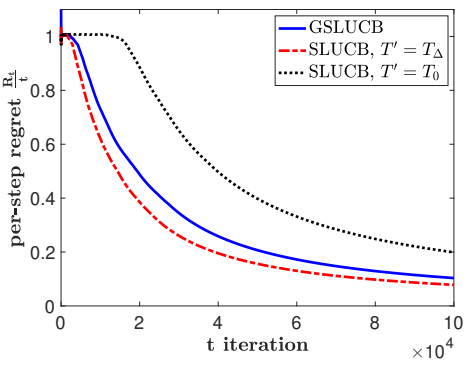
Figure 1a compares the average per-step regret of 1) Safe-LUCB with knowledge of Δ; 2) Safe-LUCB without knowledge of Δ (hence, assuming $\Delta = 0$); 3) GSLUCB without knowledge of Δ, in a simplified setting of K-armed linear bandits with strictly positive safety gap (see Appendix C). The details on the parameters of the simulations are deferred to Appendix F.

Algorithm 2 GSLUCB

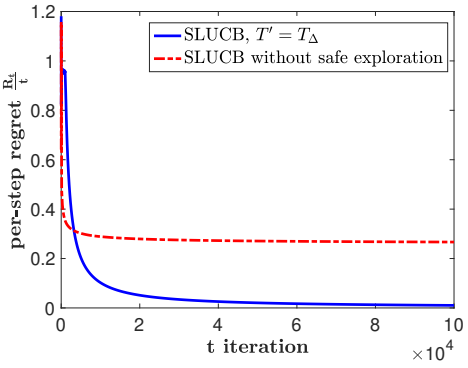
Pure exploration phase:
$t \leftarrow 1$, $T'_0 = T_0$
while $\left(t \leq \min\left(T'_{t-1}, T_0\right)\right)$
  Randomly choose $x_t \in \mathcal{D}^w$ and observe loss $\ell_t = c_t(x_t)$.
  Δ_t = Lower confidence bound on Δ at round t
  if $\Delta_t > 0$ **then** $T'_t = T_{\Delta_t}$
  else $T'_t = T_0$
  end if
  $t \leftarrow t + 1$
end while
Safe exploration exploitation phase: Lines 6 - 12 of Safe-LUCB for all remaining rounds.

5 Conclusions

We have formulated a linear stochastic bandit problem with safety constraints that depend linearly on the unknown problem parameter μ. While simplified, the model captures the additional complexity introduced in the problem by the requirement that chosen actions belong to an unknown safe set. As such, it allows us to quantify tradeoffs between learning the safe set and minimizing the regret. Specifically, we propose Safe-LUCB which is comprised of two phases: (i) a pure-exploration phase that speeds up learning the safe set; (ii) a safe exploration-exploitation phase that optimizes



(a) Average per-step regret of Safe-LUCB and GSLUCB with a decision set of K arms.



(b) Per-step regret of Safe-LUCB with and without pure exploration phase.

Figure 1: Simulation of per-step regret.

minimizing the regret. Our analysis suggests that the safety gap Δ plays a critical role. When $\Delta > 0$ we show how to achieve regret $\widetilde{\mathcal{O}}(\sqrt{T})$ as in the classical linear bandit setting. However, when $\Delta = 0$, the regret of Safe-LUCB is $\widetilde{\mathcal{O}}(T^{2/3})$. It is an interesting open problem to establish lower bounds for an arbitrary policy that accounts for the safety constraints. Our analysis of Safe-LUCB suggests that $\Delta = 0$ is a worst-case scenario, but it remains open whether the $\widetilde{\mathcal{O}}(T^{2/3})$ regret bound can be improved in that case. Natural extensions of the problem setting to multiple constraints and generalized linear bandits (possibly with generalized linear constraints) might also be of interest.

6 Acknowledgement

This research is supported by UCOP grant LFR-18-548175 and NSF grant 1847096.

References

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

Achiam, J., Held, D., Tamar, A., and Abbeel, P. (2017). Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 22–31. JMLR. org.

Agrawal, S. and Devanur, N. (2016). Linear contextual bandits with knapsacks. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 3450–3458. Curran Associates, Inc.

Akametalu, A. K., Fisac, J. F., Gillula, J. H., Kaynama, S., Zeilinger, M. N., and Tomlin, C. J. (2014). Reachability-based safe learning with gaussian processes. In *53rd IEEE Conference on Decision and Control*, pages 1424–1431.

Aswani, A., Gonzalez, H., Sastry, S. S., and Tomlin, C. (2013). Provably safe and robust learning-based model predictive control. *Automatica*, 49(5):1216–1226.

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256.

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

Badanidiyuru, A., Langford, J., and Slivkins, A. (2014). Resourceful contextual bandits. In Balcan, M. F., Feldman, V., and Szepesvári, C., editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 1109–1134, Barcelona, Spain. PMLR.

Berkenkamp, F., Krause, A., and Schoellig, A. P. (2016). Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics. *arXiv preprint arXiv:1602.04450*.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In Gordon, G., Dunson, D., and Dudík, M., editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 208–214, Fort Lauderdale, FL, USA. PMLR.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. (2010). Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594.

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

Kazerouni, A., Ghavamzadeh, M., Abbasi, Y., and Van Roy, B. (2017). Conservative contextual linear bandits. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 3910–3919. Curran Associates, Inc.

Li, L., Lu, Y., and Zhou, D. (2017). Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2071–2080. JMLR. org.

Moldovan, T. M. and Abbeel, P. (2012). Safe exploration in markov decision processes. *arXiv preprint arXiv:1205.4810*.

Ostafew, C. J., Schoellig, A. P., and Barfoot, T. D. (2016). Robust constrained learning-based nmpc enabling reliable mobile robot path tracking. *The International Journal of Robotics Research*, 35(13):1547–1563.

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

Russo, D. and Van Roy, B. (2014). Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243.

Schreiter, J., Nguyen-Tuong, D., Eberts, M., Bischoff, B., Markert, H., and Toussaint, M. (2015). Safe exploration for active learning with gaussian processes. In Bifet, A., May, M., Zadrozny, B., Gavalda, R., Pedreschi, D., Bonchi, F., Cardoso, J., and Spiliopoulou, M., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 133–149, Cham. Springer International Publishing.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 1015–1022. Omnipress.

Sui, Y., Burdick, J., Yue, Y., et al. (2018). Stagewise safe bayesian optimization with gaussian processes. In *International Conference on Machine Learning*, pages 4788–4796.

Sui, Y., Gotovos, A., Burdick, J. W., and Krause, A. (2015). Safe exploration for optimization with gaussian processes. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 997–1005. JMLR.org.

Tropp, J. A. et al. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230.

Usmanova, I., Krause, A., and Kamgarpour, M. (2019). Safe convex learning under uncertain constraints. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 2106–2114. PMLR.

Wu, H., Srikant, R., Liu, X., and Jiang, C. (2015). Algorithms with logarithmic or sublinear regret for constrained contextual bandits. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 433–441. Curran Associates, Inc.

Wu, Y., Shariff, R., Lattimore, T., and Szepesvári, C. (2016). Conservative bandits. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1254–1262. JMLR.org.