Stage-wise Conservative Linear Bandits

Ahmadreza Moradipari, Christos Thrampoulidis, Mahnoosh Alizadeh
Department of Electrical and Computer Enginnering
University of California, Santa Barbara
ahmadreza_moradipari@ucsb.edu

Abstract

We study stage-wise conservative linear stochastic bandits: an instance of bandit optimization, which accounts for (unknown) "safety constraints" that appear in applications such as online advertising and medical trials. At each stage, the learner must choose actions that not only maximize cumulative reward across the entire time horizon, but further satisfy a linear baseline constraint that takes the form of a *lower bound* on the instantaneous reward. For this problem, we present two novel algorithms, *stage-wise conservative linear Thompson Sampling* (SCLTS) and *stage-wise conservative linear UCB* (SCLUCB), that respect the baseline constraints and enjoy probabilistic regret bounds of order $\mathcal{O}(\sqrt{T}\log^{3/2} T)$ and $\mathcal{O}(\sqrt{T}\log T)$, respectively. Notably, the proposed algorithms can be adjusted with only minor modifications to tackle different problem variations, such as, constraints with bandit-feedback, or an unknown sequence of baseline rewards. We discuss these and other improvements over the state-of-the art. For instance, compared to existing solutions, we show that SCLTS plays the (non-optimal) baseline action at most $\mathcal{O}(\log T)$ times (compared to $\mathcal{O}(\sqrt{T})$). Finally, we make connections to another studied form of "safety constraints" that takes the form of an *upper bound* on the instantaneous reward. While this incurs additional complexity to the learning process as the optimal action is not guaranteed to belong to the "safe set" at each round, we show that SCLUCB can properly adjust in this setting via a simple modification.

1 Introduction

With the growing range of applications of bandit algorithms for safety critical real-world systems, the demand for safe learning is receiving increasing attention Tucker et al. (2020). In this paper, we investigate the effect of stage-wise safety constraints on the linear stochastic bandit problem. Inspired by the earlier work of Kazerouni et al. (2017); Wu et al. (2016), the type of safety constraint we consider in this paper was first introduced by Khezeli and Bitar (2019). As with the classic linear stochastic bandit problem, the learner wishes to choose a sequence of actions x_t that maximize the expected reward over the horizon. However, here the learner is also given a baseline policy that suggests an action with a guaranteed level of expected reward at each stage of the algorithm. This could be based on historical data, e.g., historical ad placement or medical treatment policies with known success rates. The safety constraint imposed on the learner requires her to ensure that the expected reward of her chosen action at every single round be no less than a predetermined fraction of the expected reward of the action suggested by baseline policy. An example that might benefit from the design of stage-wise conservative learning algorithms arises in recommender systems, where the recommender might wish to avoid recommendations that are extremely disliked by the users at any single round. Our proposed stage-wise conservative constraints ensure that at no round would the recommendation system cause severe dissatisfaction for the user, and the reward of action employed by the learning algorithm, if not better, should be close to that of baseline policy. Another example is

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

in clinical trials where the effects of different therapies on patients' health are initially unknown. We can consider the baseline policy to be treatments that have been historically employed, with known effectiveness. The proposed stage-wise conservative constraint guarantees that at each stage, the learning algorithm suggests an action (a therapy) that achieves the expected reward close to that of the baseline treatment, and as such, this experimentation does not cause harm to *any single patient's health*. To tackle this problem, Khezeli and Bitar (2019) proposed a greedy algorithm called SEGE. They use the decomposition of the regret first proposed in Kazerouni et al. (2017), and show an upper bound of order $\mathcal{O}(\sqrt{T})$ over the number of times that the learning algorithm plays the baseline actions, overall resulting in an expected regret of $\mathcal{O}(\sqrt{T}\log T)$. For this problem, we present two algorithms, SCLTS and SCLUCB, and we provide regret bounds of order $\mathcal{O}(\sqrt{T}\log^{3/2} T)$ and $\mathcal{O}(\sqrt{T}\log T)$, respectively. As it is explained in details in Section 3, we improve the result of Khezeli and Bitar (2019), i.e., we show our proposed algorithms play the (non-optimal) baseline actions at most $\mathcal{O}(\log T)$ times, while also relaxing a number of assumptions made in Khezeli and Bitar (2019). Moreover, we show that our proposed algorithms are adaptable with minor modifications to other safety-constrained variations of this problem. This includes the case where the constraint has a different unknown parameter than the reward function with bandit feedback (Section 3.1), as well as the setting where the reward of baseline action is unknown to the learner in advance (Section 4).

1.1 Conservative Stochastic Linear bandit (LB) Problem with Stage-wise Constraints

Linear Bandit. The learner is given a convex and compact set of actions $\mathcal{X} \subset \mathsf{R}^d$. At each round t, she chooses an action x_t and observes a random reward

$$y_t = \langle x_t, \theta_\star \rangle + \xi_t, \tag{1}$$

where $\theta_\star \in \mathsf{R}^d$ is *unknown* but fixed reward parameter and ξ_t is zero-mean additive noise. We let r_t be the expected reward of action x_t at round t, i.e., $r_t := \mathsf{E}[y_t] = \langle x_t, \theta_\star \rangle$.

Baseline actions and stage-wise constraint. We assume that the learner is given a baseline policy such that selecting the baseline action x_{b_t} at round t, she would receive an expected reward $r_{b_t} := \langle x_{b_t}, \theta_\star \rangle$. We assume that the learner knows the expected reward of the actions chosen by the baseline policy. We further assume that the learner's action selection rule is subject to a stage-wise conservative constraint of the form[1]

$$r_t = \langle x_t, \theta_\star \rangle \geq (1-\alpha) r_{b_t}, \tag{2}$$

that needs to be satisfied at each round t. In particular, constraint (2) guarantees that at each round t, the expected reward of the action chosen by the learner stays above the predefined fraction $1-\alpha \in (0,1)$ of the baseline policy. The parameter α, controlling the conservatism level of the learning process, is assumed known to the learner similar to Kazerouni et al. (2017); Wu et al. (2016). At each round t, an action is called *safe* if its expected reward is above the predetermined fraction of the baseline policy, i.e., $(1-\alpha)r_{b_t}$.

Remark 1.1. *It is reasonable to assume that the leaner has an accurate estimate of the expected reward of the actions chosen by baseline policy Kazerouni et al. (2017). However, in Section 4, we relax this assumption, and propose an algorithm to the case where the expected rewards of the actions chosen by baseline policy are unknown to the learner in advance.*

Regret. The *cumulative pseudo-regret* of the learner up to round T is defined as $R(T) = \sum_{t=1}^{T} \langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle$, where $x_\star$ is the optimal safe action that maximizes the expected reward,

$$x_\star = \arg\max_{x \in \mathcal{X}} \langle x, \theta_\star \rangle. \tag{3}$$

The learner's objective is to minimize the pseudo-regret, while respecting the stage-wise conservative constraint in (2). For the rest of the paper, we use regret to refer to the pseudo-regret $R(T)$.

[1] In Section 3.1, we show that our results also extend to constraints of the form $\langle x_t, \mu_\star \rangle \geq (1-\alpha) q_{b_t}$, where $\mu_\star$ is an additional unknown parameter. In this case, we assume the learner receives additional bandit feedback on the constraint after each round.

1.2 Previous work

Multi-armed Bandits. The multi-armed bandit (MAB) framework has been studied in sequential decision making problems under uncertainty. In particular, it captures the exploration-exploitation trade-off, where the learner needs to sequentially choose arms in order to maximize her reward over time while exploring to improve her estimate of the reward of each arm Bubeck and Eldan (2016). Two popular heuristics exist for MAB: Following the *optimism in face of uncertainty* (OFU) principle Auer et al. (2002); Li et al. (2017); Filippi et al. (2010), the so-called Upper Confidence Bound (UCB) based approaches choose the best feasible action- environment pair according to their current confidence regions on the unknown parameter, and Thompson Sampling (TS) Thompson (1933); Kaufmann et al. (2012); Russo and Van Roy (2016); Moradipari et al. (2018), which randomly samples the environment and plays the corresponding optimal action.

Linear Stochastic Bandits. There exists a rich literature on linear stochastic bandits. Two well-known efficient algorithms for LB are Linear UCB (LUCB) and Linear Thompson Sampling (LTS). For LUCB, Dani et al. (2008); Rusmevichientong and Tsitsiklis (2010); Abbasi-Yadkori et al. (2011) provided a regret guarantee of order $\mathcal{O}(\sqrt{T}\log T)$. For LTS, Agrawal and Goyal (2013); Abeille et al. (2017) provided a regret bound of order $\mathcal{O}(\sqrt{T}\log^{3/2} T)$ in a frequentist setting, i.e., when the unknown parameter $\theta_\star$ is a fixed parameter. We need to note that none of the aforementioned heuristics can be directly adopted in the conservative setting. However, note that the regret guarantee provided by our extensions of LUCB and LTS for the safe setting matches those stated for the original setting.

Conservativeness and Safety. The baseline model adopted in this paper was first proposed in Kazerouni et al. (2017); Wu et al. (2016) in the case of *cumulative constraints* on the reward. In Kazerouni et al. (2017); Wu et al. (2016), an action is considered feasible/safe at round t as long as it keeps the cumulative reward up to round t above a given fraction of a given baseline policy. This differs from our setting, which is focused on stage-wise constraints, where we want the expected reward of the *every single action* to exceed a given fraction of the baseline reward at each time t. This is a tighter constraint than that of Kazerouni et al. (2017); Wu et al. (2016). The setting considered in this paper was first studied in Khezeli and Bitar (2019), which proposed an algorithm called SEGE to guarantee the satisfaction of the safety constraint at each stage of the algorithm. While our paper is motivated by Khezeli and Bitar (2019), there are a few key differences: 1) We prove an upper bound of order $\mathcal{O}(\log T)$ for the number of times that the learning algorithm plays the conservative actions which is an order-wise improvement with respect to that of Khezeli and Bitar (2019), which shows an upper bound of order $\mathcal{O}(\sqrt{T})$; 2) In our setting, the action set is assumed to be a general convex and compact set in R^d. However, in Khezeli and Bitar (2019), the proof relies on the action set being a specific ellipsoid; 3) In Section 4, we provide a regret guarantee for the learning algorithm for the case where the baseline reward is unknown. However, the results of Khezeli and Bitar (2019) have not been extended to this case; 4) In Section 3.1, we also modify our proposed algorithm and provide a regret guarantee for the case where the constraint has a different unknown parameter than the one in the reward function. However, this is not discussed in Khezeli and Bitar (2019). Another difference between the two works is on the type of performance guarantees. In Khezeli and Bitar (2019), the authors bound the *expected* regret. Towards this goal, they manage to quantify the effect of the risk level δ on the regret and constraint satisfaction. However, it appears that the analysis in Khezeli and Bitar (2019) is limited to ellipsoidal action sets. Instead, in this paper, we present a bound on the regret that holds with high (constant) probability (parameterized by δ) over *all* T rounds of the algorithm. This type of results is very common in the bandit literature, e.g. Abbasi-Yadkori et al. (2011); Dani et al. (2008), and in the emerging safe-bandit literature Kazerouni et al. (2017); Amani et al. (2019); Sui et al. (2018).

Another variant of safety w.r.t a baseline policy has also been studied in Mansour et al. (2015); Katariya et al. (2018) in the multi-armed bandits framework. Moreover, there has been an increasing attention on studying the effect of safety constraints in the Gaussian process (GP) optimization literature. For example, Sui et al. (2015, 2018) study the problem of *nonlinear* bandit optimization with nonlinear constraints using GPs (as non-parametric models). The algorithms in Sui et al. (2015, 2018) come with convergence guarantees but no regret bound. Moreover, Ostafew et al. (2016); Akametalu et al. (2014) study safety-constrained optimization using GPs in robotics applications. A large body of work has considered safety in the context of model-predictive control, see, e.g., Aswani et al. (2013); Koller et al. (2018) and references therein. Focusing specifically on linear stochastic

bandits, extension of UCB-type algorithms to provide safety guarantees with provable regret bounds was considered recently in Amani et al. (2019). This work considers the effect of a linear constraint of the form $x^\top B\theta_\star \leq C$, where B and C are respectively a known matrix and positive constant, and provides a problem dependent regret bound for a safety-constrained version of LUCB that depends on the location of the optimal action in the safe action set. Notice that this setting requires the linear function $x^\top B\theta_\star$ to remain below a threshold C, as opposed to our setting which considers a lower bound on the reward. We note that the algorithm and proof technique in Amani et al. (2019) does not extend to our setting and would only work for inequalities of the given form; however, we discuss how our algorithm can be modified to provide a regret bound of order $\mathcal{O}(\sqrt{T}\log T)$ for the setting of Amani et al. (2019) in Appendix H. A TS variant of this setting has been studied in Moradipari et al. (2020); Moradipari et al. (2019).

1.3 Model Assumptions

Notation. The weighted ℓ_2-norm with respect to a positive semi-definite matrix V is denoted by $\|x\|_V = \sqrt{x^\top V x}$. The minimum of two numbers a, b is denoted $a \wedge b$. Let $\mathcal{F}_t = (\mathcal{F}_1, \sigma(x_1, \xi_1, \ldots, x_t, \xi_t))$ be the filtration (σ-algebra) that represents the information up to round t.

Assumption 1. *For all t, ξ_t is conditionally zero-mean R-sub-Gaussian noise variables, i.e., $\mathsf{E}[\xi_t|\mathcal{F}_{t\ \ 1}] = 0$, and $\mathsf{E}[e^{\lambda\xi_t}|\mathcal{F}_{t\ \ 1}] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right), \forall\lambda \in \mathsf{R}^d$.*

Assumption 2. *There exists a positive constant S such that $\|\theta_\star\|_2 \leq S$.*

Assumption 3. *The action set $\mathcal{X}$ is a compact and convex subset of R^d that contains the unit ball. We assume that $\|x\|_2 \leq L, \forall x \in \mathcal{X}$. Also, we assume $\langle x, \theta_\star\rangle \leq 1, \forall x \in \mathcal{X}$.*

Let $\kappa_{b_t} = \langle x_\star, \theta_\star\rangle - r_{b_t}$ be the difference between expected reward of the optimal and baseline actions at round t. As in Kazerouni et al. (2017), we assume the following.

Assumption 4. *There exist $0 \leq \kappa_l \leq \kappa_h$ and $0 < r_l \leq r_h$ such that, at each round t*

$$\kappa_l \leq \kappa_{b_t} \leq \kappa_h \text{ and } r_l \leq r_{b_t} \leq r_h. \tag{4}$$

We note that since these parameters are associated with the baseline policy, it can be reasonably assumed that they can be estimated accurately from data. This is because we think of the baseline policy as "past strategy", implemented before bandit-optimization, thus producing large amount of data. The lower bound $0 < r_l \leq r_{b_t}$ on the baseline reward ensures a minimum level of performance at each round. κ_h and r_h could be at most 1, due to Assumption 3. For simplicity, we assume the lower bound κ_l on the sub-optimality gap κ_{b_t} is known. If not, we can always choose $\kappa_l = 0$ by optimality of $x_\star$.

2 Stage-wise Conservative Linear Thompson Sampling (SCLTS) Algorithm

In this section we propose a TS variant algorithm in a frequentist setting referred to as *Stage-wise Conservative Linear Thompson Sampling* (SCLTS) for the problem setting in Section 1.1. Our adoption of TS is due to its well-known computational efficiency over UCB-based algorithms, since action selection via the latter involves solving optimization problems with bilinear objective functions, whereas the former would lead to linear objectives. However, this choice does not fundamentally affect our approach. In fact, in Appendix G, we propose a Stage-wise Conservative Linear UCB (SCLUCB) algorithm, and we provide the regret guarantee for it. In particular, we show a regret of order $\mathcal{O}\ \ d\sqrt{T}\log(\frac{TL^2}{\lambda\delta})$ for SCLUCB, which has the same order as the lower bound proposed for LB in Dani et al. (2008); Rusmevichientong and Tsitsiklis (2010).

At each round t, given a regularized least-square (RLS) estimate of $\hat{\theta}_t$, SCLTS samples a perturbed parameter $\tilde{\theta}_t$ with an appropriate distributional property. Then, it searches for the action that maximizes the expected reward considering the parameter $\tilde{\theta}_t$ as the true parameter while respecting the safety constraint (2). If any such action exists, it is played under certain conditions; else, the algorithm resorts to playing a perturbed version of the baseline action that satisfies the safety constraint. In order to guarantee constraint satisfaction (a.k.a safety of actions), the algorithm builds a confidence region $\mathcal{E}_t$ that contains the unknown parameter $\theta_\star$ with high probability. Then, it constructs an *estimated safe* set $\mathcal{X}_t^s$ such that all actions $x_t \in \mathcal{X}_t^s$ satisfy the safety constraint for all $v \in \mathcal{E}_t$. The summary of the SCLTS presented in Algorithm 1, and a detailed explanation follows.

Algorithm 1: Stage-wise Conservative Linear Thompson Sampling (SCLTS)

Input: $\delta, T, \lambda, \rho_1$
Set $\delta^0 = \frac{\delta}{4T}$
for $t = 1, \ldots, T$ **do**
  Sample $\eta_t \sim \mathcal{H}^{\text{TS}}$
  Compute RLS-estimate $\hat{\theta}_t$ and V_t according to (5)
  Set $\tilde{\theta}_t = \hat{\theta}_t + \beta_t V_t^{1/2} \eta_t$
  Build the confidence region $\mathcal{E}_t(\delta^0)$ in (6)
  Compute the estimated safe set $\mathcal{X}_t^s$ in (8)
  if the following optimization is feasible: $x(\tilde{\theta}_t) = \text{argmax}_{x \in \mathcal{X}_t^s} \langle x, \tilde{\theta}_t \rangle$, **then**
  Set $F = 1$, **else** $F = 0$
  if $F = 1$ **and** $\lambda_{\min}(V_t) \geq \left(\frac{2L\beta_t}{\kappa_l + \alpha r_{b_l}}\right)^2$, **then**
  Play $x_t = x(\tilde{\theta}_t)$
  else
  Play $x_t = (1 - \rho_1) x_{b_t} + \rho_1 \zeta_t$
  Observe reward y_t in (1)
end for

2.1 Algorithm description

Let $x_1, \ldots, x_t$ be the sequence of the actions and $r_1, \ldots, r_t$ be their corresponding rewards. For any $\lambda > 0$, we can obtain a regularized least-squares (RLS) estimate $\hat{\theta}_t$ of $\theta_\star$ as follows

$$\hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} y_s x_s, \text{ where } V_t = \lambda I + \sum_{s=1}^{t} x_s x_s^\top. \tag{5}$$

Algorithm 1 construct a confidence region

$$\mathcal{E}_t(\delta^0) = \mathcal{E}_t := \{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t\|_{V_t} \leq \beta_t(\delta^0)\}, \tag{6}$$

where the ellipsoid radius β_t is chosen according to the Proposition 2.1 in Abbasi-Yadkori et al. (2011) (restated below for completeness) in order to guarantee that $\theta_\star \in \mathcal{E}_t$ with high probability.

Proposition 2.1. *(Abbasi-Yadkori et al. (2011)) Let Assumptions 1, 2, and 3 hold. For a fixed* $\delta \in (0, 1)$*, and*

$$\beta_t(\delta) = R \sqrt{d \log\left(\frac{1 + \frac{tL^2}{\lambda}}{\delta}\right)} + \sqrt{\lambda} S \tag{7}$$

with probability at least $1 - \delta$*, it holds that* $\theta_\star \in \mathcal{E}_t$*.*

2.1.1 The estimated safe action set

Since $\theta_\star$ is unknown to the learner, she does not know whether an action $x \in \mathcal{X}$ is safe or not. Thus, she builds an estimated safe set such that each action $x_t \in \mathcal{X}_t^s$ satisfies the safety constraint for all $v \in \mathcal{E}_t$, i.e.,

$$\mathcal{X}_t^s := \{x \in \mathcal{X} : \langle x, v \rangle \geq (1 - \alpha) r_{b_t}, \forall v \in \mathcal{E}_t\} = \{x \in \mathcal{X} : \min_{v \in \mathcal{E}_t} \langle x, v \rangle \geq (1 - \alpha) r_{b_t}\} \tag{8}$$

$$= \{x \in \mathcal{X} : \langle x, \hat{\theta}_t \rangle - \beta_t(\delta^0) \|x\|_{V_t^{-1}} \geq (1 - \alpha) r_{b_t}\}. \tag{9}$$

Note that $\mathcal{X}_t^s$ is easy to compute since (9) involves a convex quadratic program. In order to guarantee safety, at each round t, the learner chooses her actions only from this estimated safe set in order to maximize the reward given the sampled parameter $\tilde{\theta}_t$, i.e.,

$$x(\tilde{\theta}_t) = \arg\max_{x \in \mathcal{X}_t^s} \langle x, \tilde{\theta}_t \rangle, \tag{10}$$

where $\tilde{\theta}_t = \hat{\theta}_t + \beta_t V_t^{1/2} \eta_t$, and η_t is a random IID sample from a distribution $\mathcal{H}^{\text{TS}}$ that satisfies certain distributional properties (see Abeille et al. (2017) or Defn. C.1 in Appendix C for more details). The challenge with $\mathcal{X}_t^s$ is that it contains actions which are safe with respect to all the parameters in $\mathcal{E}_t$, and not only $\theta_\star$. Hence, there may exist some rounds that $\mathcal{X}_t^s$ is empty. In order to face this problem, the algorithm proceed as follows. At round t, if the estimated action set $\mathcal{X}_t^s$ is not empty, SCLTS plays the safe action $x(\tilde{\theta}_t)$ in (10) only if the minimum eigenvalue of the Gram matrix V_t is greater than $k_t^1 = \left(\frac{2L\beta_t}{\kappa_l + \alpha r_{b_l}}\right)^2$, i.e., $\lambda_{\min}(V_t) \geq k_t^1$, where k_t^1 is of order $\mathcal{O}(\log t)$. Otherwise, it plays the conservative action which is presented next. We show in Appendix C that $\lambda_{\min}(V_t) \geq k_t^1$ ensures that for the rounds that SCLTS plays the action $x(\tilde{\theta}_t)$ in (10), the optimal action $x_\star$ belongs to the estimated safe set $\mathcal{X}_t^s$, from which we can bound the regret of Term I in (12).

2.1.2 Conservative actions

In our setting, we assume that the learner is given a baseline policy that at each round t suggests a baseline action x_{b_t}. We employ the idea proposed in Khezeli and Bitar (2019), which is merging the baseline actions with random exploration actions under stage-wise safety constraint. In particular, at each round t, SCLTS constructs a conservative action x_t^{cb} as a convex combination of the baseline action x_{b_t} and a random vector ζ_t as follows:

$$x_t^{\text{cb}} = (1 - \rho_1) x_{b_t} + \rho_1 \zeta_t, \tag{11}$$

where ζ_t is assumed to be a sequence of independent, zero-mean and bounded random vectors. Moreover, we assume that $\|\zeta_t\|_2 = 1$ almost surely and $\sigma_\zeta^2 = \lambda_{\min}(\text{Cov}(\zeta_t)) > 0$. The parameters σ_ζ and ρ_1 control the exploration level of the conservative actions. In order to ensure that the conservative actions are safe, in Lemma 2.2, we establish an upper bound on ρ_1 such that for all $\rho_1 \in (0, \bar{\rho})$, the conservative action $x_t^{\text{cb}} = (1 - \rho_1) x_{b_t} + \rho_1 \zeta_t$ is guaranteed to be safe.

Lemma 2.2. *At each round t, given the fraction α, for any $\rho \in (0, \bar{\rho})$, where $\bar{\rho} = \frac{\alpha r_l}{S + r_h}$, the conservative action $x_t^{cb} = (1 - \rho) x_{b_t} + \rho \zeta_t$ is guaranteed to be safe almost surely.*

For the ease of notation, in the rest of this paper, we simply assume that $\rho_1 = \frac{r_l}{S + r_h} \alpha$.

At round t, SCLTS plays the conservative action x_t^{cb} if the two conditions defined in Section 2.1.1 do not hold, i.e., either the estimated safe set $\mathcal{X}_t^s$ is empty or $\lambda_{\min}(V_t) < k_t^1$.

3 Regret Analysis

In this section, we provide a tight regret bound for SCLTS. In Proposition 3.1, we show that the regret of SCLTS can be decomposed into regret caused by choosing safe Thompson Sampling actions plus that of playing conservative actions. Then, we bound both terms separately. Let $N_{t\ 1}$ be the set of rounds $i < t$ at which SCLTS plays the action in (10). Similarly, $N_{t\ 1}^c = \{1, \ldots, t-1\} - N_{t\ 1}$ is the set of rounds $j < t$ at which SCLTS plays the conservative actions.

Proposition 3.1. *The regret of SCLTS can be decomposed into two terms as follows:*

$$R(T) \leq \underbrace{\sum_{t \in N_T} \left(\langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle\right)}_{\text{Term I}} + \underbrace{|N_T^c| \left(\kappa_h + \rho_1 (r_h + S)\right)}_{\text{Term II}} \tag{12}$$

The idea of bounding Term I is inspired by Abeille et al. (2017): we wish to show that LTS has a constant probability of being "optimistic", in spite of the need to be conservative. In Theorem 3.2, we provide an upper bound on the regret of Term I which is of order $\mathcal{O}(d^{3/2} \log^{1/2} d\, T^{1/2} \log^{3/2} T)$.

Theorem 3.2. *Let $\lambda, L \geq 1$. On event $\{\theta_\star \in \mathcal{E}_t, \forall t \in [T]\}$, and under Assumption 4, we can bound Term I in (12) as:*

$$\textit{Term I} \leq \left(\beta_T(\delta^0) + \gamma_T(\delta^0)\left(1 + \frac{4}{p}\right)\right) \sqrt{2Td \log\left(1 + \frac{TL^2}{\lambda}\right)} + \frac{4\gamma_T(\delta^0)}{p} \sqrt{\frac{8TL^2}{\lambda} \log \frac{4}{\delta}}, \tag{13}$$

where $\delta^0 = \frac{\delta}{6T}$, and $\gamma_t(\delta) = \beta_t(\delta^0) \left(1 + \frac{2}{C}\right) \sqrt{cd \log\left(\frac{c^0 d}{\delta}\right)}$

We note that the regret of Term I has the same bound as that of Abeille et al. (2017) in spite of the additional safety constraints imposed on the problem. As the next step, in order to bound Term II in (12), we need to find an upper bound on the number of times $|N_T^c|$ that SCLTS plays the conservative actions up to time T. We prove an upper bound on $|N_T^c|$ in Theorem 3.3.

Theorem 3.3. *Let $\lambda, L \geq 1$. On event $\{\theta_\star \in \mathcal{E}_t, \forall t \in [T]\}$, and under Assumption 4, it holds that*

$$|N_T^c| \leq \left(\frac{2L\beta_T}{\rho_1 \sigma_\zeta (\kappa_l + \alpha r_l)} + \frac{2h_1^2}{\rho_1^4 \sigma_\zeta^4} \log(\frac{d}{\delta}) + \frac{2Lh_1\beta_T}{\rho_1^3 \sigma_\zeta^3 (\kappa_l + \alpha r_l)} \sqrt{8 \log(\frac{d}{\delta})} \right)^2, \tag{14}$$

where $h_1 = 2\rho_1(1-\rho_1)L + 2\rho_1^2$ and $\rho_1 = (\frac{r_l}{S+r_h})\alpha$.

Remark 3.1. *The upper bound on the number of times SCLTS plays the conservative actions up to time T provided in Theorem 3.3 has the order $\mathcal{O}\left(\frac{L^2 d \log(\frac{T}{\delta}) \log(\frac{d}{\delta})}{\alpha^4 (r_l^2 \wedge r_l^4) \kappa_l (\sigma^2 \wedge \sigma^4)}\right)$.*

The first idea of the proof is based on the intuition that if a baseline action is played at round τ, then the algorithm does not yet have a good estimate of the unknown parameter $\theta_\star$ and the safe actions played thus far have not yet expanded properly in all directions. Formally, this translates to small $\lambda_{\min}(V_\tau)$ and the upper bound $O(\log \tau) \geq \lambda_{\min}(V_\tau)$. The second key idea is to exploit the randomized nature of the conservative actions (cf. (11)) to lower bound $\lambda_{\min}(V_\tau)$ by the number of times (N_τ^c) that SCLTS plays the baseline actions up to that round (cf. Lemma D.1 in the Appendix). Putting these together leads to the advertised upper bound $O(\log T)$ on the total number of times (N_T^c) the algorithm plays the baseline actions.

3.1 Additional Side Constraint with Bandit Feedback

We also consider the setting where the constraint depends on an unknown parameter that is different than the one in reward function. In particular, we assume the constraint of the form

$$\langle x_t, \mu_\star \rangle \geq (1-\alpha) q_{b_t}, \tag{15}$$

which needs to be satisfied by the action x_t at every round t. In (15), $\mu_\star$ is a fixed, but unknown and the positive constants $q_{b_t} = \langle x_{b_t}, \mu_\star \rangle$ are known to the learner. In Section 4, we relax this assumption and we consider the case where the learner does not know the value of q_{b_t}. Let $\nu_{b_t} = \langle x_\star, \mu_\star \rangle - \langle x_{b_t}, \mu_\star \rangle$. Similar to Assumption 4, we assume there exist constants $0 \leq \nu_l \leq \nu_h$ and $0 < q_l \leq q_h$ such that $\nu_l \leq \nu_{b_t} \leq \nu_h$ and $q_l \leq q_{b_t} \leq q_h$.

We assume that with playing an action x_t, the learner observes the following bandit feedback:

$$w_t = \langle x_t, \mu_\star \rangle + \chi_t, \tag{16}$$

where χ_t is assumed to be a zero-mean R-sub-Gaussian noise. In order to handle this case, we show how SCLTS should be modified, and we propose a new algorithm called SCLTS-BF. The details on SCLTS-BF are presented in Appendix E. In the following, we only mention the difference of SCLTS-BF with SCLTS, and show an upper bound on its regret.

The main difference is that SCLTS-BF constructs two confidence regions $\mathcal{E}_t$ in (6) and $\mathcal{C}_t$ based on the bandit feedback such that $\theta_\star \in \mathcal{E}_t$ and $\mu_\star \in \mathcal{C}_t$ with high probability. Then, based on $\mathcal{C}_t$, it constructs the estimated safe decision set denoted $\mathcal{P}_t^s = \{x \in \mathcal{X} : \langle x, v \rangle \geq (1-\alpha) q_{b_t}, \forall v \in \mathcal{C}_t\}$. We note that SCLTS-BF only plays the actions from $\mathcal{P}_t^s$ that are safe with respect to all the parameters in $\mathcal{C}_t$.

We report the details on proving the regret bound for SCLTS-BF in Appendix E. We use the decomposition in Proposition 3.1, and we upper bound Term I similar to the Theorem 3.2. Then, we show an upper bound of order $\mathcal{O}\left(\frac{L^2 d \log(\frac{T}{\delta}) \log(\frac{d}{\delta})}{\alpha^4 (q_l^2 \wedge q_l^4) \nu_l (\sigma^2 \wedge \sigma^4)}\right)$ over the number of times that SCLTS-BF plays the conservative actions.

4 Unknown Baseline Reward

Inspired by Kazerouni et al. (2017), which studies this problem in the presence of *safety constraints on the cumulative rewards*, we consider the case where the expected reward of the action chosen by baseline policy, i.e., r_{b_t} is unknown to the learner. However, we assume that the learner knows the

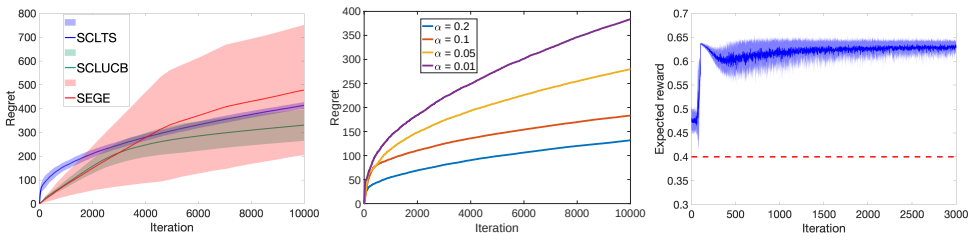

Figure 1: Left: comparison of the cumulative regret of SCLTS and SCLUCB versus SEGE algorithm in Khezeli and Bitar (2019). Middle: average regret (over 100 runs) of SCLTS algorithm for different values of α. Right: expected reward under SCLTS algorithm in the first 3000 rounds for $\alpha = 0.2$.

value of r_l in (4). We describe the required modifications on SCLTS to handle this case, and present a new algorithm called SCLTS2. Then, we prove the regret bound for SCLTS2, which has the same order as SCLTS.

Here, the learner does not know the value of r_{b_t}; however, she knows that the unknown parameter $\theta_\star$ falls in the confidence region $\mathcal{E}_t$ with high probability. Hence, we can upper bound the RHS of (2) with $\max_{v \in \mathcal{E}_t} \langle x_{b_t}, v \rangle \geq r_{b_t}$. Therefore, any action x that satisfies

$$\min_{v \in \mathcal{E}_t} \langle x(\tilde{\theta}_t), v \rangle \geq (1-\alpha) \max_{v \in \mathcal{E}_t} \langle x_{b_t}, v \rangle, \tag{17}$$

is safe with high probability. In order to ensure safety, SCLTS2 only plays the safe actions from the estimated safe actions set $\mathcal{Z}_t^s = \{x \in \mathcal{X} : \min_{v \in \mathcal{E}_t} \langle x, v \rangle \geq (1-\alpha) \max_{v \in \mathcal{E}_t} \langle x_{b_t}, v \rangle\}$. We report the details on SCLTS2 in Appendix F.

Next, we provide an upper bound on the regret of SCLTS2. To do so, we first use the decomposition in Proposition 3.1. The regret of Term I is similar to that of SCLTS (Theorem 3.2), and in Theorem 4.1, we prove an upper bound on the number of time SCLTS2 plays the conservative actions. Note that similar steps can be generalized to the setting of additional side constraints with bandit feedback.

Theorem 4.1. *Let $\lambda, L \geq 1$. On event $\{\theta_\star \in \mathcal{E}_t, \forall t \in [T]\}$, and under Assumption 4, we can upper bound the number of times SCLTS2 plays the conservative actions, i.e., $|N_T^c|$ as:*

$$|N_T^c| \leq \left(\frac{2L\beta_T(2-\alpha)}{\rho_3 \sigma_\zeta (\kappa_l + \alpha r_l)} + \frac{2h_3^2}{\rho_3^4 \sigma_\zeta^4} \log(\frac{d}{\delta}) + \frac{2Lh_3\beta_T(2-\alpha)}{\rho_3^3 \sigma_\zeta^3 (\kappa_l + \alpha r_l)} \sqrt{8 \log(\frac{d}{\delta})} \right)^2, \tag{18}$$

where $h_3 = 2\rho_3(1-\rho_3)L + 2\rho_3^2$ and $\rho_3 = (\frac{r_l}{S+1})\alpha$.

Remark 4.1. *The regret of SCLTS2 has order of $\mathcal{O}\left(\frac{L^2 d \log(\frac{T}{\delta}) \log(\frac{d}{\delta})(2-\alpha)^2}{\alpha^4 (r_l^2 \wedge r_l^4) \kappa_l (\sigma^2 \wedge \sigma^4)} \right)$, which has the same rate as that of SCLTS. Therefore, the lack of information about the reward function only hurt the regret with a constant $(2-\alpha)^2$.*

5 Numerical Results

In this section, we investigate the numerical performance of SCLTS and SCLUCB on synthetic data, and compare it with SEGE algorithm introduced by Khezeli and Bitar (2019). In all the implementations, we used the following parameters: $R = 0.1, S = 1, \lambda = 1, d = 2$. We consider the action set $\mathcal{X}$ to be a unit ball centered on the origin. The reward parameter $\theta_\star$ is drawn from $\mathcal{N}(0, I_4)$. We generate the sequence $\{\zeta_t\}_{t=1}^1$ to be IID random vectors that are uniformly distributed on the unit circle. The results are averaged over 100 realizations.

In Figure 1(left), we plot the cumulative regret of the SCLTS algorithm and SCLUCB and SEGE algorithm from Khezeli and Bitar (2019) for $\alpha = 0.2$ over 100 realizations. The shaded regions show standard deviation around the mean. In view of the discussion in Dani et al. (2008) regarding computational issues of LUCB algorithms with confidence regions specified with ℓ_2-norms, we implement a modified version of Safe-LUCB which uses ℓ_1-norms instead of ℓ_2-norms. Figure 1(left) shows that SEGE algorithm suffers a high variance of the regret over different problem instances which shows the strong dependency of the performance of SEGE algorithm on the specific problem instance. However, the regret of SCLTS and SCLUCB algorithms do not vary significantly under

different problem instances, and has a low variance. Moreover, the regret of SEGE algorithm grows faster in the beginning steps, since it heavily relies on the baseline action in order to satisfy the safety constraint. However, the randomized nature of SCLTS leads to a natural exploration ability that is much faster in expanding the estimated safe set, and hence it plays the baseline actions less frequently than SEGE algorithm even in the initial exploration stages.

In Figure 1(middle), we plot the average regret of SCLTS for different values of α over a horizon $T = 10000$. Figure 1(middle) shows that, SCLTS has a better performance (i.e., smaller regret) for the larger value of α, since for the small value of α, SCLTS needs to be more conservative in order to satisfy the safety constraint, and hence it plays more baseline actions. Moreover, Figure 1(right) illustrates the expected reward of SCLTS algorithm in the first 3000 rounds. In this setting, we assume there exists one baseline action $x_b = [0.6, 0.5]$, which is available to the learner, $\theta_\star = [0.5, 0.4]$ and the safety fraction $\alpha = 0.2$. Thus, the safety threshold is $(1 - \alpha)\langle x_b, \theta_\star \rangle = 0.4$ (shown as a dashed red line), which SCLTS respects in all rounds. In particular, in initial rounds, SCTLS plays the conservative actions in order to respect the safety constraint, which as shown have an expected reward close to 0.475. Over time as the algorithm achieves a better estimate of the unknown parameter $\theta_\star$, it is able to play more optimistic actions and as such receives higher rewards.

6 Conclusion

In this paper, we study the stage-wise conservative linear stochastic bandit problem. Specifically, we consider safety constraints that requires the action chosen by the learner at each individual stage to have an expected reward higher than a predefined fraction of the reward of a given baseline policy. We propose extensions of Linear Thompson Sampling and Linear UCB in order to minimize the regret of the learner while respecting safety constraint with high probability and provide regret guarantees for them. We also consider the setting of constraints with bandit feedback, where the safety constraint has a different unknown parameter than that of the reward function, and we propose the SCLTS-BF algorithm to handle this case. Third, we study the case where the rewards of the baseline actions are unknown to the learner. Lastly, our numerical experiments compare the performance of our algorithm to SEGE of Khezeli and Bitar (2019) and showcase the value of the randomized nature of our exploration phase. In particular, we show that the randomized nature of SCLTS leads to a natural exploration ability that is faster in expanding the estimated safe set, and hence SCLTS plays the baseline actions less frequently as theoretically shown. For future work, natural extension of the problem setting to generalized linear bandits, and possibly with generalized linear constrains might be of interest.

7 Acknowledgment

This research is supported by NSF grant 1847096. C. Thrampoulidis was partially supported by the NSF under Grant Number 1934641.

8 Broader Impact

The main goal of this paper is to design and study novel "safe" learning algorithms for safety-critical systems with provable performance guarantees. An example arises in clinical trials where the effect of different therapies on patient's health is not known in advance. We select the baseline actions to be the therapies that have been historically chosen by medical practitioners, and the reward captures the effectiveness of the chosen therapy. The stage-wise conservative constraint modeled in this paper ensures that at each round the learner should choose a therapy which results in an expected reward if not better, must be close to the baseline policy. Another example arises in societal-scale infrastructure networks such as communication/power/transportation/data network infrastructure. We focus on the case where the reliability requirements of network operation at each round depends on the reward of the selected action and certain *baseline* actions are known to not violate system constraints and achieve certain levels of operational efficiency as they have been used widely in the past. In this case, the stage-wise conservative constraint modeled in this paper ensures that at each round, the reward of action employed by learning algorithm if not better, should be close to that of baseline policy in terms of network efficiency, and the reliability requirement for network operation must not be violated by the learner. Another example is in recommender systems that at each round, we wish to avoid recommendations that are extremely disliked by the users. Our proposed stage-wise conservative constrains ensures that at no round would the recommendation system cause severe dissatisfaction for the users (consider perhaps how a really bad personal movie recommendation from a streaming platform would severely affect your view of the said platform).

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.

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

Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135.

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.

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

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.

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

Bubeck, S. and Eldan, R. (2016). Multi-scale exploration of convex functions and bandit convex optimization. In *Conference on Learning Theory*, pages 583–589.

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.

Katariya, S., Kveton, B., Wen, Z., and Potluru, V. K. (2018). Conservative exploration using interleaving. *arXiv preprint arXiv:1806.00892*.

Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pages 199–213. Springer.

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.

Khezeli, K. and Bitar, E. (2019). Safe linear stochastic bandits. *arXiv preprint arXiv:1911.09501*.

Koller, T., Berkenkamp, F., Turchetta, M., and Krause, A. (2018). Learning-based model predictive control for safe exploration. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 6059–6066. IEEE.

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.

Mansour, Y., Slivkins, A., and Syrgkanis, V. (2015). Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 565–582.

Moradipari, A., Alizadeh, M., and Thrampoulidis, C. (2020). Linear thompson sampling under unknown linear constraints. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3392–3396.

Moradipari, A., Amani, S., Alizadeh, M., and Thrampoulidis, C. (2019). Safe linear thompson sampling with side information. *arXiv*, pages arXiv–1911.

Moradipari, A., Silva, C., and Alizadeh, M. (2018). Learning to dynamically price electricity demand based on multi-armed bandits. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 917–921. IEEE.

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. (2016). An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471.

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.

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

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434.

Tucker, N., Moradipari, A., and Alizadeh, M. (2020). Constrained thompson sampling for real-time electricity pricing with grid reliability constraints. *IEEE Transactions on Smart Grid*, pages 1–1.

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

A Proof of Proposition 3.1

From the definition of regret, we can write

$$\begin{aligned} R(T) &= \sum_{t \in N_T} (\langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle) + \sum_{t \in N_T^c} (\langle x_\star, \theta_\star \rangle - \langle (1-\rho_1) x_{b_t} - \rho_1 \zeta_t, \theta_\star \rangle) \\ &= \sum_{t \in N_T} (\langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle) + \sum_{t \in N_T^c} \langle x_\star, \theta_\star \rangle - \langle x_{b_t}, \theta_\star \rangle + \rho_1 \langle x_{b_t}, \theta_\star \rangle + \rho_1 \langle \zeta_t, \theta_\star \rangle \\ &\leq \sum_{t \in N_T} (\langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle) + |N_T^c| \left(\kappa_h + \rho_1 (r_h + S) \right). \end{aligned} \tag{19}$$

B Proof of Lemma 2.2

In order to ensure that the conservative action $x_t = (1-\rho) x_{b_t} + \rho \zeta_t$ is safe, we need to show that it satisfies (2). Hence, it suffices to show that

$$\langle (1-\rho) x_{b_t} + \rho \zeta_t, \theta_\star \rangle \geq (1-\alpha) r_{b_t}. \tag{20}$$

We can lower bound the LHS of (20) as follows:

$$\langle (1-\rho) x_{b_t} + \rho \zeta_t, \theta_\star \rangle = r_{b_t} - \rho r_{b_t} + \rho \langle \zeta_t, \theta_\star \rangle \geq r_{b_t} - \rho r_{b_t} - \rho S.$$

Recall that $\|\zeta_t\|_2 = 1$ almost surely, and due to Assumption 2, we know that $\|\theta_\star\|_2 \leq S$. Hence, it suffices to show that

$$r_{b_t} - \rho r_{b_t} - \rho S \geq (1-\alpha) r_{b_t},$$

or equivalently,

$$\rho r_{b_t} + \rho S \leq \alpha r_{b_t} \tag{21}$$

From (21) we can write

$$\rho \leq \frac{\alpha r_{b_t}}{S + r_{b_t}}. \tag{22}$$

Therefore, for any ρ satisfying (22), the conservative action $x_t = (1-\rho) x_{b_t} + \rho \zeta_t$ is guaranteed to be safe almost surely. Then, we lower bound the right hand side of (22) using Assumption 4, and we establish the following upper bound on ρ,

$$\rho \leq \frac{\alpha r_l}{S + r_h}. \tag{23}$$

Therefore, for any $\rho \in (0, \bar{\rho})$, where $\bar{\rho} = \frac{\alpha r_l}{S + r_h}$, the conservative actions are safe.

C Proof of Theorem 3.2

In this section, we provide an upper bound on the regret of Term I in (12). We first rewrite Term I as follows:

$$\sum_{t \in N_T} (\langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle) \tag{24}$$

Clearly, it would be beneficial to show that (24) is non-positive. However, as stated in Abeille et al. (2017) (in the case of linear TS applied to the standard stochastic linear bandit problem with no safety constraints), this cannot be the case in general. Instead, to bound regret in the unconstrained case, Abeille et al. (2017) argues that it suffices to show that (24) is non-positive with a constant probability. But what happens in the safety-constrained scenario? It turns out that once the above stated event happens with constant probability (in our case, in the presence of safety constraints), the rest of the argument by Abeille et al. (2017) remains unaltered. Therefore, our main contribution in the proof of Theorem 3.2 is to show that (24) is non-positive with a constant probability in spite of the limitations on actions imposed because of the safety constraints. To do so, let

$$\Theta_t^{\text{opt}} = \{\theta \in \mathbb{R}^d : \langle x(\theta), \theta \rangle \geq \langle x_\star, \theta_\star \rangle\}, \tag{25}$$

be the so-called *set of optimistic parameters*, where $x(\tilde{\theta}_t) = \arg\max_{x \in \mathcal{X}_t^s} \langle x, \tilde{\theta}_t \rangle$ is the optimal safe action for the sampled parameter $\tilde{\theta}_t$ chosen from the estimated safe action set $\mathcal{X}_t^s$. LTS is considered optimistic at round t, if it samples the parameter $\tilde{\theta}_t$ from the set of optimistic parameters Θ_t^{opt} and plays the action $x(\tilde{\theta}_t)$. In Lemma C.1, we show that SCLTS is optimistic with constant probability despite the safety constraints. Before that, let us restate the distributional properties put forth in Abeille et al. (2017) for the noise $\eta \sim \mathcal{H}^{\text{TS}}$ that are required to ensure the right balance of exploration and exploitation.

Definition C.1. (Definition 1. in Abeille et al. (2017)) $\mathcal{H}^{\text{TS}}$ is a multivariate distribution on $\mathbb{R}^d$ absolutely continuous with respect to the Lebesgue measure which satisfies the following properties:

- (anti-concentration) there exists a strictly positive probability p such that for any $u \in \mathbb{R}^d$ with $\|u\|_2 = 1$,

$$\mathbb{P}_{\eta \sim \mathcal{H}^{\text{TS}}} \left(\langle u, \eta \rangle \geq 1 \right) \geq p. \tag{26}$$

- (concentration) there exists positive constants c, c^0 such that $\forall \delta \in (0,1)$

$$\mathbb{P}_{\eta \sim \mathcal{H}^{\text{TS}}} \left(\|\eta\| \leq \sqrt{cd \log(\frac{c^0 d}{\delta})} \right) \geq 1 - \delta. \tag{27}$$

Lemma C.1. *Let $\Theta_t^{opt} = \{\theta \in \mathbb{R}^d : \langle x(\theta), \theta \rangle \geq \langle x_\star, \theta_\star \rangle\}$ be the set of the optimistic parameters. For round $t \in N_T$, SCLTS samples the optimistic parameter $\tilde{\theta}_t \in \Theta_t^{opt}$ and plays the corresponding safe action $x(\tilde{\theta}_t)$ frequently enough, i.e.,*

$$\mathbb{P}(\tilde{\theta}_t \in \Theta_t^{opt}) \geq p. \tag{28}$$

Proof. We need to show that for rounds $t \in N_T$

$$\mathbb{P} \left(\langle x(\tilde{\theta}_t), \tilde{\theta}_t \rangle \geq \langle x_\star, \theta_\star \rangle \right) \geq p. \tag{29}$$

First, we show that for rounds $t \in N_T$, $x_\star$ falls in the estimated safe set, i.e., $x_\star \in \mathcal{X}_t^s$. To do so, we need to show that

$$\langle x_\star, \hat{\theta}_t \rangle - \beta_t \|x_\star\|_{V_t^{-1}} \geq (1 - \alpha) r_{b_t}, \tag{30}$$

using $\|\theta_\star - \hat{\theta}_t\|_{V_t} \leq \beta_t$, it suffices that

$$\langle x_\star, \theta_\star \rangle - 2\beta_t \|x_\star\|_{V_t^{-1}} \geq (1 - \alpha) r_{b_t}. \tag{31}$$

But we know that $\|x_\star\|_{V_t^{-1}} \leq \frac{\|x_\star\|_2}{\sqrt{\lambda_{\min}(V_t)}} \leq \frac{L}{\sqrt{\lambda_{\min}(V_t)}}$, where we also used Assumption 3 to bound $\|x_\star\|_2$. Hence, we can get

$$\langle x_\star, \theta_\star \rangle - 2\beta_t \|x_\star\|_{V_t^{-1}} \geq \langle x_\star, \theta_\star \rangle - \frac{2\beta_t L}{\sqrt{\lambda_{\min}(V_t)}}. \tag{32}$$

By substituting (32) in (31), it suffices to show that

$$\kappa_{b_t} + \alpha r_{b_t} \geq \frac{2\beta_t L}{\sqrt{\lambda_{\min}(V_t)}}, \tag{33}$$

or equivalently,

$$\lambda_{\min}(V_t) \geq \left(\frac{2L\beta_t}{\kappa_t + \alpha r_{b_t}} \right)^2. \tag{34}$$

To show (34), simply recall that $\lambda_{\min}(V_t) \geq k_t^1$, where $k_t^1 = \left(\frac{2L\beta_t}{\kappa_l + \alpha r_l} \right)^2$. Therefore, $x_\star \in \mathcal{X}_t^s$ for $t \in N_T$. Note that we are not interested in expanding the safe set in all possible directions. Instead, what aligns with the objective of minimizing regret, is expanding the safe set in the "correct" direction, that of $x_\star$. Therefore, $\lambda_{\min}(V_t) \geq \mathcal{O}(\log t)$ provides enough expansion of the safe set to bound the Term I in (12).

The rest of the proof is similar as in (Abeille et al., 2017, Lemma 3); we include in here for completeness.

For rounds $t \in N_T$, we know that

$$\langle x(\tilde{\theta}_t), \tilde{\theta}_t \rangle \geq \langle x_\star, \tilde{\theta}_t \rangle,$$

since $x(\tilde{\theta}_t) = \arg\max_{x \in \mathcal{X}_t^s} \langle x, \tilde{\theta}_t \rangle$ and we have already shown that $x_\star \in \mathcal{X}_t^s$. Therefore, it suffices to show that

$$\mathsf{P}\left\{ \langle x_\star, \tilde{\theta}_t \rangle \geq \langle x_\star, \theta_\star \rangle \right\} \geq p. \tag{35}$$

From the definition of $\tilde{\theta}_t$, we can rewrite (35) as

$$\mathsf{P}\left\{ \langle x_\star, \hat{\theta}_t \rangle + \beta_t \langle x_\star, V_t^{-1/2} \eta_t \rangle \geq \langle x_\star, \theta_\star \rangle \right\} \geq p,$$

or equivalently,

$$\mathsf{P}\left\{ \beta_t \langle x_\star, V_t^{-1/2} \eta_t \rangle \geq \langle x_\star, \theta_\star - \hat{\theta}_t \rangle \right\} \geq p. \tag{36}$$

Then, we use Cauchy-Schwarz for the LHS of (36), and given the fact that $\|\theta_\star - \hat{\theta}_t\|_{V_t} \leq \beta_t$, we get

$$\mathsf{P}\left\{ \langle x_\star, V_t^{-1/2} \eta_t \rangle \geq \|x_\star\|_{V_t^{-1}} \right\} \geq p,$$

or equivalently,

$$\mathsf{P}\left(\langle u_t, \eta_t \rangle \geq 1 \right) \geq p, \tag{37}$$

where $u_t = \frac{x_\star V_t^{-1}}{\|x_\star\|_{V_t^{-1}}}$. Therefore, $\|u_t\|_2 = 1$ by construction. At last, we know that (37) is true thanks to the anti-concentration distributional property of the parameter η_t in Definition C.1. □

As mentioned, after showing that SCLTS for rounds $t \in N_T$ samples from the set of optimistic parameters with a constant probability, the rest of the proof for bounding the regret of Term I is similar to that of Abeille et al. (2017). In particular, we conclude with the following bound

$$\text{Term I} := \sum_{t \in N_T} \left(\langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle \right) \\ \left(\beta_T(\delta^0) + \gamma_T(\delta^0)(1 + \frac{4}{p}) \right) \sqrt{2|N_T| d \log\left(1 + \frac{|N_T| L^2}{\lambda}\right) + \frac{4\gamma_T(\delta^0)}{p}} \sqrt{\frac{8|N_T| L^2}{\lambda} \log \frac{4}{\delta}}, \tag{38}$$

where $\delta^0 = \frac{\delta}{6jN_T j}$, and,

$$\gamma_t(\delta) = \beta_t(\delta^0) \left(1 + \frac{2}{C}\right) \sqrt{cd \log\left(\frac{c^0 d}{\delta}\right)}, \tag{39}$$

and since $N_T \leq T$, the proof is completed.

D Proof of Theorem 3.3

In this section, we prove an upper bound of order $\mathcal{O}(\log T)$ on the number of times that SCLTS plays the conservative actions.

Let τ be any round that the algorithm plays the conservative action, i.e., at round τ, either $F = 0$ or $\lambda_{\min}(V_\tau) < k_\tau^1 = \left(\frac{2L\beta}{\kappa + \alpha r_b}\right)^2$. By definition, if $F = 0$, we have

$$\nexists x \in \mathcal{X} : \langle x, \hat{\theta}_\tau \rangle - \beta_\tau \|x\|_{V^{-1}} \geq (1 - \alpha) r_b \ , \tag{40}$$

and since we know that $x_\star \in \mathcal{X}$, and $\theta_\star \in \mathcal{E}_t$ with high probability, we can write

$$\langle x_\star, \theta_\star \rangle - 2\beta_\tau \|x_\star\|_{V^{-1}} \leq \langle x_\star, \hat{\theta}_\tau \rangle - \beta_\tau \|x_\star\|_{V^{-1}} < (1 - \alpha) r_b \ . \tag{41}$$

From (41), we can get

$$\kappa_b \ + \alpha r_b \ < 2\beta_\tau \|x_\star\|_{V_\tau} \leq \frac{2\beta_\tau L}{\lambda_{\min}(V_\tau)}, \tag{42}$$

and hence the following upper bound on minimum eigenvalue of the Gram matrix:

$$\lambda_{\min}(V_\tau) < \left(\frac{2\beta_\tau L}{\kappa_b \ + \alpha r_b} \right)^2 \leq k_\tau^1. \tag{43}$$

Therefore, at any round τ that a conservative action is played, whether it is because $F = 0$, or because we have $\{\lambda_{\min}(V_\tau) < k_\tau\}$, we can always conclude that

$$\lambda_{\min}(V_\tau) < k_\tau^1. \tag{44}$$

The remainder of the proof builds on two auxiliary lemmas. First, in Lemma D.1, we show that the minimum eigenvalue of the Gram matrix V_t is lower bounded with the number of times SCLTS plays the conservative actions.

Lemma D.1. *Under Assumptions 1, 2, and 3, it holds that*

$$\mathsf{P}(\lambda_{min}(V_t) \leq t) \leq d \exp\left(-\frac{(\rho_1^2 |N_t^c| \sigma_\zeta^2 - t)^2}{8|N_t^c| h_1^2} \right), \tag{45}$$

where $h_1 = 2\rho_1(1-\rho_1)L + 2\rho_1^2$ *and* $\rho_1 = (\frac{r_l}{S+r_h})\alpha$.

Using (44) and applying Lemma D.1, it can be checked that with probability $1-\delta$,

$$\left(\frac{2L\beta_\tau}{\kappa_l + \alpha r_l} \right)^2 > \rho_1^2 |N_\tau^c| \sigma_\zeta^2 - \sqrt{8|N_\tau^c| h_1^2 \log(\frac{d}{\delta})}. \tag{46}$$

This gives an explicit inequality that must be satisfied by τ. Solving with respect to τ leads to the desired. In particular, we apply simple Lemma D.2 below.

Lemma D.2. *For any $a, b, c > 0$, if $ax - \sqrt{bx} < c$, then the following holds for $x \geq 0$*

$$0 \leq x < \frac{2ac + b + \sqrt{b^2 + 4abc}}{2a^2}. \tag{47}$$

Using Lemma D.2 results in the following upper bound on the $|N_\tau^c|$

$$|N_\tau^c| \leq \left(\frac{2L\beta_\tau}{\rho_1 \sigma_\zeta (\kappa_l + \alpha r_l)} \right)^2 + \frac{2h_1^2}{\rho_1^4 \sigma_\zeta^4} \log(\frac{d}{\delta}) + \frac{h_1 2L\beta_\tau}{(\kappa_l + \alpha r_l)\rho_1^3 \sigma_\zeta^3} \sqrt{8 \log(\frac{d}{\delta})}. \tag{48}$$

Therefore, we can upper bound N_T^c with the following:

$$|N_T^c| \leq \left(\frac{2L\beta_T}{\rho_1 \sigma_\zeta (\kappa_l + \alpha r_l)} \right)^2 + \frac{2h_1^2}{\rho_1^4 \sigma_\zeta^4} \log(\frac{d}{\delta}) + \frac{2Lh_1\beta_T}{\rho_1^3 \sigma^3 (\kappa_l + \alpha r_l)} \sqrt{8 \log(\tfrac{d}{\delta})}, \tag{49}$$

which has order $\mathcal{O}\left(\frac{L^2 d \log(T)}{\alpha^2 r_l^2 (\kappa_l + \alpha r_l)^2 \sigma^2} + \frac{L^2}{\alpha^2 r_l^2 \sigma^4} + d^2 \log(\frac{d}{\delta}) \right)$, as promised.

D.1 Proof of Lemma D.1

Our objective is to establish a lower bound on $\lambda_{\min}(V_t)$ for all t. It holds that

$$\begin{aligned}
V_t &= \lambda I + \sum_{s=1}^{t} x_s x_s^\top \\
&\succeq \sum_{s \in N_t^c} \left((1-\rho_1)x_{b_s} - \rho_1 \zeta_s\right)\left((1-\rho_1)x_{b_s} - \rho_1 \zeta_s\right)^\top \\
&= \sum_{s \in N_t^c} (1-\rho_1)^2 x_{b_s} x_{b_s}^\top - \rho_1(1-\rho_1) x_{b_s}\zeta_s^\top - \rho_1(1-\rho_1)\zeta_s x_{b_s}^\top + \rho_1^2 \zeta_s \zeta_s^\top \\
&\succeq \sum_{s \in N_t^c} - \rho_1(1-\rho_1) x_{b_s}\zeta_s^\top - \rho_1(1-\rho_1)\zeta_s x_{b_s}^\top + \rho_1^2 \zeta_s \zeta_s^\top \\
&= \sum_{s \in N_t^c} \rho_1^2 \mathbb{E}[\zeta_s \zeta_s^\top] - \rho_1(1-\rho_1) x_{b_s}\zeta_s^\top - \rho_1(1-\rho_1)\zeta_s x_{b_s}^\top + \rho_1^2 \zeta_s \zeta_s^\top - \rho_1^2 \mathbb{E}[\zeta_s \zeta_s^\top] \\
&\succeq \rho_1^2 \sigma_\zeta^2 |N_t^c| I + \sum_{s \in N_t^c} U_s, \qquad (50)
\end{aligned}$$

where U_s is defined as

$$U_s = - \rho_1(1-\rho_1) x_{b_s}\zeta_s^\top - \rho_1(1-\rho_1)\zeta_s x_{b_s}^\top + \rho_1^2 \zeta_s \zeta_s^\top - \rho_1^2 \mathbb{E}[\zeta_s \zeta_s^\top] \quad . \qquad (51)$$

Then, using Weyl's inequality, it follows that

$$\lambda_{\min}(V_t) \geq \rho_1^2 \sigma_\zeta^2 |N_t^c| - \lambda_{\max}\Big(\sum_{s \in N_t^c} U_s\Big).$$

Next, we apply the matrix Azuma inequality (see Theorem D.3) to find an upper bound on $\lambda_{\max}\big(\sum_{s \in N_t^c} U_s\big)$. For this, we first need to show that the sequence of matrices U_s satisfies the conditions of Theorem D.3. By definition of U_s in (51), it follows that $\mathbb{E}[U_s | \mathcal{F}_{s-1}] = 0$, and $U_s^\top = U_s$. Also, we construct the sequence of deterministic matrices A_s such that $U_s^2 \preceq A_s^2$ as follows. We know that for any matrix B, $B^2 \leq \|B\|_2^2 I$, where $\|B\|_2$ is the maximum singular value of B, i.e.,

$$\sigma_{\max}(B) = \max_{\|u\|_1 = \|v\|_2 = 1} u^\top B v.$$

Thus, we first show the following bound on the maximum singular value of the matrix U_s defined in (51):

$$\begin{aligned}
\max_{\|u\|_1 = \|v\|_2 = 1} u^\top U_s v &= -\rho_1(1-\rho_1)(u^\top x_{b_s})(v^\top \zeta_s)^\top - \rho_1(1-\rho_1)(u^\top \zeta_s)(v^\top x_{b_s})^\top + \\
&\qquad \rho_1^2 (u^\top \zeta_s)(v^\top \zeta_s)^\top - \rho_1^2 \mathbb{E}\ (u^\top \zeta_s)(v^\top \zeta_s)^\top \\
&\leq \rho_1(1-\rho_1)\|x_{b_s}\|_2 \|\zeta_s\|_2 + \rho_1(1-\rho_1)\|\zeta_s\|_2 \|x_{b_s}\|_2 + \rho_1^2 \|\zeta_s\|_2^2 + \rho_1^2 \mathbb{E}\ \|\zeta_s\|_2^2 \\
&\leq 2\rho_1(1-\rho_1)L + 2\rho_1^2, \qquad (52)
\end{aligned}$$

where we have used Cauchy-Schwarz inequality and the last inequality comes from the fact that $\|\zeta_s\|_2 = 1$ almost surely, and $\|x_{b_s}\|_2 \leq L$ by Assumption 3. From the derivations above, and choosing $A_s = h_1 I$, with $h_1 = 2\rho_1(1-\rho_1)L + 2\rho_1^2$, it almost surely holds that $U_s^2 \preceq \sigma_{\max}(U_s)^2 I \preceq h_1^2 I = A_s^2$. Moreover, using triangular inequality, it holds that

$$\Big\| \sum_{s \in N_t^c} A_s^2 \Big\| \leq \sum_{s \in N_t^c} \|A_s^2\| \leq |N_t^c| h_1^2.$$

Now we apply the the matrix Azuma inequality, to conclude that for any $c \geq 0$,

$$\mathbb{P}\Big[\lambda_{\max}\Big(\sum_{s \in N_t^c} U_s\Big) \geq c\Big] \leq d \exp\left(-\frac{c^2}{8|N_t^c| h_1^2}\right) \quad .$$

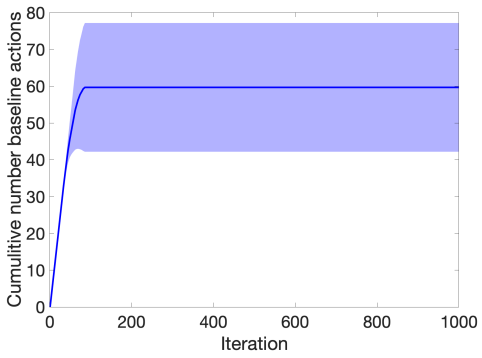

Figure 2: Cumulative number of times that the baseline actions played by SCLTS up to time t, for $t = 1 \dots, 1000$ over 100 realizations.

Therefore, it holds that with probability $1-\delta$, $\lambda_{\max}(\sum_{s2N_t^c} U_s) \leq \sqrt{8|N_t^c|h_1^2 \log(\frac{d}{\delta})}$, and hence with probability $1-\delta$,

$$\lambda_{\min}(V_t) \geq \rho^2 |N_t^c| \sigma_\zeta^2 - \sqrt{8|N_t^c|h_1^2 \log(\frac{d}{\delta})},$$

or equivalently,

$$\mathsf{P}(\lambda_{\min}(V_t) \leq t) \leq d \exp\left(-\frac{(\rho_1^2 |N_t^c| \sigma_\zeta^2 - t)^2}{8|N_t^c|h_1^2}\right),$$

where $h_1 = 2\rho_1(1-\rho_1)L + 2\rho_1^2$ and $\rho_1 = (\frac{r_l}{S+r_h})\alpha$. This completed the proof of lemma.

D.2 Matrix Azuma Inequality

Theorem D.3 (Matrix Azuma Inequality, Tropp (2012))**.** *Consider a sequence $\{Y_k\}$ of independent, random matrices adapted to the filtration $\{\mathcal{F}_k\}$. Each $\{Y_k\}$ is a self-adjoint matrix such that $\mathsf{E}[Y_k \,|\, \mathcal{F}_{k\ 1}] = 0$. Consider a fixed matrix A_k such that $Y_k^2 \preceq A_k^2$ holds almost surely. Then, for $t \geq 0$, it holds that*

$$\mathsf{P}\left(\lambda_{max}\left(\sum_{k=1} Y_k\right) \geq t\right) \leq d \exp\left(-\frac{t^2}{8\|\sum_{k=1}^{s} A_k^2\|}\right). \tag{53}$$

D.3 Numerical analysis

In order to numerically verify our results in Theorem 3.3, we plot the cumulative number of time that baseline actions played bt SCLTS until time t for $t = 1, \dots, 1000$ over 100 realizations. The solid line in Figure 2 depicts average over 100 realizations and the shaded regions show standard deviation. The figure confirms the logarithmic trend predicted by theory.

E Upper Bounding the Regret of SCLTS-BF

In this section we provide the variation of our algorithm for the case of constraints with bandit feedback, which we refer to as SCLTS-BF in Algorithm 2. We then provide a regret bound for SCLTS-BF. The summary of SCLTS-BF is presented in Algorithm 2.

In this setting, we assume that at each round t, with playing an action x_t, the learner observes the reward $y_t = \langle x_t, \theta_\star \rangle + \xi_t$ and the following bandit feedback:

$$w_t = \langle x_t, \mu_\star \rangle + \chi_t, \tag{54}$$

where χ_t is assumed to be a zero-mean R-sub-Gaussian noise.

Algorithm 2: SCLTS-BF

Input: δ, T, λ, ρ
Set $\delta^0 = \frac{\delta}{4T}$
for $t = 1, \ldots, T$ **do**
  Sample $\eta_t \sim \mathcal{H}^{\text{TS}}$
  Compute RLS-estimate $\hat{\theta}_t$ and V_t according to (5) and $\hat{\mu}_t$
  Set $\tilde{\theta}_t = \hat{\theta}_t + \beta_t V_t^{1/2} \eta_t$
  Build the confidence region $\mathcal{E}_t(\delta^0)$ in (55) and $\mathcal{C}_t(\delta^0)$ in (56)
  Compute the estimated safe set $\mathcal{P}_t^s = \{x \in \mathcal{X} : \langle x, v \rangle \geq (1-\alpha) q_{b_t}, \forall v \in \mathcal{C}_t\}$
  if the following optimization has a feasible solution: $x(\tilde{\theta}_t) = \text{argmax}_{x \in \mathcal{P}_t^s} \langle x, \tilde{\theta}_t \rangle$, **then**
  Set $F = 1$, **else** $F = 0$
  if $F = 1$ **and** $\lambda_{\min}(V_t) \geq \left(\frac{2L\beta_t}{\nu_l + \alpha q_l}\right)^2$, **then**
  Play $x_t = x(\tilde{\theta}_t)$
  else
  play $x_t = x_t^{\text{cb}}$ defined in (59)
  Observe reward r_t
end for

The main difference of SCLTS-BF with SCLTS is in the definition of the estimated safe action set. In particular, at each round t, SCLTS-BF constructs the following confidence regions:

$$\mathcal{E}_t(\delta^0) = \{\theta \in \mathbb{R} : \ \|\theta - \hat{\theta}_t\|_{V_t} \leq \beta_t(\delta^0)\}, \tag{55}$$

$$\mathcal{C}_t(\delta^0) = \{v \in \mathbb{R} : \|v - \hat{\mu}_t\|_{V_t} \leq \beta_t(\delta^0)\}, \tag{56}$$

where $\hat{\mu}_t = V_t^{-1} \sum_{s=1}^{t-1} w_s x_s$ is the RLS-estimate of $\mu_\star$. The radius in (55) and (56) is chosen according to Proposition 2.1 such that $\theta_\star \in \mathcal{E}_t$ and $\mu_\star \in \mathcal{C}_t$ with high probability. In order to ensure safety at each round t, SCLTS-BF constructs the following estimated safe action set

$$\mathcal{P}_t^s = \{x \in \mathcal{X} : \langle x, v \rangle \geq (1-\alpha) q_{b_t}, \forall v \in \mathcal{C}_t\}. \tag{57}$$

The challenge with $\mathcal{P}_t^s$ is that it contains all the actions that are safe with respect to all the parameters in $\mathcal{C}_t$. Thus, there may exist some rounds that $\mathcal{P}_t^s$ is empty. To handle this case, SCLTS-BF proceed as follows. At each round t, given the sampled parameter $\tilde{\theta}_t$, if the estimated safe action set $\mathcal{P}_t^s$ defined in (57) is not empty, SCLTS-BF plays the safe action

$$x(\tilde{\theta}_t) = \arg\max_{x \in \mathcal{P}_t^s} \langle x, \tilde{\theta}_t \rangle \tag{58}$$

only if $\lambda_{\min}(V_t) \geq k_t^2$, where $k_t^2 = \left(\frac{2L\beta_t}{\nu_l + \alpha q_l}\right)^2$. Otherwise, it plays the following conservative action

$$x_t^{\text{cb}} = (1-\rho_2) x_{b_t} + \rho_2 \zeta_t, \tag{59}$$

where $\rho_2 = \alpha(\frac{q_l}{S + q_h})$ in order to ensure that the conservative actions are safe.

Next, we provide a regret guarantee for SCLTS-BF. First, we use the following decomposition of regret:

$$\begin{aligned} R(T) &= \sum_{t=1}^{T} \langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle \\ &= \underbrace{\sum_{t \in N_T} \langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle}_{\text{Term I}} + \underbrace{\sum_{t \in N_T^c} \langle x_\star, \theta_\star \rangle - \langle (1-\rho) x_{b_t} - \rho \zeta_t, \theta_\star \rangle}_{\text{Term II}}, \end{aligned} \tag{60}$$

where N_t^c is the set of rounds $i < t$ that SCLTS-BF plays the conservative actions, and $N_t = \{1, \ldots, t\} - N_t^c$. In the following, we upper bound both Term I and Term II, separately.

Bounding Term I. Bounding Term I follows the same steps as that of Theorem 3.2. Here, we show that for SCLTS-BF, at rounds $t \in N_T$, the optimal action $x_\star$ belongs to the estimated safe safe, i.e., $x_\star \in \mathcal{P}_t^s$. Then, we conclude that regret of Term I similar to Theorem 3.2 has the order of $\mathcal{O}(d^{3/2}\log^{1/2} d\, T^{1/2}\log^{3/2} T)$.

At rounds $t \in N_T$, we know

$$\lambda_{\min}(V_t) \geq k_t^2 \geq \left(\frac{2L\beta_t}{\nu_{b_t} + \alpha q_{b_t}}\right)^2. \tag{61}$$

Then, in order to show that $x_\star \in \mathcal{X}_t^s$, we need to show

$$\langle x_\star, \hat{\mu}_t\rangle - \beta_t \|x_\star\|_{V_t^{-1}} \geq \langle x_\star, \mu_\star\rangle - 2\beta_t\|x_\star\|_{V_t^{-1}} \geq (1-\alpha)q_{b_t}. \tag{62}$$

First inequality comes from the fact that $\|\mu_\star - \hat{\mu}_t\|_{V_t} \leq \beta_t$. Therefore, it suffices to show the second inequality holds. We use the fact that $\|x_\star\|_{V_t^{-1}} \leq \frac{\|x_\star\|_2}{\sqrt{\lambda_{\min}(V_t)}} \leq \frac{L}{\sqrt{\lambda_{\min}(V_t)}}$, where we use Assumption 3 to bound $\|x_\star\|_2$. Hence, we have

$$\langle x_\star, \mu_\star\rangle - 2\beta_t\|x_\star\|_{V_t^{-1}} \geq \langle x_\star, \mu_\star\rangle - \frac{2\beta_t L}{\sqrt{\lambda_{\min}(V_t)}}. \tag{63}$$

Then, it suffices to show that

$$\nu_{b_t} + \alpha q_{b_t} \geq \frac{2\beta_t L}{\sqrt{\lambda_{\min}(V_t)}}, \tag{64}$$

From (61), we know that (64) holds, and hence, $x_\star \in \mathcal{P}_t^s$. Therefore, we can use the result of Theorem 3.2, and obtain the desired regret bound.

Bounding Term II. First, we provide the formal statement of the theorem.

Theorem E.1. *Let $\lambda, L \geq 1$. On event $\{\theta_\star \in \mathcal{E}_t, \forall t \in [T]\} \cap \{\mu_\star \in \mathcal{C}_t, \forall t \in [T]\}$, and Assumptions 4, we can upper bound the number of times SCLTS-BF plays the conservative actions, i.e., $|N_T^c|$ as:*

$$|N_T^c| \leq \left(\frac{2L\beta_T}{\rho_2\sigma_\zeta(\alpha q_l + \nu_l)}\right)^2 + \frac{2h_2^2}{\rho_2^4\sigma_\zeta^4}\log(\frac{d}{\delta}) + \frac{2Lh_2\beta_T\sqrt{8\log(\frac{d}{\delta})}}{\rho_2^3\sigma_\zeta^3(\alpha q_l + \nu_l)} \tag{65}$$

where $h_2 = 2\rho_2(1-\rho_2)L + 2\rho_2^2$ and $\rho_2 = (\frac{q_l}{S+q_h})\alpha$.

In order to prove Theorem E.1, we proceed as follows:

$$\begin{aligned}\sum_{t\in N_T^c} \langle x_\star, \theta_\star\rangle - \langle (1-\rho_2)x_{b_t} - \rho_2\zeta_t, \theta_\star\rangle &= \sum_{t\in N_T^c} \langle x_\star, \theta_\star\rangle - \langle x_{b_t}, \theta_\star\rangle + \rho_2(\langle x_{b_t} + \zeta_t, \theta_\star\rangle) \\ &\leq \sum_{t\in N_T^c} \nu_h + \rho_2(q_{b_t} + S) \leq |N_T^c|(\nu_h + \alpha q_l),\end{aligned} \tag{66}$$

where $q_h \geq q_{b_t} \geq q_l > 0$ and $\nu_h \geq \nu_{b_t} \geq \nu_l$ for all t. Therefore, in order to bound Term II, it suffices to upper bound $|N_T^c|$ which is the number of rounds that SCLTS-BF plays the conservative actions up to round T. In order to do so, we proceed as follows:

Let τ be any round that the algorithm plays the conservative action.

If $F = 0$, i.e.,

$$\nexists x \in \mathcal{X} : \langle x, \hat{\mu}_\tau\rangle - \beta_\tau\|x\|_{V^{-1}} \geq (1-\alpha)q_b\ , \tag{67}$$

and since we know that $x_\star \in \mathcal{X}$, and $\mu_\star \in \mathcal{C}_t$ with high probability, we can write

$$\langle x_\star, \mu_\star\rangle - 2\beta_\tau\|x_\star\|_{V^{-1}} \leq \langle x_\star, \hat{\mu}_\tau\rangle - \beta_\tau\|x_\star\|_{V^{-1}} < (1-\alpha)q_b\ . \tag{68}$$

Using (68), we can get

$$\nu_b \ + \alpha q_b \ < 2\beta_\tau \|x_\star\|_{V \ 1} \leq \sqrt{\frac{2\beta_\tau L}{\lambda_{\min}(V_\tau)}}, \tag{69}$$

and hence the following upper bound on minimum eigenvalue of the Gram matrix:

$$\lambda_{\min}(V_\tau) < \left(\frac{2\beta_\tau L}{\nu_b \ + \alpha q_b}\right)^2 \leq \left(\frac{2\beta_\tau L}{\nu_l + \alpha q_l}\right)^2 = k_\tau \tag{70}$$

Therefore, we show that in the cases where either the event $\{\@x \in \mathcal{X} : \langle x, \hat{\mu}_\tau \rangle - \beta_\tau \|x\|_{V \ 1} \geq (1-\alpha) q_b \ \}$ or the event $\{\lambda_{\min}(V_\tau) < k_\tau^2\}$ happen, we can conclude that at round τ

$$\lambda_{\min}(V_\tau) < k_\tau^2. \tag{71}$$

From Lemma D.1, we know that the minimum eigenvalue of the Gram matrix, i.e., $\lambda_{\min}(V_t)$ is lower bounded with the number of times that SCLTS-BF plays the conservative actions, i.e., $|N_T^c|$. Therefore, using (71), we can get

$$|N_T^c| \leq \left(\frac{2L\beta_T}{\rho_2 \sigma_\zeta (\alpha q_l + \nu_l)}\right)^2 + \frac{2h_2^2}{\rho_2^4 \sigma_\zeta^4} \log(\frac{d}{\delta}) + \frac{2Lh_2\beta_T \sqrt{2\log(\frac{d}{\delta})}}{\rho_2^3 \sigma_\zeta^3 (\alpha q_l + \nu_l)} \tag{72}$$

where $h_2 = 2\rho_2(1-\rho_2)L + 2\rho_2^2$ and $\rho_2 = \alpha(\frac{q_l}{S+q_h})$.

F Proof of Theorem 4.1

In this section, we first present the SCLTS2 algorithm, for the case where the learner does not know the reward of the actions suggested by baseline policy in advance, i.e., r_{b_t}. The summary of SCLTS2 is presented in Algorithm 3.

The algorithm relies on the fact that we can find an upper bound over the value of r_{b_t}, using the fact that $\theta_\star \in \mathcal{E}_t$, i.e.,:

$$\max_{v 2 E_t} \langle x_{b_t}, v \rangle \geq \langle x_{b_t}, \theta_\star \rangle = r_{b_t}. \tag{73}$$

Then, we can write the safety constraint as follows:

$$\min_{v 2 E_t} \langle x(\tilde{\theta}_t), v \rangle \geq (1-\alpha) \max_{v 2 E_t} \langle x_{b_t}, v \rangle. \tag{74}$$

It is easy to show that safety constraint (2) holds when (74) is true. Therefore, if we choose actions that satisfy (74), we can ensure that they are safe with respect to the safety constrain in (2).

Then we propose the estimated safe action set $\mathcal{Z}_t^s$ as:

$$\mathcal{Z}_t^s = \{x \in \mathcal{X} : \min_{v 2 E_t} \langle x, v \rangle \geq (1-\alpha) \max_{v 2 E_t} \langle x_{b_t}, v \rangle\}, \tag{75}$$

which contains actions that are safe with respect to all the parameter in $\mathcal{E}_t$. At each round t, SCLTS2 plays the safe action $x(\tilde{\theta}_t)$ from $\mathcal{Z}_t^s$ that maximizes the expected reward given the sampled parameter $\tilde{\theta}_t$, i.e.,

$$x(\tilde{\theta}_t) = \arg \max_{x 2 Z_t^s} \langle x, \tilde{\theta}_t \rangle \tag{76}$$

only if $\lambda_{\min}(V_t) \geq k_t^3$, where $k_t^3 = \left(\frac{2L\beta_t(2 \ \alpha)}{\kappa_l + \alpha r_l}\right)^2$. Otherwise it plays the conservative action $x_{b_t}^{\text{cb}}$ as:

$$x_t^{\text{cb}} = (1-\rho_3) x_{b_t} + \rho_3 \zeta_t, \tag{77}$$

where $\rho_3 = \alpha(\frac{r_l}{S+1})$ such that the conservative action x_t^{cb} is safe, where we use Assumption 3 for upper bounding the reward, i.e., $r_{b_t} \leq 1$.

In order to bound the regret of SCLTS2, we first use the decomposition defined in Proposition 3.1. The regret of Term I is similar to that of SCLTS (i.e., Theorem 3.2). Hence, it suffices to upper bound the number of time SCLTS2 plays the conservative actions, i.e., $|N_T^c|$.

Algorithm 3: SCLTS2

Input: δ, T, λ, ρ
Set $\delta^0 = \frac{\delta}{4T}$
for $t = 1, \ldots, T$ **do**
  Sample $\eta_t \sim \mathcal{H}^{\text{TS}}$
  Compute RLS-estimate $\hat{\theta}_t$ and V_t according to (5)
  Set $\tilde{\theta}_t = \hat{\theta}_t + \beta_t V_t^{1/2} \eta_t$
  Build the confidence region $\mathcal{E}_t(\delta^0)$ in (6)
  Compute the estimated safe set $\mathcal{Z}_t^s = \{x \in \mathcal{X} : \min_{v \in \mathcal{E}_t} \langle x, v \rangle \geq (1-\alpha) \max_{v \in \mathcal{E}_t} \langle x_{b_t}, v \rangle\}$
  if the following optimization is feasible: $x(\tilde{\theta}_t) = \arg\max_{x \in \mathcal{Z}_t^s} \langle x, \tilde{\theta}_t \rangle$, **then**
  Set $F = 1$, **else** $F = 0$
  if $F = 1$ **and** $\lambda_{\min}(V_t) \geq \left(\frac{2L\beta_t(2-\alpha)}{\kappa_l + \alpha r_l}\right)^2$, **then**
  Play $x_t = x(\tilde{\theta}_t)$
  else
  play $x_t = x_t^{\text{cb}}$ defined in (77)
  Observe reward y_t
end for

In order to bound $|N_T^c|$, we proceed as follows. Let τ be the round that SCLTS2 plays a conservative action. If $F = 0$, i.e.,

$$\nexists x \in \mathcal{X} : \min_{v \in \mathcal{C}} \langle x, v \rangle \geq (1-\alpha) \max_{v \in \mathcal{C}} \langle x_b, v \rangle. \tag{78}$$

Using the fact that $x_\star \in \mathcal{X}$, we can write

$$\langle x_\star, \hat{\theta}_\tau \rangle - \beta_\tau \|x_\star\|_{V_\tau^{-1}} < (1-\alpha) \langle x_b, \hat{\theta}_\tau \rangle + \beta_\tau \|x_b\|_{V_\tau^{-1}}. \tag{79}$$

Then, since $\|\theta_\star - \hat{\theta}_t\|_{V_t} \leq \beta_t$, we can upper bound the RHS and lower bound the LHS of (79), and get

$$\langle x_\star, \theta_\star \rangle - 2\beta_\tau \|x_\star\|_{V_\tau^{-1}} < (1-\alpha) \langle x_b, \theta_\star \rangle + 2\beta_\tau \|x_b\|_{V_\tau^{-1}}, \tag{80}$$

or equivalently,

$$\kappa_b + \alpha r_b < 2\beta_\tau \|x_\star\|_{V_\tau^{-1}} + 2(1-\alpha)\beta_\tau \|x_b\|_{V_\tau^{-1}}. \tag{81}$$

Then we can use the fact that $\|x_\star\|_{V_\tau^{-1}} \leq \frac{L}{\sqrt{\lambda_{\min}(V_\tau)}}$ and $\|x_b\|_{V_\tau^{-1}} \leq \frac{L}{\sqrt{\lambda_{\min}(V_\tau)}}$, where we use Assumption 3 for upper bounding $\|x_\star\|_2$. Thus, we upper bound the RHS of (81) as follows:

$$\kappa_b + \alpha r_b < 2\beta_\tau \sqrt{\frac{L}{\lambda_{\min}(V_\tau)}} + 2(1-\alpha)\beta_\tau \sqrt{\frac{L}{\lambda_{\min}(V_\tau)}}, \tag{82}$$

and hence, we can get the following upper bound $\lambda_{\min}(V_\tau)$ as follows:

$$\lambda_{\min}(V_\tau) < \left(\frac{2L\beta_T(2-\alpha)}{\kappa_b + \alpha r_b}\right)^2 \leq \left(\frac{2L\beta_T(2-\alpha)}{\kappa_l + \alpha r_l}\right)^2 = k_\tau^3. \tag{83}$$

Therefore, we show that whether the event $F = 0$ happens or $\lambda_{\min}(V_t) < k_t^3$, we can achieve the upper bound provided in (83). Then, using the result of Lemma D.1, where we show that $\lambda_{\min}(V_t)$ is lower bounded with the number of times the algorithm plays the conservative actions, we obtain the following upper bound on the $|N_\tau^c|$

$$|N_\tau^c| \leq \left(\frac{2L\beta_\tau(2-\alpha)}{\rho_3 \sigma_\zeta(\kappa_l + \alpha r_l)}\right)^2 + \frac{2h_3^2}{\rho_3^4 \sigma_\zeta^4} \log(\frac{d}{\delta}) + \frac{2Lh_3\beta_\tau(2-\alpha)}{\rho_3^3 \sigma_\zeta^3(\kappa_l + \alpha r_l)} \sqrt{2\log(\frac{d}{\delta})}, \tag{84}$$

where $h_3 = 2\rho_3(1-\rho_3)L + 2\rho_3^2$ and $\rho_3 = \alpha(\frac{r_l}{S+1})$.

G Stage-wise Conservative Linear UCB (SCLUCB) Algorithm

In this section we propose a UCB-based safe stochastic linear bandit algorithm called Stage-wise Conservative Linear-UCB (SCLUCB), which is a safe counterpart of LUCB for the stage-wise conservative bandit setting. In particular, at each round t, given the RLS-estimate $\hat{\theta}_t$ of $\theta_\star$, SCLUCB constructs the confidence region $\mathcal{E}_t$ as follows:

$$\mathcal{E}_t(\delta) = \{\theta \in \mathsf{R}^d : \|\theta - \hat{\theta}_t\|_{V_t} \leq \beta_t(\delta)\}. \tag{85}$$

The radius $\beta_t(\delta)$ is chosen as in Proposition 2.1 such that $\theta_\star \in \mathcal{E}_t(\delta)$ with probability $1-\delta$. Then, similar to SCLTS, it builds the estimated safe set $\mathcal{X}_t^s$ such that it includes actions that are safe with respect to all the parameter in $\mathcal{E}_t$, i.e.,

$$\mathcal{X}_t^s = \{x \in \mathcal{X} : \langle x, v\rangle \geq (1-\alpha) r_{b_t}, \forall v \in \mathcal{E}_t\}. \tag{86}$$

Similar to SCLTS, the challenge with $\mathcal{X}_t^s$ is that there may exist some rounds that $\mathcal{X}_t^s$ is empty. In order to face this problem, SCLUCB proceed as follows. In order to guarantee safety, at each round t, if $\mathcal{X}_t^s$ is not empty, SCLUCB plays the action $\bar{x}_t$ as

$$(\bar{x}_t, \bar{\theta}_t) = \max_{x 2 X_t^s} \max_{v 2 E_t} \langle x, v\rangle \tag{87}$$

only if $\lambda_{\min}(V_t) \geq \left(\frac{2L\beta_t}{\kappa_l + \alpha r_{b_l}}\right)^2$, otherwise it plays the conservative action x_t^{cb} defined in (11). The summary of SCLUCB is presented in Algorithm (4).

Algorithm 4: Stage-wise Conservative Linear UCB (SCLUCB)

Input: δ, T, λ, ρ
for $t = 1, \ldots, T$ **do**
  Compute RLS-estimate $\hat{\theta}_t$ and V_t according to (5)
  Build the confidence region $\mathcal{E}_t(\delta)$ in (85)
  Compute the estimated safe set $\mathcal{X}_t^s$ in (86)
  if the following optimization is feasible: $\bar{x}_t = \arg\max_{x 2 X_t^s} \max_{v 2 E_t} \langle x, v\rangle$, **then**
  Set $F = 1$, **else** $F = 0$
  if $F = 1$ **and** $\lambda_{\min}(V_t) \geq \left(\frac{2L\beta_t}{\kappa_l + \alpha r_{b_l}}\right)^2$, **then**
  Play $x_t = \bar{x}_t$
  else
  play $x_t = x_t^{\text{cb}}$ defined in (11)
  Observe reward y_t
end for

Next, we provide the regret guarantee for SCLUCB. Recall, $N_{t\ 1}$ be the set of rounds $i < t$ at which SCLUCB plays the action in (10). Similarly, $N_{t\ 1}^c = \{1, \ldots, t-1\} - N_{t\ 1}$ is the set of rounds $j < t$ at which SCLUCB plays the conservative actions.

Proposition G.1. *The regret of SCLUCB can be decomposed into two terms as follows:*

$$R(T) \leq \underbrace{\sum_{t 2 N_T} (\langle x_\star, \theta_\star\rangle - \langle x_t, \theta_\star\rangle)}_{\text{Term I}} + \underbrace{|N_T^c| (\kappa_h + \rho_1(r_h + S))}_{\text{Term II}} \tag{88}$$

In the following, we bound both terms, separately.

Bounding Term I. The first Term in (88) is the regret caused by playing the safe actions that maximize the reward given the true parameter is $\bar{\theta}_t$. The idea of bounding Term I is similar to Abbasi-Yadkori et al. (2011). We use the fact that for $t \in N_T$, $x_t = \bar{x}_t$, and start with the following decomposition of the instantaneous regret for $t \in N_T$:

$$\langle x_\star, \theta_\star\rangle - \langle x_t, \theta_\star\rangle = \underbrace{\langle x_\star, \theta_\star\rangle - \langle \bar{x}_t, \bar{\theta}_t\rangle}_{\text{Term A}} + \underbrace{\langle \bar{x}_t, \bar{\theta}_t\rangle - \langle \bar{x}_t, \theta_\star\rangle}_{\text{Term B}} \tag{89}$$

Bounding Term A. Since for round $t \in N_t$, we require that $\lambda_{\min}(V_t) \geq k_t^1$, where $k_t^1 = \left(\frac{2L\beta_t}{\kappa_l + \alpha r_{b_l}}\right)^2$, we can conclude that $x_\star \in \mathcal{X}_t^s$. Therefore, due to (87), we have $\langle \bar{x}_t, \bar{\theta}_t \rangle \geq \langle x_\star, \theta_\star \rangle$, and hence Term A is not positive.

Bounding Term B. In order to bound Term B, we use the following chain of inequalities:

$$\begin{aligned}\text{Term B} &:= \langle \bar{x}_t, \bar{\theta}_t \rangle - \langle \bar{x}_t, \theta_\star \rangle = \langle \bar{x}_t, \bar{\theta}_t \rangle - \langle \bar{x}_t, \hat{\theta}_t \rangle + \langle \bar{x}_t, \hat{\theta}_t \rangle - \langle \bar{x}_t, \theta_\star \rangle \\ &\leq \|\bar{x}_t\|_{V_t^{-1}} \|\bar{\theta}_t - \hat{\theta}_t\|_{V_t} + \|\bar{x}_t\|_{V_t^{-1}} \|\hat{\theta}_t - \theta_\star\|_{V_t} \\ &\leq 2\beta_t \|\bar{x}_t\|_{V_t^{-1}}, \end{aligned} \tag{90}$$

where the last inequality follows from Proposition 2.1. Recall, from Assumption 3, we have the following trivial bound:

$$\langle x_\star, \theta_\star \rangle - \langle \bar{x}_t, \theta_\star \rangle \leq 2. \tag{91}$$

Thus, we conclude the following

$$\text{Term B} \leq 2 \min(\beta_t \|\bar{x}_t\|_{V_t^{-1}}, 1). \tag{92}$$

Next, we state a direct application of Lemma 11 in Abbasi-Yadkori et al. (2011).

Lemma G.2. *For $\lambda > 0$, and under Assumptions 1, 2, and 3, we have*

$$\sum_{t=1}^{T} \min(\|\bar{x}_t\|^2_{V_t^{-1}}, 1) \leq 2d \log\left(1 + \frac{TL^2}{\lambda d}\right) \tag{93}$$

Therefore, from Lemma G.2, we can conclude the following bound on regret of Term B:

$$\sum_{t \in N_T} 2 \min(\beta_t \|\bar{x}_t\|_{V_t^{-1}}, 1) \leq 2\beta_T \sqrt{2d|N_T| \log(1 + \frac{|N_T|L^2}{\lambda d})}. \tag{94}$$

Next, in Theorem G.3, we provide an upper bound on the regret of Term I which is of order $\mathcal{O}\left(d\sqrt{T}\log(\frac{TL^2}{\lambda\delta})\right)$.

Theorem G.3. *On event $\{\theta_\star \in \mathcal{E}_t\}$ for a fixed $\delta \in (0,1)$, with probability $1-\delta$, it holds that:*

$$\sum_{t \in N_T} (\langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle) \leq 2\beta_T \sqrt{2dT \log(1 + \frac{TL^2}{\lambda d})} \tag{95}$$

Bounding Term II. In order to bound Term II in (88), we need to find an upper bound on the number of times that SCLUCB plays the conservative actions up to time T, i.e., $|N_T^c|$. We prove an upper bound on $|N_T^c|$ in Theorem G.4 which has the order of $\mathcal{O}\left(\frac{L^2 d \log(\frac{T}{\delta}) \log(\frac{d}{\delta})}{\alpha^4 (r_l^2 \wedge r_l^4) \kappa_l (\sigma^2 \wedge \sigma^4)}\right)$.

Theorem G.4. *Let $\lambda, L \geq 1$. On event $\{\theta_\star \in \mathcal{E}_t, \forall t \in [T]\}$, and under Assumption 4, we can upper bound the number of times SCLUCB plays the conservative actions, i.e., $|N_T^c|$ as:*

$$|N_T^c| \leq \left(\frac{2L\beta_T}{\rho_1 \sigma_\zeta (\kappa_l + \alpha r_l)}\right)^2 + \frac{2h_1^2}{\rho_1^4 \sigma_\zeta^4} \log(\frac{d}{\delta}) + \frac{2Lh_1\beta_T \sqrt{8\log(\frac{d}{\delta})}}{\rho_1^3 \sigma^3 (\kappa_l + \alpha r_l)}, \tag{96}$$

where $h_1 = 2\rho_1(1-\rho_1)L + 2\rho_1^2$ and $\rho_1 = (\frac{r_l}{S + r_h})\alpha$.

The proof is similar to that of Theorem 3.3, and we omit its proof here.

H Comparison with Safe-LUCB

In this section, we extend our results to an alternative safe bandit formulation proposed in Amani et al. (2019), where the algorithm Safe-LUCB was proposed. In order to do so, we first present the safety constraint in Amani et al. (2019), and then we show the required modification of SCLUCB to handle

this case, which we refer to as SCLUCB2. Then, we provide a problem-dependent regret bound for SCLUSB2, and we show that it matches the problem dependent regret bound of Safe-LUCB in Amani et al. (2019). We need to note that in Amani et al. (2019), they also provide a general regret bound of order $\tilde{\mathcal{O}}(T^{2/3})$ for Safe-LUCB which we do not discuss in this paper.

In Amani et al. (2019), it is assumed that the learner is given a convex and compact decision set $\mathcal{D}_0$ which contains the origin, and with playing the action x_t, she observes the reward of $y_t = x_t^\top \theta_\star + \eta_t$, where $\theta_\star$ is the fixed unknown parameter, and η_t is R-sub-Gaussian noise. Moreover, The learning environment is subject to the linear safety constraint

$$x^\top B\theta_\star \leq C, \tag{97}$$

which needs to be satisfied at all rounds t with high probability, and an action x_t is called safe, if it satisfies (97). In (97), the matrix $B \in \mathsf{R}^{d \quad d}$ and the positive constant C are known to the learner. However, the learner does not receive any bandit feedback on the value $x^\top B\theta_\star$ and her information is restricted to those she receives from the reward.

Given the above constraint, the learner is restricted to choose actions from the safe set $\mathcal{D}_0^s$ as:

$$\mathcal{D}_0^s(\theta_\star) = \{x \in \mathcal{D}_0 : x^\top B\theta_\star \leq C\}. \tag{98}$$

Since $\theta_\star$ in unknown, the safe set $\mathcal{D}_0^s$ is unknown to the learner. Then, in Amani et al. (2019), they provide the problem-dependent regret bound (for the case where $\Delta := C - x^\top B\theta_\star > 0$) of order $\mathcal{O}(\sqrt{T}\log T)$. In the following, we present the required modification of SCLUSB to handle this safe bandit formulation, and propose the new algorithm called SCLUCB2 that we prove a problem dependnt regret bound of order $\mathsf{O}(\sqrt{T}\log T)$. We need to note that Amani et al. (2019) also provide a general regret bound of order $\tilde{\mathcal{O}}(T^{2/3})$ for the case where $\Delta = 0$; however, we do not discuss this case in this paper.

At each round t, given the RLS-estimate $\hat{\theta}_t$ of $\theta_\star$, SLUCB2 builds the confidence region $\mathcal{E}_t$ as:

$$\mathcal{E}_t = \{\theta \in \mathsf{R}^d : \|\theta - \hat{\theta}_t\|_{V_t} \leq \beta_t\}, \tag{99}$$

and the radius β_t is chosen according to Proposition 2.1 such that $\theta_\star \in \mathcal{E}_t$ with high probability. The learner does not know the safe set $\mathcal{D}_0^s$; however, she knows that $\theta_\star \in \mathcal{E}_t$ with high probability. Hence, SLUCB2 constructs the estimated safe set $\mathcal{D}_t^s$ such that it contains actions that are safe with respect to all the parameter in $\mathcal{E}_t$, i.e.,

$$\begin{aligned}
\mathcal{D}_t^s &= \{x \in \mathcal{D}_0 : x^\top Bv \leq C, \forall v \in \mathcal{E}_t\} \\
&= \{x \in \mathcal{D}_0 : \max_{v 2 E_t} x^\top Bv \leq C\} \\
&= \{x \in \mathcal{D}_0 : x^\top B\hat{\theta}_t + \beta_t \|Bx\|_{V_t^{\ 1}} \leq C\}
\end{aligned} \tag{100}$$

Clearly, action $x = [0]^d$ (origin) is a safe action since $C > 0$, and also $[0]^d \in \mathcal{D}_0$. Thus, $[0]^d \in \mathcal{D}_t^s$. Since $x = [0]^d$ is a known safe action, we define the conservative action x_0^{c} as:

$$x_0^{\mathsf{c}} = (1-\rho)[0]^d + \rho\zeta_t = \rho\zeta_t, \tag{101}$$

where ζ_t is a sequence of IID random vectors such that $\|\zeta_t\|_2 = 1$ almost surly, and $\sigma_\zeta = \lambda_{\min}(\mathrm{Cov}(\zeta_t)) > 0$. We choose the constant ρ according to the Lemma H.1 in order to ensure that the conservative action x_0^{c} is safe.

Lemma H.1. *At each round t, for any $\rho \in (0, \bar{\rho})$, where*

$$\bar{\rho} = \frac{C}{\|B\|S}, \tag{102}$$

the conservative action $x_0^{\mathsf{c}} = \rho\zeta_t$ is guaranteed to be safe almost surly.

We choose $\rho = \frac{C}{kBkS}$ for the rest of this section, and hence the conservative action is

$$x_0^{\mathsf{c}} = \frac{C}{\|B\|S}\zeta_t. \tag{103}$$

Let $\Delta = C - x_\star^\top B\theta_\star$. We consider the case where $\Delta > 0$. At each t, in order to guarantee safety, SCLUCB2 only chooses its action from the estimated safe set $\mathcal{D}_t^s$. The challenge with $\mathcal{D}_t^s$ is that it includes actions that are safe with respect to all parameter in $\mathcal{E}_t$, and not only $\theta_\star$. Thus, there may exist some rounds that $\mathcal{D}_t^s$ is empty. At round t, if $\mathcal{D}_t^s$ is not empty, SCLUCB2 plays the safe action

$$\bar{x}_t = \arg\max_{x \in \mathcal{D}_t^s} \max_{v \in \mathcal{E}_t} \langle x, v \rangle \tag{104}$$

only if $\lambda_{\min}(V_t) \geq \left(\frac{2L\beta_t kBk}{\Delta}\right)^2$, otherwise it plays the conservative action x_0^c in (103). The summary of SCLUCB2 is presented in Algorithm 5.

Algorithm 5: SCLUCB2

Input: δ, T, λ, ρ
for $t = 1, \ldots, T$ **do**
  Compute RLS-estimate $\hat{\theta}_t$ and V_t according to (5)
  Build the confidence region $\mathcal{E}_t(\delta)$ in (99)
  Compute the estimated safe set $\mathcal{D}_t^s$ in (100)
  if the following optimization is feasible: $\bar{x}_t = \arg\max_{x \in \mathcal{D}_t^s} \max_{v \in \mathcal{E}_t} \langle x, v \rangle$, **then**
  Set $F = 1$, **else** $F = 0$
  if $F = 1$ **and** $\lambda_{\min}(V_t) \geq \left(\frac{2L\beta_t kBk}{\Delta}\right)^2$, **then**
  Play $x_t = \bar{x}_t$
  else
  play $x_t = x_0^c$ defined in (103)
  Observe reward y_t
end for

In the following we provide the regret guarantee for SCLUCB2. Let N_{t-1} be the set of rounds $i < t$ at which SCLUCB2 plays the action in (104). Similarly, $N_{t-1}^c = \{1, \ldots, t-1\} - N_{t-1}$ is the set of rounds $j < t$ at which SCLUCB2 plays the conservative action in (103).

First, we use the following decomposition of the regret, then we bound each term separately.

Proposition H.2. *The regret of SCLUCB2 can be decomposed to the following two terms:*

$$\begin{aligned} R(T) &= \sum_{t=1}^{T} \langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle \\ &= \sum_{t \in N_T} \langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle + \sum_{t \in N_T^c} \langle x_\star, \theta_\star \rangle - \langle x_0^c, \theta_\star \rangle , \\ &\leq \underbrace{\sum_{t \in N_T} \langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle}_{\textit{Term I}} + \underbrace{2|N_T^c|}_{\textit{Term II}} . \end{aligned} \tag{105}$$

Bounding Term I. In order to bound Term I, we proceed as follows. First, we show that at rounds $t \in N_T$, the optimal action $x_\star$ belongs to the estimated safe set $\mathcal{D}_t^s$, i.e., $x_\star \in \mathcal{D}_t^s$. To do so, we need to show that

$$x_\star^\top B\hat{\theta}_t + \beta_t \|Bx_\star\|_{V_t^{-1}} \leq C. \tag{106}$$

Since $\|\theta_\star - \hat{\theta}_t\|_{V_t} \leq \beta_t$, it suffices to show that:

$$x_\star^\top B\theta_\star + 2\beta_t \|Bx_\star\|_{V_t^{-1}} \leq C, \tag{107}$$

or equivalently

$$2\beta_t \|Bx_\star\|_{V_t^{-1}} \leq \Delta, \tag{108}$$

where $\Delta = C - x_\star^\top B\theta_\star$. It is easy to see (106) is true whenever (107) holds. Using Assumption 3, we can get $\|Bx_\star\|_{V_t^{-1}} \leq \frac{kBkkx_\star k_2}{\sqrt{\lambda_{\min}(V_t)}} \leq \frac{kBkL}{\sqrt{\lambda_{\min}(V_t)}}$. Hence, from (108), it suffices to show that

$$\frac{2\beta_t\|B\|L}{\sqrt{\lambda_{\min}(V_t)}} \leq \Delta, \tag{109}$$

or equivalently

$$\lambda_{\min}(V_t) \geq \left(\frac{2\beta_t\|B\|L}{\Delta}\right)^2 \tag{110}$$

that we know it is true for $t \in N_T$. Therefore, on event $\{\theta_\star \in \mathcal{E}_t\}$, $x_\star \in \mathcal{D}_t^s$. We can bound the regret of Term I in (105) similar to Theorem G.3, and get the regret of order $\mathcal{O}\left(d\sqrt{T}\log(\frac{TL^2}{\lambda\delta})\right)$.

Bounding Term II. We need to upper bound the number of times that SCLUCB2 plays the conservative action x_0^c, i.e., $|N_T^c|$. We prove an upper bound on $|N_T^c|$ in Theorem H.3 which has the order of $\mathcal{O}\left(\frac{L^2S^2kBk^2d\log(\frac{T}{\delta})\log(\frac{d}{\delta})}{\Delta^2(C\wedge C^2)(\sigma^2\wedge\sigma^4)}\right)$.

Theorem H.3. *Let $\lambda, L \geq 1$. On event $\{\theta_\star \in \mathcal{E}_t, \forall t \in [T]\}$, we can upper bound the number of times SCLUCB2 plays the conservative actions, i.e., $|N_T^c|$ as:*

$$|N_T^c| \leq \left(\frac{2LS\|B\|^2\beta_T}{C\Delta\sigma_\zeta}\right)^2 + \frac{32\log(\frac{d}{\delta})}{\sigma_\zeta^4} + \frac{8LS\|B\|^2\beta_T}{C\Delta\sigma_\zeta^3}\sqrt{2\log(\tfrac{d}{\delta})}. \tag{111}$$

Proof. Let τ be any round that the algorithm plays the conservative action, i.e., at round τ, either $F = 0$ or $\lambda_{\min}(V_\tau) < \left(\frac{2LkBk\beta}{\Delta}\right)^2$.

By definition, if $F = 0$, we have

$$\forall x \in \mathcal{X} : x^\top B\hat{\theta}_\tau + \beta_\tau\|Bx\|_{V_\tau^{-1}} \leq C, \tag{112}$$

and since we know that $x_\star \in \mathcal{X}$, and $\theta_\star \in \mathcal{E}_t$ with high probability, we can write

$$x_\star^\top B\theta_\star + 2\beta_\tau\|Bx_\star\|_{V_\tau^{-1}} \geq x_\star^\top B\hat{\theta}_\tau + \beta_\tau\|Bx_\star\|_{V_\tau^{-1}} > C. \tag{113}$$

Then, using the LHS and RHS of (113), we can get

$$\frac{2L\|B\|\beta_\tau}{\sqrt{\lambda_{\min}(V_\tau)}} \geq 2\beta_\tau\|x_\star\|_{V_\tau^{-1}} \geq \Delta,$$

and hence the following upper bound on minimum eigenvalue of the Gram matrix:

$$\lambda_{\min}(V_\tau) < \left(\frac{2L\|B\|\beta_\tau}{\Delta}\right)^2.$$

Therefore, at any round τ that a conservative action is played, whether it is because $\{F = 0\}$ happens or beccause we have $\{\lambda_{\min}(V_\tau) < \left(\frac{2LkBk\beta}{\Delta}\right)^2\}$, we can always conclude that

$$\lambda_{\min}(V_\tau) < \left(\frac{2L\|B\|\beta_\tau}{\Delta}\right)^2 \tag{114}$$

The remaining of the proof builds on two auxiliary lemmas. First, in Lemma H.4, we show that the minimum eigenvalue of the Gram matrix V_t is lower bounded with the number of times SCLUCB2 plays the conservative actions.

Lemma H.4. *On event $\{\theta_\star \in \mathcal{E}_t\}$, it holds that*

$$\mathbb{P}(\lambda_{min}(V_t) \leq t) \leq d\exp\left(-\frac{(\rho^2\sigma_\zeta^2|N_t^c| - t)^2}{32\rho^4|N_t^c|}\right), \tag{115}$$

where $\rho = \frac{C}{kBkS}$.

Using (114) and applying Lemma H.4, it can checked that with probability $1-\delta$

$$\frac{2L\|B\|\beta_\tau}{\Delta}^2 > \rho^2\sigma_\zeta^2|N_\tau^c| - \sqrt{32\rho^4.|N_\tau^c|\log(\frac{d}{\delta})},$$

Then using Lemma D.2, we can conclude the following upper bound

$$|N_\tau^c| \leq \frac{2LS\|B\|^2\beta_\tau}{C\Delta\sigma_\zeta}^2 + \frac{32\log(\frac{d}{\delta})}{\sigma_\zeta^4} + \frac{8LS\|B\|^2\beta_\tau}{C\Delta\sigma_\zeta^3}\sqrt{2\log(\frac{d}{\delta})}.$$

□

H.1 Proof of Lemma H.4

Our objective is to establish a lower bound on $\lambda_{\min}(V_t)$ for all t. It holds that

$$\begin{aligned}
V_t &= \lambda I + \sum_{s=1}^{\tau} x_s x_s^\top \\
&\succeq \sum_{s\in N_t^c} (\rho\zeta_s)(\rho\zeta_s)^\top \\
&= \sum_{s\in N_t^c} \rho^2\mathbb{E}[\zeta_s\zeta_s^\top] + \rho^2\zeta_s\zeta_s^\top - \rho^2\mathbb{E}[\zeta_s\zeta_s^\top] \\
&\succeq \rho^2\sigma_\zeta^2|N_t^c|I + \sum_{s\in N_t^c} G_s, \qquad (116)
\end{aligned}$$

where G_s is defined as

$$G_s = \rho^2\zeta_s\zeta_s^\top - \rho^2\mathbb{E}[\zeta_s\zeta_s^\top] \quad . \qquad (117)$$

Thus, using Weyl's inequality, it follows that

$$\lambda_{\min}(V_t) \geq \rho^2\sigma_\zeta^2|N_t^c| - \lambda_{\max}(\sum_{s\in N_t^c} G_s).$$

Next, we apply the matrix Azuma inequality (see Theorem D.3) to find an upper bound on $\lambda_{\max}(\sum_{s\in N_t^c} G_s)$. For this, we first need to show that the sequence of matrices G_s satisfies the conditions of Theorem D.3. By definition of G_s in (117), it follows that $\mathbb{E}[G_s|\mathcal{F}_{s-1}] = 0$, and $G_s^\top = G_s$. Also, we construct the sequence of deterministic matrices A_s such that $G_s^2 \preceq A_s^2$ as follows. We know that for any matrix K, $K^2 \leq \|K\|_2^2 I$, where $\|K\|_2$ is the maximum singular value of K, i.e.,

$$\sigma_{\max}(K) = \max_{\|u\|_1=\|v\|_2=1} u^\top Kv.$$

Thus, we first show the following bound on the maximum singular value of the matrix G_s defined in (117):

$$\begin{aligned}
\max_{\|u\|_1=\|v\|_2=1} u^\top G_s v &= \rho^2(u^\top\zeta_s)(v^\top\zeta_s)^\top - \rho^2\mathbb{E}\ (u^\top\zeta_s)(v^\top\zeta_s)^\top \\
&\leq \rho^2\|\zeta_s\|_2^2 + \rho^2\mathbb{E}\ \|\zeta_s\|_2^2 \\
&\leq 2\rho^2,
\end{aligned}$$

where we have used Cauchy-Schwarz inequality and the last inequality comes from the fact that $\|\zeta_s\|_2 = 1$ almost surely. From the derivations above, and choosing $A_s = 2\rho^2 I$, it almost surely holds that $G_s^2 \preceq \sigma_{\max}(G_s)^2 I \preceq 4\rho^4 I = A_s^2$. Moreover, using triangular inequality, it holds that

$$\|\sum_{s\in N_t^c} A_s^2\| \leq \sum_{s\in N_t^c} \|A_s^2\| \leq 4\rho^4|N_t^c|.$$

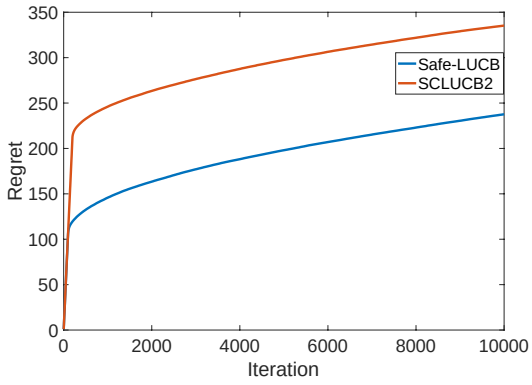

Figure 3: Cumulative regret of SCLUCB2 versus Safe-LUCB in Amani et al. (2019) averaged over 100 realizations.

Now we can apply the matrix Azuma inequality, to conclude that for any $c \geq 0$,

$$\mathbb{P}\left(\lambda_{\max}\Big(\sum_{s \in N_t^c} G_s\Big) \geq c\right) \leq d \exp\left(-\frac{c^2}{32\rho^4 |N_t^c|}\right).$$

Therefore, it holds that with probability $1-\delta$, $\lambda_{\max}\big(\sum_{s \in N_t^c} G_s\big) \leq \sqrt{32\rho^4 |N_t^c| \log(\frac{d}{\delta})}$, and hence with probability $1-\delta$,

$$\lambda_{\min}(V_t) \geq \rho^2 \sigma_\zeta^2 |N_t^c| - \sqrt{32\rho^4 |N_t^c| \log(\frac{d}{\delta})}, \tag{118}$$

or equivalently,

$$\mathbb{P}(\lambda_{\min}(V_t) \leq t) \leq d \exp\left(-\frac{(\rho^2 \sigma_\zeta^2 |N_t^c| - t)^2}{32\rho^4 |N_t^c|}\right), \tag{119}$$

where $\rho = \frac{C}{kBkS}$. This completes the proof.

H.2 Simulation Results

In order to verify our results on the regret bound of SCLUCB2, we plot the Figure 3 which plots the cumulative regret of the two algorithms averaged over 100 realizations. Therefore, the regret of SCLUCB2 matches the proposed problem-dependent upper bound in Amani et al. (2019).