LINEAR THOMPSON SAMPLING UNDER UNKNOWN LINEAR CONSTRAINTS

Ahmadreza Moradipari *Mahnoosh Alizadeh* *Christos Thrampoulidis*

University of California, Santa Barbara

ABSTRACT

We study how adding unknown linear safety constraints affects the performance of Thompson Sampling in the linear stochastic bandit problem. The additional constraints must be met at each round in spite of uncertainty about the environment requiring that the learner acts conservatively in choosing her actions. In this setting, we propose Safe-LTS, the first safe Thompson Sampling based algorithm, and we prove that it achieves no-regret learning. We obtain regrets that have the same dependence on the total number of rounds (modulo logarithmic factors) as Safe-UCB, a recently proposed safe algorithm that uses the upper confidence bound principle. Finally, we provide numerical simulations that demonstrate the efficacy of our algorithm.

***Index Terms*—** multi-armed bandit, Thompson Sampling, stochastic linear bandit, safe learning, online learning.

1. INTRODUCTION

The stochastic multi-armed bandit problem is a sequential decision-making problem where at each round, the learner chooses an action, and observes the corresponding stochastic reward [1]. The learner's goal is to maximize reward over T rounds. In the stochastic linear bandit (LB) problem, the expected value of the reward is an unknown linear function of the action. Two popular approaches have been studied for the LB setting. On the one hand, UCB-based algorithms, build a confidence region over the unknown environment and choose an action that maximizes the expected reward in the most favorable environment in the confidence region [2, 3, 4, 5, 6]. Specifically, Linear UCB (LUCB) achieves regret of order $\mathcal{O}(T^{1/2}\log T)$ [4]. On the other hand, Thompson Sampling (TS) based algorithms, define a prior over the unknown environment, and at each round choose the action that maximizes the expected reward based on a sample from that prior [7, 8, 9, 10, 11, 12, 13, 14, 15]. Specifically, linear TS (LTS) achieves regret of order $\mathcal{O}(T^{1/2}\log^{3/2} T)$ [12, 14].

This paper focuses on the effect of additional *safety constraints* on the performance of Linear TS. The concept of *safe learning* has attracted notable attention in recent years and aims to enable the use of learning algorithms in safety-critical systems with strict reliability requirements that need to be met at all rounds [16, 17, 18, 19, 20]. Focusing specifically on safety constraints in the LB setting, [21] studies a setting where the cumulative reward needs to stay above of a known baseline at each round and proposes a UCB-based algorithm. In this paper, we consider the LB problem with linear stage-wise constraints first proposed in [22, 23]. The authors of [23] propose a safe UCB-based algorithm and provide both a problem dependent regret bound of order $\widetilde{\mathcal{O}}(\sqrt{T})$ and a worst-case bound of order $\widetilde{\mathcal{O}}(T^{2/3})$. None of the aforementioned papers consider safe learning in the context of TS. Compared to UCB-based algorithms, TS-based algorithms have favorable computational features, since at each round, they involve solving a simple optimization with linear objective (rather than one with a possibly hard bilinear objective), e.g. [12]. We propose Safe-LTS, a safe TS algorithm, that provably achieves a $\widetilde{\mathcal{O}}(\sqrt{T})$ problem-dependent bound and a $\widetilde{\mathcal{O}}(T^{2/3})$ worst-case bound.

1.1. Safe stochastic linear bandit problem

We summarize the safe stochastic linear bandit problem recently introduced in [23].

Reward. The learner is given a convex and compact set of actions $\mathcal{D}_0 \in \mathbb{R}^d$. At each round t, by playing an action $x_t \in \mathcal{D}_0$, the learner observes a reward

$$r_t := x_t^\top \theta_\star + \xi_t, \tag{1}$$

which is linear in the fixed but *unknown* parameter $\theta_\star \in \mathbb{R}^d$ with additive zero-mean random noise ξ_t.

Safety constraint. The learning environment is restricted by a linear safety constraint of the form

$$x_t^\top B\theta_\star \leq z, \tag{2}$$

which must be satisfied at every round t. Here, the matrix $B \in \mathbb{R}^{d\times d}$ and the *positive* constant z are known to the learner. Note that in this setting, after playing an action x_t, the value $x^\top B\theta_\star$ is not observed. Given the unknown nature of the parameter $\theta_\star$, in order to not violate (2) with high probability, the learner must choose her actions conservatively.

Goal. The *cumulative pseudo-regret* of the learner up to round T is defined as $R(T) = \sum_{t=1}^{T} x_\star^\top \theta_\star - x_t^\top \theta_\star$, where $x_t, t \in [T]$ are the played actions at each round t, and $x_\star$ is the optimal safe action that maximizes the expected reward, i.e., $x_\star = \arg\max_{x\in\mathcal{D}_0^s} x^\top \theta_\star$, where the safe set of actions $\mathcal{D}_0^s(\theta_\star)$ is defined as

$$\mathcal{D}_0^s(\theta_\star) := \{x \in \mathcal{D}_0 : x^\top B\theta_\star \leq z\}. \tag{3}$$

For simplicity of exposition, we drop the dependence of $\mathcal{D}_0^s(\theta_\star)$ on $\theta_\star$ and refer to the cumulative pseudo-regret simply as regret. The goal of the learner is to control the growth of the pseudo-regret while choosing safe actions.

2. SAFE LINEAR THOMPSON SAMPLING ALGORITHM

We propose a safe version of the linear Thompson Sampling (LTS) that respects the safety constraint (2) at each round t. Specifically, our algorithm includes two distinct so-called *pure exploration* and *exploration-exploitation* phases. In the pure exploration phase, the

This work was supported by NSF grant #1847096 and UCOP grant LFR-18-548175. The authors are with the Department of Electrical and Computer Engineering at the University of California, Santa Barbara. ahmadreza_moradipari, alizadeh, cthrampo@ucsb.edu

algorithm randomly samples a parameter $\tilde{\theta}_t$. Then, it chooses the optimal safe action corresponding to the sampled parameter $\tilde{\theta}_t$ from a known (seed) safe subset $\mathcal{D}^w \subset \mathcal{D}_0$, and observes the corresponding reward. In the exploration-exploitation phase, at any round t, the algorithm samples a perturbed version $\tilde{\theta}_t$ of the *regularized least square* estimate (RLS-estimate) $\hat{\theta}_t$. Then, it chooses the optimal action for the sampled parameter $\tilde{\theta}_t$ while respecting the safety constraint (2). In order to ensure that the safety constraint holds at each round t, the algorithm builds a confidence region $\mathcal{C}_t$ that contains the unknown parameter $\theta_\star$ with high probability. Then, the algorithm ensures safety by choosing actions that satisfy the safety constraint $\forall v \in \mathcal{C}_t$. The summary is presented in Algorithm 1.

Algorithm 1: Safe Linear Thompson Sampling (Safe-LTS)

Input: δ, T, T', λ
Set $\delta' = \frac{\delta}{6T}$
Pure exploration phase:
for $t = 1, \dots, T'$ **do**
  Randomly sample $\tilde{\theta}_t \sim \mathcal{H}^{\text{TS}}$
  Play the following safe action: $x_t = \text{argmax}_{x \in \mathcal{D}^w} x^\top \tilde{\theta}_t$
  Observe reward r_t
end for
Safe exploration-exploitation phase:
for $t = T' + 1, \dots, T$ **do**
  Sample $\eta_t \sim \mathcal{H}^{\text{TS}}$
  Compute the RLS-estimate in (5)
  Set: $\tilde{\theta}_t = \hat{\theta}_t + \beta_t(\delta') A_t^{-\frac{1}{2}} \eta_t$
  Build the confidence region:
  $\mathcal{C}_t(\delta') = \{v \in \mathbb{R} : \left\| v - \hat{\theta}_t \right\|_{A_t} \leq \beta_t(\delta')\}$
  Compute the estimated safe set:
  $\mathcal{D}_t^s = \{x \in \mathcal{D}_0 : x^\top B v \leq z, \forall v \in \mathcal{C}_t(\delta')\}$
  Play the following safe action: $x_t = \text{argmax}_{x \in \mathcal{D}_t^s} x^\top \tilde{\theta}_t$
  Observe reward r_t
end for

2.1. Model Assumptions

Let $\mathcal{F}_t = (\mathcal{F}_1, \sigma(x_1, \dots, x_{t+1}, \xi_1, \dots, \xi_t))$ be the history at round t. We make the following standard assumptions.

Assumption 1. *For all t, given $\mathcal{F}_t$, ξ_t is a zero mean R-sub-Gaussian for a fixed constant $R \geq 0$, i.e., $\mathbb{E}[\xi_t | \mathcal{F}_{t-1}] = 0$ and $\mathbb{E}[e^{\alpha \xi_t} | \mathcal{F}_{t-1}] \leq \exp(\alpha^2 R^2 / 2)$, $\forall \alpha \in \mathbb{R}$.*

Assumption 2. *There exist positive constants S, L such that $\|\theta_\star\|_2 \leq S$ and $\|x\|_2 \leq L, \forall x \in \mathcal{D}_0$. Also, $x^\top \theta_\star \leq 1$, $\forall x \in \mathcal{D}_0$.*

Assumption 3. *The action set $\mathcal{D}_0$ is a compact and convex subset of $\mathbb{R}^d$ that contains the origin.*

2.2. Pure exploration phase

The duration of the pure exploration phase, denoted as T', is provided as an input to the algorithm. For all rounds $t \in [T']$, the algorithm randomly samples a parameter $\tilde{\theta}_t$ from an appropriate *TS distribution* $\mathcal{H}^{\text{TS}}$, which will be specified in Definition 2.1. Then, it chooses the optimal action for the sampled parameter from a *given* safe subset $\mathcal{D}^w \in \mathcal{D}_0^s$ that we define next. The safe action set $\mathcal{D}_0^s$ is unknown to the learner. However, since the unknown parameter is bounded, and the action set $\mathcal{D}_0^s$ is a compact convex body, there always exists a (sufficiently small) ϵ-ball inside $\mathcal{D}_0^s$. We assume that $\mathcal{D}^w$ is that ϵ-ball.

Purely random selection of $\tilde{\theta}_t$'s in this phase lead to random actions x_t. The purely random action-reward pairs (x_t, r_t) allow us to obtain a good estimate of $\theta_\star$. This is important since the accuracy of the estimate of $\theta_\star$ determines which actions can be considered safe. Let $Q(\mathcal{D}^w)$ denote the distribution of the random and independent actions x_t in $\mathcal{D}^w$ chosen in the first phase. For random variables $X \sim Q(\mathcal{D}^w)$, let λ_- be the minimum eigenvalue of the co-variance matrix $\mathbb{E}[XX^\top]$. Since $\mathcal{D}^w$ is an ϵ-ball, we have that $X = \tilde{\theta}/\|\tilde{\theta}\|_2$ for $\tilde{\theta} \sim \mathcal{H}^{TS}$. Hence,

$$\lambda_- := \lambda_{\min}\left(\mathbb{E}[XX^\top]\right) = \epsilon^2 \lambda_{\min}\left(\mathbb{E}\left[\tilde{\theta}\tilde{\theta}^\top / \|\tilde{\theta}\|_2^2\right]\right) > 0. \quad (4)$$

The inequality follows from the anti-concentration property of $\mathcal{H}^{TS}$ (see Definition 2.1).

2.3. Safe exploration-exploitation phase

In this paper, we follow [12, 14] and define LTS as a generic randomized algorithm constructed on the RLS-estimate of the unknown parameter $\theta_\star$ (rather than a Bayesian algorithm that updates a prior distribution every round). Specifically, at each round $t = T'+1, \dots, T$ of the safe exploration-exploitation phase, the Safe-LTS algorithm uses the previous action-observation pairs to compute the Gram matrix A_t and the RLS-estimate $\hat{\theta}_t$ of $\theta_\star$ defined as follows

$$A_t = \lambda I + \sum_{s=1}^{t-1} x_s x_s^\top, \quad \hat{\theta}_t = A_t^{-1} \sum_{s=1}^{t-1} r_s x_s. \quad (5)$$

Based on $\hat{\theta}_t$, the algorithm constructs a confidence region $\mathcal{C}_t(\delta')$ as

$$\mathcal{C}_t(\delta') = \{v \in \mathbb{R} : \left\| v - \hat{\theta}_t \right\|_{A_t} \leq \beta_t(\delta')\}, \quad (6)$$

where $\beta_t(\delta')$ is chosen according to Theorem 2 in [4] in order to ensure that $\theta_\star \in \mathcal{C}_t(\delta')$ with probability at least $1 - \delta'$. In particular, we choose $\beta_t(\delta') = R\sqrt{d \log\left(\frac{1+(t-1)L^2/\lambda}{\delta'}\right)} + \sqrt{\lambda} S$. Then, at any round $t = T' + 1, \dots, T$, the algorithm samples a perturbed parameter $\tilde{\theta}_t$ according to:

$$\tilde{\theta}_t = \hat{\theta}_t + \beta_t(\delta') A_t^{-\frac{1}{2}} \eta_t. \quad (7)$$

Here, η_t is a random sample drawn i.i.d. from the appropriately defined TS distribution $\mathcal{H}^{\text{TS}}$ and rotated by the Gram matrix A_t. Finally, the algorithm chooses the optimal action to play assuming the the true parameter is equal to $\tilde{\theta}_t$, while also respecting the safety constraint (2). As the learner does not know the safe action set $\mathcal{D}_0^s$, it is key to create a conservative inner approximation of $\mathcal{D}_0^s$ based on the confidence region $\mathcal{C}_t(\delta')$ as follows [23]:

$$\mathcal{D}_t^s := \{x \in \mathcal{D}_0 : x^\top B v \leq z, \forall v \in \mathcal{C}_t(\delta')\}. \quad (8)$$

Chosen actions belong this set. Note that, with high probability, $\mathcal{D}_t^s \subseteq \mathcal{D}_0^s$, since its actions are safe with respect to all parameter vectors in $\mathcal{C}_t(\delta')$, and not only for the true parameter $\theta_\star$. This conservative definition of the safe decision set could contribute to the growth of the overall regret. Next, we will study this effect closely.

Considering the original LTS algorithm from a frequentist point of view, [12] showed that as long as η_t is sampled from a distribution $\mathcal{H}^{\text{TS}}$ that satisfies certain anti-concentration and concentration

properties, a regret of order $\mathcal{O}(d^{3/2}\log^{1/2} d\cdot T^{1/2}\log^{3/2} T)$ can be guaranteed. The conditions on η_t provided by [12] ensure that LTS explores *enough but not too much*. However, in a safety-constrained setting, the distributional assumptions proposed by [12] no longer provide sufficient exploration to sufficiently expand the safe set for all problem instances. To address this issue, let us define the following critical parameter referred to as the *safety gap* for a given problem instance:

$$\Delta := z - x_\star^\top B\theta_\star. \tag{9}$$

Note that $\Delta \geq 0$. We provide two problem specific regret bounds that hold for $\Delta > 0$ and $\Delta = 0$, respectively. We show that the following distributional properties for $\mathcal{H}^{\text{TS}}$, provide sufficient exploration in order to bound the regret.

Definition 2.1. $\mathcal{H}^{\text{TS}}$ is a multivariate distribution on $\mathbb{R}^d$, absolutely continuous with respect to the Lebesgue measure which satisfies the following properties for $t \geq T'+1$:

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

$$\mathbb{P}_{\eta_t\sim\mathcal{H}^{\text{TS}}}\left(u^\top\eta_t \geq k_t\right) \geq p. \tag{10}$$

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

$$\mathbb{P}_{\eta_t\sim\mathcal{H}^{\text{TS}}}\left(\|\eta_t\|_2 \leq k_t\sqrt{cd\log(\frac{c'd}{\delta})}\right) \geq 1-\delta, \tag{11}$$

where

$$k_t = \begin{cases} 1 & \text{, if } \Delta > 0, \\ 1+\frac{2}{z}\|B\|_2\sqrt{\frac{\lambda+(t-1)L^2}{\lambda+(T'+\lambda_-)/2}} & \text{, if } \Delta = 0. \end{cases} \tag{12}$$

For example, it can be easily checked that the properties in Definition 2.1 are satisfied by a multivariate i.i.d Gaussian distribution with all entries having a (possibly time-dependent) variance k_t^2.

3. REGRET ANALYSIS

In this section, we study the regret of Safe-LTS. At each round t, the learner chooses her action from the estimated safe set $\mathcal{D}_t^s$ given in (8). To see how this affects the learner's regret, consider the following decomposition of the instantaneous regret $R_t, t \geq T'+1$:

$$R_t := x_\star^\top\theta_\star - x_t^\top\theta_\star = \underbrace{x_\star^\top\theta_\star - x_t^\top\tilde{\theta}_t}_{\text{Term I}} + \underbrace{x_t^\top\tilde{\theta}_t - x_t^\top\theta_\star}_{\text{Term II}} \tag{13}$$

On the one hand, controlling Term II is standard and follows previous results (e.g., [4]). On the other hand, controlling Term I, which is termed *the regret of the safety* in [23], is more challenging. The reason is that in the safe version of TS, $x_\star$ does not generally belong to $\mathcal{D}_t^s$. Our main contribution towards establishing the regret is upper bounding Term I.

Clearly, it would suffice to show that Term I is non-positive. While this is not the case in general, [12] proves that for TS algorithms it suffices to show that the term is non-positive with a constant probability. (We skip the details for brevity; see [12, Appendix D]).) Moreover, [12] proves that this indeed happens in the classical setting with no safety constraints. Our main technical contribution is extending this result to a setting with unknown constraints.

Let $\Theta_t^{\text{opt}} = \{\theta\in\mathbb{R}^d : x_t^\top\tilde{\theta}_t \geq x_\star^\top\theta_\star\}$ be the so-called *set of optimistic parameters*, where $x_t = \arg\max_{x\in\mathcal{D}_t^s} x^\top\tilde{\theta}_t$ is the optimal action for the sampled parameter $\tilde{\theta}_t$ from the conservative decision set $\mathcal{D}_t^s$. LTS is considered optimistic if it samples frequently enough from the set Θ_t^{opt}. The challenge in Safe-LTS is that $\mathcal{D}_t^s \neq \mathcal{D}_0^s$ and so we cannot directly adopt the approach used [12]. In the next lemma, we show that Safe-LTS samples $\tilde{\theta}_t$ from the optimistic set with a positive probability despite the safety constraints. Compared to [12], our proof is simpler and extends to the safe setting.

Lemma 3.1. *Let $\Theta_t^{opt} = \{\theta\in\mathbb{R}^d : x_t^\top\theta \geq x_\star^\top\theta_\star\}$ be the set of optimistic parameters. For $t = T'+1,\dots,T$ and $\tilde{\theta}_t$ defined in (7), $\mathbb{P}(\tilde{\theta}_t\in\Theta_t^{opt}) \geq p$.*

Next, we provide a proof sketch of Lemma 3.1. Simply stated, we need to show that

$$q_t = \mathbb{P}\left(x_t^\top\tilde{\theta}_t \geq x_\star^\top\theta_\star\right) \geq p,$$

where p is strictly positive. Similar to [23], consider an enlarged confidence region $\tilde{\mathcal{C}}_t$ centred on $\theta_\star$ as follows

$$\tilde{\mathcal{C}}_t(\delta') := \{v\in\mathbb{R}^d : \|v-\theta_\star\|_{A_t} \leq 2\beta_t(\delta')\}$$

and the corresponding "shrunk" safe decision set as

$$\begin{aligned}\tilde{\mathcal{D}}_t^s :&= \{x\in\mathcal{D}_0 : x^\top Bv \leq z, \forall v\in\tilde{\mathcal{C}}_t(\delta')\} \\ &= \{x\in\mathcal{D}_0 : x^\top B\theta_\star + 2\beta_t(\delta')\|Bx\|_{A_t^{-1}} \leq z\} \subset \mathcal{D}_t^s\end{aligned}$$

Further define α_t as the largest number in $(0,1]$ such that $\alpha_t x_\star \in \tilde{\mathcal{D}}_t^s$, i.e.,

$$\alpha_t\left(x_\star^\top B\theta_\star + 2\beta_t(\delta')\|Bx_\star\|_{A_t^{-1}}\right) \leq z. \tag{14}$$

By feasibility of $\alpha_t x_\star$, we have that $x_t^\top\tilde{\theta}_t \geq \alpha_t x_\star^\top\tilde{\theta}_t$. Therefore, to prove the lemma, it suffices to show that

$$\mathbb{P}\big(x_\star^\top\tilde{\theta}_t \geq \frac{1}{\alpha_t}x_\star^\top\theta_\star\big) \geq p.$$

Using the definition of $\tilde{\theta}_t = \hat{\theta}_t + \beta_t(\delta')A_t^{-\frac{1}{2}}\eta_t$, it suffices that

$$\mathbb{P}\left(\beta_t(\delta')x_\star^\top A_t^{-\frac{1}{2}}\eta_t \geq x_\star^\top(\theta_\star-\hat{\theta}_t) + (\frac{1}{\alpha_t}-1)x_\star^\top\theta_\star\right) \geq p.$$

At this point, recall from Section 2.3 that $\|\theta_\star - \hat{\theta}_t\|_{A_t} \leq \beta_t(\delta')$. Also, by Assumption 2, $x_\star^\top\theta_\star \leq 1$. All these combined, it remains to show that for any vector $\|u\|_2 = 1$ it holds

$$\mathbb{P}\left(u^\top\eta_t \geq 1 + \frac{1/\alpha_t - 1}{\beta_t(\delta')\|x_\star\|_{A_t^{-1}}}\right) \geq p. \tag{15}$$

In order to show (15), we need to control the term $1/\alpha - 1$. We accomplish this by controling the minimum eigenvalue of the Gram matrix $\lambda_{\min}(A_t)$. In order to do so, we use the fact that the Gram matrix forms a non-decreasing sequence $\lambda_{\min}(A_t) \geq \lambda_{\min}(A_{T'+1}), t \geq T'+1$, and rely on the pure exploration phase to bound $\lambda_{\min}(A_{T'+1})$. In essence, the pure exploration phase helps us develop a sufficiently accurate estimate of $\mathcal{D}_0^s$ for $t \geq T'+1$ [23].

Recall the definition of λ_- in (4). The matrix Chernoff inequality [23, Lemma 1] shows that appropriately choosing T' can $\lambda_{\min}(A_{T'+1}) \geq \lambda + \frac{\lambda_- T'}{2}$. Next, we use this fact to show that (15) holds. We consider separately the cases $\Delta > 0$ and $\Delta = 0$.

3.1. Problem dependent upper bound

For the case where $\Delta > 0$, it can be shown that by choosing an appropriate T', we can guarantee that $x_\star \in \mathcal{D}_t^s$ for all $t = T'+1, \dots, T$. Specifically, [23, Lemma 2] shows that choosing T' of order $\mathcal{O}(\log T)$, ensures that $x_\star \in \mathcal{D}_t^s$ with high probability. In reference to our earlier discussion this means for $t \geq T'+1$, $\alpha_t = 1$. Therefore, the second term inside the probability in (15) vanishes and the desired lower bound follows directly from the anti-concentration property (10) for $k_t = 1$. This completes the proof of Lemma 3.1 for the case $\Delta > 0$.

As mentioned previously, once Lemma 10 is shown to be true, we can adapt the results of [12] to prove a total regret for Safe-LTS of order $\widetilde{\mathcal{O}}(\sqrt{T})$. We summarize the end result next.

Theorem 3.2. *Let Assumptions 1, 2 and 3 holds. For a $\delta \in (0,1)$, with probability $1-\delta$, we have that*

$$R(T) \leq \quad 2T' + (\beta_T(\delta') + \gamma_T(\delta')(1+\frac{4}{p}))\sqrt{2d(T-T')\log(\frac{2TL^2}{d(2\lambda+\lambda_- T')})} + \frac{4\gamma_T(\delta')}{p}\sqrt{\frac{16(T-T')L^2}{2\lambda+\lambda_- T'}\log(\frac{4}{\delta})}, \tag{16}$$

where $\delta' = \frac{\delta}{6T}$, $\gamma_t(\delta) = \beta_t(\delta')\sqrt{cd\log(\frac{c'd}{\delta})}$.

The first part is a trivial bound over the pure exploration phase, since from Assumption 2, we have $x^\top \theta_\star \leq 1, \forall x \in \mathcal{D}_0$. The second part is a regret of safety which is of order $\widetilde{\mathcal{O}}(\sqrt{T})$ when T' is chosen as $\mathcal{O}(\log T)$.

3.2. General upper bound

As it is stated in [23], when the safety gap $\Delta = 0$, there is no guarantee that $x_\star \in \mathcal{D}_t^s$ for $t > T'$. Thus, in order to show that for $t = T'+1, \dots, T$, Safe-LTS samples from the optimistic set with constant probability, we need to appropriately modify the distributional properties for $\mathcal{H}^{\text{TS}}$ in [12] as it is stated in (10) and (11). In order to prove (15) recall that $\alpha_t, t > T'+1$ is such that

$$\frac{1}{\alpha_t} = 1 + \frac{2}{z}\beta_t(\delta')\left\|Bx_\star\right\|_2 \tag{17}$$

Substituting this in (15), it suffices that

$$q_t \geq \mathbb{P}\left(u^\top \eta_t \geq 1 + \frac{2}{z}\frac{\left\|Bx_\star\right\|_{A_t^{-1}}}{\left\|x_\star\right\|_{A_t^{-1}}}\right) \geq p. \tag{18}$$

Thanks to the pure exploration phase, we have

$$\frac{\left\|Bx_\star\right\|_{A_t^{-1}}}{\left\|x_\star\right\|_{A_t^{-1}}} \leq \frac{\left\|Bx_\star\right\|_2\sqrt{\lambda_{\max}(V_t^{-1})}}{\left\|x_\star\right\|_2\sqrt{\lambda_{\min}(V_t^{-1})}} \leq \left\|B\right\|_2\sqrt{\frac{\lambda_{\max}(V_t)}{\lambda_{\min}(V_{T'+1})}} \leq \left\|B\right\|_2\sqrt{\frac{\lambda+(t-1)L^2}{\lambda+(\lambda_- T')/2}}.$$

The last inequality comes from upper bounding the $\lambda_{\max}(V_t)$ and lower bounding $\lambda_{\min}(V_t)$. Specifically, from Lemma 1 in [23], we can lower bound $\lambda_{\min}(V_t)$ with $\lambda_{\min}(V_{T'+1})$. Moreover, by Assumption 2, $\|x\|_2 \leq L$, we can upper bound $\lambda_{\max}(V_t)$ with $\lambda+(t-1)L^2$. From the anti-concentration inequality (10), we can get $q_t \geq p$.

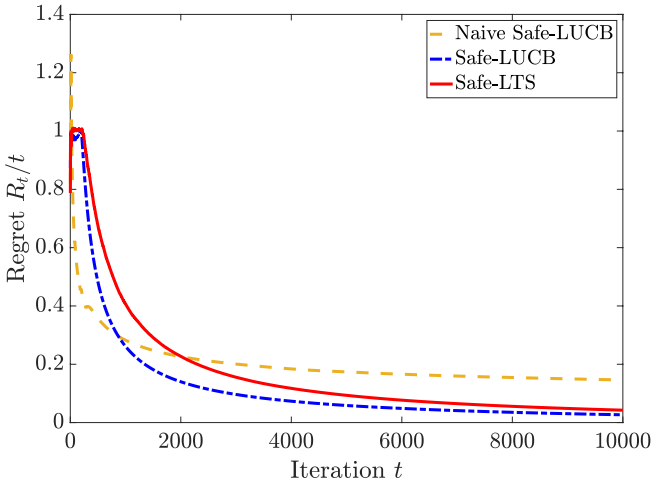

Fig. 1. Comparison of the average per-step regret of Safe-LTS with Safe-LUCB and Naive Safe-LUCB.

Therefore, from [12, Theorem 1] and [23, Theorem 3] we know that with choosing T' of order $\widetilde{\mathcal{O}}(T^{2/3})$, we are able show that when $\Delta = 0$, we can obtain the regret of order $\widetilde{\mathcal{O}}(T^{2/3})$. With replacing $\gamma_T(\delta')$ in (16) with $\psi_T(\delta') = \beta_T(\delta')k_T\sqrt{cd\log(\frac{c'd}{\delta})}$, we can obtain the formal regret bound for the case when $\Delta = 0$.

4. NUMERICAL RESULTS

We evaluate the performance of Safe-LTS by comparing it against 1) Safe-LUCB presented in [23] and 2) Naive Safe-LUCB, which is a modification of the LUCB algorithm presented in [3, 4] that simply requires actions to be chosen from the estimated safe set (but with no pure exploration phase involved). Fig. 1 compares the average per step regret $\frac{R_t}{t}$ of Safe-LTS against that of Safe-LUCB and of Naive Safe-LUCB over 20 problem realizations. The result verifies that Safe-LTS is a no-regret learning algorithm. Also, Safe-LTS and Safe-LUCB are seen to have similar general regret of order $\widetilde{\mathcal{O}}(T^{2/3})$. Moreover, Fig. 1 demonstrates the importance of pure exploration phase considering the performance of Naive Safe-LUCB. For the simulation we consider a time horizon $T = 10000$, $\delta = 1/4T$, and $R = 0.1$. The reward and parameter $\theta_\star$ are sampled from $\mathcal{N}(0, I_4)$, and z is sampled uniformly from $[0,1]$. We consider a decision set $\mathcal{X}_0 = [-1,1]^4$ in $\mathbb{R}^4$. As discussed in [3] LUCB-based algorithms have computational issues with confidence regions defined with 2-norms. Instead, we use Safe-LUCB with modified confidence regions according to the 1-norm; see [23]. On the other hand, Safe-LTS does not suffer from this issue.

5. CONCLUSION

In this paper, we study a linear stochastic bandit problem with unknown safety constraints. These constraints depend on the unknown parameter $\theta_\star$, and must be satisfied at each round. We propose a Thompson Sampling-based algorithm called Safe-LTS, which consists of two phases: a pure exploration phase and a TS-based exploration-exploitation phase. We show regret bounds that depend on a problem-specific parameter referred to as the *safety gap*, Δ. Specifically, when $\Delta > 0$, we show that Safe-LTS has a general regret of order $\widetilde{\mathcal{O}}(\sqrt{T})$, and, when $\Delta = 0$, a regret of order $\widetilde{\mathcal{O}}(T^{2/3})$. An interesting direction for future work would is to investigate whether the worst-case bound for $\Delta = 0$ can be improved.

6. REFERENCES

[1] S. Bubeck, N. Cesa-Bianchi *et al.*, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[3] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," 2008.

[4] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems*, 2011, pp. 2312–2320.

[5] P. Rusmevichientong and J. N. Tsitsiklis, "Linearly parameterized bandits," *Mathematics of Operations Research*, vol. 35, no. 2, pp. 395–411, 2010.

[6] W. Chu, L. Li, L. Reyzin, and R. Schapire, "Contextual bandits with linear payoff functions," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, G. Gordon, D. Dunson, and M. Dudík, Eds., vol. 15. Fort Lauderdale, FL, USA: PMLR, 11–13 Apr 2011, pp. 208–214.

[7] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling," *Mathematics of Operations Research*, vol. 39, no. 4, pp. 1221–1243, 2014.

[8] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *Conference on Learning Theory*, 2012, pp. 39–1.

[9] S. Dong and B. Van Roy, "An information-theoretic analysis for thompson sampling with many actions," in *Advances in Neural Information Processing Systems*, 2018, pp. 4157–4165.

[10] I. Osband and B. Van Roy, "Bootstrapped thompson sampling and deep exploration," *arXiv preprint arXiv:1507.00300*, 2015.

[11] S. Dong, T. Ma, and B. Van Roy, "On the performance of thompson sampling on logistic bandits," *arXiv preprint arXiv:1905.04654*, 2019.

[12] M. Abeille, A. Lazaric *et al.*, "Linear thompson sampling revisited," *Electronic Journal of Statistics*, vol. 11, no. 2, pp. 5165–5197, 2017.

[13] A. Gopalan, S. Mannor, and Y. Mansour, "Thompson sampling for complex online problems," in *International Conference on Machine Learning*, 2014, pp. 100–108.

[14] S. Agrawal and N. Goyal, "Thompson sampling for contextual bandits with linear payoffs," in *International Conference on Machine Learning*, 2013, pp. 127–135.

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

[16] Y. Sui, J. Burdick, Y. Yue *et al.*, "Stagewise safe bayesian optimization with gaussian processes," in *International Conference on Machine Learning*, 2018, pp. 4788–4796.

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

[18] C. J. Ostafew, A. P. Schoellig, and T. D. Barfoot, "Robust constrained learning-based nmpc enabling reliable mobile robot path tracking," *The International Journal of Robotics Research*, vol. 35, no. 13, pp. 1547–1563, 2016.

[19] I. Usmanova, A. Krause, and M. Kamgarpour, "Safe convex learning under uncertain constraints," in *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019, pp. 2106–2114.

[20] N. Tucker, A. Moradipari, and M. Alizadeh, "Constrained thompson sampling for real-time electricity pricing with grid reliability constraints," *arXiv preprint arXiv:1908.07964*, 2019.

[21] A. Kazerouni, M. Ghavamzadeh, Y. Abbasi, and B. Van Roy, "Conservative contextual linear bandits," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 3910–3919. [Online]. Available: http://papers.nips.cc/paper/6980-conservative-contextual-linear-bandits.pdf

[22] A. Moradipari, S. Amani, M. Alizadeh, and C. Thrampoulidis, "Safe linear thompson sampling," *arXiv preprint arXiv:1911.02156*, 2019.

[23] S. Amani, M. Alizadeh, and C. Thrampoulidis, "Linear stochastic bandits under safety constraints," *arXiv preprint arXiv:1908.05814*, 2019.