Safe Linear Thompson Sampling with Side Information

Ahmadreza Moradipari[1], Sanae Amani[1], Mahnoosh Alizadeh[1], Christos Thrampoulidis[1,2]
[1]Department of Electrical and Computer Enginnering, University of California, Santa Barbara
[2]Department of Electrical and Computer Enginnering, University of British Columbia

***Abstract*—The design and performance analysis of bandit algorithms in the presence of stage-wise safety or reliability constraints has recently garnered significant interest. In this work, we consider the linear stochastic bandit problem under additional unknown linear safety constraints that need to be satisfied at each round. For this problem, we present and analyze a new safe algorithm based on linear Thompson Sampling (TS). Our analysis shows that, with high probability, the algorithm chooses actions that are safe at each round and achieve cumulative regret of order $\mathcal{O}(d^{3/2}\log^{1/2} d \cdot T^{1/2}\log^{3/2} T)$. Remarkably, this matches the regret bound provided by [1], [2] for the standard linear TS algorithm in the absence of safety constraints. Also, our analysis highlights how the inherently randomized nature of the TS selection rule suffices to properly expand the set of safe actions that the algorithm has access to at each round. In particular, we compare this behavior to alternative safe algorithms, which typically require distinct rounds of randomization that are dedicated to learning the unknown constraints.**

Index Terms–. Multi-armed bandits, Linear Stochastic Bandits, Safe Learning, Bandits with Safety Constraint.

I. Introduction

The application of stochastic bandit optimization algorithms to safety-critical systems has received significant attention in the past few years. In such cases, the learner repeatedly interacts with a system with an uncertain reward function and operational constraints. In spite of this uncertainty, the learner needs to ensure that her actions do not violate the operational constraints *at any round of the learning process*. As such, especially in the earlier rounds, there is a need to choose actions with caution, while at the same time making sure that the chosen action provide sufficient learning opportunities about the set of safe actions. Notably, the actions deemed safe by the algorithm might not originally include the optimal action. This uncertainty about safety and the resulting conservative behavior means the learner could experience additional regret in such constrained environments.

This paper focuses on linear stochastic bandits (LB) where each action is associated with a feature vector x, and the expected reward of playing each action is equal to the inner product of its feature vector and an unknown parameter vector $\theta^\star$. There exists several variants of LB that study the finite or infinite [3], [4], [5] or time-varying [6], [7] set of actions. Two efficient approaches have been developed: *linear UCB* (LUCB) and *linear Thompson Sampling* (LTS). For LUCB, [5] provides a regret bound of order $\mathcal{O}(d\cdot T^{1/2}\log T)$. For LTS [1], [2] adopt a frequentist view and show regret $\mathcal{O}(d^{3/2}\log^{1/2} d \cdot T^{1/2}\log^{3/2} T)$. Here we provide an LTS algorithm that respects *linear safety constraints* and study its performance. We formally define the problem setting before summarizing our contributions.

This work is supported by NSF grant 1847096 and UCOP grant LFR-18-548175. C. Thrampoulidis was supported in part by NSF grant 1934641. Email addresses: ahmadreza_moradipari@ucsb.edu, samani@ucla.edu, alizadeh@ucsb.edu, cthrampo@ucsb.edu

A. Safe Stochastic Linear Bandit Model

Reward function. The learner is given a convex and compact set of actions $\mathcal{D}_0 \subset \mathbb{R}^d$. At each round t, playing an action $x_t \in \mathcal{D}_0$ results in observing reward $r_t := x_t^\top \theta_\star + \xi_t$, where $\theta_\star \in \mathbb{R}^d$ is a fixed, but *unknown*, parameter and ξ_t is a zero-mean additive noise.

Safety constraint. We further assume that the environment is subject to a linear constraint:

$$x_t^\top \mu_\star \leq C, \tag{1}$$

which needs to be satisfied by the action x_t at every round t, to guarantee safe operation of the system. Here, C is a *positive* constant that is *known* to the learner, while $\mu_\star$ is a fixed, but *unknown* vector parameter. Let us denote the set of "safe actions" that satisfy the constraint (1) as follows:

$$\mathcal{D}_0^s(\mu_\star) := \{x \in \mathcal{D}_0 : x^\top \mu_\star \leq C\}. \tag{2}$$

By having $C > 0$ and further assuming that $0 \in \mathcal{D}_0$, we know that the action 0 is always as safe action. However, beyond that $\mathcal{D}_0^s(\mu_\star)$ is unknown to the learner, since $\mu_\star$ is itself unknown. We consider a bandit-feedback setting in which, at every round t, the learner receives *side information* about the safety set via noisy measurements:

$$w_t = x_t^\top \mu_\star + \zeta_t, \tag{3}$$

where ζ_t is zero-mean additive noise. During the learning process, the learner needs a mechanism that allows her to use the side measurements in (3) for determining the safe set $\mathcal{D}_0^s(\mu_\star)$. This is critical, since it is required (at least with high-probability) that $x_t \in \mathcal{D}_0^s(\mu_\star)$ for all rounds t.

Regret. The *cumulative pseudo-regret* for T rounds is $R(T) = \sum_{t=1}^T x_\star^\top \theta_\star - x_t^\top \theta_\star$, where $x_\star = \arg\max_{x\in\mathcal{D}_0^s(\mu^*)} x^\top \theta_\star$ is the optimal *safe* action that maximizes the expected reward over $D_0^s(\mu_\star)$.

Learning goal. The learner's objective is to control the growth of the pseudo-regret. Moreover, we require that the chosen actions $x_t, t \in [T]$ are safe (i.e., they belong to $\mathcal{D}_0^s(\mu_\star)$ in (2)),

with high probability over T rounds. As is common, we use regret to refer to the pseudo-regret $R(T)$.

Example. We do believe that, albeit simple, linear models for safety constraints could be directly relevant in traditionally advocated applications of bandit problems such as medical trials applications [8], recommendation systems [9], and ad placement [10]. Even in more complex settings where linear models are not directly applicable, we still believe that this is the appropriate first step towards a principled study of the performance of safe algorithms.

As a concrete motivation example of our setting, consider medical trials, a problem traditionally advocated as an application area for linear bandits, where the effect of different therapies is unknown a-priori to the doctors and can only be determined through clinical trials. Free exploration is not possible, since it may lead to actions that cause harm to the patient, an outcome to be avoided at all times. To model this, we pick the unknown parameter $\mu_\star$ so as to represent the patients' response, and the known parameter C so as to represent a safety threshold that doctors need to account for. The hazard-threshold C can be assumed known as it is the same for all patients (and can be estimated from existing data). In this example, actions x_t represent selected therapies at time t (e.g., drug-dosage) and we assume that a (conservative) safe seed set of harmless (but, plausibly not efficient) therapies is known to the doctor. Overall, while doctors try to select therapies (x_t) with high reward (which could be a signal that shows improvement in patient's health condition), they should not violate the safety constraint $x_t^\top \mu_\star \leq C$ at any time.

B. Contributions

- We provide the first *safe* Linear Thompson Sampling (Safe-LTS) algorithm with provable regret guarantees for the linear bandit problem with linear safety constraints.
- Our analysis shows that Safe-LTS achieves the *same* order $\mathcal{O}(d^{3/2}\log^{1/2} d \cdot T^{1/2}\log^{3/2} T)$ of regret as the original LTS (without safety constraints) [2]. Hence, the dependence of our regret bound on the time horizon T *cannot* be improved modulo logarithmic factors (see lower bounds in [3], [4]).
- We compare Safe-LTS to existing safe versions of LUCB. We show that our algorithm has: better regret in the worst-case, fewer parameters to tune and superior empirical performance.
- We propose a heuristic modification to our Safe-LTS algorithm that adapts a *dynamic noise-distribution scheme* and is shown empirically to outperform the latter. This idea might also be relevant in the unconstrained linear bandit setting.

On a technical level, to derive Safe-LTS and its regret bound, need to properly account for the fact that the optimal safe action $x_\star$ is *not* necessarily in the estimated safe decision set (see Eqn. (8) for formal definition) at each round t. This is because, at each time step, we only have an estimate of the unknown parameter μ_*, thus the estimated set is only a conservative inner approximation of the actual safe set in (2). Consequently, we need to design an action selection rule that is simultaneously: (i) *Frequently optimistic in spite of limitations on actions imposed because of safety*. Here, we achieve this by appropriately tuning the randomization of Thompson Sampling. Specifically, through a careful analysis, essentially controlling the distance of the optimal action $x_\star$ from the estimated safe set, we find that the appropriate tuning involves scaling with a simple function of the problem parameters including the safety constant C. (ii) Guarantees a *proper expansion of the estimated safe set* so as to not exclude *good* actions for a long time, leading to large *regret of safety*. Here, we show that it is the **randomized** nature of LTS that achieves this second goal, and this is exactly where the LUCB action selection rule seems to fail to provide the same guarantees.

C. Other Related Work

Multi-armed Bandits (MAB) - Two popular algorithms have been studied in MAB in order to capture the trade-off between exploration and exploitation in sequential decision making problems: 1) Upper confidence bound (UCB), which consists of choosing the optimal action according to the upper-confidence bounds on the true parameter (i.e., $\theta_\star$) [11]; 2) Thompson Sampling (TS), which samples the true parameter from a prior distribution, and selects the optimal action with respect to the sampled parameter [12]. Moreover, [13] considers a new approach to the MAB problem based on Deterministic Sequencing of Exploration and Exploitation (DSEE). In particular, they divide time horizon to the pure exploration phase and pure exploitation phase. In the former, the player plays all arms in a round-robin fashion. In the latter, the player plays the arm with the largest sample mean. [14], [15] study the MAB problem in the multiplayer settings where a team of agents cooperate on a network in order to maximize their collective reward. In [16], [17], they study the multi-objective MAB problem where the components of the reward signal correspond to different objectives. They evaluate the performance of their algorithm with notions of 2-D regret and Pareto regret. Other lines of works have studied best-arm identification problem in MAB that aims to identify the arm with the largest expected regret [18] as well as cascading bandits where the goal is to learn arms in order to rank them based on the users preferences such as recommendation systems [19]. In [20], [21], they study the MAB problem given adversarial attacks, where the adversary can change the action selected by the learner, and they propose a robust algorithm for the case that the total attack cost is given. Also, [22] studies the MAB problem in the case that the statistical rewards of different arms may be correlated. In particular, they study the regional bandits problem where the arms belong to different groups such the expected reward of the arms in a same group is a function of the common parameter, and the parameters are independent across different groups. Another line of work focuses on the design of risk-sensitive algorithms [23], [24], [25]. In particular, in economic and finance applications, the learner may be more interested in reducing the uncertainty (i.e., risk) in the outcome, rather than achieving the highest cumulative reward [26], [27].

Safety - A diverse body of related works on stochastic optimization and control have considered the effect of safety constraints that need to be met during the run of the algorithm [28], [29] and references therein. Closely related to our work, [30], [31] study *nonlinear* bandit optimization with *nonlinear* safety

constraints using Gaussian processes (GPs) as non-parametric models for both the reward and the constraint functions. Their algorithms have shown great promises in robotics applications [32], [33]. Without the GP assumption, [8] proposes and analyzes a safe variant of the Frank-Wolfe algorithm to solve a smooth optimization problem with an unknown convex objective function and unknown *linear* constraints (with side information, similar to our setting). All the above algorithms come with provable convergence guarantees, but *no* regret bounds. To the best of our knowledge, the first work that derived an algorithm with provable regret guarantees for bandit optimization with stage-wise safety constraints, as the ones imposed on the aforementioned works, is [34]. While [34] restricts attention to a *linear* setting, their results reveal that the presence of the safety constraint –even though linear– can have a non-trivial effect on the performance of LUCB-type algorithms. Specifically, the proposed Safe-LUCB algorithm comes with a problem-dependent regret bound that depends critically on the location of the optimal action in the safe action set – increasingly so in problem instances for which the safety constraint is active. In [34], the linear constraint function involves the same unknown vector (say, $\theta_\star$) as the one that specifies the linear reward. Instead, in Section I-A we allow the constraint to depend on a new parameter vector (say, $\mu_\star$) to which the learner get access via side-information measurements (3). This latter setting is the direct *linear* analogue to that of [30], [31], [8] and we demonstrate that an appropriate Safe-LTS algorithm enjoys regret guarantees of the same order as the original LTS *without* safety constraints. A more elaborate comparison to [34] is provided in Section IV-C. In contrast to the previously mentioned references, another recent work [9] defines safety as the requirement of ensuring that the *cumulative* (linear) reward up to each round stays *above* a given percentage of the performance of a known *baseline* policy. A "stage-wise" variant of this type of constraints was recently studied in another interesting work [35]. Compared to [9], the setting of [35] is closer to ours, but there are still some key differences. Most notably, in contrast to [35], the constraint studied here is such that the optimal action $x_\star$ is not guaranteed to be in the estimated safe-set (especially at early rounds t). Because of this, the analysis of [35] is not directly applicable here. On a technical side, [35] proves a bound on the expected reward (but they restrict actions to an ellipsoidal). Instead, we present a high-probability bound on the regret similar to [9], [34].
Also, it is worth mentioning that the algorithms presented in [34], [35] require distinct rounds of randomization that are dedicated to learning the unknown constraints. Instead, our analysis shows that the inherent randomization of the TS action selection rule suffices for this purpose. As a closing remark, [34], [9], [8], [35], [36] show that simple linear models for safety constraints might be directly relevant to several applications such as medical trials, recommendation systems or managing the customers' demand in power-grid systems. Moreover, even in more complex settings where linear models do not directly apply (e.g., [32], [33]), we believe that this simplification is an appropriate first step towards a principled study of regret behavior of safe algorithms in sequential decision settings.

Thompson Sampling - Even though TS-based algorithms [37] are computationally easier to implement than UCB-based algorithms and have shown great empirical performance, they were largely ignored by the academic community until a few years ago, when a series of papers [38], [2], [12], [39] showed that TS achieves optimal performance in both frequentist and Bayesian settings. Most of the literature focused on the analysis of the Bayesian regret of TS for general settings such as linear bandits or reinforcement learning (see e.g., [40]). More recently, [41], [42], [43] provided an information-theoretic analysis of TS. Additionally, [44] provides regret guarantees for TS in the finite and infinite MDP setting. Another notable paper is [45], which studies the stochastic MAB problem in complex action settings providing a regret bound that scales logarithmically in time with improved constants. None of these papers study the performance of TS for LB with safety constraints.

II. Safe Linear Thompson Sampling

Our proposed algorithm is a safe variant of Linear Thompson Sampling (LTS). At any round t, given a regularized least-squares (RLS) estimate $\hat{\theta}_t$, the algorithm samples a perturbed parameter $\tilde{\theta}_t$ that is appropriately distributed to guarantee sufficient exploration. Considering this sampled $\tilde{\theta}_t$ as the true environment, the algorithm chooses the action with the highest possible reward while making sure that the safety constraint (1) holds. The presence of the safety constraint complicates the learner's choice of actions. In order to ensure that actions remain safe at all rounds, the algorithm uses the side-information (3) to construct a confidence region $\mathcal{C}_t$, which contains the unknown parameter $\mu_\star$ with high probability. With this, it forms an *inner* approximation $\mathcal{D}_t^s$ of the safe set, which is composed by all actions x_t that satisfy the safety constraint *for all* $v \in \mathcal{C}_t$. The summary is presented in Algorithm 1 and a detailed description follows.

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

Input: δ, T, λ. Set $\delta' = \frac{\delta}{6T}$
for $t = 1, \dots, T$ **do**
  Sample $\eta_t \sim \mathcal{H}^{\text{TS}}$
  Set $V_t = \lambda I + \sum_{s=1}^{t-1} x_s x_s^\top$ and compute RLS-estimates $\hat{\theta}_t$ and $\hat{\mu}_t$
  Set: $\tilde{\theta}_t = \hat{\theta}_t + \beta_t(\delta') V_t^{-\frac{1}{2}} \eta_t$
  Build the confidence region: $\mathcal{C}_t(\delta') = \{v \in \mathbb{R} : \|v - \hat{\mu}_t\|_{V_t} \le \beta_t(\delta')\}$
  Compute the estimated safe set: $\mathcal{D}_t^s = \{x \in \mathcal{D}_0 : x^\top v \le C, \forall v \in \mathcal{C}_t(\delta')\}$
  Play the following action: $x_t = \arg\max_{x \in \mathcal{D}_t^s} x^\top \tilde{\theta}_t$
  Observe reward r_t and measurement w_t

A. Model assumptions

Notation. $[n]$ denotes the set $\{1, 2, \dots, n\}$. The Euclidean norm of a vector x is denoted by $\|x\|_2$. Its weighted ℓ_2-norm with respect to a positive semidefinite matrix V is

denoted by $\|x\|_V = \sqrt{x^\top V x}$. We also use the standard $\widetilde{\mathcal{O}}$ notation that ignores poly-logarithmic factors. Finally, for ease of notation, from now on-wards we refer to the safe set in (2) by $\mathcal{D}_0^s$ and drop the dependence on $\mu_\star$. Let $\mathcal{F}_t = (\mathcal{F}_1, \sigma(x_1, \ldots, x_t, \xi_1, \ldots, \xi_t, \zeta_1, \ldots, \zeta_t))$ denote the filtration representing the accumulated information up to round t. We also introduce standard assumptions on the problem as follows.

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

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

Assumption 3. *The action set $\mathcal{D}_0$ is a star-convex subset of $\mathbb{R}^d$ and contains the origin. We assume $\|x\|_2 \leq L$, $\forall x \in \mathcal{D}_0$.*

It is straightforward to generalize our results when the sub-Gaussian constants of ξ_t and ζ_t and/or the upper bounds on $\|\theta_\star\|_2$ and $\|\mu_\star\|_2$ are different. Throughout, we assume they are equal, for brevity.

B. Algorithm description and discussion

Let $\{x_i\}_{i\in[t]}$ be the sequence of actions and $\{r_i\}_{i\in[t]}$, $\{w_i\}_{i\in[t]}$ be the corresponding rewards and side-information measurements. For any $\lambda > 0$, the RLS-estimates $\hat{\theta}_t$ of $\theta_\star$ and $\hat{\mu}_t$ of $\mu_\star$ are $\hat{\theta}_t = V_t^{-1}\sum_{s=1}^{t-1} r_s x_s$, $\hat{\mu}_t = V_t^{-1}\sum_{s=1}^{t-1} w_s x_s$, where $V_t = \lambda I + \sum_{s=1}^{t-1} x_s x_s^\top$. Based on $\hat{\theta}_t$ and $\hat{\mu}_t$, we construct two confidence regions $\mathcal{E}_t := \mathcal{E}_t(\delta')$ and $\mathcal{C}_t := \mathcal{C}_t(\delta')$ as follows:

$$\mathcal{E}_t := \{\theta \in \mathbb{R}^d : \left\|\theta - \hat{\theta}_t\right\|_{V_t} \leq \beta_t(\delta')\}, \tag{4}$$

$$\mathcal{C}_t := \{v \in \mathbb{R}^d : \|v - \hat{\mu}_t\|_{V_t} \leq \beta_t(\delta')\}. \tag{5}$$

Both $\mathcal{E}_t$ and $\mathcal{C}_t$ depend on δ', but we will often suppress notation for simplicity. The ellipsoid radius β_t is properly chosen as in [5] in order to guarantee that $\theta_\star \in \mathcal{E}_t$ and $\mu_\star \in \mathcal{C}_t$ with high probability.

Theorem II.1. *Let Assumptions 1-2 hold. For $\delta \in (0,1)$, and $\beta_t(\delta) = R\sqrt{d\log\left(\frac{1+tL^2}{\delta}\right)} + \sqrt{\lambda}S$, with probability at least $1-\delta$, it holds that $\theta_\star \in \mathcal{E}_t(\delta)$ and $\mu_\star \in \mathcal{C}_t(\delta)$, $\forall t \geq 1$.*

1) Background on LTS: a frequently optimistic algorithm: Our algorithm inherits the frequentist view of LTS first introduced in [1], [2], which is essentially defined as a *randomized* algorithm over the RLS-estimate $\hat{\theta}_t$ of the unknown parameter $\theta_\star$. Specifically, at any round t, the randomized algorithm of [1], [2] samples a parameter $\tilde{\theta}_t$ centered at $\hat{\theta}_t$:

$$\tilde{\theta}_t = \hat{\theta}_t + \beta_t(\delta') V_t^{-\frac{1}{2}} \eta_t, \tag{6}$$

and chooses the action x_t that is best with respect to the new sampled parameter, i.e., maximizes the objective $x_t^\top \tilde{\theta}_t$. The key idea of [1], [2] on how to select the random perturbation $\eta_t \in \mathbb{R}^d$ to guarantee good regret performance is as follows. On the one hand, $\tilde{\theta}_t$ must stay close enough to the RLS-estimate $\hat{\theta}_t$ so that $x_t^\top \tilde{\theta}_t$ is a good proxy for the true (but unknown) reward $x_t^\top \theta_\star$. Thus, η_t must satisfy an appropriate *concentration* property. On the other hand, $\tilde{\theta}_t$ must also favor exploration in a sense that it leads –often enough– to actions x_t that are *optimistic*, i.e., they satisfy

$$x_t^\top \tilde{\theta}_t \geq x_\star^\top \theta_\star \tag{7}$$

Thus, η_t must satisfy an appropriate *anti-concentration* property. Algorithm 1 also builds on these two key ideas, but the safe setting imposes additional challenges that we need to address.

2) Addressing challenges in the safe setting: Compared to the classical linear bandit setting [1], [2], the presence of the safety constraint raises the following two questions: *(i)* How to guarantee actions played at each round are safe? *(ii)* In the face of the safety restrictions, how can optimism (cf. (7)) be maintained? In the rest of this section, we explain the mechanisms that Safe-LTS employs to address both of these challenges.

Safety - First, the chosen action x_t at each round need not only maximize $x_t^\top \tilde{\theta}_t$, but also, it needs to be safe. Since the learner does not know the safe action set $\mathcal{D}_0^s$, Algorithm 1 performs conservatively and guarantees safety as follows. After creating the confidence region $\mathcal{C}_t$ around the RLS-estimate $\hat{\mu}_t$, it forms the so-called *safe decision set at round t* denoted as $\mathcal{D}_t^s$:

$$\mathcal{D}_t^s = \{x \in \mathcal{D}_0 : x^\top v \leq C, \forall v \in \mathcal{C}_t\}. \tag{8}$$

Then, the chosen action is optimized over only the subset $\mathcal{D}_t^s$, i.e.,

$$x_t = \arg\max_{x \in \mathcal{D}_t^s} x^\top \tilde{\theta}_t. \tag{9}$$

We make the following two remarks about $\mathcal{D}_t^s$. On a positive note, $\mathcal{D}_t^s$ is easy to compute:

$$\mathcal{D}_t^s := \{x \in \mathcal{D}_0 : x^\top v \leq C, \forall v \in \mathcal{C}_t\} \tag{10}$$

$$= \{x \in \mathcal{D}_0 : \max_{v \in \mathcal{C}_t} x^\top v \leq C\} \tag{11}$$

$$= \{x \in \mathcal{D}_0 : x^\top \hat{\mu}_t + \beta_t(\delta') \|x\|_{V_t^{-1}} \leq C\}. \tag{12}$$

Indeed, the optimization in (9) is an efficient convex quadratic program. Yet, the challenge remains that $\mathcal{D}_t^s$ contains actions which are safe with respect to *all* the parameters in $\mathcal{C}_t$, and not only $\mu_\star$. As such, it is only an *inner* approximation of the true safe set $\mathcal{D}_0^s$. As we will see next, this fact complicates the requirement for optimism.

Optimism in the face of safety - The fact that $\mathcal{D}_t^s$ is only an inner approximation of $\mathcal{D}_0^s$ makes it harder to maintain optimism of x_t as defined in (7). To see this, note that in the classical setting, the algorithm of [2] would choose x_t as the action that maximizes $\tilde{\theta}_t$ over the *entire* set $\mathcal{D}_0$. In turn, this would imply that $x_t^\top \tilde{\theta}_t \geq x_\star^\top \tilde{\theta}_t$ because $x_\star$ belongs to the feasible set $\mathcal{D}_0$. This observation is the critical first argument in proving that x_t is optimistic often enough, i.e., (7) holds with fixed probability $p > 0$. Unfortunately, in the presence of safety constraints, x_t is a maximizer over only the subset $\mathcal{D}_t^s$. Since $x_\star$ may *not* lie within $\mathcal{D}_t^s$, there is no guarantee that $x_t^\top \tilde{\theta}_t \geq x_\star^\top \tilde{\theta}_t$ as before. So, how does then one guarantee optimism?

Intuitively, at the first rounds, the estimated safe set $\mathcal{D}_t^s$ is only a small subset of the true $\mathcal{D}_0^s$. Thus, $x_t \in \mathcal{D}_t^s$ is a vector

of small norm compared to that of $x_\star \in \mathcal{D}_0^s$. Thus, for (7) to hold, it must be that $\tilde{\theta}_t$ is not only in the direction of $\theta_\star$, but it also has larger norm than that. To satisfy this latter requirement, the random vector η_t must be large; hence, it will "anti-concentrate more". As the algorithm progresses, and –thanks to side-information measurements– the set $\mathcal{D}_t^s$ becomes an increasingly better approximation of $\mathcal{D}_0^s$, the requirements on anti-concentration of η_t become the same as if no safety constraints were present. Overall, at least intuitively, we might hope that optimism is possible in the face of safety, but only provided that η_t is set to satisfy a stronger (at least at the first rounds) anti-concentration property than that required by [2] in the classical setting.

At the heart of Algorithm 1 and its proof of regret lies an analytic argument that materializes the intuition described above. Specifically, we will prove that optimism is possible in the presence of safety at the cost of a stricter anti-concentration property compared to that specified in [2]. While the proof of this fact is deferred to Section III-A, we now summarize the appropriate distributional properties that provably guarantee good regret performance of Algorithm 1 in the safe setting.

Definition II.1. In Algorithm 1, the random vector η_t is sampled IID at each t from a distribution $\mathcal{H}^{\text{TS}}$ on $\mathbb{R}^d$ that is absolutely continuous with respect to the *Lebesgue* measure and satisfies:

Anti-concentration: There exists constant $p > 0$ such that for any $u \in \mathbb{R}^d$ with $\|u\|_2 = 1$,

$$\mathbb{P}\big(u^\top \eta_t \geq 1 + \frac{2}{C} LS\big) \geq p. \tag{13}$$

Concentration: There exists positive constants $c, c' > 0$ such that $\forall \delta \in (0,1)$,

$$\mathbb{P}\big(\|\eta_t\|_2 \leq \big(1 + \frac{2}{C} LS\big) \sqrt{cd \log\big(\frac{c'd}{\delta}\big)} \big) \geq 1 - \delta. \tag{14}$$

In particular, the difference to the distributional assumptions required by [2] in the classical setting is the extra term $\frac{2}{C} LS$ in (13) (naturally, the same term affects the concentration property (14)). Our proof of regret in Section III shows that this extra term captures an appropriate notion of the distance between the approximation $\mathcal{D}_t^s$ (where x_t lives) and the true safe set $\mathcal{D}_0^s$ (where $x_\star$ lives), and provides enough exploration for the sampled parameter $\tilde{\theta}_t$ so that actions in $\mathcal{D}_t^s$ can be optimistic. While this intuition can possibly explain the need for an additive term in Definition II.1, it is insufficient when it comes to determining its "correct" value. This is determined by our analytic treatment in Section III-A.

Finally, we remark that it is not hard to construct distributions that simultaneously satisfy the two conditions in (13) and (14). For example, a multivariate zero-mean IID Gaussian distribution with all entries having a (possibly time-dependent) variance $(1 + \frac{2}{C} LS)^2$ satisfies the Definition II.1 and can be chosen to sample η_t in Algorithm 1 from it.

III. Regret Analysis

Here, we present a tight regret bound for Safe-LTS by proving that its action selection rule is simultaneously: 1) frequently optimistic, and, 2) guarantees a proper expansion of the estimated safe set. Our main result Theorem III.1 is perhaps surprising: in spite of the additional safety constraints, Safe-LTS has regret $\mathcal{O}(d^{3/2} \log^{1/2} d \cdot T^{1/2} \log^{3/2} T)$ that is order-wise the same as that in the classical setting [1], [2].

Theorem III.1 (Regret of Safe-LTS)**.** *Let $\lambda \geq 1$ and Assumptions 1, 2, 3 hold. Fix $\delta \in (0,1)$. Then, with probability at least $1 - \delta$, Safe-LTS is safe and its regret is upper bounded as follows:*

$$R(T) \leq \big(\beta_T(\delta') + \gamma_T(\delta')(1 + \frac{4}{p})\big) \sqrt{2Td \log\big(1 + \frac{TL^2}{\lambda}\big)} + \frac{4\gamma_T(\delta')}{p} \sqrt{\frac{8TL^2}{\lambda} \log \frac{4}{\delta}}, \tag{15}$$

where $\delta' = \frac{\delta}{6T}$, $\beta_t(\delta')$ as in Theorem II.1 and, $\gamma_t(\delta') = \beta_t(\delta')\big(1 + \frac{2}{C} LS\big) \sqrt{cd \log\big(\frac{c'd}{\delta'}\big)}$.

The theorem above provides guarantees both on the safety of the actions chosen by Safe-LTS Algorithm 1, as well as, on its regret.

First, we comment on the safety of the actions, which is ensured by construction of the algorithm as discussed in Section II-B. Formally, fix a desired δ and set $\delta' = \frac{\delta}{6T}$. Consider any time $t \in [T]$. On the one hand, from Theorem II.1, it holds that $\mathbb{P}(\mu_\star \in \mathcal{C}_t(\delta')) \geq 1 - \delta'$. On the other hand, by construction (lines 7-8, Algorithm 1), Safe-LTS guarantees that x_t at time t belongs to $\mathcal{D}_t^s$, i.e., $x_t^\top v \leq C, \forall v \in C_t(\delta')$. Putting these two together shows that $\mathbb{P}(x_t^\top \mu_\star \leq C) \geq 1 - \delta'$. Then, a union bound (see Lemma VI.4) over all time steps from 1 to T proves that $\mathbb{P}(\forall t \in [T] : x_t^\top \mu_\star \leq C) \geq 1 - T\delta' \geq 1 - \frac{\delta}{6}$, i.e., Safe-LTS is with high probability at least $1 - \delta$ safe at all rounds.

Next, we discuss the regret bound of Theorem III.1, which requires a careful analysis. The detailed proof is in given in App. VII. In the rest of the section, we highlight the key changes compared to [1], [2] that occur due to the safety constraint. To begin, let us consider the following standard decomposition of the cumulative regret

$$R(T) = \sum_{t=1}^{T} \big(\underbrace{x_\star^\top \theta_\star - x_t^\top \tilde{\theta}_t}_{\text{Term I}} \big) + \sum_{t=1}^{T} \big(\underbrace{x_t^\top \tilde{\theta}_t - x_t^\top \theta_\star}_{\text{Term II}} \big). \tag{16}$$

Regarding Term II, the concentration property of $\mathcal{H}^{TS}$ guarantees that $\tilde{\theta}_t$ is close to $\hat{\theta}_t$, and consequently, close to $\theta_\star$ thanks to Theorem II.1. Therefore, controlling Term II can be done similar to previous works e.g., [5], [2]; see App. VII-B for more details. Next, we focus on Term I.

To see how the safety constraints affect the proofs let us first review the treatment of Term I in the classical setting. For UCB-type algorithms, Term I is always non-positive since the pair $(\tilde{\theta}_t, x_t)$ is optimistic at each round t by design [3], [4], [5]. For LTS, Term I can be positive; that is, (7) may not hold at every round t. However, [1], [2] proved that thanks to the anti-concentration property of η_t, this optimistic property occurs often enough. Our main technical contribution, detailed in the next section, is to show that the properly modified anti-concentration property in Definition II.1 together with

the construction of approximated safe sets as in (12) can yield frequently optimistic actions even in the face of safety. Specifically, it is the extra term $\frac{2}{C}LS$ in (13) that allows enough exploration to the sampled parameter $\tilde{\theta}_t$ in order to compensate for safety limitations on the chosen actions, and because of that we are able to show Safe-LTS obtains the same order of regret as that of [2]. After that, in Section III-B, we show that we can bound the overall regret of Term I with the V_τ norm of the optimistic actions.

As a closing remark, we note that our proof of optimism in the face of safety directly applies as is above to a scenario where the constraint and the reward function are parameterized by the same vector $\theta_\star$, i.e., the constraint is of the form $x_t^\top \theta_\star \leq C$. In this case, obviously, no side information is required and we can show the same order of regret as in Theorem III.1. Please see Section IV-C for a discussion on how this result improves upon that of [34] who studied constraints parameterized by $\theta_\star$.

A. Proof sketch: Optimism despite safety constraints

We prove that $\tilde{\theta}_t$ is optimistic with constant probability (see App. VI for formal statement and proof).

Lemma III.2. *(Optimism in the face of safety; Informal) For any $t \geq 1$, Safe-LTS samples parameter $\tilde{\theta}_t$ and chooses action x_t such that the pair $(\tilde{\theta}_t, x_t)$ is optimistic frequently enough, i.e., $\mathbb{P}\left(x_t^\top \tilde{\theta}_t \geq x_\star^\top \theta_\star\right) \geq p$, where $p > 0$ is the probability of the anti-concentration property (13).*

The challenge in the proof is that x_t is chosen from $\mathcal{D}_t^s$, which does not necessarily contain all feasible actions and hence, *may not contain* $x_\star$. Thus, we need a mechanism to control the distance of $x_\star$ from the optimistic actions that can only lie within the subset $\mathcal{D}_t^s$ (*distance* is defined here in terms of an inner product with the optimistic parameters $\tilde{\theta}_t$). Unfortunately, we do not have a direct control on this distance term and so at the heart of the proof lies the idea of identifying a "good" feasible action $\tilde{x}_t \in \mathcal{D}_t^s$ whose distance to $x_\star$ is easier to control. To be concrete, we show that it suffices to choose the good feasible point in the direction of $x_\star$, i.e., $\tilde{x}_t = \alpha_t x_\star$, where the key parameter $\alpha_t \in (0, 1]$ must be set to satisfy $\tilde{x}_t \in \mathcal{D}_t^s$. Naturally, the value of α_t is determined by the approximated safe set $\mathcal{D}_t^s$ as defined in (12). The challenge though is that we do not know how the value of $x_*^\top \hat{\mu}_t$ compares to the constant C. We circumvent this issue by introducing an enlarged confidence region centered at $\mu_\star$ as $\tilde{\mathcal{C}}_t := \{v \in \mathbb{R}^d : \|v - \mu_\star\|_{V_t} \leq 2\beta_t(\delta')\}$, and the corresponding shrunk safe decision set as

$$\begin{aligned}\tilde{\mathcal{D}}_t^s &:= \{x \in \mathcal{D}_0 : x^\top v \leq C, \forall v \in \tilde{\mathcal{C}}_t\} \\ &= \{x \in \mathcal{D}_0 : x^\top \mu_\star + 2\beta_t(\delta') \|x\|_{V_t^{-1}} \leq C\} \subseteq \mathcal{D}_t^s. \quad (17)\end{aligned}$$

$\tilde{\mathcal{D}}_t^s$ is defined with respect to an ellipsoid centered at $\mu_\star$ (rather than at $\hat{\mu}_t$). This is convenient since $x_\star^\top \mu_\star \leq C$. Using this, it can be easily checked that $\alpha_t = \left(1 + \frac{2}{C}\beta_t(\delta') \|x_\star\|_{V_t^{-1}}\right)^{-1}$ ensures $\alpha_t x_\star \in \tilde{\mathcal{D}}_t^s \subseteq \mathcal{D}_t^s$. From this, and optimality of $x_t = \arg\max_{x \in \mathcal{D}_t^s} x^\top \tilde{\theta}_t$ we have that

$$x_t^\top \tilde{\theta}_t \geq \alpha_t x_\star^\top \tilde{\theta}_t. \quad (18)$$

Using (18), it suffices to prove that $p \leq \mathbb{P}\big(\alpha_t x_\star^\top \tilde{\theta}_t \geq x_\star^\top \theta_\star\big) = \mathbb{P}\big(x_\star^\top \tilde{\theta}_t \geq x_\star^\top \theta_\star + \frac{2}{C}\beta_t(\delta') \|x_\star\|_{V_t^{-1}} x_\star^\top \theta_\star\big)$, where, the equality follows by definition of α_t. To continue, recall that $\tilde{\theta}_t = \hat{\theta}_t + \beta_t V_t^{-\frac{1}{2}} \eta_t$. Thus, the probability we want to lower bound can be equivalently rewritten as

$$\mathbb{P}\big(\beta_t(\delta') x_\star^\top V_t^{-\frac{1}{2}} \eta_t \geq x_\star^\top (\theta_\star - \hat{\theta}_t) + \frac{2}{C}\beta_t(\delta') \|x_\star\|_{V_t^{-1}} x_\star^\top \theta_\star\big).$$

To simplify the above, we use (i) $|x_\star^\top \theta_\star| \leq \|x_\star\|_2 \|\theta_\star\|_2 \leq LS$; (ii) $x_\star^\top (\theta_\star - \hat{\theta}_t) \leq \|x_\star\|_{V_t^{-1}} \|\theta_\star - \hat{\theta}_t\|_{V_t} \leq \beta_t(\delta') \|x_\star\|_{V_t^{-1}}$, because of Cauchy-Schwartz and Theorem II.1. Put together, we need that $p \leq \mathbb{P}\big(\beta_t(\delta') x_\star V_t^{-\frac{1}{2}} \eta_t \geq \beta_t(\delta') \|x_\star\|_{V_t^{-1}} + \frac{2}{C} LS \beta_t(\delta') \|x_\star\|_{V_t^{-1}}\big)$, or equivalently,

$$p \leq \mathbb{P}\big(u_t^\top \eta_t \geq 1 + (2/C)LS\big), \quad (19)$$

where we have defined $u_t = V_t^{-\frac{1}{2}} x_\star / \|x_\star\|_{V_t^{-1}}$. By definition of u_t, note that $\|u_t\|_2 = 1$. Hence, the desired (19) holds due to the anti-concentration property of the $\mathcal{H}^{\text{TS}}$ distribution in (13).

The key differences to the proof of optimism in the classical setting in [2, Lemma 3] are as follows. First, we present an algebraic version of the basic machinery introduced in [2, Sec. 5] that we show is convenient to extend to the safe setting. Second, we employ the idea of relating x_t to a "better" feasible point $\alpha_t x_\star$ and show optimism for the latter. Third, even after introducing α_t, the fact that $1/\alpha_t - 1$ is proportional to $\|x_\star\|_{V_t^{-1}}$ is critical for the seemingly simple algebraic steps that follow (18). In particular, in deducing (19) from the expression above, note that we have divided both sides in the probability term by $\|x_\star\|_{V_t - 1}$. It is only thanks to the proportionality observation that we made above that the term $\|x_\star\|_{V_t - 1}$ cancels throughout and we can conclude with (19) without a need to lower bound the minimum eigenvalue of the Gram matrix V_t (which is known to be hard).

B. Proof sketch: Why frequent optimism is enough to bound Term I

As discussed in Section III, the presence of the safety constraints complicates the requirement for optimism. We show in Section III-A that Safe-LTS is optimistic with constant probability in spite of safety constraints. Based on this, we complete the sketch of the proof here by showing that we can bound the overall regret of Term I in (16) with the V_τ-norm of optimistic (and in our case, safe) actions. Let us first define the set of the optimistic parameters as

$$\Theta_t^{\text{opt}}(\delta') = \{\theta \in \mathbb{R}^d : \max_{x \in \mathcal{D}_t^s} x^\top \theta \geq x_\star^\top \theta_\star\}. \quad (20)$$

In Section III-A, we show that Safe-LTS samples from this set i.e., $\tilde{\theta}_t \in \Theta_t^{\text{opt}}$, with constant probability. Note that, if at round t Safe-LTS samples from the set of optimistic parameters, Term I at that round is non-positive. In the following, we show that selecting the optimal arm corresponding to any optimistic parameter can control the overall regret of Term I. The argument below is adapted from [2] with required modifications.

For the purpose of this proof sketch, we assume that at each round t, the safe decision set contains the previous safe action

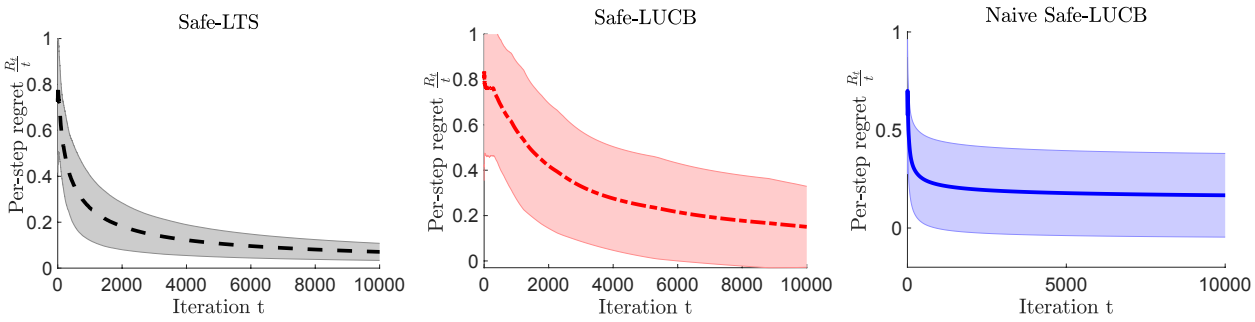

Figure 1. Comparison of mean per-step regret for Safe-LTS, Safe-LUCB, and Naive Safe-LUCB. The shaded regions show one standard deviation around the mean. The results are averages over 30 problem realizations.

that the algorithm played, i.e., $x_{t-1} \in \mathcal{D}_t^s$. However, for the formal proof in App. VII-A, we do not need such an assumption. Let τ be a time such that $\tilde{\theta}_\tau \in \Theta_t^{\text{opt}}$, i.e., $x_\tau^\top \tilde{\theta}_\tau \geq x_\star^\top \theta_\star$. Then, for any $t \geq \tau$ we have

$$\begin{aligned}\text{Term I} := R_t^{\text{TS}} &= x_\star^\top \theta_\star - x_t^\top \tilde{\theta}_t \\ &\leq x_\tau^\top \tilde{\theta}_\tau - x_t^\top \tilde{\theta}_t \leq x_\tau^\top \left(\tilde{\theta}_\tau - \tilde{\theta}_t\right). \end{aligned} \tag{21}$$

The last inequality comes from the assumption that at each round t, the safe decision set contains the previous played safe actions for rounds $s \leq t$; hence, $x_\tau^\top \tilde{\theta}_t \leq x_t^\top \tilde{\theta}_t$. To continue from (21), we use Cauchy-Schwarz, and obtain

$$\begin{aligned} R_t^{\text{TS}} &\leq \|x_\tau\|_{V_\tau^{-1}} \left\|\tilde{\theta}_\tau - \tilde{\theta}_t\right\|_{V_\tau} \\ &\leq \left(\left\|\tilde{\theta}_\tau - \theta_\star\right\|_{V_\tau} + \left\|\theta_\star - \tilde{\theta}_t\right\|_{V_\tau}\right) \|x_\tau\|_{V_\tau^{-1}} \\ &\leq \left(\left\|\tilde{\theta}_\tau - \theta_\star\right\|_{V_\tau} + \left\|\theta_\star - \tilde{\theta}_t\right\|_{V_t}\right) \|x_\tau\|_{V_\tau^{-1}}. \end{aligned} \tag{22}$$

The last inequality comes from the fact that the Gram matrices construct a non-decreasing sequence ($V_\tau \preceq V_t$, $\forall t \geq \tau$). Then, we define the ellipsoid $\mathcal{E}_t^{\text{TS}}(\delta')$ such that

$$\mathcal{E}_t^{\text{TS}}(\delta') := \{\theta \in \mathbb{R}^d : \left\|\theta - \hat{\theta}_t\right\|_{V_t} \leq \gamma_t(\delta')\}, \tag{23}$$

where

$$\gamma_t(\delta') = \beta_t(\delta')\big(1 + \frac{2}{C}LS\big)\sqrt{cd\log\big(\frac{c'd}{\delta}\big)}. \tag{24}$$

It is not hard to see by combining Theorem II.1 and the concentration property that $\tilde{\theta}_t \in \mathcal{E}_t^{\text{TS}}(\delta')$ with high probability. Hence, we can bound (22) using triangular inequality such that:

$$R_t^{\text{TS}} \leq \Big(\gamma_\tau(\delta') + \beta_\tau(\delta') + \gamma_t(\delta') + \beta_t(\delta')\Big) \|x_\tau\|_{V_\tau^{-1}} \tag{25}$$

$$\leq 2\Big(\gamma_T(\delta') + \beta_T(\delta')\Big) \|x_\tau\|_{V_\tau^{-1}} \tag{26}$$

The last inequality comes from the fact that $\beta_t(\delta')$ and $\gamma_t(\delta')$ are non-decreasing in t by construction. Therefore, following the intuition of [2], we can upper bound Term I with respect to the V_τ-norm of the optimal safe action at time τ (see App. VII-A for formal proof). Bounding the term $\|x_\tau\|_{V_\tau^{-1}}$ is standard based on the analysis provided in [5] (see Proposition VI.1 in the Appendix).

IV. Numerical Results and Comparison to State of the Art

We present details of our numerical experiments on synthetic data. First, we show how the presence of safety constraints affects the performance of LTS in terms of regret. Next, we evaluate Safe-LTS by comparing it against safe versions of LUCB. Then, we compare Safe-LTS to [34]'s Safe-LUCB. In all the implementations, we used: $T = 10000, \delta = 1/4T$, $R = 0.1$ and $\mathcal{D}_0 = [-1, 1]^2$. Unless otherwise specified, the reward and constraint parameters $\theta_\star$ and $\mu_\star$ are drawn from $\mathcal{N}(0, I_2)$ each; C is drawn uniformly from $[0, 1]$. Throughout, we have implemented a modified version of Safe-LUCB which uses ℓ_1-norms instead of ℓ_2-norms, due to computational considerations (e.g., [3], [34]). This highlights a well-known benefit associated with TS-based algorithms, namely that they are easier to implement and more computationally-efficient than UCB-based algorithms. In particular, the action selection rule in UCB-based algorithms involves solving optimization problems with bilinear objective functions, whereas, for TS-based algorithms, it would lead to linear objectives (see [2]).

A. The effect of safety constraints on LTS

In Fig. 2(left), we compare the average cumulative regret of Safe-LTS to the standard LTS with *oracle access* to the true safe set $\mathcal{D}_0^s$. The results are averages over 20 problem realizations. As shown, even though Safe-LTS requires that chosen actions belong to the conservative inner-approximation set $\mathcal{D}_t^s$, it still achieves a regret of the same order as the oracle reaffirming the prediction of Theorem III.1. Also, the comparison to the oracle reveals that the action selection rule of Safe-LTS is indeed such that it guarantees fast safe-set expansion so as to not exclude optimistic actions for a long time. Fig. 2(left) also shows the performance Safe-LTS with dynamic noise distribution. In order for Safe-LTS to be frequently optimistic, our theory requires that the random perturbation η_t satisfies (13) *for all rounds*. Specifically, we need the extra $\frac{2}{C}LS$ factor compared to [2] in order to ensure safe set expansion. While this result is already sufficient for the tight regret guarantees of Theorem III.1, it does not fully capture our intuition (see also Sec. II-B2) that as the algorithm progresses and $\mathcal{D}_t^s$ gets closer to $\mathcal{D}_0^s$, exploration (and thus, the requirement on anti-concentration) does not need

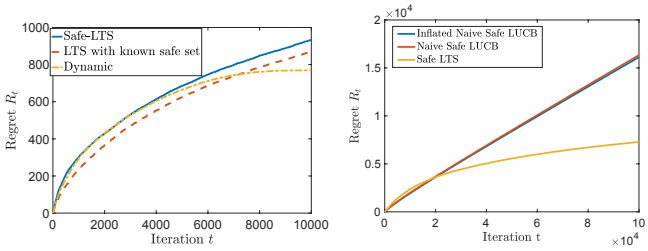

Figure 2. Left: Average cumulative regret of Safe-LTS vs standard LTS with oracle access to the safe set and Safe-LTS with a dynamic noise distribution described in Section IV-A. Right: Cumulative regret of Safe-LTS, Naive Safe-LUCB and Inflated Naive Safe-LUCB for a specific problem instance.

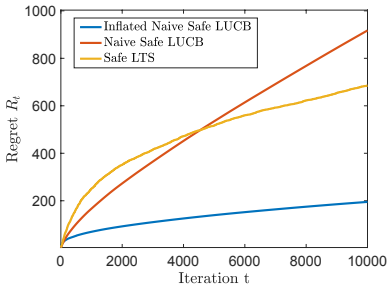

Figure 3. Comparison of the cumulative regret of Safe-LTS and Naive Safe-LUCB and Inflated Naive Safe-LUCB algorithms over randomly generated instances.

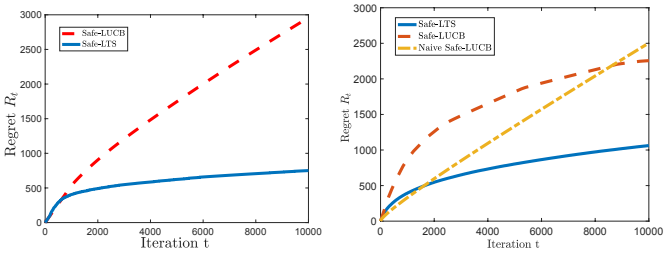

Figure 4. Left: Regret of Safe-LUCB vs Safe-LTS, for a single problem instance with active safety constraint. Right: Average cumulative regret of Safe-LTS vs two safe LUCB algorithms.

to be so aggressive. Based on this intuition, we propose the following heuristic modification, in which Safe-LTS uses a perturbation with the following *dynamic* property:

$$\mathbb{P}_{\eta\sim\mathcal{H}^{\text{TS}}}\left(u^\top\eta \geq k(t)\right) \geq p, \tag{27}$$

for $k(t)$ a linearly-decreasing function $k(t) = (1+\frac{2}{C}LS)^2(1-t/T)$. In particular, this can be implemented by sampling each entry of $\eta_t, t \in [T]$ i.i.d from $\mathcal{N}(0, k(t))$. Fig. 2(left) shows empirical evidence of the superiority of the heuristic.

B. Comparison to the safe version of LUCB

Here, we compare the performance of our algorithm with the safe version of LUCB, as follows. We implement a natural extension of the classical LUCB algorithm in [3], which we call "Naive Safe-LUCB" and which respects safety constraints by choosing actions from the estimated safe set in (8). We consider an improved version, which we call "Inflated Naive Safe-LUCB" and which is motivated by our analysis of Safe-LTS. Specifically, in light of Lemma III.2, we implement the improved LUCB algorithm with an inflated confidence ellipsoid by a fraction $1+\frac{2}{C}LS$ in order to favor optimistic exploration. In Fig. 2(right), we employ these two algorithms for a specific problem instance showing that both fail to provide the $\widetilde{\mathcal{O}}(\sqrt{T})$ regret of Safe-LTS, in general. Specifically, we chooose $\theta_* = \begin{bmatrix} 0.5766 \\ -0.1899 \end{bmatrix}$, $\mu_* = \begin{bmatrix} 0.2138 \\ -0.0020 \end{bmatrix}$, and $C = 0.0615$. Further numerical simulations suggest that while Safe-LTS always outperforms Naive Safe-LUCB, the Inflated Naive Safe-LUCB can have superior performance to Safe-LTS in many problem instances (see Fig. 3). Unfortunately, not only is this not always the case (cf. Fig. 2(right)), but also we are not aware of an appropriate modification to our proofs to show this problem-dependent performance. Further investigations in this direction might be of interest.

C. Comparison to Safe-LUCB

We compare our algorithm to the Safe-LUCB algorithm of [34]. In [34], the linear safety constraint involves the *same* unknown parameter vector $\theta_\star$ of the linear reward function and –in our notation– it takes the form $x^\top B\theta_\star \leq C$, for some *known* matrix B. As such, *no* side-information measurements are needed. First, while our proof does not show a regret of $\widetilde{\mathcal{O}}(\sqrt{T})$ for the setting of [34] in the general case, it does so for special cases. For example, it is not hard to see that our proofs readily extend to their setting when $B = I$. This already improves upon the $\widetilde{\mathcal{O}}(T^{2/3})$ guarantee provided by [34]. Indeed, for $B = I$, there are non-trivial instances where $C - x_*^\top\theta_* = 0$ (i.e., the safety constraint is active), in which Safe-LUCB suffers from a $\widetilde{\mathcal{O}}(T^{2/3})$ bound [34]. Second, while our proof adapts to a special case of [34]'s setting, the other way around is *not* true, i.e., it is not obvious how one would modify the proof of [34] to obtain a $\widetilde{\mathcal{O}}(\sqrt{T})$ guarantee even in the presence of side information. This point is highlighted by Fig. 4(left) that numerically compares the two algorithms for a specific problem instance with side information: $\theta_* = [0.9, 0.23]^\top$, $\mu_* = [0.55, 0.31]^T$, and $C = 0.11$ (note that the constraint is active at the optimal). Also, see Section IV-E for a numerical comparison of the estimated safe-sets' expansion for the two algorithms. Fig. 4(right) compares Safe-LTS against Safe-LUCB and Naive Safe-LUCB over 30 problem realizations. As already pointed out in [34], Naive Safe-LUCB generally leads to poor regret, since the LUCB action selection rule alone does not provide sufficient exploration towards safe set expansion. In contrast, Safe-LUCB is equipped with a pure exploration phase over a given seed safe set, which is shown to lead to proper safe set expansion. Our paper reveals that the inherent randomized nature of Safe-LTS is alone capable to properly expand the safe set without the need for an explicit initialization phase (during which regret grows linearly).

D. Standard deviations

Figure 1 shows the sample standard deviation of regret around the average per-step regret for each one of the curves depicted in Figure 4(right). We remark on the strong

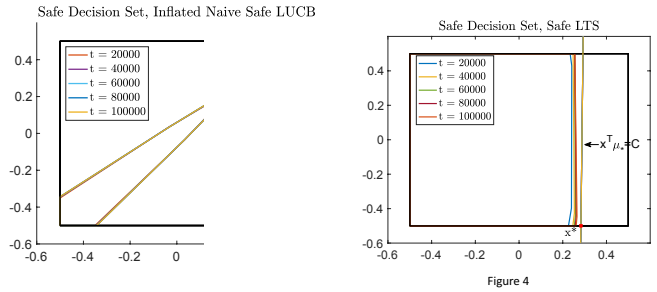

Figure 5. Comparison of expansion of safe decision sets for Safe-LTS, and Inflated Naive Safe-LUCB.

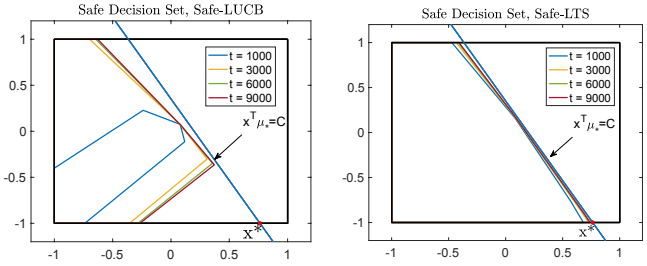

Figure 6. Comparison of expansion of a safe decision sets for Safe-LUCB and Safe-LTS, for a single problem instance.

dependency of the performance of LUCB-based algorithms on the specific problem instance, whereas the performance of Safe-LTS does not vary significantly under the same instances.

E. Safe-set expansion

We also plot the expansion of the estimated safe set $\mathcal{D}_t^s$ in time for different problem instances for Saf-LTS and "Inflated Naieve Safe-LUCB" and Safe-LUCB in [34]. In particular, Fig. 6 highlights the gradual expansion of the safe decision set for Safe-LUCB in [34] and Safe-LTS for a problem instance in which the safety constraint is active for parameters $\theta_* = \begin{bmatrix} 0.9 \\ 0.23 \end{bmatrix}$, $\mu_* = \begin{bmatrix} 0.55 \\ -0.31 \end{bmatrix}$, and $C = 0.11$. Similarly, Fig. 5 illustrates the expansion of the safe decision set for "Inflated Naive Safe-LUCB" and Safe-LTS for a problem instance with parameters $\theta_* = \begin{bmatrix} 0.5766 \\ -0.1899 \end{bmatrix}$, $\mu_* = \begin{bmatrix} 0.2138 \\ -0.0020 \end{bmatrix}$, and $C = 0.0615$ in which the former provides poor (almost linear) regret. These empirical experiments reinforce the main message of our paper that the inherent randomized nature of TS is crucial for properly expanding the safe action set.

Next, we comment on the dependence of the regret of Safe-LTS on the size of the safe set. Note that the size of the safe action set depends on the safety constant C as well as on the unknown parameter $\mu_\star$. Recall that S is an upper bound on the norm of $\mu_\star$, and, also $\|x\|_2 \leq L$ for any action vector $x \in \mathcal{D}_0$. Since the constraint is of the form $x^\top \mu_\star \leq C$, the size of the set of safe actions depends on the values L, S, C. We will also assume that $LS > C$, since otherwise it follows by Cauchy-Schwartz that all actions in $\mathcal{D}_0$ are safe and the regret is no different compared to the unconstrained case. Intuitively, for smaller values of C (compared to LS), the "smaller" the safe set around zero. This means that the algorithm can only take actions in a very conservative manner to guarantee that actions remain safe. At an intuitive level, we would then expect an increase on the regret. This intuition is in fact captured by our regret bound in Theorem III.1 showing that the bound increases with increasing values of the ratio $\frac{LS}{C}$. Thus, the smaller C, the larger our regret bound.

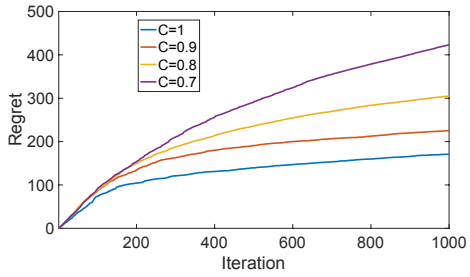

Figure 7. Comparison of the cumulative regret of Safe-LTS for different values of the safety constant C.

In Figure 7 we showcase the effect of decreasing C on the regret. Specifically, we have chosen $\theta_\star = [0.3; 0.8]$, $\mu_\star = [0.2; 0.7]$, $\mathcal{D}_0 = [-1, 1]^2$, $S = \sqrt{2}$ and $L = \sqrt{2}$ and we have plotted the regret of Safe-LTS for different values of $C = 0.7, 0.8, 0.9$ and 1. We see that the regret increases for smaller values of C as suggested by our bound of Theorem III.1.

As a closing remark, while we make no claim that our bound captures sharply the effect of the size of the safe set (perhaps measured in terms of some geometric quantity such as volume), we showed that our bound captures the effect of the size in the summary term LS/C, which also appears to agree with the empirical results of Figure 7.

V. Conclusion

In this paper, we study a linear stochastic bandit (LB) problem in which the environment is subject to unknown linear safety constraints that need to be satisfied at each round. As such, the learner must make necessary modifications to ensure that the chosen actions belong to the unknown safe set. We propose Safe-LTS, which to the best of our knowledge, is the first safe linear TS algorithm with provable regret guarantees for this problem. Moreover, we show that the Safe-LTS achieves the same frequentist regret of order $\mathcal{O}(d^{3/2}\log^{1/2} d \cdot T^{1/2}\log^{3/2} T)$ as the original LTS problem studied in [2]. We also compare Safe-LTS with several several UCB-type safe algorithms. We show that our algorithm has: better regret in the worst-case ($\widetilde{\mathcal{O}}(T^{1/2})$ vs. $\widetilde{\mathcal{O}}(T^{2/3})$), fewer parameters to tune and often superior empirical performance. Interesting directions for future work include gaining a theoretical understanding of the regret of the algorithm when the TS distribution satisfies the dynamic property in (27), which empirically leads regret of smaller order as well as, investigating TS-based alternatives to the GP-UCB-type algorithms of [30], [31]. Additionally, it is interesting to study extensions of our theory on linear constraints to the more general setting in which constraints are modeled as Gaussian Processes. This would also allow more complex settings in which the safe regions may even be disconnected.

REFERENCES

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

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

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

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

[5] 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.

[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] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 661–670.

[8] I. Usmanova, A. Krause, and M. Kamgarpour, "Safe convex learning under uncertain constraints," in *Proceedings of Machine Learning Research*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and M. Sugiyama, Eds., vol. 89. PMLR, 16–18 Apr 2019, pp. 2106–2114. [Online]. Available: http://proceedings.mlr.press/v89/usmanova19a.html

[9] 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.

[10] S. Daulton, S. Singh, V. Avadhanula, D. Dimmery, and E. Bakshy, "Thompson sampling for contextual bandit problems with auxiliary safety constraints," *arXiv preprint arXiv:1911.00638*, 2019.

[11] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2-3, pp. 235–256, May 2002. [Online]. Available: https://doi.org/10.1023/A:1013689704352

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

[13] S. Vakili, K. Liu, and Q. Zhao, "Deterministic sequencing of exploration and exploitation for multi-armed bandit problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 5, pp. 759–767, 2013.

[14] C. Tekin and M. van der Schaar, "Distributed online learning via cooperative contextual bandits," *IEEE Transactions on Signal Processing*, vol. 63, no. 14, pp. 3700–3714, 2015.

[15] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multiplayer multiarmed bandits," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2331–2345, 2014.

[16] C. Tekin and E. Turğay, "Multi-objective contextual multi-armed bandit with a dominant objective," *IEEE Transactions on Signal Processing*, vol. 66, no. 14, pp. 3799–3813, 2018.

[17] C. Tekin and E. Turgay, "Multi-objective contextual bandits with a dominant objective," in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2017, pp. 1–6.

[18] S. Shahrampour, M. Noshad, and V. Tarokh, "On sequential elimination algorithms for best-arm identification in multi-armed bandits," *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pp. 4281–4292, 2017.

[19] C. Gan, R. Zhou, J. Yang, and C. Shen, "Cost-aware cascading bandits," *IEEE Transactions on Signal Processing*, vol. 68, pp. 3692–3706, 2020.

[20] G. Liu and L. Lai, "Action-manipulation attacks against stochastic bandits: Attacks and defense," *IEEE Transactions on Signal Processing*, vol. 68, pp. 5152–5165, 2020.

[21] ——, "Action-manipulation attacks on stochastic bandits," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3112–3116.

[22] Z. Wang, R. Zhou, and C. Shen, "Regional multi-armed bandits with partial informativeness," *IEEE Transactions on Signal Processing*, vol. 66, no. 21, pp. 5705–5717, 2018.

[23] S. Vakili and Q. Zhao, "Risk-averse online learning under mean-variance measures," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 1911–1915.

[24] A. Cassel, S. Mannor, and A. Zeevi, "A general approach to multi-armed bandits under risk criteria," *arXiv preprint arXiv:1806.01380*, 2018.

[25] A. Sani, A. Lazaric, and R. Munos, "Risk-aversion in multi-armed bandits," in *Advances in Neural Information Processing Systems*, 2012, pp. 3275–3283.

[26] S. Vakili and Q. Zhao, "Risk-averse multi-armed bandit problems under mean-variance measure," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 6, pp. 1093–1111, 2016.

[27] O.-A. Maillard, "Robust risk-averse stochastic multi-armed bandits," in *International Conference on Algorithmic Learning Theory*. Springer, 2013, pp. 218–233.

[28] A. Aswani, H. Gonzalez, S. S. Sastry, and C. Tomlin, "Provably safe and robust learning-based model predictive control," *Automatica*, vol. 49, no. 5, pp. 1216–1226, 2013.

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

[30] Y. Sui, A. Gotovos, J. W. Burdick, and A. Krause, "Safe exploration for optimization with gaussian processes," in *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15. JMLR.org, 2015, pp. 997–1005. [Online]. Available: http://dl.acm.org/citation.cfm?id=3045118.3045225

[31] 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.

[32] 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.

[33] A. K. Akametalu, J. F. Fisac, J. H. Gillula, S. Kaynama, M. N. Zeilinger, and C. J. Tomlin, "Reachability-based safe learning with gaussian processes," in *53rd IEEE Conference on Decision and Control*, Dec 2014, pp. 1424–1431.

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

[35] K. Khezeli and E. Bitar, "Safe linear stochastic bandits," *arXiv preprint arXiv:1911.09501*, 2019.

[36] A. Moradipari, C. Thrampoulidis, and M. Alizadeh, "Stage-wise conservative linear bandits," *arXiv preprint arXiv:2010.00081*, 2020.

[37] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.

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

[39] E. Kaufmann, N. Korda, and R. Munos, "Thompson sampling: An asymptotically optimal finite-time analysis," in *International Conference on Algorithmic Learning Theory*. Springer, 2012, pp. 199–213.

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

[41] D. Russo and B. Van Roy, "An information-theoretic analysis of thompson sampling," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2442–2471, 2016.

[42] 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.

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

[44] A. Gopalan and S. Mannor, "Thompson sampling for learning parameterized markov decision processes," in *Conference on Learning Theory*, 2015, pp. 861–898.

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

[46] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

VI. APPENDIX A

We first state the standard results that plays an important role in most proofs for linear bandits problems.

Proposition VI.1. *(from[5]) Let $\lambda \geq 1$. For any arbitrary sequence of actions $(x_1, \ldots, x_t) \in \mathcal{D}^t$, let V_t be the corre-*

sponding Gram matrix, then

$$\sum_{s=1}^{t} \|x_s\|^2_{V_s^{-1}} \leq 2\log\frac{det(V_{t+1})}{det(\lambda I)} \leq 2d\log\left(1+\frac{tL^2}{\lambda}\right). \quad (28)$$

In particular, we have

$$\sum_{s=1}^{T} \|x_s\|_{V_s^{-1}} \leq \sqrt{T}\left(\sum_{s=1}^{T}\|x_s\|^2_{V_s^{-1}}\right)^{\frac{1}{2}} \leq \sqrt{2Td\log\left(1+\frac{TL^2}{\lambda}\right)}. \quad (29)$$

Also, we recall the Azuma's concentration inequality for super-martingales.

Proposition VI.2. *(Azuma's inequality [46]) If a super-martingale $(Y_t)_{t\geq 0}$ corresponding to a filtration $\mathcal{F}_t$ satisfies $|Y_t - Y_{t-1}| < c_t$ for some positive constant c_t, for all $t = 1,\dots,T$, then, for any $u > 0$,*

$$\mathbb{P}\left(Y_T - Y_0 \geq u\right) \leq 2e^{-\frac{u^2}{2\sum_{t=1}^{T} c_t^2}}. \quad (30)$$

Next, we define the high probability confidence regions for the RLS-estimates that er use in the rest of the proof.

Definition VI.1. Let $\delta \in (0,1)$, $\delta' = \frac{\delta}{6T}$, and $t \in [T]$. We define the following events:

- $\hat{E}_t$ is the event that the RLS-estimate $\hat{\theta}$ concentrates around $\theta_\star$ for all steps $s \leq t$, i.e., $\hat{E}_t = \{\forall s \leq t, \left\|\hat{\theta}_s - \theta_\star\right\|_{V_s} \leq \beta_s(\delta')\}$;
- $\hat{Z}_t$ is the event that the RLS-estimate $\hat{\mu}$ concentrates around $\mu_\star$, i.e., $\hat{Z}_t = \{\forall s \leq t, \|\hat{\mu}_s - \mu_\star\|_{V_s} \leq \beta_s(\delta')\}$. Moreover, define Z_t such that
$$Z_t = \hat{E}_t \cap \hat{Z}_t.$$
- $\tilde{E}_t$ is the event that the sampled parameter $\tilde{\theta}_t$ concentrates around $\hat{\theta}_t$ for all steps $s \leq t$, i.e., $\tilde{E}_t = \{\forall s \leq t, \left\|\tilde{\theta}_s - \hat{\theta}_s\right\|_{V_s} \leq \gamma_s(\delta')\}$. Let E_t be such that $E_t = \tilde{E}_t \cap Z_t$.

Lemma VI.3. *Under Assumptions 1, 2, we have $\mathbb{P}(Z) = \mathbb{P}(\hat{E}\cap\hat{Z}) \geq 1 - \frac{\delta}{3}$ where $\hat{E} = \hat{E}_T \subset \cdots \subset \hat{E}_1$, and $\hat{Z} = \hat{Z}_T \subset \cdots \subset \hat{Z}_1$.*

Proof. The proof is similar to the one in Lemma 1 of [2] and is ommited for brevity. □

Lemma VI.4. *Under Assumptions 1, 2, we have $\mathbb{P}(E) = \mathbb{P}(\tilde{E}\cap Z) \geq 1 - \frac{\delta}{2}$, where $\tilde{E} = \tilde{E}_T \subset \cdots \subset \tilde{E}_1$.*

Proof. We show that $\mathbb{P}(\tilde{E}) \geq 1 - \frac{\delta}{6}$. Then, from Lemma VI.3 we know that $\mathbb{P}(Z) \geq 1 - \frac{\delta}{3}$, thus we can conclude that $\mathbb{P}(E) \geq 1 - \frac{\delta}{2}$. Bounding $\tilde{E}$ comes directly from concentration inequality (14). Specifically, for $1 \leq t \leq T$

$$\mathbb{P}\left(\left\|\tilde{\theta}_t - \hat{\theta}_t\right\|_{V_t} \leq \gamma_t(\delta')\right) = \mathbb{P}\left(\|\eta_t\|_2 \leq \frac{\gamma_t(\delta')}{\beta_t(\delta')}\right)$$
$$= \mathbb{P}\left(\|\eta_t\|_2 \leq \left(1+\frac{2}{C}LS\right)\sqrt{cd\log\left(\frac{c'd}{\delta'}\right)}\right) \geq 1-\delta'.$$

Applying union bound on this ensures that $\mathbb{P}(\tilde{E}) \geq 1 - T\delta' = 1 - \frac{\delta}{6}$. □

Now we are ready to provide the formal proof of Lemma III.2. First, we provide a formal statement and a detailed proof of Lemma III.2. Here, we need several modifications compared to [2] that are required because in our setting, actions x_t belong to inner approximations of the true safe set $\mathcal{D}_0^s$. Moreover, we follow an algebraic treatment that is perhaps simpler compared to the geometric viewpoint in [2].

Lemma VI.5. *Let $\Theta_t^{\text{opt}} = \{\theta \in \mathbb{R}^d : \max_{x\in\mathcal{D}_t^s} x^\top\theta \geq x_\star^\top\theta_\star\} \cap \mathcal{E}_t^{TS}$ be the set of optimistic parameters, $\tilde{\theta}_t = \hat{\theta}_t + \beta_t(\delta')V_t^{-\frac{1}{2}}\eta_t$ with $\eta_t \sim \mathcal{D}^{TS}$, then $\forall t \geq 1$, $\mathbb{P}\left(\tilde{\theta}_t \in \Theta_t^{\text{opt}} | \mathcal{F}_t, Z_t\right) \geq \frac{p}{2}$.*

Proof. First, we provide the shrunk version $\tilde{\mathcal{D}}_t^s$ of $\mathcal{D}_t^s$ as follows:

A shrunk safe decision set $\tilde{\mathcal{D}}_t^s$. Consider the enlarged confidence region $\tilde{\mathcal{C}}_t$ centered at $\mu_\star$ as

$$\tilde{\mathcal{C}}_t := \{v \in \mathbb{R}^d : \|v - \mu_\star\|_{V_t} \leq 2\beta_t(\delta')\}. \quad (31)$$

We know that $\mathcal{C}_t \subseteq \tilde{\mathcal{C}}_t$, since $\forall v \in \mathcal{C}_t$, we know that $\|v-\mu_\star\|_{V_t} \leq \|v-\hat{\mu}_t\|_{V_t} + \|\hat{\mu}_t - \mu_\star\|_{V_t} \leq 2\beta(t)$. From the definition of enlarged confidence region, we can get the following definition for shrunk safe decision set:

$$\begin{aligned}\tilde{\mathcal{D}}_t^s :=& \{x \in \mathcal{D}_0 : x^\top v \leq C, \forall v \in \tilde{\mathcal{C}}_t\}\\ =& \{x \in \mathcal{D}_0 : \max_{v\in\tilde{\mathcal{C}}_t} x^\top v \leq C\}\\ =& \{x \in \mathcal{D}_0 : x^\top\mu_\star + 2\beta_t(\delta')\|x\|_{V_t^{-1}} \leq C\}, \end{aligned} \quad (32)$$

and note that $\tilde{\mathcal{D}}_t^s \subseteq \mathcal{D}_t^s$, and they are not empty, since they include zero due to Assumption 3.

Then, we define the parameter α_t such that the vector $z_t = \alpha_t x_\star$ in direction $x_\star$ belongs to $\tilde{\mathcal{D}}_t^s$ and is closest to $x_\star$. Hence, we have:

$$\alpha_t := \max\left\{\alpha \in [0,1] : z_t = \alpha x_\star \in \tilde{\mathcal{D}}_t^s\right\}. \quad (33)$$

Since $\mathcal{D}_0$ is convex by Assumption 3 and both $0, x_\star \in \mathcal{D}_0$, we have

$$\alpha_t = \max\left\{\alpha \in [0,1] : \alpha\left(x_\star^\top\mu_\star + 2\beta_t(\delta')\|x_\star\|_{V_t^{-1}}\right) \leq C\right\}. \quad (34)$$

From constraint (1), we know that $x_\star^\top\mu_\star \leq C$. We choose α_t such that

$$1 + \frac{2}{C}\beta_t(\delta')\|x_\star\|_{V_t^{-1}} = \frac{1}{\alpha_t}. \quad (35)$$

We need to study the probability that a sampled $\tilde{\theta}_t$ drawn from $\mathcal{H}^{\text{TS}}$ distribution at round t is optimistic, i.e.,

$$p_t = \mathbb{P}\left((x_t(\tilde{\theta}_t))^\top\tilde{\theta}_t \geq x_\star^\top\theta_\star \;\middle|\; \mathcal{F}_t, Z_t\right).$$

Using the definition of α_t in (34), we have

$$(x_t(\tilde{\theta}_t))^\top\tilde{\theta}_t = \max_{x\in\mathcal{D}_t^s} x^\top\tilde{\theta}_t \geq \alpha_t x_\star^\top\tilde{\theta}_t. \quad (36)$$

Hence, we can write

$$p_t \geq \mathbb{P}\left(\alpha_t x_\star^\top \tilde{\theta}_t \geq x_\star^\top \theta_\star \;\middle|\; \mathcal{F}_t, Z_t\right)$$
$$= \mathbb{P}\left(x_\star^\top \left(\hat{\theta}_t + \beta_t(\delta') V_t^{-\frac{1}{2}} \eta_t\right) \geq \frac{x_\star^\top \theta_\star}{\alpha_t} \;\middle|\; \mathcal{F}_t, Z_t\right)$$

Then, we use the value that we chose for α_t in (35), and we have

$$= \mathbb{P}\Bigg(x_\star^\top \hat{\theta}_t + \beta_t(\delta') x_\star^\top V_t^{-\frac{1}{2}} \eta_t \geq x_\star^\top \theta_\star + \frac{2}{C}\beta_t(\delta') \left\|x_\star\right\|_{V_t^{-1}} x_\star^\top \theta_\star | \mathcal{F}_t, Z_t\Bigg)$$

we know that $|x_\star^\top \theta_\star| \leq \|x_\star\|_2 \|\theta_\star\|_2 \leq LS$. Hence,

$$p_t \geq \mathbb{P}\Bigg(\beta_t(\delta') x_\star^\top V_t^{-\frac{1}{2}} \eta_t \geq x_\star^\top (\theta_\star - \hat{\theta}_t) + \frac{2}{C} LS \beta_t(\delta') \left\|x_\star\right\|_{V_t^{-1}} \mid \mathcal{F}_t, Z_t\Bigg)$$

From Cauchy–Schwarz inequality and (5), we have

$$\mid x_\star^\top \left(\theta_\star - \hat{\theta}_t\right) \mid \leq \|x_\star\|_{V_t^{-1}} \left\|\theta_\star - \hat{\theta}_t\right\|_{V_t} \leq \beta_t(\delta') \left\|x_\star\right\|_{V_t^{-1}}.$$

Therefore, we can write

$$p_t \geq \mathbb{P}\Bigg(x_\star^\top V_t^{-\frac{1}{2}} \eta_t \geq \|x_\star\|_{V_t^{-1}} + \frac{2}{C} LS \left\|x_\star\right\|_{V_t^{-1}} \mid \mathcal{F}_t, Z_t\Bigg) \tag{37}$$

We define $u^\top = \frac{x_\star^\top V_t^{-\frac{1}{2}}}{\|x_\star\|_{V_t^{-1}}}$, and hence $\|u\|_2 = 1$. It follows from (37) that

$$p_t \geq \mathbb{P}\left(u^\top \eta_t \geq 1 + \frac{2}{C} LS\right) \geq p, \tag{38}$$

where the last inequality follows the concentration inequality (14) of the TS distribution. We also need to show that the high probability concentration inequality event does not effect the TS of being optimistic. This is because the chosen confidence bound $\delta' = \frac{\delta}{6T}$ is small enough compared to the anti-concentration property (13). Moreover, we assume that $T \geq \frac{1}{3p}$ which implies that $\delta' \leq \frac{p}{2}$. We know that for any events A ans B, we have

$$\mathbb{P}(A \cap B) = 1 - \mathbb{P}(A^c \cup B^c) \geq \mathbb{P}(A) - \mathbb{P}(B^c). \tag{39}$$

We apply (39) with $A = \{J_t(\tilde{\theta}_t) \geq J(\theta_\star)\}$ and $B = \{\tilde{\theta}_t \in \mathcal{E}_t^{\text{TS}}\}$ which leads to

$$\mathbb{P}\left(\tilde{\theta}_t \in \Theta_t^{\text{opt}} \;\middle|\; \mathcal{F}_t, Z_t\right) \geq p - \delta' \geq \frac{p}{2}.$$

□

VII. Appendix B

The proof presented below follows closely the proof of [2] and is primarily presented here for completeness. Specifically, we have identified that the only critical change that needs to be made to account for safety is the proof of actions being frequently optimistic in the face of constraints thanks to the modified anti-concentration property 13. This was handled in the previous section VI. For completeness, we also prove in Lemma VII.1 that the first action of Safe-LTS is always safe under our assumptions.

We use the following decomposition for bounding the regret:

$$R(T) \leq \sum_{t=1}^{T} \left(x_\star^\top \theta_\star - x_t \theta_\star\right) \mathbb{1}\{E_t\} =$$
$$\sum_{t=1}^{T} \left(\underbrace{x_\star^\top \theta_\star - x_t^\top \tilde{\theta}_t}_{\text{Term I}}\right) \mathbb{1}\{E_t\} + \sum_{t=1}^{T} \left(\underbrace{x_t^\top \tilde{\theta}_t - x_t^\top \theta_\star}_{\text{Term II}}\right) \mathbb{1}\{E_t\}. \tag{40}$$

A. Bounding Term I.

For any θ, we denote $x_t(\theta) = \arg\max_{x \in \mathcal{D}_t^s} x^\top \theta$. On the event E_t, $\tilde{\theta}_t$ belongs to $\mathcal{E}_t^{\text{TS}}$ which leads to

$$\text{(Term I)}\mathbb{1}\{E_t\} := R_t^{\text{TS}} \mathbb{1}\{E_t\} \leq \left(x_\star^\top \theta_\star - \inf_{\theta \in \mathcal{E}_t^{\text{TS}}} (x_t(\theta))^\top \theta\right) \mathbb{1}\{Z_t\}. \tag{41}$$

Here and onwards, we use $\mathbb{1}\{\mathcal{E}\}$ as the indicator function applied to an event $\mathcal{E}$. We have also used the fact that E_t is a subset of Z_t. Next, we can also bound (41) by the expectation over any random choice of $\tilde{\theta} \in \Theta_t^{\text{opt}}$ (recall (20)) that leads to

$$R_t^{\text{TS}} \leq \mathbb{E}\left[\left((x_t(\tilde{\theta}))^\top \tilde{\theta} - \inf_{\theta \in \mathcal{E}_t^{\text{TS}}} (x_t(\theta))^\top \theta\right) \mathbb{1}\{Z_t\} \;\middle|\; \mathcal{F}_t, \tilde{\theta} \in \Theta_t^{\text{opt}}\right].$$

Equivalently, we can write

$$R_t^{\text{TS}} \leq \mathbb{E}\left[\sup_{\theta \in \mathcal{E}_t^{\text{TS}}} \left(\left(x_t(\tilde{\theta})\right)^\top \tilde{\theta} - \left(x_t(\theta)\right)^\top \theta\right) \mathbb{1}\{Z_t\} \;\middle|\; \mathcal{F}_t, \tilde{\theta} \in \Theta_t^{opt}\right], \tag{42}$$

Then, using Cauchy–Schwarz and the definition of $\gamma_t(\delta')$ in (24)

$$\mathbb{E}\left[\sup_{\theta \in \mathcal{E}_t^{\text{TS}}} \left(x_t(\tilde{\theta})\right)^\top \left(\tilde{\theta} - \theta\right) \mathbb{1}\{Z_t\} \;\middle|\; \mathcal{F}_t, \tilde{\theta} \in \Theta_t^{opt}\right]$$
$$\leq \mathbb{E}\left[\left\|x_t(\tilde{\theta})\right\|_{V_t^{-1}} \sup_{\theta \in \mathcal{E}_t^{\text{TS}}} \left\|\tilde{\theta} - \theta\right\|_{V_t} \;\middle|\; \mathcal{F}_t, \tilde{\theta} \in \Theta_t^{opt}, Z_t\right] \mathbb{P}(Z_t)$$
$$\leq 2\gamma_t(\delta') \mathbb{E}\left[\left\|x_t(\tilde{\theta})\right\|_{V_t^{-1}} \;\middle|\; \mathcal{F}_t, \tilde{\theta} \in \Theta_t^{opt}, Z_t\right] \mathbb{P}(Z_t).$$

This property shows that the regret R_t^{TS} is upper bounded by V_t^{-1}-norm of the optimal safe action corresponding to the any optimistic parameter $\tilde{\theta}$. Hence, we need to show that TS samples from the optimistic set with high frequency. We prove in Lemma VI.5 that TS is optimistic with a fixed probability $(\frac{p}{2})$ which leads to bounding R_t^{TS} as follows:

$$R_t^{\text{TS}} \frac{p}{2} \leq 2\gamma_t(\delta') \mathbb{E}\left[\left\|x_t(\tilde{\theta}_t)\right\|_{V_t^{-1}} \;\middle|\; \mathcal{F}_t, \tilde{\theta}_t \in \Theta_t^{opt}, Z_t\right] \mathbb{P}(Z_t) \frac{p}{2} \leq \tag{43}$$
$$2\gamma_t(\delta') \mathbb{E}\left[\left\|x_t(\tilde{\theta}_t)\right\|_{V_t^{-1}} \;\middle|\; \mathcal{F}_t, \tilde{\theta}_t \in \Theta_t^{opt}, Z_t\right] \mathbb{P}(Z_t) \mathbb{P}\left(\tilde{\theta}_t \in \Theta_t^{\text{opt}} \;\middle|\; \mathcal{F}_t, Z_t\right)$$
$$\leq 2\gamma_t(\delta') \mathbb{E}\left[\left\|x_t(\tilde{\theta}_t)\right\|_{V_t^{-1}} \;\middle|\; \mathcal{F}_t, Z_t\right] \mathbb{P}(Z_t). \tag{44}$$

By reintegrating over the event Z_t we get

$$R_t^{\text{TS}} \leq \frac{4\gamma_t(\delta')}{p}\mathbb{E}\left[\left\|x_t(\tilde{\theta}_t)\right\|_{V_t^{-1}} \mathbb{1}\{Z_t\} \;\middle|\; \mathcal{F}_t\right]. \quad (45)$$

Recall that $E_t \subset Z_t$, hence

$$\begin{aligned} R^{\text{TS}}(T) &\leq \sum_{t=1}^{T} R_t^{\text{TS}} \mathbb{1}\{E_t\} \\ &\leq \frac{4\gamma_T(\delta')}{p} \sum_{t=1}^{T} \mathbb{E}\left[\left\|x_t(\tilde{\theta}_t)\right\|_{V_t^{-1}} \;\middle|\; \mathcal{F}_t\right]. \end{aligned} \quad (46)$$

For bounding this term, we rewrite the RHS above as:

$$\begin{aligned} R^{\text{TS}}(T) \leq &\sum_{t=1}^{T} \|x_t\|_{V_t^{-1}} + \\ &\sum_{t=1}^{T} \left(\mathbb{E}\left[\left\|x_t(\tilde{\theta}_t)\right\|_{V_t^{-1}} \;\middle|\; \mathcal{F}_t\right] - \|x_t\|_{V_t^{-1}}\right). \end{aligned} \quad (47)$$

We can now bound the first expression using Proposition VI.1. For the second expression we proceed as follows:

- First, the sequence

$$Y_t = \sum_{s=1}^{t} \left(\mathbb{E}\left[\left\|x_s(\tilde{\theta}_s)\right\|_{V_s^{-1}} \;\middle|\; \mathcal{F}_s\right] - \|x_s\|_{V_s^{-1}}\right)$$

  is a martingale by construction.
- Second, under Assumption 3, $\|x_t\|_2 \leq L$, and since $V_t^{-1} \leq \frac{1}{\lambda} I$, we can write

$$\mathbb{E}\left[\left\|x_s(\tilde{\theta}_s)\right\|_{V_s^{-1}} \;\middle|\; \mathcal{F}_s\right] - \|x_s\|_{V_s^{-1}} \leq \frac{2L}{\sqrt{\lambda}}, \forall t \geq 1. \quad (48)$$

- Third, for bounding Y_T, we use Azuma's inequality, and we have that with probability $1 - \frac{\delta}{2}$,

$$Y_T \leq \sqrt{\frac{8TL^2}{\lambda} \log \frac{4}{\delta}}. \quad (49)$$

Putting these together, we conclude that with probability $1 - \frac{\delta}{2}$,

$$R^{\text{TS}}(T) \leq \frac{4\gamma_T(\delta')}{p}\left(\sqrt{2Td\log\left(1 + \frac{TL^2}{\lambda}\right)} + \sqrt{\frac{8TL^2}{\lambda}\log\frac{4}{\delta}}\right).$$

B. Bounding Term II

We can bound on Term II using the general result of [5]. In fact, we can use the following general decomposition:

$$\begin{aligned} &\sum_{t=1}^{T} (\text{Term II}) \mathbb{1}\{E_t\} := R^{\text{RLS}}(T) \\ &= \sum_{t=1}^{T} \left(x_t^\top \tilde{\theta}_t - x_t^\top \theta_\star\right) \mathbb{1}\{E_t\} \\ &\leq \sum_{t=1}^{T} \mid x_t^\top (\tilde{\theta}_t - \hat{\theta}_t) \mid \mathbb{1}\{E_t\} + \sum_{t=1}^{T} \mid x_t^\top (\hat{\theta}_t - \theta_\star) \mid \mathbb{1}\{E_t\}. \end{aligned} \quad (50)$$

By Definition VI.1, we have $E_t \subseteq Z_t$ and $E_t \subseteq \tilde{E}_t$, and hence

$$\begin{aligned} \mid x_t^\top (\tilde{\theta}_t - \hat{\theta}_t) \mid \mathbb{1}\{E_t\} &\leq \|x\|_{V_t^{-1}} \gamma_t(\delta') \\ \mid x_t^\top (\hat{\theta}_t - \theta_\star) \mid \mathbb{1}\{E_t\} &\leq \|x\|_{V_t^{-1}} \beta_t(\delta'). \end{aligned}$$

Therefore, from Proposition VI.1, we have with probability $1 - \frac{\delta}{2}$

$$R^{\text{RLS}}(T) \leq (\beta_T(\delta') + \gamma_T(\delta'))\sqrt{2Td\log\left(1 + \frac{TL^2}{\lambda}\right)}. \quad (51)$$

C. Overall Regret Bound

Recall that from (16), $R(T) \leq R^{TS}(T) + R^{\text{RLS}}(T)$. As shown previously, each term is bounded separately with probability $1 - \frac{\delta}{2}$. Using union bound over two terms, we get the following expression:

$$\begin{aligned} R(T) \leq &(\beta_T(\delta') + \gamma_T(\delta')(1 + \frac{4}{p}))\sqrt{2Td\log(1 + \frac{TL^2}{\lambda})} \\ &+ \frac{4\gamma_T(\delta')}{p}\sqrt{\frac{8TL^2}{\lambda}\log\frac{4}{\delta}}, \end{aligned} \quad (52)$$

holds with probability $1 - \delta$ where $\delta' = \frac{\delta}{6T}$.

For completeness we show below that action x_1 is safe. Having established that, it follows that the rest of the actions $x_t, t > 1$ are also safe with probability at least $1 - \delta'$. This is by construction of the feasible sets $\mathcal{D}_t^s$ and by the fact that $\mu_\star \in \mathcal{C}_t(\delta')$ with the same probability for each t.

Lemma VII.1. *The first action that Safe-LTS chooses is safe, that is* $x_1^\top \mu_\star \leq C$.

Proof. At round $t = 1$, the RLS-estimate $\hat{\mu}_1 = 0$ and $V_1 = \lambda I$. Thus, Safe-LTS chooses the action which maximizes the expected reward while satisfying $x_1^\top \hat{\mu}_1 + \beta_1(\delta') \|x_1\|_{V_1^{-1}} \leq C$. Hence, x_1 satisfies:

$$\beta_1(\delta') \|x_1\|_{V_1^{-1}} \leq C.$$

From Theorem II.1 and $V_1^{-1} = (1/\lambda)I$ leads to $S\|x_1\|_2 \leq C$ which completes the proof. □