Stochastic Linear Bandits with Unknown Safety Constraints and Local Feedback

Anonymous Authors[1]

Abstract

In many real-world decision-making tasks, e.g. clinical trials, the agents must satisfy a diverse set of unknown safety constraints at all times while getting feedback only on the safety constraints relevant to the chosen action, e.g. the ones close to violation. In this work, we study stochastic linear bandits with such unknown safety constraints and local safety feedback. The agent's goal is to maximize the cumulative reward while satisfying *multiple unknown affine or nonlinear* safety constraints. At each time step, the agent receives noisy feedback on a particular safety constraint *only if* the chosen action belongs to the associated constraint set, i.e. local safety feedback. For this setting, we design upper confidence bound and Thompson Sampling-based algorithms. In the design of these algorithms, we carefully prescribe an additional exploration incentive that guarantees the selection of high-reward actions that are also safe and ensures sufficient exploration in the relevant constraint sets to recover the optimal safe action. We show that for M distinct constraints, both of these algorithms attain $\tilde{\mathcal{O}}(\sqrt{MT})$ regret after T time steps without any safety violations. We empirically study the performance of the proposed algorithms under various safety constraints and with a real-world credit dataset. We show that both algorithms safely explore and quickly recover the optimal safe actions.

1. Introduction

Stochastic linear bandits (SLB) is a sequential decision-making framework where at each time step the agent aims to choose an action that maximizes the reward which is stochastic and its expected value is an unknown linear function of the action (Lattimore & Szepesvári, 2020). The goal of this problem is to minimize regret in decision-making, which denotes the difference between the cumulative reward obtained by the agent and the optimal cumulative reward if the reward function is known a priori. Since the underlying reward function is unknown, the main challenge is to balance the exploration-exploitation trade-off. In particular, besides maximizing the instantaneous reward, the agent needs to explore the action space to improve the accuracy of its estimate of the reward function to achieve a higher cumulative reward. Despite its simplicity, the SLB framework is able to capture the crux of the challenge in many decision-making under uncertainty applications such as recommendation systems (Li et al., 2010; Balakrishnan et al.), path planning (Dani et al., 2007; György et al., 2007), and wireless networks (Maghsudi & Hossain, 2016).

Safe Decision-Making: However, in many real-world decision-making problems, the agents require to satisfy some safety/operational constraints while aiming to maximize the cumulative reward. Thus, the tools developed for unconstrained SLB framework do not directly apply to real-world safety-critical decision-making tasks such as clinical trials (Villar et al., 2015). There have been several new frameworks with different forms of safety constraints proposed to model these tasks. Some of these frameworks include safety constraints through stage-wise reward (Moradipari et al., 2020; Khezeli & Bitar, 2020), while some of them focus on cumulative or policy-based constraints (Kazerouni et al., 2017; Pacchiano et al., 2021; Liu et al., 2021).

Prior Work on Safety-Critical SLB: Another line of work considers a more challenging setting of hard constraints on the actions, where the safety constraint needs to be met at every time step (stage-wise) (Amani et al., 2019; Moradipari et al., 2021). This setting is more suitable for safety-critical tasks where executing even one unsafe action may lead to catastrophic results. However, prior works in this setting only consider very simple models where there is a single unknown linear constraint depending on the reward function that the agent observes feedback from at every time step. Despite giving an initial understanding of safety in the SLB, these works do not capture the complex constraint and feedback structure of real-world decision-making tasks. The following considers a safety-critical decision-making scenario in which the prior works fail to model.

Modeling Case Study - Loan Approval: Consider an organization that gives out loans. The goal is to approve individuals who are likely to repay the loan and maximize the profit, *i.e.*, good credit risk, while avoiding extremely risky deals to ensure the safety of the organization's assets. Each individual is usually described as a diverse set of

feature representations, *i.e.*, actions, that the agent needs to interact with and select for approval of the loan, such that it learns which individuals are likely to have good credit risk. However, there can be various safety constraints that determine the extreme risk within the feature space. For example, the organization may want to ensure that it does not select "risky" individuals who have high credit or are retired and are classified as bad credit risk individuals. An increase in the rate of missed payments can provide feedback to adjust modeled safety constraints to avoid safety violations. This mechanism helps to ensure that the organization maximizes its return while strictly avoiding extreme risks. Previous safe SLB frameworks, which consider only a single stage-wise linear constraint, fail to model such a scenario. To address this challenge, multiple stage-wise safety constraints and a local feedback mechanism based on features are needed.

Our contributions

1. We study a novel SLB framework with unknown safety constraints where we model safety constraints as either multiple unknown affine or nonlinear functions, which generalizes the constraints considered in prior works. Given a set of actions, the goal of the decision-making agent is to maximize its reward by selecting safe actions defined by these constraints at every time step. Since the safety functions are unknown, the agent needs to learn them through feedback.

To design a realistic feedback mechanism, we model these safety constraints locally in the action space and associate feedback sets for each constraint. The agent receives a noisy observation of the constraint function only when it picks actions from the corresponding constraint feedback set. This local feedback mechanism captures the feedback structure in many applications where choosing actions outside of a known safe set is subject to safety constraints.

2. We first consider the framework with multiple unknown affine safety constraints. For this setting, we propose two novel safe SLB algorithms. The first algorithm is the safe version of the linear upper confidence bound algorithm (Abbasi-Yadkori et al., 2011) (LinUCB): Safe-LinUCB. In the design of Safe-LinUCB, we decouple the exploration for learning the reward parameter and the safety constraints. This is in contrast to prior works which rely on the same exploration strategy for both reward and safety which fails in the affine safety function setting. The main technical challenge in the design of Safe-LinUCB is to carefully prescribe an additional exploration within the UCB framework which guarantees the selection of optimistic safe actions and ensures sufficient exploration of the relevant constraint sets. For M distinct unknown affine constraints, we prove that Safe-LinUCB attains $\tilde{\mathcal{O}}(\sqrt{MT})$ regret after T time steps without violating any safety constraints.

3. We propose the safe version of the well-known Thompson Sampling algorithm (Abeille & Lazaric, 2017) (LinTS): Safe-LinTS. Safe-LinTS is a computationally efficient alternative to Safe-LinUCB which can possibly have computational challenges in finding optimistic actions similar to the other UCB algorithms. In the design of Safe-LinTS, unlike prior works, we also decouple the exploration for the reward parameter and safety functions such that the agent chooses the optimal action with respect to the sampled reward parameter from the estimated safe action sets at every time step.

The main technical challenge in the design of Safe-LinTS is to lower bound the probability of being optimistic for the sampled reward parameter while satisfying the safety constraints. To this end, we carefully design the sampling distributions for the reward parameter and safety functions such that the sampled parameters satisfy certain concentration and anti-concentration properties, and give a novel lower bound for this probability tailored for our SLB framework. For M distinct unknown affine constraints, we also show that Safe-LinTS attains $\tilde{\mathcal{O}}(\sqrt{MT})$ regret.

4. We study the setting of multiple unknown nonlinear constraints. We extend Safe-LinUCB and Safe-LinTS for this setting via a novel initial exploration strategy. We propose to learn Taylor approximations of the underlying safety constraints and design a new initial exploration phase that uses a priori known one safe action per constraint to achieve uniform exploration, *i.e.*, the persistence of excitation. We show that this exploration strategy allows the error in the estimates of the safety functions to be well-controlled and guarantees the identification of a safety set that contains the optimal safe action with high probability. We eventually show that the proposed method also attains $\tilde{\mathcal{O}}(\sqrt{T})$ regret.

5. We empirically study these algorithms on both synthetic and real-world datasets. On the synthetic dataset with various safety constraints, we observe that both algorithms achieve sublinear regret without any safety violations, concurring with our theoretical results. We then modify a credit classification task on the German Credit dataset (Keogh et al., 1998) into the loan approval SLB setting with two safety constraints by featurizing each individual as discussed in the case study above. We demonstrate the benefit of additional exploration in achieving improved regret while maintaining zero safety violations.

Our results subsume and generalize the state-of-the-art algorithms for SLB with stage-wise safe action constraints, see Table 1 for comparison. To summarize our contributions:

- We study a new SLB framework with multiple unknown affine or nonlinear safety constraints with a local safety feedback structure to capture real-world decision-making tasks.

- We propose two bandit algorithms for the proposed framework with affine constraints: Safe-LinUCB & Safe-LinTS, and prove $\tilde{\mathcal{O}}(\sqrt{T})$ regret upper bounds for both.

Table 1. Comparison with prior works on safe SLB with $\tilde{\mathcal{O}}(\sqrt{T})$ Regret. These works achieve this result using different methods for different safety aspects with different constraint types and for different numbers of constraints.

Work	Safety Aspect	Constraint Type	# of Constraints	Method
(Kazerouni et al., 2017)	Reward	Cumulative	1	UCB
(Khezeli & Bitar, 2020)	Reward	Stage-wise - Linear	1	UCB
(Moradipari et al., 2020)	Reward	Stage-wise - Linear	1	UCB + TS
(Pacchiano et al., 2021)	Policy	Stage-wise - Linear	Multiple	UCB
(Amani et al., 2019)	Action	Stage-wise - Linear	1	UCB
(Moradipari et al., 2021)	Action	Stage-wise - Linear	1	TS
Our Work	**Action**	**Stage-wise - Affine/Nonlinear**	**Multiple**	**UCB + TS**

• For nonlinear safety constraint setting, we extend the proposed algorithms via a novel initial exploration strategy to provide uniform exploration and show $\tilde{\mathcal{O}}(\sqrt{T})$ regret.

• We empirically study the proposed algorithms on real-world data and demonstrate that both algorithms explore efficiently without any safety violations and attain low regret.

2. Problem Formulation

Notation. For $x, y \in \mathbb{R}^d$, $\langle x, y\rangle = x^\top y$ and $\langle x, y\rangle_A = x^\top A y$, for a positive definite matrix $A \in \mathbb{R}^{d\times d}$. For $x \in \mathbb{R}^d$, we denote $\|x\| = \sqrt{x^\top x}$ and $\|x\|_A = \sqrt{x^\top A x}$. The maximum and the minimum eigenvalue of A is denoted by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ respectively. The maximum and minimum of two numbers α and β is denoted by $\alpha \vee \beta$ and $\alpha \wedge \beta$, respectively. For a positive integer n, $[n]$ denotes the set $\{1, \dots n\}$. $\tilde{\mathcal{O}}(\cdot)$ presents the order up to logarithmic terms.

Reward Model. We study the stochastic linear bandits (SLB) setting. At each time step t, the agent plays an action $x_t \in D_0$, where D_0 denotes the fixed decision set. Subsequently, the agent observes the reward $r_t = \mu^\top x_t + \eta_t^r$, where $\mu \in \mathbb{R}^d$ is unknown and η_t^r is random noise.

Safety Constraints. The environment is subjected to M distinct safety constraints, where $\mathbf{M} := [M]$ is the index set of the constraints. We model these constraints as affine functions unknown to the agent (they will be modeled as nonlinear functions in Section 5). We consider localized safety constraints, where we define associated constraint feedback sets $\Gamma_i \subseteq D_0, \forall i \in \mathbf{M}$. At each time step, the agent needs to satisfy all the constraints corresponding to the feedback sets that the chosen action x_t belongs to. More precisely, if $x_t \in \Gamma_i$, the agent needs to have

$$\gamma_i^\top x_t + c_i \le \tau, \quad \forall t, \tag{1}$$

for some γ_i and c_i are *unknown* and τ known to the agent $\forall i \in \mathbf{M}$. These constraints, therefore, form a region of safe actions $D_0^{\text{safe}} \subseteq D_0$, where

$$D_0^{\text{safe}} := \bigcup\nolimits_{i\in M} \left\{x \in \Gamma_i : \gamma_i^\top x_t + c_i \le \tau\right\}. \tag{2}$$

In this work, we study the setting where the agent is subject to hard constraints, *i.e.*, the agent needs to play actions that belong to D_0^{safe} with high probability at all time steps. This safety constraint formulation captures many safety-critical real-world decision-making applications. Since the safety constraints are unknown to the agent, and the agent needs to learn them via feedback and conservatively pick actions to ensure that the constraints are satisfied. In particular, we consider localized feedback such that the agent gets noisy observations of the constraint functions only when it picks actions from their corresponding constraint feedback set, *i.e.*,

$$\tilde{y}_t^i = \gamma_i^\top x_t + c_i + \eta_t^i \quad \text{if} \quad x_t \in \Gamma_i. \tag{3}$$

Figure 1 illustrates an example safety constraint structure.

Regret. We study the (pseudo) regret of the agent

$$R_T = \sum\nolimits_{t=1}^{t=T} \mu^\top x^* - \mu^\top x_t$$

for T time steps, where $x^* = \arg\max_{x\in D_0^{\text{safe}}} \mu^\top x$, *i.e.*, the optimal safe action. The goal of the agent is to minimize the regret over time and achieve sublinear regret while satisfying the safety constraints at all time steps. Let F_t denote the σ-algebra (history) up to time t, such that x_t is F_{t-1} measurable and the noise terms, *i.e.*, η_t^r and $\eta_t^{i_t}$, are F_t measurable. Before describing our first algorithm for this setting, we adopt some technical assumptions, which are standard in the literature, (Abbasi-Yadkori et al., 2011; Abeille & Lazaric, 2017; Amani et al., 2019; Khezeli & Bitar, 2020).

Assumption 2.1 (Subgaussian Noise)**.** For all $t \in [T]$ and $i \in \mathbf{M}$, η_t^r, η_t^i are conditionally R-sub-Gaussian where $R \ge 0$ is a fixed constant, *i.e.*, $\forall \lambda \in \mathbb{R}$, $\mathbb{E}[e^{\lambda \eta_t^r} | F_{t-1}] \le \exp(\lambda^2 R^2/2), \mathbb{E}[e^{\lambda \eta_t^i} | F_{t-1}] \le \exp(\lambda^2 R^2/2)$.

Assumption 2.2 (Boundedness)**.** $s < \|\mu\|_2, \|\gamma_i\|_2 < S, \|x\|_2 < L$, $\|x - x_i^s\|_2 < L^c \le L$, if $x \in \Gamma_i$ and $\mu^\top x \in [-1, 1], \forall x \in D_0$ for some $s, S, L, L^c > 0$.

Assumption 2.3 (Known safe actions)**.** For every constraint $i \in \mathbf{M}$, the agent knows a safe action $x_i^s \in \Gamma_i$ such that $x_i^s \in D_0^{\text{safe}}$ and $\gamma_i^\top x_i^s + c_i = \tau_i^s < \tau$ where τ, τ_i^s are known.

Note that Assumption 2.3 holds in many real-world decision-making tasks such as robotics and clinical trials where there are known safe actions. Note that the known safe actions do not need to be unique. If τ_i^ss are unknown, the agent can sample the known safe actions to estimate the values of τ_i^ss.

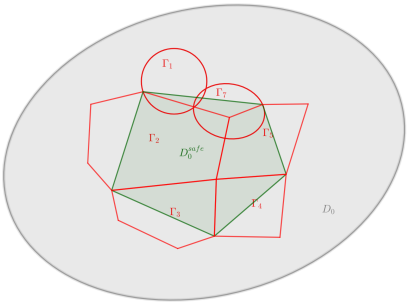

Figure 1. Illustration of the safety constraints: D_0 represents all actions. Γ_i represents the constraint feedback regions where the affine constraints (1) need to be satisfied. D_0^{safe} is the safe set of actions formed by the union of the safe regions from each Γ_i.

3. Safe-LinUCB

In this section, we propose Safe-LinUCB, the safe version of the well-known linear upper confidence bound algorithm studied in the literature (Abbasi-Yadkori et al., 2011), (also named OFUL) for the SLB framework presented in Section 2. Similar to LinUCB, Safe-LinUCB deploys the optimism in the face of uncertainty (OFU) principle to balance the exploration vs. exploitation trade-off. This algorithmic approach proposes to construct confidence sets for the underlying parameter μ using the history of actions and rewards and to play the optimal action for the most optimistic model within these sets. However, unlike the unconstrained setting of LinUCB, the agent in our SLB framework needs to satisfy the unknown safety constraints at every time step.

To address this, Safe-LinUCB conservatively explores starting around the known safe actions to learn the safety constraints as well as the underlying reward parameter while avoiding safety violations. During the course of interaction, besides the confidence set for the underlying reward parameter μ, it forms confidence sets for the unknown safety functions, i.e. affine parameters γ_i, and includes this information to safely expand its estimate of the safety set D_0^{safe}. In deploying the OFU principle, it includes an additional exploration to tolerate the uncertainty in the safety set estimate which enforces the algorithm to pick conservatively to avoid safety violations. Safe-LinUCB is given in Algorithm 1. Safe-LinUCB consists of 3 key elements: Parameter estimation, Safety construction, and Acting optimistically.

Parameter Estimation. At each time step t, Safe-LinUCB uses the history of action-reward pairs to obtain a ℓ_2-regularized (for some $\lambda > 0$) least squares (RLS) estimate of the underlying reward parameter μ via

$$\hat{\mu}_t = V_t^{-1} \sum_{k=1}^{t-1} r_k x_k, \tag{4}$$

where $V_t = \lambda I + \sum_{k=1}^{t-1} x_k x_k^\top$. Safe-LinUCB then builds a confidence set around this RLS estimate

$$\mathcal{C}_t = \left\{ v \in \mathbb{R}^d : \|v - \hat{\mu}_t\|_{V_t} \leq \beta_t \right\}, \tag{5}$$

Algorithm 1 Safe-LinUCB

Input: $\tau_i^s, x_i^s, \tau, L, S, R$
for t=1,. . . ,T **do**
  Compute $\hat{\mu}_t$ via (4) & $\hat{\gamma}_{i,t}$ via (6)
  Construct β_t in (5) & $\beta_t^i \; \forall i \in \mathbf{M}$ in (7)
  Construct D_t^{safe} according to (8)
  Find $x_t = \text{argmax}_{i \in \mathbf{M}, x \in \hat{\Gamma}_{i,t}} \mathbf{ucb}(x, i, t-1)$ via (9)
  Play x_t and observe reward r_t

where $\beta_t = R\sqrt{d \log\left((1+(t-1)L^2/\lambda)/\delta\right)} + \sqrt{\lambda} S$, for $\delta \in (0,1)$. The choice of β_t follows from Theorem 2 of Abbasi-Yadkori et al. (2011), such that $\mu \in \mathcal{C}_t$ with probability at least $1-\delta$, for all $t > 0$. Thus, Safe-LinUCB guarantees that the event $\mathcal{E}_\mu = \{\mu \in \mathcal{C}_t\}$ holds with high probability.

Similarly, Safe-LinUCB estimates the unknown safety functions, *i.e.*, parameters γ_i for all $i \in \mathbf{M}$, via RLS as

$$\hat{\gamma}_{i,t} = A_{i,t}^{-1} \sum_{k=1}^{N_i(t)} y_k^i (x_k - x_i^s), \tag{6}$$

for $y_t^i = \tilde{y}_t^i - \tau_i^s, \forall t$, where $A_{i,t} = \lambda I + \sum_{k=1}^{N_i(t)} (x_k - x_i^s)(x_k - x_i^s)^\top$ and $N_i(t)$ denotes the number of times the agent has gotten feedback from the constraint set Γ_i until time t. It also builds confidence sets around these estimates

$$\mathcal{C}_t^i = \left\{ v \in \mathbb{R}^d : \|v - \hat{\gamma}_{i,t}\|_{A_t^i} \leq \beta_t^i \right\}, \tag{7}$$

with $\beta_t^i = R\sqrt{d \log\left(|M|(1+N_i(t)L^2/\lambda)/\delta\right)} + \lambda^{1/2} S_\gamma$, such that the event $\mathcal{E}_{\gamma_i} = \{\gamma_i \in \mathcal{C}_t^i\}$ holds with probability at least $1-\delta$, for all $t > 0$ and $i \in \mathbf{M}$.

Safety Construction. Next, conditioned in the joint event $\mathcal{E} := \mathcal{E}_\mu \cup \bigcup_{i \in M} \mathcal{E}_{\gamma_i}$, Safe-LinUCB aims to satisfy the unknown safety constraints when picking actions. To achieve this, it conservatively constructs a safe set of actions $\hat{\Gamma}_{i,t} := \{x \in \Gamma_i : \hat{\gamma}_{i,t}^\top (x - x_i^s) + \beta_t^i \|x - x_i^s\|_{A_{i,t}^{-1}} \leq \tau - \tau_i^s\}$, where

$$D_t^{\text{safe}} = \bigcup_{i \in M} \hat{\Gamma}_{i,t}. \tag{8}$$

For this constructed safety set, we have the following result.

Lemma 3.1. *Conditioned on $\mathcal{E}$, $D_t^{safe} \subseteq D_0^{safe}$, for all $t > 0$.*

The proof is given in Appendix A.1, where we show that conditioned on $\mathcal{E}$, $\hat{\gamma}_{i,t}^\top (x - x_i^s) + \beta_t^i \|x - x_i^s\|_{A_{i,t}^{-1}}$ is an upper bound on $\gamma_i^\top (x - x_i^s)$, $\forall i \in M$. This ensures that D_t^{safe} is a conservative estimate of D_0^{safe}, such that Safe-LinUCB satisfies the safety constraints with high probability.

Acting Optimistically. At the final step, Safe-LinUCB picks an action x_t from the constructed safe set D_t^{safe} which maximizes the Upper Confidence Bound (**ucb**) defined as

$$\mathbf{ucb}(x, i, t) = \hat{\mu}_t^\top x + \beta_t \|x\|_{V_t^{-1}} + k_i \beta_t^i \|x - x_i^s\|_{A_{i,t}^{-1}}, \tag{9}$$

$\forall i \in \mathbf{M}$, where $k_i \geq 2LS/(\tau - \tau_i^s)$. In the following, we show that this construction of **ucb** ensures sufficient exploration of the safety constraint set in order to balance exploration vs. exploitation via optimistic action selection.

3.1. Theoretical guarantees for Safe-LinUCB

Before presenting the theoretical guarantees, we place the following technical assumption on the safety feedback sets that the optimal safe action belongs to, denoted as Γ_{i^*}.

Assumption 3.2 (Star convex optimal constraint sets)**.** Γ_{i^*} is star convex around the safe known action $x_{i^*}^s$ such that the convex combination $\alpha x^* + (1-\alpha)x_i^s \in \Gamma_{i^*}, \forall \alpha \in [0,1]$.

Note that since the constraint sets are localized around a particular safe action x_i^s, this assumption is reasonable in the safe SLB framework, and weaker than the prior work, e.g. (Amani et al., 2021), where the entire space of actions is considered to be star convex. In the regret analysis of Safe-LinUCB, we follow the standard analysis of UCB and decompose the regret R_T into two terms: **(i)** $\sum_{t=1}^{t=T}(\mu^\top x^* - \textbf{ucb}(x_t, i_t, t))$ and **(ii)** $\sum_{t=1}^{t=T}(\textbf{ucb}(x_t, i_t, t) - \mu^\top x_t)$.

In the unconstrained setting, the optimism principle is satisfied by construction, since the optimal action belongs to the decision set D_0, yielding **(i)** to be non-positive. However, in the safe SLB framework, the optimal safe action x^* may not belong to the constructed safe set D_t^{safe} where optimistic action selection happens. Thus, we first show that the new construction of **ucb** in (9) still provides optimistic actions.

Theorem 3.3 (Optimism)**.** *For all $i \in \mathbf{M}$, setting $k_i \geq 2LS/(\tau - \tau_i^s)$ guarantees optimism with high probability:*

$$\max_{i \in M, x \in \hat{\Gamma}_{i,t}} \textbf{\textit{ucb}}(x, i, t) \geq \mu^\top x^* \quad \forall t.$$

The proof is given in Appendix A.2. To sketch the proof idea, we consider two cases of whether $x^* \in D_t^{\text{safe}}$ or not. If yes, via standard UCB arguments, we guarantee that Safe-LinUCB selects optimistic actions. If not, we show that the additional exploration bonus $k_i \beta_t^i ||x - x_i^s||_{A_t^{i-1}}$ ensures optimistic action selection for the given choice of k_i. This shows that adjusting the additional exploration bonus around the known safe actions ensures that the relevant constraint regions are well-explored, *i.e.*, x^* eventually belongs to D_t^{safe}.

The choice of k_i highlights the key challenge in our proposed SLB framework. In contrast to prior works, the agent gets feedback from a constraint only if it plays an action within the associated feedback set. Therefore, while aiming to learn the underlying reward function, Safe-LinUCB needs to cautiously choose actions from the constraint sets where it wants to learn the constraints at the cost of not receiving any feedback from the non-active constraints. The new **ucb** term in (9) captures this trade-off and selecting k_i as in Theorem 3.3 balances it effectively. In particular, we see that this exploration bonus is inversely proportional to the gap between the safety threshold and the value of the known safe action. Intuitively, this means that if the known safe action is close to violation, Safe-LinUCB needs to explore more/act more optimistically to learn the optimal safe action. We pay an extra price in regret due to this additional effort.

Theorem 3.4 (Regret Bound)**.** *Suppose Assumptions 2.1-2.3 and 3.2 hold. Then for any $\delta \in (0,1)$ and $k_i = 2LS/(\tau - \tau_i^s)$, with probability at least $1-2\delta$, the regret of Safe-LinUCB is $R_T \leq R_\mu + R_\gamma$, where $R_\mu = 2\beta_T \sqrt{2Td\log((1+TL^2/(d\lambda))/\delta)}$ and $R_\gamma = (k_{i_{max}}\beta_T^{i_{max}} + 2)\sqrt{2|M|Td\log((1+TL^2/(d\lambda))/\delta)}$, for $\beta_T^{i_{max}} = \max_{j \in \mathbf{M}} \beta_T^j$ and $k_{i_{max}} = \max_{j \in \mathbf{M}} k_j$.*

The proof is given in Appendix A.3. In the proof, since **(i)** is non-positive via Theorem 3.3, we study **(ii)** and decompose it into 2 terms. R_μ results from learning the unknown reward parameter and R_γ is due to learning M different constraints. Notice that R_γ scales with the hardest, *i.e.*, the most exploration needed, constraint through $\beta_T^{i_{max}}$ and $k_{i_{max}}$. Moreover, the regret rate of Safe-LinUCB matches the prior unconstrained UCB results (Abbasi-Yadkori et al., 2011) and single linear constrained UCB results (Amani et al., 2019; Pacchiano et al., 2021), where the additional price of learning under M distinct constraints with local safety feedback, which generalizes the prior work, appears as $\sqrt{M}$.

Extensions: 1) Many Constraints. For a significantly large number of constraints, R_γ dominates the overall regret. Since the ultimate goal is to pick actions with highest reward, Safe-LinUCB should focus further to learn the constraints around the optimal action. In Appendix B, we propose a modified version of Safe-LinUCB, where we add a pure exploration phase in which the agent learns about the general direction of the unknown reward parameter. This information helps the agent to recognize the important constraint sets and improves the efficiency of Safe-LinUCB.

2) Featurized Constraints. In many real-world scenarios, the safety constraints can be affine in (un)known feature maps of actions. These feature maps can be complex and obtained via a deep neural network. Safe-LinUCB can be easily extended to these scenarios. In Section 6, we study an SLB problem on a real-world credit dataset with such featurization. Note that these feature maps are only needed to hold locally, i.e., within feedback sets, which is more general than considering the same featurization for the entire space.

4. Safe-LinTS

In many scenarios, solving the bilinear optimization problem of UCB-type algorithms, *i.e.*, line 6 of Algorithm 1, can be computationally challenging. To this end, Thompson Sampling (TS)-based methods are proposed, e.g. LinTS (Agrawal & Goyal, 2013; Abeille & Lazaric, 2017).

These approaches sample a model within the constructed confidence set of plausible models and find the optimal action with respect to this sampled model. Therefore, they consider a linear optimization problem for decision-making, which can be solved efficiently. Because of this computational efficiency, simplicity, and possibly better empirical performance, they are adopted in many decision-making scenarios (Abeille & Lazaric, 2018; Kargin et al., 2022). In this section, we propose Safe-LinTS, the safe version of LinTS. The pseudocode of Safe-LinTS is given in Algorithm 2.

The construction of Safe-LinTS follows similarly to Safe-LinUCB regarding the estimation of the reward parameter and safety parameters and safety construction (Lines 3-5). After this, it draws two random perturbations $\eta_t \in \mathbb{R}^d$ and $\eta_t^c \in \mathbb{R}^d$ from i.i.d. distributions $\mathcal{P}^{TS}$ and $\mathcal{P}_c^{TS}$ respectively (will be characterized shortly). Among these perturbations, while Safe-LinTS uses η_t in a standard way to sample a reward parameter, it uses η_t^c in a novel way to expand the estimated safe set to satisfy optimistic action selection.

The main novelty in the design of Safe-LinTS lies in this decoupling of the exploration for the reward parameter and the safety functions. In particular, the prior work in safe linear bandits (Moradipari et al., 2021) relies on using the same Gram matrix to learn both the safety and reward parameters simultaneously. However, learning in the affine setting involves separate Gram matrices, thus, Safe-LinTS explicitly balances the exploration trade-off between learning the unknown reward parameter and the safety parameters, ensuring safety and optimism for the entire horizon.

To this end, $\mathcal{P}^{TS}$ and $\mathcal{P}_c^{TS}$ are chosen to satisfy certain concentration and anti-concentration properties. In particular, for some $\delta \in (0,1)$ and constants c, c', Safe-LinTS selects $\mathcal{P}^{TS}$ such that $\mathbb{P}(\|\eta_t\|_2 \leq \sqrt{cd\log(c'd/\delta)}) \geq 1-\frac{\delta}{2}$, and $\mathbb{P}(u^\top \eta_t \geq 1) = p_1 > 0$, for any $u \in \mathbb{R}^d$ with $\|u\| = 1$. Similarly, $\mathcal{P}_c^{TS}$ is chosen such that $\mathbb{P}(\|\eta_t^c\|_2 \leq \frac{2LS^\gamma}{\tau-\tau_*^s}\sqrt{cd\log(c'd/\delta)})) \geq 1-\frac{\delta}{2}$ and $\mathbb{P}(u^\top \eta_t^c \geq \frac{2LS^\gamma}{\tau-\tau_*^s}) = p_2 > 0$, where $S^\gamma > \max_{i\in\mathbf{M}} \|\gamma_i\|$ and $\tau_*^s = \max_{i\in\mathbf{M}} \tau_i^s$. These requirements imply that these distributions with high probability should concentrate, yet, still provide a certain amount of exploration (anti-concentration), which is crucial in achieving low regret. Natural candidates for $\mathcal{P}^{TS}$ and $\mathcal{P}_c^{TS}$ are $\mathcal{N}(0, I)$ and $\mathcal{N}(0, \frac{2LS^\gamma}{\tau-\tau_*^s} I)$, respectively.

Safe-LinTS uses $\eta_t \sim \mathcal{P}^{TS}$ to sample $\tilde{\mu}_t$ around $\hat{\mu}_t$ which provides the balance between exploration and exploitation while learning the unknown reward parameter, *i.e.*, $\tilde{\mu}_t = \hat{\mu}_t + \beta_t V_t^{-1/2}\eta_t$. It then uses $\eta_t^c \sim \mathcal{P}_c^{TS}$ to sample $\tilde{\omega}_{i,t} = \beta_t^i A_{i,t}^{-1/2}\eta_t^c, \forall i \in \mathbf{M}$, which will be used to provide the exploration needed to expand the estimated safe set to include higher rewarding actions, *i.e.*, optimistic actions.

At the final step, Safe-LinTS picks an action x_t from D_t^{safe} by maximizing $\tilde{\mu}_t^\top x + \tilde{\omega}_{i,t}^\top (x - x_i^s)$. Note that this is a linear objective with transparent exploration goals. In particular, the reward exploration is similar to LinTS in Abeille & Lazaric (2017), whereas the second term adds exploration along the safety constraints using the known safe actions. Notice that this approach generalizes the algorithm proposed in (Moradipari et al., 2021) whose setting is a special case of the SLB framework considered in this work.

Algorithm 2 Safe-LinTS

1: **Input:** $\tau_i^s, x_i^s, \tau, L, S, R, \mathcal{P}^{TS}, \mathcal{P}_c^{TS}$
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Compute $\hat{\mu}_t$ via (4) & $\hat{\gamma}_{i,t}$ via (6)
4: Construct β_t in (5) & $\beta_t^i \; \forall i \in M$ in (7)
5: Construct D_t^{safe} according to (8)
6: Sample $\eta_t \sim \mathcal{P}^{TS}$ and $\eta_t^c \sim \mathcal{P}_c^{TS}$
7: Compute $\tilde{\mu}_t = \hat{\mu}_t + \beta_t V_t^{-1/2}\eta_t$
8: Compute $\tilde{\omega}_{i,t} = \beta_t^i A_{i,t}^{-1/2}\eta_t^c, \forall i \in \mathbf{M}$
9: Find $x_t = \text{argmax}_{i\in\mathbf{M}, x\in\hat{\Gamma}_{i,t}} \tilde{\mu}_t^\top x + \tilde{\omega}_{i,t}\top (x - x_i^s)$
10: Play x_t and observe reward r_t

Theorem 4.1. *Suppose Assumptions 2.1-2.3 and 3.2 hold. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, the regret of Safe-LinTS is $R_T = \tilde{\mathcal{O}}(d^{3/2}\sqrt{|M|T})$.*

The proof and the exact expressions are given in Appendix C. In the proof, we first show that Safe-LinTS selects safe (via Lemma 3.1) optimistic actions with at least $p_1 p_2$ probability by showing that η_t and η_t^c provide sufficient exploration (Appendix C.1). Finally, we use the regret decomposition in (Abeille & Lazaric, 2017) to give the regret upper bound. Notably, this result matches the regret upper bound in (Moradipari et al., 2021) for their setting. In the exact regret expression, the leading term has $1/(p_1 p_2)$ *i.e.*, the inverse of optimistic action probability. This relation is similar to that of $k_{i_{max}}$ in Safe-LinUCB. In particular, similar to Safe-LinUCB, for a smaller worst-case safety gap of known safe actions, Safe-LinTS needs to explore more to learn the optimal safe action which results in increased regret through p_2.

5. Linear Bandits with Nonlinear Constraints

In this section, we consider the most general setting of multiple nonlinear safety constraints, which captures the most diverse class of decision-making scenarios.

Safety Constraints: The environment is subject to M distinct nonlinear safety constraints, such that if $x_t \in \Gamma_i$, the agent needs to have $f_i(x_t) \leq \tau, \forall t$ for some unknown f_i and known $\tau, \forall i \in \mathbf{M}$. The region of safe actions corresponds to $D_0^{\text{safe}} := \bigcup_{i\in\mathbf{M}} \{x \in \Gamma_i : f_i(x) \leq \tau\}$. Similar to the affine case, we consider localized feedback for the agent: $\tilde{y}_t^i = f_i(x_t) + \eta_t^i$ if $x_t \in \Gamma_i$. Moreover, without loss of generality, we consider $\Gamma_i = \{x \in \mathbb{R}^d : \|x - x_i^s\|_2 \leq \delta_f\}$ for some $\delta_f > 0$ for all $i \in \mathbf{M}$. In parallel with Assumption

2.3, we assume that the agent knows a safe action for each constraint, $x_i^s \in D_0^{\text{safe}}, \forall i \in \mathbf{M}$, as well as their safety values $f_i(x_i^s) = \tau_i^s < \tau$. Finally, we adopt the following simple regularity assumption on the nonlinear constraints.

Assumption 5.1 (Smooth & Lipschitz Safety Constraints)**.** $f_i(x)$ is ζ-smooth and S-Lipschitz, $\forall x \in \Gamma_i$, $\forall i \in \mathbf{M}$.

The local smoothness assumption is a significantly weak assumption (Bartlett et al., 2019), while the local Lipschitzness is the nonlinear analog to Assumption 2.2 with affine constraints. The setting characterized above subsumes and generalizes the affine case in Section 2. Using the first-order Taylor expansion about the known safe actions, we obtain $f_i(x) = f_i(x_i^s) + \nabla f_i(x_i^s)^\top (x - x_i^s) + \epsilon_i(x)$, where $\epsilon_i(x)$ represents the remainder terms. Notice that for small enough δ_f, this expansion behaves very similarly to affine functions studied in previous sections, which motivates the following our algorithm design. To avoid any further structural assumptions and keep the setting as general as possible, while keeping the problem tractable, we assume some safety gap for the optimal safe action to account for the function approximation errors.

Assumption 5.2 (Safety gap for optimal action)**.** The optimal safe action x^* has at least Δ safety gap from constraint boundary, *i.e.*, $f_{i^*}(x^*) \leq \tau - \Delta$, such that $\Delta > \zeta \delta_f^2$.

This is a mild assumption since for a nonlinear function the optimal action need not be at the boundary, unlike linear constraints. Moreover, this assumption holds in many safe decision-making tasks, where the optimal safe action might be a significantly safe one, yet, to learn this action one might need to consider a higher threshold in the learning process.

5.1. Safe-LinUCB/LinTS with Pure Exploration

We propose an extension of our prior algorithms to achieve safe and effective decision-making for the SLB with multiple nonlinear safety constraints. Due to Assumption 5.1, we know that there exists a safe ball of actions around each $x_i^s, \forall i \in \mathbf{M}$, *i.e.*, $f_i(x_t) \leq \tau$ if $x_t \in \{x \in \Gamma_i : ||x - x_i^s||_2 < \delta_r\}$ for $\delta_r \leq (\tau - \tau_i^s)/(S + \zeta \delta_f)$. The existence of this ball helps the agent to estimate the gradient of the nonlinear function around the known safe actions x_i^s. The main idea in our algorithm design is to learn the first-order function approximation in each Γ_i while taking into account the estimation error so that the agent can eventually get to the optimal action x^* without violating safety. The algorithm consists of two phases: (i) Pure Exploration and (ii) Safe-LinUCB/LinTS. The pseudocode is given in Algorithm 3.

Pure Exploration. In this phase, the agent samples T' actions from each constraint set Γ_i. It uniformly excites all the directions by playing $x_t = \arg\max_{x \in D_i^w} \|x - x_i^s\|_{A_{i,t}^{-1}}$ for T' steps, where D_i^w is the $d-1$ dimensional boundary surface of the δ_r-ball around the known safe actions x_i^s

Algorithm 3 Safe-LinUCB/LinTS with Pure Exploration

1: **Input:** $\tau_i^s, x_i^s, \tau, \zeta, S, \Delta, \delta_f$
2: **for** $i \in \mathbf{M}$ **do**
3: **for** $t = 1, 2, \ldots, T'$ **do**
4: Play $x_t = \arg\max_{x \in D_i^w} \|x - x_i^s\|_{A_{i,t}^{-1}}$
5: Construct $D_{MT'}^{\text{safe}}$
6: Run Safe-LinUCB/LinTS for the remainder with $D_{MT'}^{\text{safe}}$

defined as $D_i^w = \{x \in \Gamma_i : ||x - x_i^s||_2 = \delta_r\}$, and $A_{i,t} = \lambda I + \sum_{k=1}^{N_i(t)} (x_k - x_i^s)(x_k - x_i^s)^\top$. By construction of D_i^w, the agent achieves safe exploration. Moreover, this exploration strategy ensures that the agent always picks the direction of the smallest eigenvalue, resulting in persistent excitation in all directions since actions in D_i^w have the same norm.

At the end of this phase, the algorithm estimates the gradient of the constraint functions using RLS such that $\nabla \hat{f}_{it} = A_{i,t}^{-1} \sum_{k=1}^{N_i(t)} y_k^i (x_k - x_i^s)$ for $y_t^i = \tilde{y}_t^i - \tau_i^s, \forall t$. Note that $N_i(t)$ is equal to T' for all i at the end of this phase. Next, the algorithm further decomposes $\nabla \hat{f}_{it} = \nabla \hat{f}_{it}^{LS} + \hat{\epsilon}_{it}$, where $\nabla \hat{f}_{it}^{LS} = A_{i,t}^{-1} \sum_{\tau=1}^{N_i(t)} (x_\tau - x_i^s)(\nabla f(x_i^s)^\top (x_\tau - x_i^s) + \eta_t^i)$ and $\hat{\epsilon}_{it} = A_{i,t}^{-1} \sum_{\tau=1}^{N_i(t)} (x_\tau - x_i^s) \epsilon_i(x_\tau)$. Notice that the expression for $\nabla \hat{f}_{it}^{LS}$ is the nonlinear analog of (6). Thus, the algorithm builds confidence sets around the estimates $\nabla \hat{f}_{it}^{LS}, \forall i \in \mathbf{M}$: $\mathcal{C}_t^i = \{v \in \mathbb{R}^d : \|v - \nabla \hat{f}_{it}^{LS}\|_{A_t^i} \leq \beta_t^i\}$, with $\beta_t^i = R\sqrt{d \log\left(|M|(1 + T'L^2/\lambda)/\delta\right)} + \lambda^{1/2} S$. It also defines the event $\mathcal{E}_{\nabla f_i} = \{\nabla f_i(x_i^s) \in \mathcal{C}_t^i\}$ which holds with probability at least $1 - \delta$, for all $t > 0$ and $i \in \mathbf{M}$.

Safety Construction. Next, conditioned in the joint event $\mathcal{E}_{\nabla f_t} := \bigcup_{i \in M} \mathcal{E}_{\nabla f_{it}}$, the algorithm aims to satisfy safety constraints when picking actions. To achieve this, it conservatively constructs a safe set of actions $\hat{\Gamma}_t^i = \{x \in \Gamma_i : \nabla \hat{f}_{it}^\top (x - x_i^s) + \frac{\Delta}{2} \leq \tau - \tau_i^s\}$, where $D_{MT}^{\text{safe}} = \bigcup_{i \in M} \hat{\Gamma}_{i,t}$.

Theorem 5.3. *Suppose Assumptions 5.1 & 5.2 hold. For any $\delta \in (0, 1)$, after T' time steps of pure exploration per constraint set, we have i) $x^* \in D_{MT'}^{safe}$ and ii) $D_{MT'}^{safe} \subseteq D_0^{safe}$ with probability at least $1 - \delta$, if $\frac{T'}{\log^2 T'} \geq \left(2d \frac{4\delta_f^2}{(\Delta - \zeta\delta_f^2)^2}\right)^2$.*

The proof is in Appendix D.1. The main idea of the proof is to show that we can control the error from non-linearity using smoothness and simultaneously learn the gradient at that point by uniformly playing actions around and close to the known safe actions. We then build $D_{|M|T'}^{\text{safe}}$ using $\hat{\nabla} f_i(x_i^s)$ and add error margin to compensate for smoothness approximation error, away from x_i^s. After this phase, the agent executes the previously proposed algorithms using $D_{|M|T'}^{\text{safe}}$.

Corollary 5.4 (Regret Bound)**.** *Suppose 5.1 & 5.2 hold. Then for the given duration of T' in Theorem 5.3, for any $\delta \in (0, 1)$, with probability at least $1 - 2\delta$, the regret of Algorithm 3 the above algorithm is $\tilde{\mathcal{O}}(|M|T' + \sqrt{T})$.*

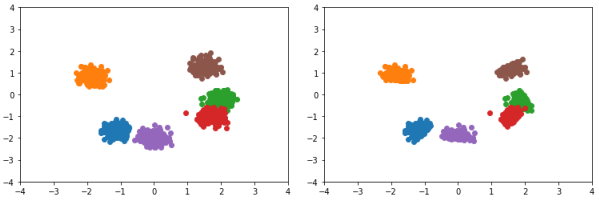

Figure 2. D_0 and D_0^{safe} respectively for affine constraints. Different colors represent different feedback sets Γ_i.

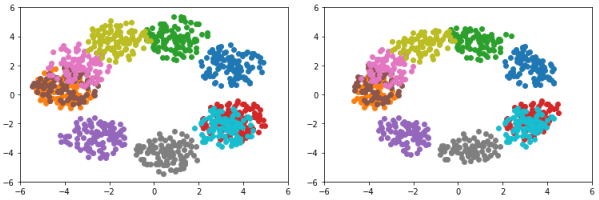

Figure 3. D_0 and D_0^{safe} respectively for nonlinear (ℓ_2 norm bound) constraints. Different colors represent different feedback sets Γ_i.

6. Experiments

6.1. Illustrative 2D Simulations

We first empirically study the proposed algorithms in $2D$ action space. In the setting with 6 unknown affine constraints and feedback regions, we perform 5 independent runs of Safe-LinUCB and Safe-LinTS for 2000 time steps and report their performance. An example of the decision set D_0 with different (color) feedback regions and the region of safe actions determined by the affine constraints are shown in Figure 2. The cumulative regret of the algorithms in this setting is given in the first plot of Figure 4. We observe that both of the algorithms achieve competitive, *i.e.*, sublinear, regret without any safety violations. We show that Safe-LinTS achieves improved practical performance in this setting with optimized exploration parameters η_t and η_t^c, which further motivates the use of sampling-based methods in practice.

Next, we study the setting with 10 unknown nonlinear constraints and feedback regions. We model the constraints as ℓ_2-norm bound constraints. An example of D_0 and D_0^{safe} are given in Figure 3. We consider an optimal action with a safety gap in parallel with Assumption 5.2. We implement Algorithm 3 using Safe-LinUCB and provide the cumulative regret in Figure 4. As predicted by the theory, algorithm attains linear regret during its orthogonal pure exploration phase. However, this phase allows sufficient exploration of the safety sets and unknown reward function such that Algorithm 3 discovers a safe action that achieves at least as high reward as the optimal action, yielding constant regret after pure exploration. This shows that the novel initial exploration strategy in Section 5 is effective in uniformly exploring the decision set without any safety violations.

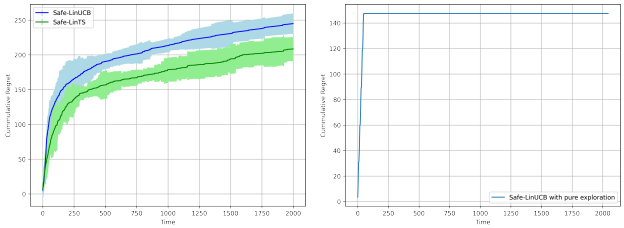

Figure 4. **Left:** Cumulative regret of Safe-LinUCB and Safe-LinTS for the setting in Fig. 2 (Solid line is the average, shaded region is one std), **Right:** Cumulative regret of Algorithm 3 (Safe-LinUCB with initial pure exploration) for the setting in Fig. 3.

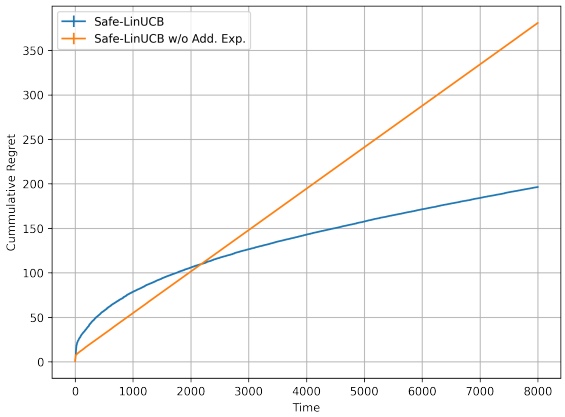

Figure 5. Cumulative regret in loan approval problem

6.2. Loan Approval as a Safe SLB Problem

We consider the German Credit Dataset from Keogh et al. (1998). The data classify customers as good or bad for credit for loan approval and provide 24 attributes per user. To turn this into a safe SLB problem, We featurize the user attributes using a neural network by posing it as a regression problem with affine safety constraints in the feature space. We impose two safety violations as picking bad customers with 1) high credit and 2) with high age, where the last one is a surrogate to retirement discussed in the case study at the beginning. We compare Safe-LinUCB with a naive version which does not include the additional exploration bonus needed to ensure optimism under safety.

Figure 5 gives the cumulative regret comparison. In the beginning, Safe-LinUCB attains higher regret than the naive version due to additional exploration incentive as expected. However, this additional exploration provides the sufficient exploration needed in the relevant constraint regions and allows Safe-LinUCB to achieve lower cumulative regret in the long run with no safety violations, concurring with the theory. The naive method, on the other hand, does not select optimistic actions and fails to explore efficiently, resulting in sub-optimal actions. This result highlights the importance of the carefully tuned exploration bonus under safety constraints to recover underlying reward parameter.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Abeille, M. and Lazaric, A. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pp. 176–184. PMLR, 2017.

Abeille, M. and Lazaric, A. Improved regret bounds for thompson sampling in linear quadratic control problems. In *International Conference on Machine Learning*, pp. 1–9, 2018.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pp. 127–135. PMLR, 2013.

Amani, S., Alizadeh, M., and Thrampoulidis, C. Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Balakrishnan, A., Bouneffouf, D., Mattei, N., and Rossi, F. Using contextual bandits with behavioral constraints for constrained online movie recommendation.

Bartlett, P. L., Gabillon, V., and Valko, M. A simple parameter-free and adaptive approach to optimization under a minimal local smoothness assumption. In *Algorithmic Learning Theory*, pp. 184–206. PMLR, 2019.

Dani, V., Kakade, S. M., and Hayes, T. The price of bandit information for online optimization. *Advances in Neural Information Processing Systems*, 20, 2007.

György, A., Linder, T., Lugosi, G., and Ottucsák, G. The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8(10), 2007.

Kargin, T., Lale, S., Azizzadenesheli, K., Anandkumar, A., and Hassibi, B. Thompson sampling achieves $\tilde{O}\sqrt{T}$ regret in linear quadratic control. In *Conference on Learning Theory*, pp. 3235–3284. PMLR, 2022.

Kazerouni, A., Ghavamzadeh, M., Abbasi Yadkori, Y., and Van Roy, B. Conservative contextual linear bandits. *Advances in Neural Information Processing Systems*, 30, 2017.

Keogh, E., Blake, C., and Merz, C. J. Uci repository of machine learning databases. *Irvine, CA: Uni of California, Department of Information and Computer Science*, 1998.

Khezeli, K. and Bitar, E. Safe linear stochastic bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10202–10209, 2020.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

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

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

Maghsudi, S. and Hossain, E. Multi-armed bandits with application to 5g small cells. *IEEE Wireless Communications*, 23(3):64–73, 2016.

Moradipari, A., Thrampoulidis, C., and Alizadeh, M. Stage-wise conservative linear bandits. *Advances in neural information processing systems*, 33:11191–11201, 2020.

Moradipari, A., Amani, S., Alizadeh, M., and Thrampoulidis, C. Safe linear thompson sampling with side information. *IEEE Transactions on Signal Processing*, 69:3755–3767, 2021.

Pacchiano, A., Ghavamzadeh, M., Bartlett, P., and Jiang, H. Stochastic bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, pp. 2827–2835. PMLR, 2021.

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

Villar, S. S., Bowden, J., and Wason, J. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

Appendix

In Appendix A, we provide the proofs for the regret analysis of Safe-LinUCB. In Appendix B, we discuss the ways to extend Safe-LinUCB to tolerate significantly many safety constraints and prove the regret guarantees for the modified algorithm. We provide the theoretical guarantees of Safe-LinTS in Appendix C. In Appendix D, we analyze the novel pure exploration strategy and give the regret guarantees for Algorithm 3. Appendix E contains auxiliary theorems and lemmas used in the proofs.

A. Proofs of Section 3 - Safe-LinUCB

A.1. Proof of Lemma 3.1

Lemma A.1. *Conditioned on $\mathcal{E}$, D_t^{safe} is a conservative estimate of D_0^{safe} i.e to say $D_t^{safe} \subseteq D_0^{safe}$*

Proof. If $x \in D_t^{\text{safe}}$, we have following two cases:

Case 1 : $x \in D^w$

Trivially $x \in D_0^{\text{safe}}$

Case 2 : $x \in \Gamma_i$

then by definition

$$\begin{aligned}\tau_i - \tau_i^s &\geq \hat{\gamma}_{i,t}^\top (x - x_i^s) + \beta_t^i ||x - x_i^s||_{A_{i,t}^{-1}} \\ &= \gamma_i^\top (x - x_i^s) + (\hat{\gamma}_{i,t} - \gamma_i)^\top (x - x_i^s) + \beta_t^i ||x - x_i^s||_{A_{i,t}^{-1}} \\ &\geq \gamma_i^\top (x - x_i^s) \quad \text{(Conditioned on } \mathcal{E}_\mu \text{ and Lemma E.1)}\end{aligned} \tag{10}$$

Therefore $x \in D_0^{\text{safe}}$. □

A.2. Proof of Theorem 3.3

Theorem A.2 (Optimism)**.** . *For optimism to hold i.e*

$$\boldsymbol{ucb}(x_t, i_t, t) \geq \mu^\top x^* \quad \forall t$$

we require

$$k_i \geq \frac{2LS}{\tau_i - \tau_i^s}$$

Proof. Recall

$$\textbf{ucb}(x, i, t) = \hat{\mu}_t^\top x + \beta_t ||x||_{V_t^{-1}} + k_i \beta_t^i ||x - x_i^s||_{A_{i,t}^{-1}}$$

We consider the following two cases :

Case(1) : $x^* \in D_t^{\text{safe}}$

$$\begin{aligned}\max_{i \in M, x \in \hat{\Gamma}_{i,t}} \textbf{ucb}(x, i, t) &\geq \textbf{ucb}(x^*, i^*, t) \\ &\geq \hat{\mu}^\top x^* + \beta_t ||x^*||_{V_t^{-1}} \quad \text{(Since } k_i \beta_t^i ||x - x_i^s||_{A_{i,t}^{-1}} \geq 0) \\ &= < \mu, x^* > + < \hat{\mu}_t - \mu, x^* > + \beta_t ||x^*||_{V_t^{-1}} \\ &\geq < \mu, x^* > + (1-1)\beta_t ||x^*||_{V_t^{-1}} \quad \text{(Conditioned on } \mathcal{E}_\mu \text{ and Lemma E.1)} \\ &\geq \mu^\top x^*\end{aligned} \tag{11}$$

Case(2) : $x^* \notin D_t^{\text{safe}}$

We consider the constraint set Γ_{i^*} in which x^* belongs to and define

$$\alpha_t = \max\{\alpha \in [0,1] : \alpha\hat{\gamma}_{i^*t}^\top(x^* - x_{i^*}^s) + \alpha\beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}} = \tau - \tau_{i^*}^s\}$$

This definition ensures $z_t = \alpha_t x^* + (1-\alpha_t)x_{i^*}^s \in D_t^{\text{safe}}$

Now we have

$$\begin{aligned}
\max_{i\in M, x\in\hat{\Gamma}_{i,t}} \textbf{ucb}(x,i,t) &\geq \mu^\top z_t + (\hat{\mu}_t - \mu)^\top z_t + \beta_t||z_t||_{V_t^{-1}} + k_{i^*}\beta_t^{i^*}||z_t - x_{i^*}^s||_{A_{i^*,t}^{-1}} \\
&\geq \mu^\top z_t + k_{i^*}\beta_t^{i^*}||z_t - x_{i^*}^s||_{A_{i^*,t}^{-1}} \quad \text{(Conditioned on } \mathcal{E}_\mu \text{ and Lemma E.1)} \\
&\geq \alpha_t\mu^\top(x^* - x_{i^*}^s) + \mu^\top x_{i^*}^s + \alpha_t k_{i^*}\beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}} \\
&\geq \alpha_t[\mu^\top(x^* - x_{i^*}^s) + k_{i^*}\beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}}] + \mu^\top x_{i^*}^s
\end{aligned} \tag{12}$$

Define $B = \hat{\gamma}_{i^*t}^\top(x^* - x_{i^*}^s) + \beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}}$, by assumption $x^* \notin D_t^{\text{safe}}$ we have $B \geq \tau - \tau_{i^*}^s$

By definition of α_t we have

$$\alpha_t B = \tau - \tau_{i^*}^s$$

In order to lower bound α_t we first upper bound B

$$\begin{aligned}
B &= \hat{\gamma}_{i^*t}^\top(x^* - x_{i^*}^s) + \beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}} \\
&= \gamma_{i^*}^\top(x^* - x_{i^*}^s) + (\hat{\gamma}_{i^*t} - \gamma_{i^*})^\top(x^* - x_{i^*}^s) + \beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}} \\
&\leq \gamma_{i^*}^\top(x^* - x_{i^*}^s) + 2\beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}} \quad \text{(Conditioned on } \mathcal{E}_\mu \text{ and Lemma E.1)} \\
&\leq \tau - \tau_{i^*}^s + 2\beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}}
\end{aligned} \tag{13}$$

Therefore we have

$$\alpha_t \geq \frac{\tau - \tau_{i^*}^s}{\tau - \tau_{i^*}^s + 2\beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}}}$$

If we choose k_i such that optimism is ensured for this lower bound, overall optimism is guaranteed.

$$\begin{aligned}
\max_{i\in M, x\in\hat{\Gamma}_{i,t}} \textbf{ucb}(x,i,t) &\geq \alpha_t[\mu^\top(x^* - x_{i^*}^s) + k_{i^*}\beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}}] + \mu^\top x_{i^*}^s \\
&\geq \frac{\tau - \tau_{i^*}^s}{\tau - \tau_{i^*}^s + 2\beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}}}[\mu^\top(x^* - x_{i^*}^s) + k_{i^*}\beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}}] + \mu^\top x_{i^*}^s
\end{aligned} \tag{14}$$

To show

$$\begin{aligned}
\frac{\tau - \tau_{i^*}^s}{\tau - \tau_{i^*}^s + 2\beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}}}[\mu^\top(x^* - x_{i^*}^s) + k_{i^*}\beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}}] + \mu^\top x_{i^*}^s &\geq \mu^\top(x^* - x_{i^*}^s) + \mu^\top x_{i^*}^s \\
(\tau - \tau_{i^*}^s)k_{i^*}\beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}} &\geq 2\mu^\top(x^* - x_{i^*}^s)\beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}} \\
k_{i^*} &\geq \frac{2LS}{\tau - \tau_{i^*}^s}
\end{aligned} \tag{15}$$

□

A.3. Proof of Theorem 3.4

Theorem A.3 (Regret Bound)**.** *The regret for the above algorithm is*

$$R_T \leq 2\beta_T\sqrt{2Td\log(\frac{d\lambda+TL^2}{d\lambda})} + (k_{i_{max}}\beta_T^{i_{max}}+2)\sqrt{2|M|Td\log(\frac{d\lambda+TL^2}{d\lambda})}$$

Proof.

$$\begin{aligned}
R_T &= \sum_{t=1}^{T}\delta_t = \sum_{t=1}^{\top}(\mu^\top x^* - \mu^\top x_t) \\
&\leq \sum_{t=1}^{T}(\textbf{ucb}(x_t,i_t,t) - \mu^\top x_t)\wedge 2 \qquad (16)\\
&= \sum_{t=1}^{T}(<\hat{\mu}_t-\mu, x_t> + \beta_t||x_t||_{V_t^{-1}} + k_{i_t}\beta_t^{i_t}||x-x_{i_t}^s||_{A_{i_t,t}^{-1}})\wedge 2 \\
&\leq \sum_{t=1}^{T}(2\beta_t||x_t||_{V_t^{-1}} + k_{i_{max}}\beta_t^{i_t}||x-x_{i_t}^s||_{A_{i_t,t}^{-1}})\wedge 2 \qquad (17)\\
&\leq \sum_{t=1}^{T}(2\beta_t||x_t||_{V_t^{-1}})\wedge 2 + \sum_{t=1}^{T}(k_{i_{max}}\beta_t^{i_t}||x-x_{i_t}^s||_{A_{i_t,t}^{-1}})\wedge 2
\end{aligned}$$

Here (16) follows from optimism and (17) follows from Lemma E.1 conditioned on $\mathcal{E}_\mu$.

Next we analyse these self normalised summations using standard technique in (Abbasi-Yadkori et al., 2011)

$$\begin{aligned}
\sum_{t=1}^{\top}||x_t||_{V_t^{-1}} &\leq \sqrt{T\sum_{t=1}^{\top}||x_t||^2_{V_t^{-1}}} \quad \text{(Cauchy Schwartz)}\\
&\leq \sqrt{2T\log(\frac{\det(A_T)}{\det(A_1)})} \qquad (18)\\
&\leq \sqrt{2Td\log(\frac{d\lambda+TL^2}{d\lambda})} \qquad (19)
\end{aligned}$$

In inequality (18), we used the standard argument in regret analysis of linear bandits (Abbasi-Yadkori et al., 2011) (Lemma 11) as follows:

$$\sum_{t=1}^{n}\min\left(\|\mathbf{y}_t\|^2_{\mathbf{V}_t^{-1}},1\right) \leq 2\log\frac{\det\mathbf{V}_{n+1}}{\det\mathbf{V}_1} \quad \text{where} \quad \mathbf{V}_n = \mathbf{V}_1 + \sum_{t=1}^{n-1}\mathbf{y}_t\mathbf{y}_t^\top.$$

In inequality (19), we used Assumption 2.2 and the fact that $\det(\mathbf{A}) = \prod_{i=1}^{d}\lambda_i(\mathbf{A}) \leq (\text{trace}(\mathbf{A})/d)^d$. Combining all these, we have with probability at least $1-2\delta$

$$\begin{aligned}
R_T &\leq 2\beta_T\sqrt{2Td\log(\frac{d\lambda+TL^2}{d\lambda})} + (k_{i_{max}}\beta_T^{i_{max}}+2)\Sigma_{i\in M}\sqrt{2N_i(T)d\log(\frac{d\lambda+N_i(T)L^2}{d\lambda})}\\
&\leq 2\beta_T\sqrt{2Td\log(\frac{d\lambda+TL^2}{d\lambda})} + (k_{i_{max}}\beta_T^{i_{max}}+2)\sqrt{2|M|Td\log(\frac{d\lambda+TL^2}{d\lambda})}
\end{aligned}$$

Last step from AM-QM inequality. □

B. Extension of Safe-LinUCB to Many Affine Constraints Setting

Additional assumptions

Assumption B.1 (Safe ball of actions)**.** There exists a safe ball of actions D_b^w of radius $r > 0$, known apriori to the agent such that

$$D_b^w = \{x \in R^d : ||x||_2 \leq r\} \subseteq D^w$$

Assumption B.1 is a standard assumption where the agent knows the minimum radius of the known safe set D_b^w. Having a known surface of safe actions of fixed norm makes the analysis much cleaner. In order to efficiently restrict learning all the constraint sets, we need some assumptions on the structure of the environment.

Assumption B.2 (Similar direction of optimal action and reward parameter)**.** We assume that the optimal action is α-close in angle to the reward parameter μ i.e

$$\frac{\mu^\top x^*}{||\mu||_2||x^*||_2} \geq \cos\alpha$$

Assumption B.2 states that the direction of the optimal arm x^* is similar to the unknown reward parameter μ. Since the optimal action maximises the projection onto μ, one can expect that in many scenarios that we have x^* aligned in similar direction as μ.

Since we have many constraints, a good way to model them is by using a constraint density function. If we consider a vector x and a θ-cone around it defined as

$$C(x, \theta) = \{v \in D_0 : \frac{v^\top x^*}{||v||_2||x^*||_2} \geq \cos\theta\}$$

Then the set of constraints intersected by this cone is given by

$$B(x, \theta) = \{i \in \mathbf{M} : \Gamma_i \cap C(x, \theta) \neq \emptyset\}$$

Since these constraint sets mostly occur on the boundaries of the constraint, the surface area on the boundary that the cone captures should be proportional to the number of constraint sets it intersects. It is a well known fact that the surface area of a d-dimensional sphere of radius R captured by the cone $C(x, \theta)$ is $O((R\sin\theta)^{d-1})$. Therefore we model our constraint density as follows

Assumption B.3 (Constraint density)**.** For any vector $x \in R^d$ and for any $\theta < \frac{\pi}{2}$ we have the following bound on the number of constraints intersected by the cone $C(x, \theta)$

$$\rho_l(\sin\theta)^{d-1} \leq |B(x, \theta)| \leq \rho_u(\sin\theta)^{d-1} + 1,$$

for $\rho_l, \rho_u > 0$.

Algorithm 4

for $t = 1, 2, \ldots, T'$ **do**
  Randomly choose $x_t \in \mathcal{D}_b^w$ and observe loss $\ell_t = c_t(x_t)$.
Construct important constraints set B using (21)
for $t = T' + 1, T' + 2, \ldots, T$ **do**
  Compute $\hat{\mu}_t$ as (4)
  Compute $\hat{\gamma_{i,t}}$ as (6)
  Compute β_t and β_t^i $\forall i \in \mathbf{M}$ using (5) and (7) respectively
  Construct D_t^{safe} according to (23)
  Pick $x_t = \arg\max_{x \in \mathcal{D}_t^{\text{safe}}} \mathbf{ucb}(x, i, t-1)$ and observe reward r_t. (where $\mathbf{ucb}(x, i, t)$ is defined in (24))

B.1. Algorithm Description

B.1.1. Pure Exploration Phase

In the initial pure exploration phase the agent randomly samples safe actions from D^w for T' steps. Since D^w is known apriori to the agent, actions sampled from this set satisfy the safety condition. The agent uses this feedback data to have a good estimate of the unknown reward parameter μ. This additional information helps the agent in restricting itself from learning safety constraints that are far away from the optimal action x^*. Therefore the agent exploits the fact that it doesn't need to learn all the constraints in order to minimize its regret.

To be more precise, for $t \in [0, T']$ we sample i.i.d $x_t \sim \tilde{D}_b^w$, where $\tilde{D}_b^w$ is the $d-1$ dimensional boundary surface of the ball D_b^w defined as

$$\tilde{D}_b^w = \{x \in R^d : ||x||_2 = r\} \subseteq D_b^w$$

This ensures uniform persistent excitation in all directions and the covariance matrix of the actions $\Sigma := E[x_t x_t^\top] = \frac{r^2}{d} I$. This implies that the minimum eigen value of covariance matix $\lambda_- := \lambda_{min}(\Sigma) = \frac{r^2}{d}$.

B.1.2. Exploration Exploitation Phase

The Exploration Exploitation Phase is very similar to the previous algorithm. The main difference is that we have a reduced set of constraints which the agent might need to explore. Using the feedback history from the pure exploration phase, the agent constructs a reduced set of constraints $B \subseteq \mathbf{M}$.

So the agent now estimates the reward parameter using linear regression using the previous data of actions and noisy feedback till $t = T'$ using the following update rule

$$\hat{\mu}_t = V_t^{-1} \sum_{\tau=1}^{t-1} r_\tau x_\tau \quad where \quad V_t = \lambda I + \sum_{\tau=1}^{t-1} x_\tau x_\tau^\top \tag{20}$$

The next step is to build a high confidence interval around $\hat{\mu}_{T'}$ which contains true reward parameter μ with high probability.

$$\mathcal{C}_{T'} = \left\{ v \in \mathbb{R}^d : \|v - \hat{\mu}_t\|_{A_{T'}} \leq \beta_{T'} \right\}$$

we define the event of μ belonging to this high confidence region as

$$\mathcal{E}_\mu = \{\mu \in \mathcal{C}_t\}$$

We choose $\beta_{T'}$ according to Theorem E.2 from (Abbasi-Yadkori et al., 2011), which ensure that $\mathcal{E}_\mu$ holds with high probability.

The previous pure exploration phase establishes the following confidence region on μ

Lemma B.4. *Conditioned on $\mathcal{E}_\mu$, if agent has done pure exploration for T' steps then for any $\delta \in [0, 1]$ it holds with probability $1 - \delta$ that*

$$||\mu - \hat{\mu}_{T'}||_2 \leq \frac{\beta_{T'}}{\sqrt{\lambda + \frac{r^2 T'}{2d}}}$$

provided

$$T' \geq t_\delta = \frac{8L^2}{\lambda_-} \log \frac{d}{\delta} = \frac{8dL^2}{r^2} \log \frac{d}{\delta}$$

Since the explorations of actions is uniform in all directions, the high confidence region around $\hat{\mu}_{T'}$ is a l_2-ball, See Appendix B.3 for more details. This geometry makes it easy to analyse the new restricted set of constraints to which contain x^*.

Conditioned on $\mathcal{E}_\mu$, this establishes a high confidence cone $C(\hat{\mu}_{T'}, \theta)$ around $\hat{\mu}_{T'}$ which contains μ

Lemma B.5. *Conditioned on $\mathcal{E}_\mu$, if $\mu \in C(\hat{\mu}_{T'}, \theta)$ then*

$$\sin\theta \geq \frac{R}{s - R} = \sin\theta(T')$$

where s is the lower bound on μ and r is defined as:

$$R = \frac{\beta_{T'}}{\sqrt{\lambda + \frac{r^2 T'}{2d}}}$$

We report proof in Appendix B.4. The above lemmas in conjunction with assumption B.2 imply that conditioned on $\mathcal{E}_\mu$, the optimal action $x^* \in C(\hat{\mu}_{T'}, \alpha + \theta(T'))$. Therefore we construct the restricted constraint set as

$$B := B(\hat{\mu}_{T'}, \alpha + \theta(T')) = \{i \in M : \Gamma_i \cap C(\hat{\mu}_{T'}, \alpha + \theta(T')) \neq \phi\} \tag{21}$$

Using assumption B.3 we have an upper bound on the number of important constraints as

$$|B| \leq \rho_u \sin\left(\alpha + \theta(T')\right)^{d-1} + 1$$

Therefore this exponentially decreases the number of constraints under consideration with d and can make the algorithm quite efficient in for high dimensional problems.

Next we build confidence regions for each γ_i as

$$\mathcal{C}_t^i = \left\{v \in \mathbb{R}^d : \|v - \hat{\gamma}_{it}\|_{A_t^i} \leq \beta_t^i\right\}$$

we define the event of γ_i belonging to this high confidence region as

$$\mathcal{E}_{\gamma_i} = \{\gamma_i \in \mathcal{C}_t^i\}$$

Again using Theorem E.2 we choose

$$\beta_t^i = R\sqrt{d \log\left(|B| \frac{1 + N_i(t)L^2/\lambda}{\delta}\right)} + \lambda^{1/2} S_\gamma \tag{22}$$

This ensures that the event $\mathcal{E}_{\gamma_i}$ happens with probability $1 - \frac{\delta}{|B|}$. Next conditioned in the joint event $\mathcal{E} := \mathcal{E}_\mu \cup \bigcup_{i \in M} \mathcal{E}_{\gamma_i}$ we want the agent to satisfy safety constraints when picking actions. To achieve this the algorithm conservatively constructs safe set of actions as follows

$$D_t^{\text{safe}} = \bigcup_{i \in B} \{x \in \Gamma_i : \hat{\gamma}_{i,t}^\top (x - x_i^s) + \beta_t^i ||x - x_i^s||_{A_t^{i-1}} \leq \tau_i - \tau_i^s\} \tag{23}$$

Here $\hat{\gamma}_{i,t}^\top (x - x_i^s) + \beta_t^i ||x - x_i^s||_{A_t^{i-1}}$ is a high probability upper bound on $\gamma_i^\top (x - x_i^s)\ \forall i \in B$. This ensures that by picking any action from D_t^{safe}, we satisfy the safety constraints as shown by Lemma 3.1.

The next step of the algorithm is to pick action x_t from the constructed safe set D_t^{safe} which maximises the Upper Confidence Bound (UCB) defined as:

$$\textbf{ucb}(x, i, t) = \hat{\mu}_t^\top x + \beta_t ||x||_{V_t^{-1}} + k_{i_t} \beta_t^{i_t} ||x - x_{i_t}^s||_{A_t^{i_t - 1}} = \max_{i \in B} \hat{\mu}_t^\top x + \beta_t ||x||_{V_t^{-1}} + \mathbb{1}\{x \in \Gamma_i\} k_i \beta_t^i ||x - x_i^s||_{A_{i,t}^{-1}} \tag{24}$$

where $k_i \geq \frac{2LS}{\tau_i - \tau_i^s}$.

The second phase of the algorithm follows OFU policy similar to the previous case, with the replacement of M with $B \subseteq M$, as now we only consider these reduced set of constraints.

Theorem B.6 (Regret Bound)**.** *If Assumptions 2.1-2.3 and Assumptions B.1-B.3 hold and then for any $\delta \in [0, \frac{1}{3}]$ if we define β_T and β_T^i according to (5) and (7) and $k_i = \frac{2}{\tau_i - \tau_i^s}$,with probability $1 - 3\delta$, the regret for the above algorithm is*

$$R_T \leq 2T' + 2\beta_T \sqrt{2Td \log(\frac{d\lambda + TL^2}{d\lambda})} + (\rho_u(\sin\left(\alpha + \theta(T')\right))^{d-1} + 1)(k_{i_{max}} \beta_T^{i_{max}} + 2)\sqrt{2Td \log(\frac{d\lambda + TL^2}{d\lambda})}$$

where $\beta_T^{i_{max}} = \max_{j \in B} \beta_T^j$ and $k_{i_{max}} = \max_{j \in B} k_j$

The regret analysis is quite similar to Theorem 3.4 with the addition of pure exploration time and is shown in Appendix B.5

B.2. Many Constraints Proofs

B.3. Proof of Lemma B.4

Lemma B.7. *Conditioned on $\mathcal{E}_\mu$, if agent has done pure exploration for T' steps then for any $\delta \in [0,1]$ it holds with probability $1-\delta$ that*

$$||\mu - \hat{\mu}_{T'}||_2 \leq \frac{\beta_{T'}}{\sqrt{\lambda + \frac{r^2 T'}{2d}}}$$

provided

$$T' \geq t_\delta = \frac{8L^2}{\lambda_-} \log \frac{d}{\delta} = \frac{8dL^2}{r^2} \log \frac{d}{\delta}$$

Proof. Recall, for $t \in [0, T']$ we sample i.i.d $x_t \sim \tilde{D}_b^w$, where $\tilde{D}_b^w$ is the $d-1$ dimensional boundary surface of the ball D_b^w defined as

$$\tilde{D}_b^w = \{x \in R^d : ||x||_2 = r\} \subseteq D_b^w$$

Clearly from this construction we get $\Sigma := E[x_t x_t^\top] = \frac{r^2}{d} I$. Also define $\lambda_- = \lambda_{min}(\Sigma) = \frac{r^2}{d}$

In order to bound the minimum eigenvalue of the Gram matrix at round $T'+1$, we use the Matrix Chernoff Inequality in Theorem E.3.

We use a similar analysis as (Amani et al., 2019). Let $X_t = x_t x_t^\top$ for $t \in [T']$, such that each X_t is a symmetric matrix with $\lambda_{\min}(X_t) \geq 0$ and $\lambda_{\max}(X_t) \leq L^2$. In this notation, $A_{T'} = \lambda I + \sum_{t=1}^{T'} X_t$. In order to apply the above Theorem, we compute:

$$\mu_{\min} := \lambda_{\min}\left(\sum_{t=1}^{T'} \mathbb{E}[X_t]\right) = \lambda_{\min}\left(\sum_{t=1}^{T'} \mathbb{E}\left[x_t x_t^\top\right]\right) = \lambda_{\min}(T'\Sigma) = \lambda_- T'.$$

Thus, the theorem implies the following for any $\epsilon \in [0,1)$:

$$\Pr\left[\lambda_{\min}\left(\sum_{t=1}^{T'} X_t\right) \leq \epsilon \lambda_- T'\right] \leq d \cdot \exp\left(-(1-\epsilon)^2 \frac{\lambda_- T'}{2L^2}\right).$$

To complete the proof of the lemma, simply choose $\epsilon = 0.5$ (say) and $T' \geq \frac{8L^2}{\lambda_-} \log\left(\frac{d}{\delta}\right)$. This gives $\Pr\left[\lambda_{\min}(A_{T'+1}) \geq \lambda + \frac{\lambda_- T'}{2}\right] \geq 1-\delta$. This implies with probability $1-\delta$

$$\lambda_{\min}(A_{T'+1}) \geq \lambda + \frac{\lambda_- T'}{2} = \lambda + \frac{r^2 T'}{2d}$$

For any $T' < t \leq T$, conditioned on $\mathcal{E}_\mu$ we have

$$\beta_{T'} \geq \|\mu - \hat{\mu}_{T'}\|_{A_{T'}^{-1}} \geq \|\mu - \hat{\mu}_{T'}\|_2 \sqrt{\lambda_{\min}(A_{T'})}$$

This gives us the desired bound

$$||\mu - \hat{\mu}_{T'}||_2 \leq \frac{\beta_{T'}}{\sqrt{\lambda_{\min}(A_{T'})}} \leq \frac{\beta_{T'}}{\sqrt{\lambda + \frac{r^2 T'}{2d}}}$$

□

B.4. Proof of Lemma B.5

Lemma B.8. *Conditioned on $\mathcal{E}_\mu$, if $\mu \in C(\hat{\mu}_{T'}, \theta)$ then*

$$\sin\theta \geq \frac{R}{s-R} = \sin\theta(T')$$

where s is the lower bound on μ and r is defined as:

$$R = \frac{\beta_{T'}}{\sqrt{\lambda + \frac{r^2 T'}{2d}}}$$

Proof. As a consequence of Lemma B.4, the high confidence region of μ is a ball of radius R defines as:

$$||\mu - \hat{\mu}_{T'}||_2 \leq \frac{\beta_{T'}}{\sqrt{\lambda + \frac{r^2 T'}{2d}}} = R$$

The smallest cone $C(\hat{\mu}_t, \alpha)$ along $\hat{\mu}_t$ that intersects this entire ball of radius R, has its surface tangential to this ball at the point they touch. This implies that:

$$\sin\alpha = \frac{R}{||\hat{\mu}_t||_2} \leq \frac{R}{||\mu||_2 - R}$$

Since by Assumption 2.2 $\mu \geq s$ we can further upper bound $\sin\alpha$ as

$$\sin\alpha \leq \frac{R}{||\mu||_2 - R} \leq \frac{R}{s-R} = \sin\theta(T')$$

Since $C(\hat{\mu}_t, \alpha) \subseteq C(\hat{\mu}_t, \theta T')$, $\mu \in C(\hat{\mu}_t, \theta T')$ □

B.5. Proof of Theorem B.6

Theorem B.9 (Regret Bound)**.** *The regret for the above algorithm is*

$$R_T \leq 2T' + 2\beta_T\sqrt{2Td\log(\frac{d\lambda + TL^2}{d\lambda})} + |B|(k_{i_{max}}\beta_T^{i_{max}} + 2)\sqrt{2Td\log(\frac{d\lambda + TL^2}{d\lambda})}$$

Proof. The proof analysis is very similar to Theorem 3.4, so we briefly go over it for completeness.

$$\begin{aligned}
R_T &= \sum_{t=1}^{\top}\delta_t = \sum_{t=1}^{\top}(\mu^\top x^* - \mu^\top x_t) \\
&\leq 2T' + \sum_{t=T'+1}^{\top}(\textbf{ucb}(x_t, i_t, t) - \mu^\top x_t) \wedge 2 \quad \text{(optimism)} \\
&= 2T' + \sum_{t=T'+1}^{\top}(< \hat{\mu}_t - \mu, x_t > + \beta_t||x_t||_{V_t^{-1}} + \sum_{i\in B}\mathbb{1}\{x \in \Gamma_i\}k_i\beta_t^i||x - x_i^s||_{A_{i,t}^{-1}}) \wedge 2 \\
&\leq 2T' + \sum_{t=T'+1}^{\top}(2\beta_t||x_t||_{V_t^{-1}} + k_{i_{max}}\sum_{i\in B}\mathbb{1}\{x \in \Gamma_i\}\beta_t^i||x - x_i^s||_{A_{i,t}^{-1}}) \wedge 2 \\
&\leq 2T' + \sum_{t=T'+1}^{\top}(2\beta_t||x_t||_{V_t^{-1}}) \wedge 2 + \sum_{t=T'+1}^{\top}(k_{i_{max}}\sum_{i\in B}\mathbb{1}\{x \in \Gamma_i\}\beta_t^i||x - x_i^s||_{A_{i,t}^{-1}}) \wedge 2 \\
&\leq 2T' + 2\beta_T\sqrt{2Td\log(\frac{d\lambda + TL^2}{d\lambda})} + |B|(k_{i_{max}}\beta_T^{i_{max}} + 2)\sqrt{2Td\log(\frac{d\lambda + TL^2}{d\lambda})}
\end{aligned} \tag{25}$$

The last step follows from standard analysis of stochastic linear bandits (Abbasi-Yadkori et al., 2011) and 3.4 □

C. Proofs of Section 4 - Safe-LinTS

Theorem C.1. *Suppose Assumptions 2.1-2.3 and 3.2 hold. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, the regret of Safe-LinTS is $R_T = \tilde{\mathcal{O}}(d^{3/2}\sqrt{|M|T})$.*

C.1. Optimism

Proof. Recall

$$\tilde{\mu}_t = \hat{\mu}_t + \beta_t (V_t)^{-\frac{1}{2}} \eta_t$$

$$\tilde{\omega}_t^i = \beta_t^i (A_{i,t})^{-\frac{1}{2}} \eta_t^c$$

Anti-Concentration

$$\mathbb{P}(u^\top \eta_t \geq 1) = p_1$$

$$\mathbb{P}(u^\top \eta_t^c \geq \frac{2}{\tau - \tau_*^s} L S^\gamma) = p_2$$

Concentration

$$\mathbb{P}(\|\eta_t\|_2 \leq \sqrt{cd\log(\frac{c'd}{\delta})}) \geq 1 - \frac{\delta}{2}$$

$$\mathbb{P}(\|\eta_t^c\|_2 \leq \frac{2LS^\gamma}{\tau - \tau_*^s}\sqrt{cd\log(\frac{c'd}{\delta})}) \geq 1 - \frac{\delta}{2}$$

We prove sub-linear regret by first showing that the algorithm is optimistic with constant probability. To be more precise we define the following high probability events:

Let $\delta \in (0,1)$, $\delta' = \frac{\delta}{6T}$ then

- $\mathcal{E}_{\mu,t}$ is the event that the RLS estimate $\hat{\mu}_t$ concentrates around μ for all $s \leq t$ defined as

  $$\mathcal{E}_{\mu,t} = \{\forall s \leq t : \|\hat{\mu}_s - \mu\|_{V_s} \leq \beta_s(\delta')\}$$

  then $\mathbb{P}(\mathcal{E}_{\mu,T}) \geq 1 - \frac{\delta}{6}$

- $\mathcal{E}_{\gamma,t}$ is the event that the RLS estimate $\hat{\gamma}_{i,t}$ concentrates around γ_i for all $s \leq t$ and for all $i \in \mathbf{M}$ defined as

  $$\mathcal{E}_{\gamma,t} = \{\forall s \leq t, \forall i \in M : \|\hat{\gamma}_{i,s} - \gamma_i\|_{A_{i,s}} \leq \beta_s^i(\frac{\delta'}{|M|})\}$$

  then $\mathbb{P}(\mathcal{E}_{\gamma,T}) \geq 1 - \frac{\delta}{6}$

- $\tilde{\mathcal{E}}$ be the event that such the sampled η_t and η_t^c are bounded for all $t \leq T$

  $$\tilde{\mathcal{E}} = \{\forall t \leq T, \|\eta_t\|_2 \leq \sqrt{cd\log(\frac{12Tc'd}{\delta})}\} \cap \{\forall t \leq T, \|\eta_t^c\|_2 \leq \sqrt{cd\log(\frac{12Tc'd}{\delta})}\}$$

  then $\mathbb{P}(\tilde{\mathcal{E}}) \geq 1 - \frac{\delta}{6}$

- Let

  $$Z_t = \mathcal{E}_{\gamma,t} \cap \mathcal{E}_{\mu,t}$$

  then $\mathbb{P}(Z_T) \geq 1 - \frac{\delta}{3}$

- Let
$$E_t = \tilde{\mathcal{E}} \cap \mathcal{E}_{\gamma,t} \cap \mathcal{E}_{\mu,t}$$
then $\mathbb{P}(E_T) \geq 1 - \frac{\delta}{2}$

Recall $\hat{\Gamma}_{i,t}$ is defined as:
$$\hat{\Gamma}_{i,t} = \{x \in \Gamma_i : \nabla \hat{\gamma}_{i,t}^\top (x - x_i^s) + \beta_t^i ||x - x_i^s||_{A_{i,t}^{-1}} \leq \tau - \tau_i^s\}$$
Let
$$\alpha_t^* := \max\{\alpha \in [0,1] : z_t = \alpha x^* + (1-\alpha) x_{i^*}^s \in \hat{\Gamma}_{i^*,t}\}$$
then we can show that there exists $\alpha_t \leq \alpha_t^*$ such that
$$\alpha_t \gamma_{i^*}^\top (x^* - x_i^s) + 2\alpha_t \beta_t^i ||x - x_i^s||_{A_t^{i-1}} = \tau - \tau_i^s$$
rearranging we get
$$\frac{1}{\alpha_t} = 1 + \frac{2}{\tau - \tau_{i^*}^s} \beta_t^{i^*} \|x^* - x_{i^*}^s\|_{A_t^{i^*-1}}$$
The goal is to show that playing the safe action $z_t = \alpha_t x^* + (1-\alpha_t) x_{i^*}^s$ is optimistic with constant probability. Now we define
$$J_t(\eta, \eta^c, i, x) = \tilde{\mu}_t^\top x + \tilde{\omega}_{i,t}^\top (x - x_i^s)$$
$$J_t(\eta, \eta^c, i) = \max_{x \in \hat{\Gamma}_{i,t}} J_t(\eta, \eta^c, i, x)$$
$$J_t(\eta, \eta^c) = \max_{i \in M} J_t(\eta, \eta^c, i)$$
and we analyse the probability with which the sampled parameters are optimistic i.e.
$$J_t(\eta_t, \eta_t^c) \geq \mu^\top x^*$$
Let
$$p_t = \mathbb{P}(J_t(\eta_t, \eta_t^c) \geq \mu^\top x^* | \mathcal{F}_t, Z_t)$$
then
$$\begin{aligned} p_t &= \mathbb{P}(J_t(\eta_t, \eta_t^c) \geq \mu^\top x^* | \mathcal{F}_t, Z_t) \\ &\geq \mathbb{P}(J_t(\eta_t, \eta_t^c, i^*, \alpha_t x^* + (1-\alpha_t) x_{i^*}^s) \geq \mu^\top x^* | \mathcal{F}_t, Z_t) \\ &= \mathbb{P}(\tilde{\mu}_t^\top z_t + \alpha_t \tilde{\omega}_{i^*,t}^\top (x^* - x_{i^*}^s) \geq \mu^\top x_{i^*}^s + \mu^\top (x^* - x_{i^*}^s) | \mathcal{F}_t, Z_t) \end{aligned}$$
where $z_t = \alpha x^* + (1-\alpha) x_{i^*}^s$. Consider:
$$\begin{aligned} \tilde{\mu}_t^\top z_t &= \hat{\mu}_t^\top z_t + z_t^\top \beta_t (V_t)^{-\frac{1}{2}} \eta_t \\ &\geq \mu^\top z_t - \beta_t \|z_t\|_{V_t^{-1}} + z_t^\top \beta_t (V_t)^{-\frac{1}{2}} \eta_t \end{aligned}$$
By construction of η_t we have:
$$\mathbb{P}(z_t^\top \beta_t (V_t)^{-\frac{1}{2}} \eta_t \geq \beta_t \|z_t\|_{V_t^{-1}} | \mathcal{F}_t, Z_t) = \mathbb{P}(u^\top \eta_t \geq 1) = p_1$$

Using the fact that η_t, η_t^c are independent and then substituting in α_t we have :

$$\begin{aligned}
p_t &\geq p_1 \mathbb{P}(\mu^\top z_t + \alpha_t \tilde{\omega}_{i^*,t}^\top (x^* - x_{i^*}^s) \geq \mu^\top x_{i^*}^s + \mu^\top (x^* - x_{i^*}^s) | \mathcal{F}_t, Z_t) \\
&\geq p_1 \mathbb{P}(\mu^\top x_{i^*}^s + \alpha_t \mu^\top (x^* - x_{i^*}^s) + \alpha_t \tilde{\omega}_{i^*,t}^\top (x^* - x_{i^*}^s) \geq \mu^\top x_{i^*}^s + \mu^\top (x^* - x_{i^*}^s) | \mathcal{F}_t, Z_t) \\
&= p_1 \mathbb{P}(\tilde{\omega}_{i^*,t}^\top (x^* - x_{i^*}^s) \geq \frac{1-\alpha_t}{\alpha_t} \mu^\top (x^* - x_{i^*}^s) | \mathcal{F}_t, Z_t) \\
&\geq p_1 \mathbb{P}(\tilde{\omega}_{i^*,t}^\top (x^* - x_{i^*}^s) \geq \frac{2LS}{\tau - \tau_{i^*}} \beta_t^{i^*} \|x^* - x_{i^*}^s\|_{A_{i^*,t}^{-1}} | \mathcal{F}_t, Z_t) \\
&= p_1 \mathbb{P}(\beta_t^{i^*} (A_t^{i^*})^{-\frac{1}{2}} \eta_t^{c\top} (x^* - x_{i^*}^s) \geq \frac{2LS}{\tau - \tau_{i^*}} \beta_t^{i^*} \|x^* - x_{i^*}^s\|_{A_{i^*,t}^{-1}} | \mathcal{F}_t, Z_t) = p_1 \mathbb{P}(u^\top \eta_t^c \geq \frac{2}{\tau - \tau_i} LS) \geq p_1 p_2
\end{aligned}$$

Next we need to show that Conditioned on E_T , algorithm is still optimistic. This is because the chosen confidence bound $\delta' = \frac{\delta}{6T}$ is small enough compared to the anti-concentration property. Moreover, we assume that $T \geq \frac{1}{3p_1p_2}$ which implies that $\delta' \leq \frac{p_1p_2}{2}$. We know that for any events A ans B, we have

$$\mathbb{P}(A \cap B) = 1 - \mathbb{P}\left(A^c \cup B^c\right) \geq \mathbb{P}(A) - \mathbb{P}\left(B^c\right)$$

choosing $A = J_t(\eta_t, \eta_t^c) \geq \mu^\top x^*$ and $B = E_T$ we get

$$\mathbb{P}(J_t(\eta_t, \eta_t^c) \geq \mu^\top x^* | \mathcal{F}_t, Z_t) \geq p_1 p_2 - \delta' \geq \frac{p_1 p_2}{2}$$

C.2. Regret

$$R(T) = \sum_{t=1}^{T} \left(\underbrace{x^{*\top} \mu - J_t(\eta_t, \eta_t^c)}_{\text{Term I}} \right) + \sum_{t=1}^{T} \left(\underbrace{J_t(\eta_t, \eta_t^c) - x_t^\top \mu}_{\text{Term II}} \right). \tag{16}$$

Term 1 $R^{TS}(T)$:

$$\begin{aligned}
R_t^{TS} &= x^{*T} \mu - J_t(\eta_t, \eta_t^c) \\
&\leq \mathbb{E}[J_t(\eta, \eta^c) - J_t(\eta_t, \eta_t^c) | (\eta, \eta^c) \in \Theta] \\
&\leq \mathbb{E}[J_t(\eta, \eta^c, i, x) - J_t(\eta_t, \eta_t^c, i, x) | (\eta, \eta^c, i, x) \in \Theta] \\
&\leq \mathbb{E}[(\tilde{\mu} - \tilde{\mu}_t)^\top x + (\tilde{\omega}^i - \tilde{\omega}_t^i)^\top (x - x_i^s) | (\eta, \eta^c, i, x) \in \Theta] \\
&\leq \mathbb{E}[\|\tilde{\mu} - \tilde{\mu}_t\|_{A_t} \|x\|_{V_t^{-1}} + \|\tilde{\omega}^i - \tilde{\omega}_t^i\|_{A_t^i} \|x - x_i^s\|_{A_{i,t}^{-1}} | (\eta, \eta^c, i, x) \in \Theta] \\
&\leq 2\sigma_t(\delta) \mathbb{E}[\|x\|_{V_t^{-1}} | (\eta, \eta^c, i, x) \in \Theta] + 2\sigma_t^i(\delta) \mathbb{E}[\|x - x_i^s\|_{A_{i,t}^{-1}} | (\eta, \eta^c, i, x) \in \Theta] \\
&\leq \frac{4}{p_1 p_2} \{\sigma_t(\delta) \mathbb{E}[\|x\|_{V_t^{-1}}] + \sigma_t^i(\delta) \mathbb{E}[\|x - x_i^s\|_{A_{i,t}^{-1}}]\}
\end{aligned}$$

where

$$\sigma_t(\delta) = \beta_t(\delta) \sqrt{cd \log(\frac{c'd}{\delta})}$$

$$\sigma_t^i(\delta) = \beta_t^i(\delta) \frac{2L^c S^c}{\tau - \tau_i} \sqrt{cd \log(\frac{c'd}{\delta})}$$

and $(\eta, \eta^c, i, x) \in \Theta$ denotes optimistic parameters.

Next consider the sum

$$R(T) = \Sigma_{t=1}^T \mathbb{E}[\|x\|_{V_t^{-1}}] = \sum_{t=1}^{T} \|x\|_{V_t^{-1}} + \sum_{t=1}^{T} (\mathbb{E}[\|x\|_{V_t^{-1}}] - \|x\|_{V_t^{-1}})$$

The second summation is a martingale sum, so we use Azuma's Inequality to get the following bound with probability $1-\frac{\delta}{2}$

$$\sum_{t=1}^{T}(\mathbb{E}[\|x\|_{V_t^{-1}}] - \|x\|_{V_t^{-1}}) \leq \sqrt{\frac{8Tl^2}{\lambda}\log\frac{4}{\delta}}$$

Since $\|x_t\|_2 \leq L$ and $V_t^{-1} \leq \frac{1}{\lambda}I$ so

$$\mathbb{E}[\|x\|_{A_s^{-1}}] - \|x\|_{A_s^{-1}} \leq \frac{2L}{\sqrt{\lambda}}$$

Now using standard analysis from previous sections and previous inequality we get

$$R^{TS}(T) \leq \frac{4\sigma_t(\delta)}{p_1p_2}(\sqrt{2dT\log(1+\frac{TL^2}{\lambda})}+\sqrt{\frac{8TL^2}{\lambda}\log\frac{8}{\delta}})+\frac{4\sigma_t^{i_{max}}(\delta)+2}{p_1p_2}(\sqrt{2d|M|T\log(1+\frac{TL^2}{\lambda})}+\sqrt{\frac{8TL^2}{\lambda}\log\frac{8}{\delta}})$$

Term 2 $R^{RLS}(T)$:

$$\begin{aligned}
R_t^{RLS} &= J_t(\eta_t,\eta_t^c,i_t,x_t) - \mu^\top x_t \\
&= \tilde{\mu}_t^\top x + \tilde{\omega}_t^{iT}(x-x_i^s) - \mu^\top x_t \\
&\leq (\hat{\mu}_t-\mu)^\top x_t + \beta_t A_t^{-\frac{1}{2}}\eta_t^\top x_t + \beta_t^{i_t}A_t^{i_t-\frac{1}{2}}\eta_t^{cT}(x_t-x_{i_t}^s) \\
&\leq \beta_t\|x_t\|_{V_t^{-1}} + \sigma_t\|x_t\|_{V_t^{-1}} + \sigma_t^{i_t}\|x_t-x_{i_t}^s\|_{A_t^{i_t-1}}
\end{aligned}$$

So

$$R^{RLS}(T) \leq (\beta_T+\sigma_T)\sqrt{2dT\log(1+\frac{TL^2}{\lambda})} + \sigma_T^{i_{max}}\sqrt{2d|M|T\log(1+\frac{TL^2}{\lambda})}$$

$$R \leq \tilde{\mathcal{O}}(d^{3/2}\sqrt{|M|T}) \tag{26}$$

We provide the regret proof for completeness, detailed analysis of the proof technique is done in (Moradipari et al., 2021).

□

D. Proofs of Section 5 Nonlinear Constraints

Theorem D.1. *Suppose Assumptions 5.1 & 5.2 hold. For any $\delta \in (0,1)$, after T' time steps of pure exploration per constraint set, we have i) $x^* \in D_{MT'}^{safe}$ and ii) $D_{MT'}^{safe} \subseteq D_0^{safe}$ with probability at least $1-\delta$, if $\frac{T'}{\log^2 T'} \geq \left(2d\frac{4\delta_f^2}{(\Delta-\zeta\delta_f^2)^2}\right)^2$.*

First, we prove the following helper lemma

Lemma D.2. *If we consider a δ_f-ball around a point, the approximation error for the first order Taylor expansion for a ζ-smooth function is bounded as*

$$|f(x) - f(a) - \nabla f(a)^\top(x-a)| \leq \frac{\zeta\delta^2}{2}$$

Then least squares parameter of this approximation error is bounded as

$$\|\hat{\epsilon}_{i,T}\|_2 \leq 2\zeta\delta_r\sqrt{2d\log T} = O(\zeta\delta_r\sqrt{d\log T})$$

Proof. Recall

$$\hat{\epsilon}_{it} = A_{i,t}^{-1}\sum_{\tau=1}^{N_i(t)}(x_\tau - x_i^s)\epsilon_i(x_\tau)$$

where

$$\epsilon_i(x_t) = f(x_t) - (f(x_i^s) + \nabla f(x_i^s)^\top (x_t - x_i^s)$$

Define $Y_{\epsilon,t}$ as the column vector enumerating the approximation errors $y_{\epsilon,\tau} = \epsilon_i(x_\tau)$ for $0 < \tau \leq t$, and X_t corresponds to the matrix enumerating the shifted actions $x_\tau - x_i^s$ for $0 < \tau \leq t$.

By definition we have

$$\hat{\epsilon}_{i,t} = arg\min_\theta \|Y_{\epsilon,t} - X_t^\top \theta\|_2^2$$

Next, define

$$T_1(\theta) := \|Y_{\epsilon,t} - X_t^\top \theta\|_2^2 = (Y_{\epsilon,t}^\top Y_{\epsilon,t} - 2Y_{\epsilon,t}^\top X_t^\top \theta + \theta^\top X_t X_t^\top \theta)$$

and

$$T_2(\theta) := \|y_{\epsilon,t+1} - x_{t+1}^\top \theta\|_2^2 = (y_{\epsilon,t+1}^2 - 2y_{\epsilon,t+1} x_{t+1}^\top \theta + \theta^\top x_{t+1} x_{t+1}^\top \theta)$$

Now consider $\hat{\epsilon}_{i,t+1}$

$$\begin{aligned}\hat{\epsilon}_{i,t+1} &= arg\min_\theta T_1(\theta) + T_2(\theta) \\ &= arg\min_\theta (Y_{\epsilon,t}^\top Y_{\epsilon,t} - 2Y_{\epsilon,t}^\top X_t^\top \theta + \theta^\top X_t X_t^\top \theta) + (y_{\epsilon,t+1}^2 - 2y_{\epsilon,t+1} x_{t+1}^\top \theta + \theta^\top x_{t+1} x_{t+1}^\top \theta)\end{aligned}$$

at the minimiser we have $\nabla T_1 + \nabla T_2 = 0$ and

$$\nabla T_1 = 2X_t X_t^\top \theta - 2X_t Y_{\epsilon,t}$$

$$\nabla T_2 = 2x_{t+1} x_{t+1}^\top \theta - 2x_{t+1} y_{\epsilon,t+1}$$

If we re-parameterise $\theta = \hat{\epsilon}_{i,t} + w$, then at minima we have

$$2X_t X_t^\top w = 2x_{t+1}(y_{\epsilon,t+1} - x_{t+1}^\top \theta)$$

$$(X_t X_t^\top + x_{t+1} x_{t+1}^\top) w = x_{t+1}(y_{\epsilon,t+1} - x_{t+1}^\top \hat{\epsilon}_{i,t})$$

$$w = (X_t X_t^\top + x_{t+1} x_{t+1}^\top)^{-1} x_{t+1}(y_{\epsilon,t+1} - x_{t+1}^\top \hat{\epsilon}_{i,t})$$

Recall that in pure exploration we picks action such that

$$x_t = \max_{x \in \tilde{D}_i^w} \|x - x_i^s\|_{A_{i,t}^{-1}}$$

as consequence we pick orthogonal vectors in subsequent turns, which ensures that x_t is always an eigen vector, which implies

$$w = \frac{x_{t+1}}{\lambda_{t+1}}(y_{\epsilon,t+1} - x_{t+1}^\top \hat{\epsilon}_{i,t})$$

Note that there is no rotation as the exploration strategy ensures that x_{t+1} are eigen vector of $X_t X_t^\top + x_{t+1} x_{t+1}^\top$ with eigen value $\lambda_{t+1} \geq \delta_r^2$.

Next we upper bound the magnitude difference at each step. By definition we have

$$\|\hat{\epsilon}_{i,t+1}\|^2 = \|\hat{\epsilon}_{i,t} + w\|^2$$

substituting w and rearranging we get

$$\|\hat{\epsilon}_{i,t+1}\|_2^2 - \|\hat{\epsilon}_{i,t}\|_2^2 = \frac{\|x_{t+1}\|^2}{\lambda_{t+1}^2}(y_{\epsilon,t+1} - x_{t+1}^\top \hat{\epsilon}_{i,t})^2 + \frac{2x_{t+1}^\top \hat{\epsilon}_{i,t}}{\lambda_{t+1}}(y_{\epsilon,t+1} - x_{t+1}^\top \theta_{t+1}) \tag{27}$$

Now re-parameterise as : $\|\hat{\epsilon}_{i,t}\| = B$, $\|x_{t+1}\| = \delta_r$, $\cos\alpha = \frac{<x_{t+1}, \hat{\epsilon}_{i,t}>}{B\delta}$ and $L = \|\hat{\epsilon}_{i,t+1}\|_2^2 - \|\hat{\epsilon}_{i,t}\|_2^2$, expanding L we get

$$L = \frac{\delta_r^2}{\lambda_{t+1}}(y_{\epsilon,t+1} - B\delta_r \cos(\alpha))^2 + \frac{2}{\lambda_{t+1}}(y_{\epsilon,t+1} - B\delta_r \cos(\alpha))B\delta_r \cos(\alpha)$$

rearranging gives us

$$L = (\frac{B^2\delta^4}{\lambda_{t+1}^2} - \frac{2B^2\delta_r^2}{\lambda_{t+1}})\cos^2(\alpha) + (\frac{2y_{\epsilon,t+1}B\delta_r}{\lambda_{t+1}} - \frac{2y_{\epsilon,t+1}B\delta_r^3}{\lambda_{t+1}^2})\cos(\alpha) + \frac{\delta_r^2 y_{\epsilon,t+1}^2}{\lambda_{t+1}^2}.$$

Now, to upper-bound L, we maximise L over α, to do this we set

$$\frac{\partial L}{\partial \alpha} = 0$$

which gives us the following condition

$$\frac{B^2\delta_r^2}{\lambda_{t+1}}(2 - \frac{\delta_r^2}{\lambda_{t+1}})\cos(\alpha)\sin(\alpha) - \frac{y_{\epsilon,t+1}B\delta_r}{\lambda_{t+1}}(1 - \frac{\delta_r^2}{\lambda_{t+1}})\sin(\alpha) = 0$$

$$\frac{B\delta_r}{\lambda_{t+1}}\sin(\alpha)[\frac{2\lambda_{t+1} - \delta_r^2}{\lambda_{t+1}}B\delta_r\cos(\alpha) - \frac{\lambda_{t+1} - \delta_r^2}{\lambda_{t+1}}y_{\epsilon,t+1}] = 0$$

Case 1 : $(\sin\alpha = 0)$

So $\cos(\alpha) = \pm 1$, which implies the increment w is along the direction of $\hat{\epsilon}_{i,t}$, now recall that

$$w = \frac{x_{t+1}}{\lambda_{t+1}}(y_{\epsilon,t+1} - B\delta_r\cos(\alpha))$$

because $\cos(\alpha) = \pm 1$, we get the following equality.

$$|\hat{\epsilon}_{i,t+1}| = |\hat{\epsilon}_{i,t}| + \frac{\delta_r\cos(\alpha)}{\lambda_{t+1}}(y_{\epsilon,t+1} - B\delta_r\cos(\alpha))$$

if $B \geq \frac{\zeta\delta_r}{2}$ then using our smoothness assumption, we have $y_{\epsilon,t+1} \leq \frac{\zeta\delta_r^2}{2}$, so as a consequence we always have:

$$|\hat{\epsilon}_{i,t+1}| - |\hat{\epsilon}_{i,t}| \leq 0$$

otherwise when $B < \frac{\zeta\delta_r}{2}$, we get the following bound

$$|\hat{\epsilon}_{i,t+1}| \leq \frac{\zeta\delta_r}{2} + \frac{\zeta\delta_r^3}{\lambda_{t+1}}$$

Recall $\lambda_{t+1} > \delta_r^2$ by construction, so

$$|\hat{\epsilon}_{i,t+1}| \leq \frac{3\zeta\delta_r}{2} \leq 2\zeta\delta_r$$

Case 2 : $(B\delta_r\cos(\alpha) = y_{\epsilon,t+1}\frac{\lambda_{t+1} - \delta_r^2}{2\lambda_{t+1} - \delta_r^2})$

Substituting we get

$$L^* = \frac{\delta_r^2 y_{\epsilon,t+1}^2}{\lambda_t^2}(1 - \frac{\lambda_{t+1} - \delta_r^2}{2\lambda_{t+1} - \delta_r^2})^2 + \frac{2y_{\epsilon,t+1}^2}{\lambda_{t+1}}(1 - \frac{\lambda_{t+1} - \delta_r^2}{2\lambda_{t+1} - \delta_r^2})(\frac{\lambda_{t+1} - \delta_r^2}{2\lambda_{t+1} - \delta_r^2})$$

which simplifies to

$$L^* = \frac{y_{\epsilon,t+1}^2}{2\lambda_{t+1} - \delta_r^2} \leq \frac{y_{\epsilon,t+1}^2}{\lambda_{t+1}}$$

last inequality is because $\lambda_{t+1} \geq \delta_r^2$.

Taking both cases into consideration and adding the telescopic series (27) we get

$$\begin{aligned}
\|\hat{\epsilon}_{i,T}\|_2^2 &\leq 4\zeta^2\delta_r^2 + \Sigma_{s=1}^{T}(\|\theta_s\|_2^2 - \|\theta_{s-1}\|_2^2) \\
&\leq 4\zeta^2\delta_r^2 + \Sigma_{s=1}^{T}\frac{y_{\epsilon,s}^2}{\lambda_s} \\
&\leq 4\zeta^2\delta_r^2 + \Sigma_{s=1}^{\frac{T}{d}}\frac{y_{\epsilon,s}^2}{\delta_r^2}\frac{d}{s} \qquad (28) \\
&\leq 4\zeta^2\delta_r^2 + \frac{d}{4}\zeta^2\delta_r^2\log T \qquad (29)
\end{aligned}$$

where (28), comes because we need to pick d orthogonal vectors before all the eigen values become equal again, and (29) is a standard bound on harmonic sum and because $y_{i,t}$ is bounded by $\frac{\zeta\delta_r^2}{2}$ by smoothness assumption.

Therefore

$$\|\hat{\epsilon}_{i,T}\|_2 \leq 2\zeta\delta_r\sqrt{2d\log T} = O(\zeta\delta_r\sqrt{d\log T})$$

□

D.1. Optimal Action in the Estimated Safe Set

Next we show i) $x^* \in D_{MT'}^{\text{safe}}$ and ii) $D_{MT'}^{\text{safe}} \subseteq D_0^{\text{safe}}$ with probability at least $1-\delta$, if $\frac{T'}{\log^2 T'} \geq \left(2d\frac{4\delta_f^2}{(\Delta-\zeta\delta_f^2)^2}\right)^2$.

Proof. Recall

$$\Gamma_t^i = \{x \in \Gamma_i : \nabla\hat{f}_{it}^\top(x - x_i^s) + \frac{\Delta}{2} \leq \tau - \tau_i^s\}$$

There exists $T'(\Delta)$ such that

$$\Delta \geq 2\zeta\delta_f^2\sqrt{2d\log T'(\Delta)} + S\beta_T\sqrt{\frac{d}{T'(\Delta)}}$$

where $T'(\Delta)$ is defines as

$$T'(x) = \min\{t > 0 : \Delta \geq 2\zeta\delta_f^2\sqrt{2d\log t} + S\beta_T\sqrt{\frac{d}{t}}\}$$

Now Consider

$$\begin{aligned}
\nabla\hat{f}_{i^*t}^\top(x^* - x_{i^*}^s) &= \nabla\hat{f}_{i^*}^{LS\top}(x^* - x_{i^*}^s) + \hat{\epsilon}_{it}^\top(x^* - x_{i^*}^s) \\
&\leq \nabla f_{i^*}^\top(x^* - x_{i^*}^s) + \beta_t^{i^*}||x - x_{i^*}^s||_{A_{i^*,t}^{-1}} + \hat{\epsilon}_{it}^\top(x^* - x_{i^*}^s) \\
&\leq \nabla f_{i^*}^\top(x^* - x_{i^*}^s) + \beta_t^{i^*}||x - x_{i^*}^s||_{A_{i^*,t}^{-1}} + 2\zeta\delta_r\delta_f\sqrt{2d\log T} \\
&\leq f_{i^*}(x^*) - f_{i^*}(x_{i^*}^s) + \frac{\zeta\delta_f^2}{2} + \beta_t^{i^*}||x - x_{i^*}^s||_{A_{i^*,t}^{-1}} + 2\zeta\delta_r\delta_f\sqrt{2d\log T}
\end{aligned}$$

Since we pick actions as :

$$x_t = \max_{x \in \tilde{D}_i^w} \|x - x_i^s\|_{A_{i,t}^{-1}}$$

any d consecutive actions are orthogonal and uniformly expand the eigen spectrum i.e

$$\sum_{t=s}^{t=s+d} x_t x_t^\top = \delta_r^2 \mathbf{I}$$

as a consequence

$$\beta_{T'}^{i^*} ||x - x_{i^*}^s||_{A_{T'}^{i^*-1}} \leq \beta_{T'}^{i^*} \frac{\sqrt{d}}{\delta_r \sqrt{T'}} \|x - x_{i^*}^s\| \leq \beta_{T'}^{i^*} \frac{\sqrt{d}}{\delta_r \sqrt{T'}} \delta_f$$

which gives us the following upper bound:

$$\begin{aligned}\nabla \hat{f}_{i^* T'}^\top (x^* - x_{i^*}^s) &\leq f_{i^*}(x^*) - f_{i^*}(x_{i^*}^s) + \frac{\zeta \delta_f^2}{2} + \beta_{T'}^{i^*} \frac{\sqrt{d}}{\delta_r \sqrt{T'}} \delta_f + 2\zeta \delta_r \delta_f \sqrt{2d \log T} \\ &= f_{i^*}(x^*) - f_{i^*}(x_{i^*}^s) + \frac{\Delta}{2}\end{aligned}$$

Last inequality follows when T' is large enough such that

$$\frac{\Delta}{2} \geq \frac{\zeta \delta_f^2}{2} + 2\zeta \delta_r \delta_f \sqrt{2d \log T'} + \beta_T \sqrt{\frac{d}{T'}} \frac{\delta_f}{\delta_r}$$

further we can scale $\delta_r \sim (\frac{1}{T'})^{0.25}$ to get

$$\begin{aligned}\frac{\Delta}{2} &\geq \frac{\zeta \delta_f^2}{2} + 2\zeta \delta_f \sqrt{\frac{2d \log T'}{\sqrt{T'}}} + \beta_T \delta_f \sqrt{\frac{d}{\sqrt{T'}}} \\ &\sim O(1)\end{aligned}$$

To find T', we show that upper bound of RHS is less than LHS, show the following relation:

$$\frac{\zeta \delta_f^2}{2} + 2\zeta \delta_f \sqrt{\frac{2d \log T'}{\sqrt{T'}}} + \beta_T \delta_f \sqrt{\frac{d}{\sqrt{T'}}} \leq \frac{\zeta \delta_f^2}{2} + \delta_f \sqrt{\frac{2d \log T'}{\sqrt{T'}}} (2\zeta + \beta_T) \leq \frac{\Delta}{2}$$

rearranging the above we get

$$\frac{T'}{\log^2 T'} \geq \left(2d \frac{4\delta_f^2}{(\Delta - \zeta \delta_f^2)^2} \right)^2$$

Therefore for any $\Delta > \zeta \delta_f^2$, we can arbitrarily find large enough exploration time T' to satisfy safety gap.

Now plugging into the definition of $\hat{\Gamma}_{i,t}$, we get

$$f_{i^*}(x^*) - f_{i^*}(x_{i^*}^s) + \frac{\Delta}{2} + \frac{\Delta}{2} \leq \tau - \tau_i^s$$

which gives

$$f_{i^*}(x^*) \leq \tau - \Delta$$

as desired.

Next we use similar argument to show that if $x \in \hat{\Gamma}_{i,t}$, then $f_i(x) \leq \tau$. To this we consider the following lower bound:

$$\begin{aligned}\nabla \hat{f}_{i,T'}^{\top}(x - x_{i*}^s) &\geq f_i(x) - f_i(x_i^s) - \frac{\zeta\delta_f^2}{2} - \beta_{T'}^i \frac{\sqrt{d}}{\delta_r\sqrt{T'}}\delta_f + 2\zeta\delta_r\delta_f\sqrt{2d\log T} \\ &= f_i(x) - f_i(x_i^s) - \frac{\Delta}{2}\end{aligned}$$

Plugging into definition of $\hat{\Gamma}_{i,t}$, the $\frac{\Delta}{2}$ terms cancel out to give

$$f_i(x) \leq \tau$$

as desired. □

E. Helpful Lemmas

Lemma E.1 (Cauchy Schwarz)**.** *If*

$$||v - \mu||_A \leq \beta$$

then

$$(v-\mu)^{\top}x \leq ||v-\mu||_A||x||_{A^{-1}} \leq \beta||x||_{A^{-1}}$$

Theorem E.2 (Confidence Region)**.** . *Let Assumptions 2.2 and 2.3 hold. Fix any $\delta \in (0,1)$ and let β_t in (6) be chosen as follows,*

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

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

Proof. This result is from Theorem 4.1 in (Abbasi-Yadkori et al., 2011). □

Theorem E.3. *(Matrix Chernoff Inequality, (Tropp et al., 2015)). Consider a finite sequence $\{X_k\}$ of independent, random, symmetric matrices in $\mathbb{R}^d$. Assume that $\lambda_{\min}(X_k) \geq 0$ and $\lambda_{\max}(X_k) \leq L$ for each index k. Introduce the random matrix $Y = \sum_k X_k$. Let μ_{min} denote the minimum eigenvalue of the expectation $\mathbb{E}[Y]$,*

$$\mu_{\min} = \lambda_{\min}(\mathbb{E}[Y]) = \lambda_{\min}\left(\sum_k E[X_k]\right).$$

Then, for any $\epsilon \in (0,1)$, it holds,

$$\Pr\left(\lambda_{\min}(Y) \leq \epsilon\mu_{\min}\right) \leq d \cdot \exp\left(-(1-\epsilon)^2\frac{\mu_{\min}}{2L}\right).$$