Directional Optimism for Safe Linear Bandits

Spencer Hutchinson **Berkay Turan** **Mahnoosh Alizadeh**
University of California, Santa Barbara

Abstract

The safe linear bandit problem is a version of the classical stochastic linear bandit problem where the learner's actions must satisfy an uncertain constraint at all rounds. Due its applicability to many real-world settings, this problem has received considerable attention in recent years. By leveraging a novel approach that we call *directional optimism*, we find that it is possible to achieve improved regret guarantees for both well-separated problem instances and action sets that are finite star convex sets. Furthermore, we propose a novel algorithm for this setting that improves on existing algorithms in terms of empirical performance, while enjoying matching regret guarantees. Lastly, we introduce a generalization of the safe linear bandit setting where the constraints are convex and adapt our algorithms and analyses to this setting by leveraging a novel convex-analysis based approach.

1 INTRODUCTION

The stochastic linear bandit setting Dani et al. (2008); Rusmevichientong and Tsitsiklis (2010); Abbasi-Yadkori et al. (2011) is a sequential decision-making problem where, at each round, a learner chooses a vector action and subsequently receives a reward that, in expectation, is a linear function of the action. This problem has found broad applications in fields ranging from online recommendation engines to ad placement systems, to clinical trials. In the rich literature that has emerged, it is often assumed that any constraints on the learner's actions are *known*. In the real world, however, there are often constraints that are both *uncertain* and need to be met at *all rounds*, such as toxicity limits in clinical trials or sensitive topics for recommendation engines. As a result, linear bandit problems with uncertain and roundwise constraints have received considerable attention in recent years from works such as Amani et al. (2019), Khezeli and Bitar (2020), Pacchiano et al. (2021), Moradipari et al. (2021) and Varma et al. (2023).

A natural formulation of the safe linear bandit problem, initially studied by Moradipari et al. (2021), imposes a linear constraint on every action x_t of the form $a^\top x_t \leq b$ where a is unknown, b is known and the learner gets noisy feedback on $a^\top x_t$. To address this problem, algorithms have been proposed with $\tilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ (Moradipari et al. (2021)) and $\tilde{\mathcal{O}}(d\sqrt{T})$ (Pacchiano et al. (2021); Amani and Thrampoulidis (2021)) regret. These algorithms operate by choosing actions from a pessimistically safe set using versions of Thompson sampling or upper confidence bound where the confidence set for the reward parameter is scaled by a *fixed* constant in both cases. In this work, we introduce an algorithm with matching $\tilde{\mathcal{O}}(d\sqrt{T})$ regret that avoids the use of this fixed scaling by implementing optimism with respect to *directions*, and find that, when compared to the above-mentioned approach, our algorithm enjoys improved performance in problem instances with less restrictive constraints. Leveraging this intuition, we then give algorithms that enjoy improved regret guarantees in terms of the problem dimension for well-separated problem instances and settings with finite star convex action sets.

In fact, this approach is part of a broader perspective for the safe linear bandit problem in which we understand this setting as fundamentally a problem of choosing directions (rather than actions). Since the set of feasible actions is unknown in safe linear bandits, the uncertainty in the problem comes from both the uncertainty in the reward and the uncertainty in the diameter of the feasible action set in each direction. Accordingly, any algorithm for this setting should appropriately quantify both of these uncertainties to ensure low regret. This understanding facilitates our contributions in both geometry-dependent regret guarantees and empirical performance.

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

Algorithm	General	Problem-dependent	Finite-star convex action set	Linked convex constraint
Safe-LTS [1]	$\tilde{\mathcal{O}}(d^{3/2}\sqrt{T})$	-	-	-
GenOP [2],[3]	$\tilde{\mathcal{O}}(d\sqrt{T})$	$\tilde{\mathcal{O}}\left(\frac{d^2}{\Delta}+\sqrt{T}\right)$ Appendix E	-	$\tilde{\mathcal{O}}(d\sqrt{T})$ Appendix D
ROFUL (Alg. 1)	$\tilde{\mathcal{O}}(d\sqrt{T})$ Theorem 1	$\tilde{\mathcal{O}}\left(\frac{d^2}{\Delta}+\sqrt{T}\right)$ Corollary 1	-	$\tilde{\mathcal{O}}(d\sqrt{T})$ Appendix D
Safe-PE (Alg. 3)	-	-	$\tilde{\mathcal{O}}(\sqrt{dT})$ Theorem 3	$\tilde{\mathcal{O}}(\sqrt{dT})$ Appendix D

Table 1: Algorithms and regret bounds developed in existing works and this paper for the safe linear bandit problem where T is horizon and d is problem dimension. Existing work is shown in gray, where the references are [1] for Moradipari et al. (2021), [2] for Pacchiano et al. (2021) and [3] for Amani and Thrampoulidis (2021). Due to variations in problem settings in existing work, we use the name GenOP to refer to a generic upper confidence bound-based algorithm that uses the expanded confidence set approach from [2], [3] (see Section 3.3 for details).

Contributions Our contributions are summarized in Table 1 and in the following:

- We propose a novel UCB-based algorithm, ROFUL, which enjoys $\tilde{\mathcal{O}}\left(d\sqrt{T}\right)$ regret. We provide some intuition and empirical evidence as to when ROFUL is preferred over existing approaches. (Section 3)
- We introduce a notion of well-separated problem instance in safe linear bandits, and show that it is possible to achieve $\tilde{\mathcal{O}}\left(\frac{d^2}{\Delta}+\sqrt{T}\right)$ regret in this setting. (Section 3.2)
- We study the case when the action set is a finite star convex set and introduce a phased elimination-based algorithm, Safe-PE, which is proven to enjoy $\tilde{\mathcal{O}}\left(\sqrt{dT}\right)$ regret. (Section 4)
- We introduce a generalization of the safe linear bandit problem, which we call *linked convex constraints*, where each action x_t needs to satisfy $Ax_t \in \mathcal{G}$ for all $t \in [T]$ where $\mathcal{G}$ is an arbitrary convex set. We extend the ROFUL and Safe-PE algorithms and their analyses to this setting with a novel convex analysis-based approach. (Section 5)
- Simulation results provide validation for the theoretical guarantees and numerical comparison to existing approaches. (Section 6)

Related Work Uncertain constraints have been considered in various learning and optimization problems, often under the umbrella of "safe learning". This includes constrained Markov decision processes (CMDP), where the constraints take the form of limits on auxiliary cost functions (Achiam et al. (2017); Wachi and Sui (2020); Liu et al. (2021a); Amani et al. (2021); Bura et al. (2022); Lindner et al. (2023)). There have also been works that study convex optimization with uncertain constraints that are linear (Usmanova et al. (2019); Fereydounian et al. (2020)), and safe bandit optimization with Gaussian process priors on the objective and constraints (Sui et al. (2015, 2018)). Although the Gaussian process bandit framework is able to capture a wider class of reward and constraint functions than linear bandits, safe Gaussian process bandit works typically make the stronger assumption that the constraint is not tight on the optimal action. Some recent literature has also studied safe exploration of bandits (Wang et al. (2022)) as well as best arm identification under safety constraints (Wan et al. (2022); Lindner et al. (2022); Camilleri et al. (2022)). These works consider objectives other than regret minimization, i.e. accurate estimation of policy value or finding the best arm, and are therefore distinct from the regret minimization setting that we study here.

For the bandit setting in particular, various types of constraints have been considered, including knapsacks, cumulative constraints and conservatism constraints. In knapsack bandits, pulling each arm yields both a reward and a resource consumption with the objective being to maximize the reward before the resource runs out (Badanidiyuru et al. (2013, 2014); Agrawal and Devanur (2016); Agrawal et al. (2016); Cayci et al. (2020)). There are also works that consider various types of cumulative constraints on the actions, including ones with fairness constraints (Joseph et al. (2018); Grazzi et al. (2022)), budget constraints (Combes et al. (2015); Wu et al. (2015)) and general nonlinear constraints for which the running total is constrained (Liu et al. (2021b)). Similarly, there are works that bound the cumulative constraint violation in the multi-armed

(Chen et al. (2022b)) and linear (Chen et al. (2022a)) settings. Similar to us, Chen et al. (2022a) uses an optimistic action set, although their algorithm does not ensure constraint satisfaction at each round and instead aims for sublinear constraint violation. These types of cumulative constraints differ from the setting we study, where constraints are roundwise and must hold at each round. In the conservative bandit literature, the running total of the reward needs to stay close to the baseline reward (Wu et al. (2016); Kazerouni et al. (2017)).

Various works have also studied linear bandits with roundwise constraints. In particular, Amani et al. (2019) studies a stochastic linear bandit setting with a linear constraint, where the constraint parameter is the linearly transformed reward parameter and there is no feedback on the constraint value. Also, Khezeli and Bitar (2020) and Moradipari et al. (2020) study a conservative bandit setting where the reward at each round needs to stay close to a baseline. Pacchiano et al. (2021) studies a setting where the learner chooses a distribution over the actions in each round and the constraint needs to be satisfied in expectation with respect to this distribution. Although this is a slightly different type of constraint than we consider, their approach can be adapted to our setting which we discuss further in Section 3.3.

Most relevantly, several works have studied linear bandit problems with an auxiliary constraint function that the learner observes noisy feedback of and needs to ensure is always below a threshold. Moradipari et al. (2021) studied such a setting with a linear constraint function and proposed the Safe-LTS algorithm. Also, Amani and Thrampoulidis (2021) studied a decentralized version of the same problem where the agents collaborate over a communication network. Lastly, the recent work by Varma et al. (2023) considers a safe linear bandit problem where different constraints apply to different parts of the domain and the learner only receives feedback on a given constraint when she selects an action from the applicable part of the domain. However, all of these works use algorithms that choose actions from a pessimistic action set using either linear UCB or linear TS with a confidence set that is scaled by a fixed constant, which significantly differs from our proposed algorithms as detailed in Section 3.3. Also, they do not achieve improved regret guarantees for well-separated and finite star-convex settings as we do.

2 PRELIMINARIES

Notation We use $\mathcal{O}(\cdot)$ to refer to big-O notation and $\tilde{\mathcal{O}}(\cdot)$ for the same except ignoring log factors. To refer to the p-norm ball and sphere of radius one, we use the notation $\mathbb{B}_p$ and $\mathbb{S}_p$ respectively, where $\mathbb{B}$ and $\mathbb{S}$ refers to the 2-norm ball and sphere. For some $n \in \mathbb{N}$, we use $[n]$ to refer to the set $\{1, 2, ..., n\}$. For a matrix M, its transpose is denoted by $M^\top$. For a positive definite matrix M and vector x, the notation for the weighted norm is $\|x\|_M = \sqrt{x^\top M x}$. For a real number x, the ceiling function is denoted by $\lceil x \rceil$.

Problem Setup We study a stochastic linear bandit problem with a constraint that must be satisfied at all rounds (at least with high probability). At each round $t \in [T]$, the learner chooses an action x_t from the closed set $\mathcal{X}$. She subsequently receives the reward $y_t = \theta^\top x_t + \epsilon_t$ and the noisy constraint observation $z_t = a^\top x_t + \eta_t$, where the reward vector $\theta \in \mathbb{R}^d$ and constraint vector $a \in \mathbb{R}^d$ are unknown, and ϵ_t and η_t are noise terms. Critically, the learner must ensure that $a^\top x_t \leq b$ for all $t \in [T]$, where $b > 0$ is known. We will refer to the feasible set of actions as $\mathcal{Y} := \{x \in \mathcal{X} : a^\top x \leq b\}$.

In addition to guaranteeing constraint satisfaction, the learner also aims to minimize the pseudo-regret,

$$R_T := \sum_{t=1}^{T} \theta^\top (x_* - x_t),$$

where $x_* = \arg\max_{x \in \mathcal{Y}} \theta^\top x$ is the optimal constraint-satisfying action. Going forward, we will use the term regret to refer to pseudo-regret.

We use the following assumptions.

Assumption 1. *The action set $\mathcal{X}$ is star-convex. Also, it holds that $\|x\| \leq 1$ for all $x \in \mathcal{X}$ and that $\theta^\top x_* > 0$.*

Assumption 2. *There exists positive real numbers S_a and S_θ such that $\|a\| \leq S_a$ and $\|\theta\| \leq S_\theta$. Let $S := \max(S_a, S_\theta)$. Also, it holds that $\nu := \frac{b}{S_a} \leq 1$.*

Remark 1. *If $\nu > 1$, then it is known that the constraint is loose and therefore the problem can be treated as a conventional linear bandit problem.[1] Therefore, our assumption that $\nu \leq 1$ avoids this trivial setting and allows for cleaner presentation of results.*

Technical Approach Our approach to this problem is based on the perspective that it is fundamentally a problem of choosing *directions* rather than actions and therefore any solution approach should be focused on choosing directions that will result in low regret. This perspective comes from the understanding that the only viable solutions are actions that are in the maximally-scaled part of the feasible set (i.e. the set of $x \in \mathcal{Y}$ such that $\zeta x \notin \mathcal{Y}$ for all $\zeta > 1$).[2] Therefore, the

[1] If $\nu > 1$, then for all $x \in \mathcal{X}$ it holds that $a^\top x \leq \|a\|\|x\| \leq S_a \nu < b$ given that $\|x\| \leq 1 < \nu$ for all $x \in \mathcal{X}$.

[2] To see that the optimal action must be in the maximally scaled part of the set, suppose that it is not (i.e. that there

Algorithm 1: Restrained OFUL (ROFUL)

Input: $\mathcal{X}, \nu, b, \beta_t, \delta \in (0,1), \lambda \geq 1$

1 **for** $t = 1$ **to** T **do**

2 Update $\hat{a}_t := V_t^{-1} \sum_{k=1}^{t-1} x_k z_k$ and $\hat{\theta}_t := V_t^{-1} \sum_{k=1}^{t-1} x_k y_k$, where $V_t = \sum_{k=1}^{t-1} x_k x_k^\top + \lambda I$.

3 Update $\mathcal{Y}_t^p := \left\{ x \in \mathcal{X} : \hat{a}_t^\top x + \beta_t \|x\|_{V_t^{-1}} \leq b \right\}$ and $\mathcal{Y}_t^o := \left\{ x \in \mathcal{X} : \hat{a}_t^\top x - \beta_t \|x\|_{V_t^{-1}} \leq b \right\}$.

4 Find a $\tilde{x}_t \in \arg\max_{x \in \mathcal{Y}_t^o} \left(\hat{\theta}_t^\top x + \beta_t \|x\|_{V_t^{-1}} \right)$.

5 Set $\tilde{b}_t = \begin{cases} \min\left(\frac{\nu}{\|\tilde{x}_t\|}, 1 \right) & \text{if } \tilde{x}_t \neq \mathbf{0}, \\ 1 & \text{else.} \end{cases}$

6 Set $\mu_t = \max \{ \mu \in [0,1] : \mu \tilde{x}_t \in \mathcal{Y}_t^p \}$ and $\gamma_t = \max\left(\tilde{b}_t, \mu_t \right)$.

7 Play $x_t = \gamma_t \tilde{x}_t$ and observe y_t, z_t.

8 **end**

challenge lies in identifying the optimal direction given that the maximum scaling of this direction is the only viable solution in that direction. Unlike the conventional linear bandit setting, however, the feasible set is unknown and therefore the uncertainty in the problem comes from both the uncertain reward parameter *and* the uncertainty in the maximum scaling in each direction (i.e. $\zeta = \max\{\alpha \geq 0 : \alpha u \in \mathcal{Y}\}$ for each unit vector $u \in \mathbb{S}$). As such, our solutions to the problem will aim to explicitly characterize both these uncertainties in order to choose directions that will result in low regret. This will be realized via both an upper confidence bound-based algorithm (Section 3) and a phased elimination-based algorithm (Section 4) which are each suited for different action set geometries.

3 RESTRAINED OPTIMISM ALGORITHM

In this section, we first propose the algorithm *Restrained Optimism in the Face of Uncertainty for Linear bandits* (ROFUL, Algorithm 1) to address the stated problem, and then provide general and problem-dependent regret analyses for ROFUL in Sections 3.1 and 3.2, respectively. Additionally, we provide a detailed comparison with existing algorithms in Section 3.3.

exists $\zeta > 1$ such that $\zeta x_* \in \mathcal{Y}$). It follows that the point ζx_* has larger reward than x_*, i.e. $\theta^\top(\zeta x_*) > \theta^\top x_*$, and therefore x_* cannot be the optimal action (where we use $\theta^\top x_* > 0$ from Assumption 1).

Optimistic Direction Selection The key idea behind the ROFUL algorithm is that it uses an *optimistic* action set ($\mathcal{Y}_t^o$) to find which direction should be played to efficiently balance exploration and exploitation, while using a *pessimistic* action set ($\mathcal{Y}_t^p$) to find the scaling of this direction that will ensure constraint satisfaction. In each round, the algorithm first finds the action $\tilde{x}_t$ which maximizes the upper confidence bound over the optimistic set (line 4), and then finds the largest scalar γ_t such that $\gamma_t \tilde{x}_t$ is known to be in the pessimistic set (line 6). The optimistic set overestimates the feasible set and the upper-confidence bound overestimates the reward, so $\tilde{x}_t$ can be viewed as the optimistic action with respect to both the reward and the constraint. As such, the algorithm uses $\tilde{x}_t$ to determine which direction to play. However, the action $\tilde{x}_t$ may not satisfy the constraints, so it needs to be scaled down until it is within the pessimistic set and will therefore satisfy the constraints.

Confidence Sets for Unknown Parameters In order to construct the optimistic and pessimistic action sets as well as the upper confidence bound for the reward, we use confidence sets for the unknown parameters θ, a. To specify these confidence sets, we need to impose some structure on the noise terms. In particular, the following assumption specifies that the noise terms ϵ_t, η_t are ρ-subgaussian conditioned on the history up to the point that y_t, z_t are observed.

Assumption 3. *For all* $t \in [T]$, *it holds that* $\mathbb{E}[\epsilon_t | x_1, \epsilon_1, ..., \epsilon_{t-1}, x_t] = 0$ *and* $\mathbb{E}[\exp(\lambda\epsilon_t) | x_1, \epsilon_1, ..., \epsilon_{t-1}, x_t] \leq \exp(\frac{\lambda^2\rho^2}{2}), \forall \lambda \in \mathbb{R}$. *The same holds replacing* ϵ_t *with* η_t.

The specific confidence set that we use is from Abbasi-Yadkori et al. (2011) and is given in the following.

Lemma 1 (Theorem 2 in Abbasi-Yadkori et al. (2011))**.** *Let Assumptions 1, 2 and 3 hold. Also, let*

$$\beta_t := \rho \sqrt{d \log\left(\frac{1 + (t-1)/\lambda}{\delta/2} \right)} + \sqrt{\lambda} S. \tag{1}$$

Then with probability at least $1 - \delta$, *it holds that both* $|x^\top(\hat{\theta}_t - \theta)| \leq \beta_t \|x\|_{V_t^{-1}}$ *and* $|x^\top(\hat{a}_t - a)| \leq \beta_t \|x\|_{V_t^{-1}}$ *for all* $x \in \mathbb{R}^d$ *and all* $t \geq 1$.

It follows from Lemma 1 that, with high probability, the optimistic and pessimistic action sets contain and are contained by the true feasible set $\mathcal{Y}$, respectively. Since ROFUL only chooses actions from the pessimistic action set (or those with norm less than ν), the actions chosen by the algorithm satisfy the constraints at all rounds with high probability.

3.1 General Analysis

The ROFUL algorithm (Algorithm 1) is proven to enjoy sublinear regret and maintain constraint satisfaction in the following theorem.

Theorem 1. *Let Assumptions 1, 2 and 3 hold. Then, with probability at least $1-\delta$, the regret of ROFUL (Algorithm 1) satisfies*

$$R_T \leq 2\frac{\|\theta\| + S_a}{b}\beta_T\sqrt{2dT\log\left(1+\frac{T}{\lambda d}\right)}, \tag{2}$$

and $a^\top x_t \leq b$ for all $t \in [T]$.

Inspecting the bound in Theorem 1, we can see that the regret is $\mathcal{O}\left(\frac{1}{b}d\sqrt{T}\log(T)\right)$, only considering T, d and b. This matches the orderwise regret of other safe upper-confidence bound approaches, as discussed in Section 3.3. In the next section, we find that it is possible to achieve improved problem-dependent regret guarantees.

Remark 2. *The ROFUL algorithm and Theorem 1 easily extend to the setting where the action set $\mathcal{X}$ and constraint limit b are allowed to vary in each round.*

3.2 Problem-dependent Analysis

We also study the case where the optimal reward is well-separated from the reward of any feasible action that is not in the same direction as the optimal action. To make this concrete, let the *reward gap* be defined as[3]

$$\Delta := \inf_{x\in\mathcal{Y}:\ x\neq\alpha x_*\ \forall\alpha>0} \theta^\top(x_* - x), \tag{3}$$

We study the case where $\Delta > 0$. Note that the typical notion of a "reward gap" in linear bandits, such as that used by the problem-dependent analysis in Dani et al. (2008) and Abbasi-Yadkori et al. (2011), is not particularly useful in the safe linear bandit setting because it relies on the optimal reward being separated from the reward of any other action that the learner might play. This could occur in the conventional linear bandit setting either when the feasible set is finite, which would not be a star convex set (except for the trivial case), or when the feasible set has finite extrema, which will not ensure that the played actions are well-separated in safe linear bandits given that the feasible set is unknown. Nonetheless, when the constraint is loose (i.e. $\nu > 1$), a well-separated problem in our setting ($\Delta > 0$) implies a well-separated problem in the conventional linear bandit setting.

[3]Unlike the problem-dependent analysis in Amani et al. (2019), our notion of a reward gap does not depend on how tight the constraints are on the optimal action.

Wrong Directions are Rarely Selected We find that when $\Delta > 0$, we can establish a polylogarithmic bound on the number of times that ROFUL chooses the wrong direction, which is denoted by

$$B_T := \sum_{t=1}^{T} \mathbb{1}\{\nexists\ \alpha > 0 : x_t = \alpha x_*\}.$$

Specifically, the following theorem shows that ROFUL chooses $\mathcal{O}\left(\frac{1}{b^2\Delta^2}d^2\log^2(T)\right)$ wrong directions when $\Delta > 0$.

Theorem 2. *Let Assumptions 1, 2 and 3 hold. If $\Delta > 0$, then the number of wrong directions chosen by ROFUL (Algorithm 1) satisfies*

$$B_T \leq \frac{32S^2\beta_T^2 d}{b^2\Delta^2}\log\left(1+\frac{T}{\lambda d}\right)$$

with probability at least $1-\delta$.

Nearly Dimension-free Regret Leveraging Theorem 2, we can devise a version of ROFUL that achieves improved regret guarantees when $\Delta > 0$ and known. In particular, Theorem 2 implies that the optimal direction can be identified in a polylogarithmic number of rounds. Once the optimal direction has been identified, the problem becomes one-dimensional and therefore does not suffer any dimensional dependence.

Corollary 1. *Let Assumptions 1, 2 and 3 hold. If $\Delta > 0$, consider the algorithm PD-ROFUL:* [4]

1. *Play ROFUL until any single direction has been played more than $\bar{B} := \frac{32S^2\beta_T^2 d}{b^2\Delta^2}\log\left(1+\frac{T}{\lambda d}\right)$ times. Let this direction be denoted by u_*.*
2. *For the remaining rounds, play ROFUL (after restarting) for the 1-dimensional safe linear bandit problem of choosing $\xi_t \in \mathbb{R}_+$ and then playing $\xi_t u_*$.*

Then, with probability at least $1-2\delta$,

$$R_T \leq \frac{4S}{b}\beta_{2\bar{B}+1}\sqrt{2d(2\bar{B}+1)\log\left(1+\frac{2\bar{B}+1}{\lambda d}\right)} + \frac{4S}{b}\tilde{\beta}_T\sqrt{2T\log\left(1+\frac{T}{\lambda d}\right)}$$

where $\tilde{\beta}_T$ is β_T with $d=1$.

Corollary 1 indicates that when $\Delta > 0$ and known, it is possible to achieve $\tilde{\mathcal{O}}(\frac{d^2}{b^2\Delta} + \frac{1}{b}\sqrt{T})$ regret. When T is large and $\frac{1}{\Delta}$ is $\mathcal{O}(1)$, this improves on the general regret

[4]Detailed pseudo-code of PD-ROFUL is given in Algorithm 2 in Appendix B.

bound in Theorem 1 because the second term dominates. Concretely, as T goes to infinity, $R_T/\sqrt{T}\log(T)$ goes to $\mathcal{O}(\frac{1}{b})$ whereas in the general case (i.e. Theorem 1), it goes to $\mathcal{O}(d\frac{1}{b})$.

Remark 3. *This problem-dependent analysis approach yields similar guarantees for existing safe linear bandit algorithms as shown in Appendix E.*

3.3 Comparison with Existing Algorithms

In this section, we discuss the key differences between ROFUL and existing safe linear bandit algorithms. Compared to ROFUL, which uses an optimistic action set to identify low-regret actions, existing safe linear bandit algorithms often choose actions directly from the pessimistic action set using either linear UCB (Pacchiano et al. (2021); Amani and Thrampoulidis (2021)) or linear TS (Moradipari et al. (2021)) where an expanded confidence set is used in both cases. In our specific setting, the linear UCB version of the expanded confidence set approach can be written as

$$x_t \in \arg\max_{x\in\mathcal{Y}_t^p} \left(\hat{\theta}_t^\top x + \kappa\beta_t\|x\|_{V_t^{-1}}\right), \tag{4}$$

with an appropriately chosen parameter $\kappa \geq 1$. The specific choice of κ ensures that optimism holds, i.e. that $\hat{\theta}_t^\top x_t + \kappa\beta_t\|x_t\|_{V_t^{-1}} \geq \theta^\top x_*$, which is critical to ensuring that the algorithm enjoys sublinear regret. We call this generic algorithm GenOP (as in Generic Optimism-Pessimism) in reference to the concept of optimism-pessimism that is often used in safe linear bandits (e.g. Pacchiano et al. (2021)). Note that the choice of κ used in existing UCB-based algorithms is not appropriate for our setting because such algorithms were developed for slightly different settings (i.e. decentralized Amani and Thrampoulidis (2021), local constraints Varma et al. (2023), or constraints in expectation Pacchiano et al. (2021)) so we show in Appendix D.2 that it is sufficient to choose $\kappa = 1+\frac{2S_\theta}{b}$, to get

$$R_T \leq (1+\kappa)\beta_T\sqrt{2dT\log\left(1+\frac{T}{\lambda d}\right)}. \tag{5}$$

Note that because GenOP uses a *fixed* κ parameter that must be chosen ahead of time, it is necessarily defined using worst-case quantities (such as S_θ). Conversely, ROFUL uses the optimistic action set and safe scaling γ_t which are updated with empirical quantities in each round and therefore improve as more data is collected. This suggests that ROFUL is preferable in “easier” problem instances in which worst-case quantities are loose on the true empirical quantities.

We can gain additional insight into the respective benefits of either algorithm by comparing the regret bounds. Specifically, it follows from (2) and (5) that ROFUL enjoys a tighter regret bound than GenOP when

$$S_\theta - \|\theta\| > S_a - b. \tag{6}$$

The quantity on the left-hand side represents how loose the assumed bound on the reward parameter is on the true value (given that Assumption 2 specifies that $\|\theta\| \leq S_\theta$), while the right-hand side represents how loose the assumed bound on the constraint limit (b) is (given that Assumption 2 specifies that $\nu = b/S_a \leq 1$ and therefore that $b \leq S_a$). Therefore, (6) suggests that ROFUL is preferred over GenOP when the bound on the reward parameter is loose and the bound on the constraint limit is tight. Our numerical experiments support this intuition as ROFUL outperforms GenOP on average when b is large (and therefore S_a is tighter on b), while the two algorithms perform similarly in the settings when b is small (and therefore S_a is looser on b).

4 SAFE PHASED ELIMINATION ALGORITHM

In this section, we propose the algorithm *Safe Phased Elimination* (Safe-PE) for the case when the action set is a finite star-convex set. We provide a high-level description of Safe-PE here and give the full algorithm in Appendix C. The assumption that the action set is a finite star-convex set means that it can be represented as

$$\mathcal{X} = \bigcup_{i\in[k]} \{\alpha u_i : \alpha \in [0,\alpha_i]\}, \tag{7}$$

where $u_1, ..., u_k \in \mathbb{S}$ are unit vectors and $\alpha_1, ..., \alpha_k \in \mathbb{R}_{++}$ are the maximum scalings for each unit vector. We find that in such a setting, it is possible to reduce the dependence on the problem dimension when $k \ll 2^d$. The key insight is that a confidence set at a single action applies to all scalings of that action without the need for a union bound over a cover (or related technique). This insight allows us to leverage the reduced dimension dependence offered by phased elimination algorithms in the safe linear bandit setting. Nonetheless, it also introduces additional challenges due to the fact that the pessimistic action set varies from phase to phase.

Algorithm Description Our Safe-PE algorithm operates in phases $j = 1, 2, ...$ that grow exponentially in duration, and maintains a set of viable directions $\mathcal{A}$ and a pessimistic set of actions $\mathcal{Y}^p$ that are updated in each round. In particular, each phase j proceeds as:

1. For 2^{j-1} rounds, play the action with the largest confidence set width $\|\cdot\|_{V_t^{-1}}$ in each round.

2. Eliminate directions from $\mathcal{A}$ that have low estimated reward.

3. Update $\mathcal{Y}^p$ by scaling the directions in $\mathcal{A}$ as large as possible while still being verifiably safe.

This algorithm builds on existing phased elimination algorithms, including those from Auer (2002), Chu et al. (2011) and, specifically, Valko et al. (2014) and Kocák et al. (2020). However, Safe-PE differs in that it eliminates directions, instead of distinct actions, and maintains a set of safe actions to ensure constraint satisfaction. Furthermore, it requires a looser criterion when eliminating directions to ensure that the optimal direction is not eliminated.

Regret Analysis As is commonly used for phased elimination algorithms Auer (2002); Chu et al. (2011); Lattimore et al. (2020), we assume that the noise terms are independent subgaussian random variables.

Assumption 4. *The noise sequences* $(\epsilon_t)_{t=1}^T$ *and* $(\eta_t)_{t=1}^T$ *are sequences of independent* ρ*-subgaussian random variables.*

With this, we state the regret guarantees for the Safe-PE algorithm in Theorem 3.

Theorem 3. *Let Assumptions 1, 2 and 4 hold. When the action set is a finite star-convex set, the regret of Safe-PE (Algorithm 3 in Appendix C) is* $\tilde{\mathcal{O}}(\frac{1}{b^2}\sqrt{dT})$.

Theorem 3 shows that, for the case when the action set is a finite star convex set, the regret of Safe-PE is $\tilde{\mathcal{O}}(\sqrt{dT})$ in terms of d and T. Note that the regret only depends on the number of directions (k) in log factors and therefore this improves on the regret of ROFUL in terms of d when $k \ll 2^d$. For example, if the directions are the coordinate directions, i.e. $u_i = e_i$, then $k = 2d$ and therefore the regret bound of Safe-PE is $\tilde{\mathcal{O}}(\sqrt{dT})$ since d only appears in log factors. However, if the directions are the corners of the hypercube, then $k = 2^d$ and the regret bound of Safe-PE is $\tilde{\mathcal{O}}(d\sqrt{T})$. Also, note that the regret bound of Safe-PE depends on $\frac{1}{b^2}$, whereas the regret bound of ROFUL depends on $\frac{1}{b}$. As such, the regret bound of ROFUL is still tighter than that of Safe-PE in some settings, e.g. when b is small and $d = 1$.

5 EXTENSION TO LINKED CONVEX CONSTRAINTS

In this section, we generalize the design and analysis of the algorithms ROFUL, Safe-PE and GenOP to a novel setting which we call *linked convex constraints*, where the output of the constraint function is multi-dimensional and must lie in an arbitrary convex set.

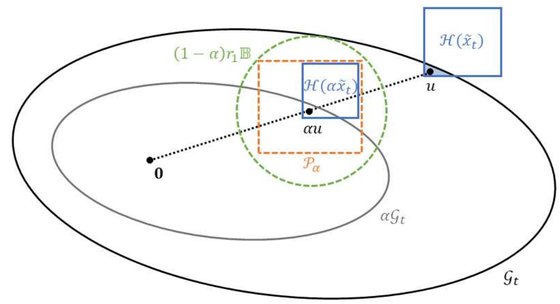


Figure 1: Graphical representation of the approach for lower bounding γ_t (in ROFUL) for the setting with linked convex constraints.

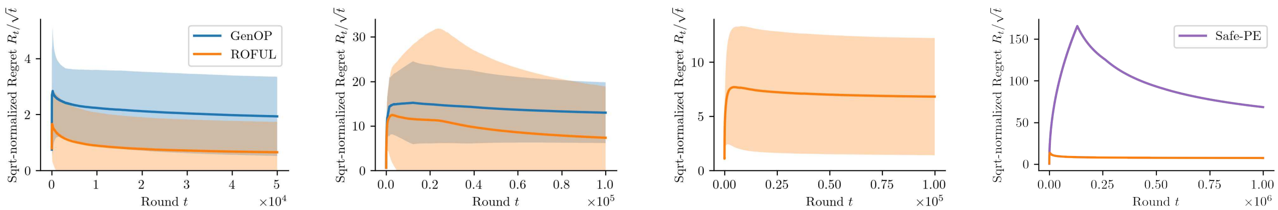
The key challenge in this setting is characterizing how far a point in the optimistic action set is from the pessimistic action set. To address this, we leverage a theoretical tool from the zero-order optimization literature. We only provide a description of key contributions in this section and leave the details of the algorithms to Appendix D.

Problem Description The problem setting is specified as follows. At each round $t \in [T]$, the learner observes $z_t = Ax_t + \eta_t$, where $A \in \mathbb{R}^{n\times d}$ is the unknown constraint matrix and $\eta_t \in \mathbb{R}^n$ is a vector noise term. The learner must ensure that Ax_t is in the known convex set $\mathcal{G}$ for all $t \in [T]$. The reward function and feedback mechanism are the same as the original setting described in Section 2. We assume that there exists $r > 0$ such that $r\mathbb{B} \subseteq \mathcal{G}$. Lastly, we assume that each element of η_t satisfies the assumptions on the noise used for ROFUL (Assumption 3) or Safe-PE (Assumption 4).

Analysis Techniques Although the design of the algorithms trivially extends to this setting, the regret analysis is more challenging. In particular, it is difficult to characterize the distance from any point in the optimistic action set to the pessimistic action set. To address this, we use an analysis tool that is popular in the zero-order optimization literature (e.g. Flaxman et al. (2005)). This tool, given in Fact 1, allows us to consider a shrunk version of the constraint set in order to bound the scaling that is required to take any point in $\mathcal{Y}_t^o$ to $\mathcal{Y}_t^p$, i.e. γ_t in ROFUL. We use a similar approach to bound the scaling required to take any point in $\mathcal{Y}$ to $\mathcal{Y}_t^p$ for GenOP and Safe-PE.

Fact 1. *Let* $\mathcal{D}$ *be a convex set such that* $r\mathbb{B} \subseteq \mathcal{D}$ *for some* $r \geq 0$. *Then, for any* $\alpha \in [0, 1]$ *and* $x \in \alpha\mathcal{D}$, *it holds that* $x + (1 - \alpha)r\mathbb{B} \subseteq \mathcal{D}$.

With this in hand, we can then describe our approach for lower bounding γ_t, which is illustrated in Figure 1. Recall the definition of $\tilde{x}_t$ in ROFUL, where it is known



(a) Linear constraints with large b.

(b) Linear constraints with small b.

(c) Linked convex constraints.

(d) Star-convex multi-armed bandit.

Figure 2: Simulation results of our proposed algorithms (ROFUL, Safe-PE) and generic expanded confidence set algorithm GenOP (see Section 3.3).

that $\tilde{x}_t$ is in $\mathcal{Y}_t^o$. Then, the overall objective is to find some positive scaling α such that $\alpha\tilde{x}_t$ is in $\mathcal{Y}_t^p$ and then it follows that $\gamma_t \geq \alpha$. To do so, we define the uncertainty set for the constraint function at point x as the box $\mathcal{H}(x) := \hat{a}_t^\top x + \beta_t \|x\|_{V_t^{-1}} \mathbb{B}_\infty$ and note that $\mathcal{Y}_t^p$ and $\mathcal{Y}_t^o$ are precisely the set of $x \in \mathcal{X}$ such that $\mathcal{H}(x)$ has nonempty intersection with $\mathcal{G}$ and the set of $x \in \mathcal{X}$ such that $\mathcal{H}(x)$ is contained in $\mathcal{G}$, respectively. First, we consider a point u in the intersection of $\mathcal{G}$ and $\mathcal{H}(\tilde{x}_t)$. Such a point exists given that $\tilde{x}_t$ is in $\mathcal{Y}_t^o$. Next, we scale u by some non-negative scalar α. Note that αu is in $\mathcal{H}(\alpha\tilde{x}_t)$ given that $\mathcal{H}$ is positive homogeneous, i.e. $\alpha\mathcal{H}(x) = \mathcal{H}(\alpha x)$ for any x. In order to show that $\alpha\tilde{x}_t$ is in $\mathcal{Y}_t^p$, we need to show that $\mathcal{H}(\alpha\tilde{x}_t)$ is contained in $\mathcal{G}$. To do so, we first consider a set $\mathcal{P}_\alpha$ that is centered at αu but has twice the radius of $\mathcal{H}(\alpha\tilde{x}_t)$ and therefore contains $\mathcal{H}(\alpha\tilde{x}_t)$ (this is illustrated in Figure 1). We then use Fact 1 to reason that, because u is in $\mathcal{G}$, the ball $\alpha u + (1-\alpha)r_1\mathbb{B}$ is contained in $\mathcal{G}$. Therefore, we choose α such that $(1-\alpha)r_1\mathbb{B} = 2\alpha\sqrt{n}\beta_t\|\tilde{x}_t\|_{V_t^{-1}}$, where the $\sqrt{n}$ is necessary to bound an infinity-norm ball with a 2-norm ball. Some simple algebra shows that $\gamma_t \geq 1 - \frac{2\sqrt{n}}{r}\beta_t\|x_t\|_{V_t^{-1}}$.

6 NUMERICAL EXPERIMENTS

In this section, we numerically validate the theoretical guarantees and assess the performance of the proposed algorithms. Note that we only give a high-level description of the simulations in this section. The details of the experimental settings and additional results are given in Appendix F.

First, we consider a setting with a linear constraint. We study the case when b is large (Figure 2a) and when b is small (Figure 2b). We simulate ROFUL and GenOP for 30 trials for each case, where b is uniformly sampled in the interval $[0.25, 1]$ for the first case and in the interval $[0.05, 0.25]$ for the second case. The action set is taken to be a finite star-convex set with 10 directions that are randomly sampled in each trial. Furthermore, the reward vector θ, constraint vector a, constraint limit b and noise realizations ϵ_t, η_t are also randomly sampled in each trial. The average and standard deviation of the regret normalized by $\sqrt{t}$ are shown in Figures 2a and 2b. These plots suggest that, when b is large, ROFUL outperforms GenOP in the aggregate. When b is small, the average performance of the two algorithms is similar, although GenOP enjoys a smaller standard deviation than ROFUL.

Next, we consider a setting with convex constraints. In particular, we study the case where the constraint set is a ball, i.e. $\mathcal{G} = b\mathbb{B}$ for scalar b. We simulated ROFUL for 30 trials in each setting, where constraint set radius b, the reward vector θ, constraint vector a and noise realizations ϵ_t, η_t are all randomly sampled. The average and standard deviation of the regret normalized by $\sqrt{t}$ is shown in Figure 2c. In this plot, ROFUL converges to constant $\sqrt{t}$ regret. We provide additional results for the case when $\mathcal{G}$ is an infinity-norm ball in Figure 3 in Appendix F.

Lastly, we consider a star-convex multi-armed bandit with results shown in Figure 2d. In this setting, the action set only includes the coordinate directions. We simulate both ROFUL and Safe-PE in this setting with $d = 10$. The regret normalized by $\sqrt{t}$ is shown in Figure 2d. From this plot, it is clear that ROFUL outperforms Safe-PE despite the fact that Safe-PE enjoys a tighter regret bound in terms of the problem dimension. In fact, it is well known that UCB-based algorithms often empirically outperform elimination-based algorithms despite the orderwise tighter regret bound, as discussed by Valko et al. (2014) and Chu et al. (2011). Simulation results for PD-ROFUL (specified in Corollary 1) are given in Appendix F.

7 CONCLUSION

In this work, we take a novel approach to the safe linear bandit problem in which we view it as fundamentally a problem of choosing directions rather than actions. We find that this approach leads to improvements in empirical performance in certain problem instances as well as tighter geometry-dependent regret bounds. An interesting direction for future work is to investigate if this approach yields similar gains when applied to related safe learning problems, such as constrained MDPs or safe Gaussian process optimization.

Acknowledgements

This work was supported by NSF grant #1847096.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pages 22–31. PMLR, 2017.

Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 29, 2016.

Shipra Agrawal, Nikhil R Devanur, and Lihong Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Conference on Learning Theory*, pages 4–18. PMLR, 2016.

Sanae Amani and Christos Thrampoulidis. Decentralized multi-agent linear bandits with safety constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6627–6635, 2021.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

Sanae Amani, Christos Thrampoulidis, and Lin Yang. Safe reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 243–253. PMLR, 2021.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 207–216. IEEE, 2013.

Ashwinkumar Badanidiyuru, John Langford, and Aleksandrs Slivkins. Resourceful contextual bandits. In *Conference on Learning Theory*, pages 1109–1134. PMLR, 2014.

Archana Bura, Aria HasanzadeZonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-Francois Chamberland. Dope: Doubly optimistic and pessimistic exploration for safe reinforcement learning. *Advances in Neural Information Processing Systems*, 35:1047–1059, 2022.

Romain Camilleri, Andrew Wagenmaker, Jamie H Morgenstern, Lalit Jain, and Kevin G Jamieson. Active learning with safety constraints. *Advances in Neural Information Processing Systems*, 35:33201–33214, 2022.

Semih Cayci, Atilla Eryilmaz, and Rayadurgam Srikant. Budget-constrained bandits over general cost and reward distributions. In *International Conference on Artificial Intelligence and Statistics*, pages 4388–4398. PMLR, 2020.

Tianrui Chen, Aditya Gangrade, and Venkatesh Saligrama. A doubly optimistic strategy for safe linear bandits. *arXiv preprint arXiv:2209.13694*, 2022a.

Tianrui Chen, Aditya Gangrade, and Venkatesh Saligrama. Strategies for safe multi-armed bandits with logarithmic regret and risk. In *International Conference on Machine Learning*, pages 3123–3148. PMLR, 2022b.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

Richard Combes, Chong Jiang, and Rayadurgam Srikant. Bandits with budgets: Regret lower bounds and optimal algorithms. *ACM SIGMETRICS Performance Evaluation Review*, 43(1):245–257, 2015.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

Mohammad Fereydounian, Zebang Shen, Aryan Mokhtari, Amin Karbasi, and Hamed Hassani. Safe learning under uncertain objectives and constraints. *arXiv preprint arXiv:2006.13326*, 2020.

Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394, 2005.

Riccardo Grazzi, Arya Akhavan, John IF Falk, Leonardo Cella, and Massimiliano Pontil. Group meritocratic fairness in linear contextual bandits. *Advances in Neural Information Processing Systems*, 35:24392–24404, 2022.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Meritocratic fairness for infinite and contextual bandits. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 158–163, 2018.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Tomáš Kocák, Rémi Munos, Branislav Kveton, Shipra Agrawal, and Michal Valko. Spectral bandits. *The Journal of Machine Learning Research*, 21(1):9003–9046, 2020.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR, 2020.

David Lindner, Sebastian Tschiatschek, Katja Hofmann, and Andreas Krause. Interactively learning preference constraints in linear bandits. In *International Conference on Machine Learning*, pages 13505–13527. PMLR, 2022.

David Lindner, Xin Chen, Sebastian Tschiatschek, Katja Hofmann, and Andreas Krause. Learning safety constraints from demonstrations with unknown rewards. *arXiv preprint arXiv:2305.16147*, 2023.

Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained mdps. *Advances in Neural Information Processing Systems*, 34:17183–17193, 2021a.

Xin Liu, Bin Li, Pengyi Shi, and Lei Ying. An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. *Advances in Neural Information Processing Systems*, 34:24075–24086, 2021b.

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

Ahmadreza Moradipari, Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Safe linear thompson sampling with side information. *IEEE Transactions on Signal Processing*, 69:3755–3767, 2021.

Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. In *International conference on artificial intelligence and statistics*, pages 2827–2835. PMLR, 2021.

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

Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with gaussian processes. In *International conference on machine learning*, pages 997–1005. PMLR, 2015.

Yanan Sui, Vincent Zhuang, Joel Burdick, and Yisong Yue. Stagewise safe bayesian optimization with gaussian processes. In *International conference on machine learning*, pages 4781–4789. PMLR, 2018.

Ilnura Usmanova, Andreas Krause, and Maryam Kamgarpour. Safe convex learning under uncertain constraints. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2106–2114. PMLR, 2019.

Michal Valko, Rémi Munos, Branislav Kveton, and Tomáš Kocák. Spectral bandits for smooth graph functions. In *International Conference on Machine Learning*, pages 46–54. PMLR, 2014.

K Nithin Varma, Sahin Lale, and Anima Anandkumar. Stochastic linear bandits with unknown safety constraints and local feedback. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023.

Akifumi Wachi and Yanan Sui. Safe reinforcement learning in constrained markov decision processes. In *International Conference on Machine Learning*, pages 9797–9806. PMLR, 2020.

Runzhe Wan, Branislav Kveton, and Rui Song. Safe exploration for efficient policy evaluation and comparison. In *International Conference on Machine Learning*, pages 22491–22511. PMLR, 2022.

Zhenlin Wang, Andrew J Wagenmaker, and Kevin Jamieson. Best arm identification with safety constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 9114–9146. PMLR, 2022.

Huasen Wu, Rayadurgam Srikant, Xin Liu, and Chong Jiang. Algorithms with logarithmic or sublinear regret for constrained contextual bandits. *Advances in Neural Information Processing Systems*, 28, 2015.

Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In *International Conference on Machine Learning*, pages 1254–1262. PMLR, 2016.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/**No**/Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]
   (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**/No/Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/**Not Applicable**]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**/No/Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**/No/Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/**Not Applicable**]
   (b) The license information of the assets, if applicable. [Yes/No/**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/**Not Applicable**]
   (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

Contents

A Proof of Theorem 1

In this section, we prove the general regret bound given in Theorem 1. First, we introduce some notation. Let the event that the confidence sets hold be defined as

$$\mathcal{E}_{\text{conf}} := \{|x^\top(\hat{\theta}_t - \theta)| \leq \beta_t \|x\|_{V_t^{-1}}, |x^\top(\hat{a}_t - a)| \leq \beta_t \|x\|_{V_t^{-1}}, \forall x \in \mathbb{R}^d, \forall t \geq 1\}, \tag{8}$$

and note that $\mathbb{P}(\mathcal{E}_{\text{conf}}) \geq 1 - \delta$ by Lemma 1.

We start by giving a key lemma that lower bounds γ_t.

Lemma 2. *If Assumptions 1 and 2, and $\mathcal{E}_{\text{conf}}$ holds, then*

$$\gamma_t \geq \max\left(1 - \frac{2}{b}\beta_t \|x_t\|_{V_t^{-1}}, \nu\right)$$

for all $t \in [T]$.

Proof. We will find lower bounds individually for μ_t (line 6) and $\tilde{b}_t$ (line 5) in the following.

Lower bound on μ_t: Since $\mu_t = \max\{\mu \in [0,1] : \mu\tilde{x}_t \in \mathcal{Y}_t^p\}$, we find a lower bound on μ_t by finding an $\alpha \in \{\mu \in [0,1] : \mu\tilde{x}_t \in \mathcal{Y}_t^p\}$. Specifically, we will show that α can be chosen as

$$\alpha = \frac{b}{b + 2\beta_t \|\tilde{x}_t\|_{V_t^{-1}}}.$$

For this to be a valid choice for α, we need that (i) $\alpha \in [0,1]$, (ii) $\alpha \tilde{x}_t \in \mathcal{X}$ and (iii) $\hat{a}_t^\top(\alpha \tilde{x}_t) + \beta_t \|\alpha \tilde{x}_t\|_{V_t^{-1}} \leq b$. Point (i) follows by definition. Point (ii) holds because $\tilde{x}_t \in \mathcal{X}$, $\mathcal{X}$ is star-convex and $\alpha \in [0,1]$, so $\alpha \tilde{x}_t \in \mathcal{X}$. Then, to show point (iii), we have that

$$\begin{aligned}
\hat{a}_t^\top(\alpha \tilde{x}_t) + \beta_t \|\alpha \tilde{x}_t\|_{V_t^{-1}} &= \alpha(\hat{a}_t^\top \tilde{x}_t + \beta_t \|\tilde{x}_t\|_{V_t^{-1}}) \\
&= \alpha(\hat{a}_t^\top \tilde{x}_t - \beta_t \|\tilde{x}_t\|_{V_t^{-1}} + 2\beta_t \|\tilde{x}_t\|_{V_t^{-1}}) \\
&\leq \alpha(b + 2\beta_t \|\tilde{x}_t\|_{V_t^{-1}}) \\
&= b,
\end{aligned}$$

where the inequality uses the fact that $\tilde{x}_t \in \mathcal{Y}_t^o$ and therefore $\hat{a}_t^\top \tilde{x}_t - \beta_t \|\tilde{x}_t\|_{V_t^{-1}} \leq b$, and the last equality uses the choice of α. Therefore,

$$\mu_t \geq \alpha = \frac{b}{b + 2\beta_t \|\tilde{x}_t\|_{V_t^{-1}}}.$$

Lower bound on $\tilde{b}_t$: Recall that,

$$\tilde{b}_t = \begin{cases} \min\left(\frac{\nu}{\|\tilde{x}_t\|}, 1\right) & \text{if } \tilde{x}_t \neq \mathbf{0}, \\ 1 & \text{else.} \end{cases}$$

We consider each case separately. If $\tilde{x}_t = \mathbf{0}$, then $\tilde{b}_t = 1 \geq \nu$ given that $\nu \leq 1$ by Assumption 2. Alternatively, if $\tilde{x}_t \neq \mathbf{0}$, then

$$\tilde{b}_t = \min\left(\frac{\nu}{\|\tilde{x}_t\|}, 1\right) \geq \min(\nu, 1) = \nu, \tag{9}$$

where the inequality holds because $\tilde{x}_t \in \mathcal{X}$, and therefore, $\|\tilde{x}_t\| \leq 1$ by Assumption 1 which implies that $\frac{\nu}{\|\tilde{x}_t\|} \geq \nu$. The last equality holds because $\nu \leq 1$ by Assumption 2. Therefore, it holds that $\tilde{b}_t \geq \nu$ in either case.

Completing the proof: With the above, we have shown that

$$\gamma_t = \max\left(\tilde{b}_t, \mu_t\right) \geq \max\left(\nu, \frac{b}{b + 2\beta_t \|\tilde{x}_t\|_{V_t^{-1}}}\right) \tag{10}$$

In order to complete the proof, we need a bound on γ_t that is in terms of $\|x_t\|_{V_t^{-1}}$ instead of $\|\tilde{x}_t\|_{V_t^{-1}}$. To get this, first note that $\gamma_t \tilde{x}_t = x_t$ and $\gamma_t \geq 0$, and therefore

$$\gamma_t \|\tilde{x}_t\|_{V_t^{-1}} = \|\gamma_t \tilde{x}_t\|_{V_t^{-1}} = \|x_t\|_{V_t^{-1}}.$$

Using this, we can rearrange (10) to get that

$$\gamma_t \geq \frac{b}{b + 2\beta_t \|\tilde{x}_t\|_{V_t^{-1}}} \iff \gamma_t b + 2\beta_t \|x_t\|_{V_t^{-1}} \geq b \iff \gamma_t \geq 1 - \frac{2}{b}\beta_t \|x_t\|_{V_t^{-1}}. \tag{11}$$

Finally, combining (10) and (11), we get that

$$\gamma_t \geq \max\left(1 - \frac{2}{b}\beta_t \|x_t\|_{V_t^{-1}}, \nu\right),$$

completing the proof. □

Then, we turn our attention to the instantaneous regret. In particular, we will utilize the decomposition,

$$r_t := \theta^\top(x_* - x_t) = \underbrace{\theta^\top(x_* - \tilde{x}_t)}_{\text{Term I}} + \underbrace{\theta^\top(\tilde{x}_t - x_t)}_{\text{Term II}}.$$

Term I can be understood as the regret due to the optimistic actions, while Term II can be understood as the cost incurred by maintaining constraint satisfaction. In the following two lemmas, we bound Term I and Term II seperately.

Lemma 3. *Conditioned on $\mathcal{E}_{\text{conf}}$, it holds that*

$$\text{Term I} = \theta^\top (x_* - \tilde{x}_t) \leq \frac{2}{\nu}\beta_t \|x_t\|_{V_t^{-1}}.$$

Proof. We condition on $\mathcal{E}_{\text{conf}}$ without further reference. First, it holds for all $t \in [T]$ that

$$\theta^\top x_* = \max_{x\in\mathcal{Y}} \theta^\top x \leq \max_{x\in\mathcal{Y}} \left(\hat{\theta}_t^\top x + \beta_t\|x\|_{V_t^{-1}}\right) \leq \max_{x\in\mathcal{Y}_t^o} \left(\hat{\theta}_t^\top x + \beta_t\|x\|_{V_t^{-1}}\right) = \hat{\theta}_t^\top \tilde{x}_t + \beta_t\|\tilde{x}_t\|_{V_t^{-1}}. \tag{12}$$

Also, note that $\gamma_t \geq \nu > 0$ by Lemma 2 and therefore

$$\|\tilde{x}_t\|_{V_t^{-1}} = \left\|\frac{x_t}{\gamma_t}\right\|_{V_t^{-1}} = \frac{1}{\gamma_t}\|x_t\|_{V_t^{-1}} \leq \frac{1}{\nu}\|x_t\|_{V_t^{-1}}. \tag{13}$$

Therefore, it holds that

$$\begin{aligned}
\text{Term I} = \theta^\top x_* - \theta^\top \tilde{x}_t &\leq \hat{\theta}_t^\top \tilde{x}_t + \beta_t\|\tilde{x}_t\|_{V_t^{-1}} - \theta^\top \tilde{x}_t \\
&= (\hat{\theta}_t^\top - \theta)^\top \tilde{x}_t + \beta_t\|\tilde{x}_t\|_{V_t^{-1}} \\
&\leq 2\beta_t\|\tilde{x}_t\|_{V_t^{-1}} \\
&\leq \frac{2}{\nu}\beta_t\|x_t\|_{V_t^{-1}},
\end{aligned}$$

where the first inequality uses (12), the second inequality uses the definition of $\mathcal{E}_{\text{conf}}$ and the third inequality uses (13). ◻

Lemma 4. *Let Assumption 1 hold. Then, conditioned on $\mathcal{E}_{\text{conf}}$, it holds that*

$$\text{Term II} = \theta^\top(\tilde{x}_t - x_t) \leq \|\theta\|\frac{2\beta_t}{b}\|x_t\|_{V_t^{-1}}$$

for all $t \in [T]$.

Proof. Conditioned on $\mathcal{E}_{\text{conf}}$, it holds that

$$\begin{aligned}
\text{Term II} = \theta^\top(\tilde{x}_t - x_t) &= \theta^\top(\tilde{x}_t - \gamma_t\tilde{x}_t) \\
&= (1-\gamma_t)\theta^\top \tilde{x}_t \\
&\leq (1-\gamma_t)\|\theta\|\|\tilde{x}_t\| \\
&\leq (1-\gamma_t)\|\theta\| \\
&\leq \|\theta\|\frac{2\beta_t}{b}\|x_t\|_{V_t^{-1}},
\end{aligned}$$

where the second inequality uses the fact that $\tilde{x}_t \in \mathcal{X}$ which implies that $\|\tilde{x}_t\| \leq 1$ by Assumption 1, and the third inequality uses Lemma 2. ◻

Finally, we turn our attention to the cumulative regret. To bound this, we will need the so-called elliptic potential, which is standard in the stochastic linear bandit literature.

Lemma 5 (Lemma 11 in Abbasi-Yadkori et al. (2011))**.** *Consider a sequence $(w_k)_{k\in\mathbb{N}}$ where $w_k \in \mathbb{R}^d$ and $\|w_k\| \leq 1$ for all $k \in \mathbb{N}$. Let $W_k = \lambda I + \sum_{s=1}^{k-1} w_s w_s^\top$ for some $\lambda \geq 1$. Then, it holds that*

$$\sum_{k=1}^{K} \|w_k\|^2_{W_k^{-1}} \leq 2d\log\left(1 + \frac{K}{\lambda d}\right).$$

With this, we complete the proof of Theorem 1 in the following.

Theorem 4 (Duplicate of Theorem 1)**.** *Let Assumptions 1, 2 and 3 hold. Then, with probability at least* $1-\delta$*, the regret of ROFUL (Algorithm 1) satisfies*

$$R_T \leq 2\left(\frac{\|\theta\|}{b}+\frac{1}{\nu}\right)\beta_T\sqrt{2dT\log\left(1+\frac{T}{\lambda d}\right)},$$

and $a^\top x_t \leq b$ *for all* $t \in [T]$.

Proof. We condition on $\mathcal{E}_{\text{conf}}$ without further reference. We give the regret guarantee and then the safety guarantee in the following.

Regret guarantee: Using Lemmas 3 and 4, it holds that

$$\begin{aligned}
r_t &= \theta^\top(x_* - x_t)\\
&= \theta^\top(x_* - \tilde{x}_t) + \theta^\top(\tilde{x}_t - x_t)\\
&= \text{Term I} + \text{Term II}\\
&\leq 2\left(\frac{\|\theta\|}{b}+\frac{1}{\nu}\right)\beta_t\|x_t\|_{V_t^{-1}}.
\end{aligned}$$

We can then study the sum of the squared instantaneous regret,

$$\begin{aligned}
\sum_{t=1}^T r_t^2 &\leq \sum_{t=1}^T 4\left(\frac{\|\theta\|}{b}+\frac{1}{\nu}\right)^2\beta_t^2\|x_t\|^2_{V_t^{-1}}\\
&= 4\left(\frac{\|\theta\|}{b}+\frac{1}{\nu}\right)^2\sum_{t=1}^T\beta_t^2\|x_t\|^2_{V_t^{-1}}\\
&\leq 4\left(\frac{\|\theta\|}{b}+\frac{1}{\nu}\right)^2\beta_T^2\sum_{t=1}^T\|x_t\|^2_{V_t^{-1}}\\
&\leq 8\left(\frac{\|\theta\|}{b}+\frac{1}{\nu}\right)^2 d\beta_T^2\log\left(1+\frac{T}{\lambda d}\right),
\end{aligned}$$

where the second inequality uses the fact that β_t is monotone in t and the third inequality uses Lemma 5. Then, by Cauchy-Schwarz, it holds that

$$\begin{aligned}
R_T &= \sum_{t=1}^T r_t\\
&\leq \sqrt{T\sum_{t=1}^T r_t^2}\\
&\leq \sqrt{8T\left(\frac{\|\theta\|}{b}+\frac{1}{\nu}\right)^2 d\beta_T^2\log\left(1+\frac{T}{\lambda d}\right)}\\
&= 2\left(\frac{\|\theta\|}{b}+\frac{1}{\nu}\right)\beta_T\sqrt{2dT\log\left(1+\frac{T}{\lambda d}\right)}.
\end{aligned}$$

Safety guarantee: In order to show that $a^\top x_t \leq b$, we note that $\gamma_t = \max(\tilde{b}_t, \mu_t)$ and therefore it holds that either $\gamma_t = \tilde{b}_t$ or $\gamma_t = \mu_t$. If $\gamma_t = \tilde{b}_t$ and $\tilde{x}_t \neq \mathbf{0}$, then using the quantity $\nu = b/S_a$ as defined in Assumption 2, it holds that

$$a^\top x_t \leq \|a\|\|x_t\| \leq S_a\|x_t\| = S_a\|\tilde{b}_t\tilde{x}_t\| = S_a\tilde{b}_t\|\tilde{x}_t\| = S_a\min\left(\frac{\nu}{\|\tilde{x}_t\|},1\right)\|\tilde{x}_t\| \leq S_a\frac{\nu}{\|\tilde{x}_t\|}\|\tilde{x}_t\| = S_a\nu = b,$$

where we use Assumption 2 in the second inequality. If $\tilde{x}_t = \mathbf{0}$, then $\tilde{b}_t = 1$ and therefore $x_t = \tilde{x}_t = \mathbf{0}$ which implies that $a^\top x_t = 0 < b$.

Alternatively, if $\gamma_t = \mu_t$, it holds that

$$x_t = \mu_t \tilde{x}_t \in \mathcal{Y}_t^p \subseteq \mathcal{Y}.$$

Therefore, it holds for both cases that $a^\top x_t \leq b$ for all $t \in [T]$. □

B Proof of Theorem 2

In this section, we prove the problem-dependent analysis given in Theorem 2 and then Corollary 1. In order to do so, we first restate the definition of Δ as follows

$$\Delta := \inf_{x \in \mathcal{Y}:\ x \neq \alpha x_* \ \ \forall \alpha > 0} \theta^\top (x_* - x),$$

and then restate the definition of B_T,

$$B_T := \sum_{t=1}^{T} \mathbb{1}\{\nexists\ \alpha > 0 : x_t = \alpha x_*\}.$$

Then, we give a lemma with some useful facts.

Lemma 6. *Let Assumptions 1 and 2 hold, and let $\mathcal{E}_{\mathrm{conf}}$ hold. Also, let*

$$\zeta_t := \max\{\zeta \geq 0 : \zeta \tilde{x}_t \in \mathcal{Y}\}, \tag{14}$$

and $v_t = \zeta_t \tilde{x}_t$. Then, it follows that:

1. $\zeta_t \in [\gamma_t, 1]$
2. $\theta^\top (\tilde{x}_t - v_t) \leq \frac{2S}{b} \beta_t \|x_t\|_{V_t^{-1}}$
3. *If there exists $\alpha > 0$ such that $x_t = \alpha x_*$, then $v_t = x_*$.*
4. *If there does not exists $\alpha > 0$ such that $x_t = \alpha x_*$, then $\theta^\top (x_* - v_t) \geq \Delta$.*

Proof. We condition on $\mathcal{E}_{\mathrm{conf}}$ throughout the proof without further reference. We will first give some useful facts. In particular, it holds that,

$$\hat{\theta}_t^\top \tilde{x}_t + \beta_t \|\tilde{x}_t\|_{V_t^{-1}} \geq \theta^\top x_* > 0 \tag{15}$$

where the first inequality is from (12) and the second is Assumption 2. It follows from (15) that $\tilde{x}_t \neq \mathbf{0}$ and therefore the set $\{\zeta \geq 0 : \zeta \tilde{x}_t \in \mathcal{Y}\}$ is compact. Also, note that $\{\zeta \geq 0 : \zeta \tilde{x}_t \in \mathcal{Y}\}$ contains 0 and is therefore nonempty, so ζ_t is well-defined. Next, we prove each item individually in the following.

1: First, we show that $\zeta_t \leq 1$. If this were not the case, i.e. $\zeta_t > 1$, then there exists $\zeta' \in \{\zeta \geq 0 : \zeta \tilde{x}_t \in \mathcal{Y}\}$ such that $\zeta' > 1$. Then, from the definition of $\tilde{x}_t$ in line 4 and the fact that $u = \zeta' \tilde{x}_t$ is in $\mathcal{Y} \subseteq \mathcal{Y}_t^o$,

$$\hat{\theta}_t^\top \tilde{x}_t + \beta_t \|\tilde{x}_t\|_{V_t^{-1}} = \max_{x \in \mathcal{Y}_t^o} \left(\hat{\theta}_t^\top x + \beta_t \|x\|_{V_t^{-1}} \right) \geq \hat{\theta}_t^\top u + \beta_t \|u\|_{V_t^{-1}}. \tag{16}$$

At the same time, it follows from (15) that $\hat{\theta}_t^\top \tilde{x}_t + \beta_t \|\tilde{x}_t\|_{V_t^{-1}}$ is positive and therefore,

$$\hat{\theta}_t^\top \tilde{x}_t + \beta_t \|\tilde{x}_t\|_{V_t^{-1}} < \zeta' \left(\hat{\theta}_t^\top \tilde{x}_t + \beta_t \|\tilde{x}_t\|_{V_t^{-1}} \right) = \hat{\theta}_t^\top u + \beta_t \|u\|_{V_t^{-1}}. \tag{17}$$

Since (16) and (17) cannot simultaneously be true, it must hold that $\zeta_t \leq 1$.

Then, we show that $\zeta_t \geq \gamma_t$. Since $x_t = \gamma_t \tilde{x}_t \in \mathcal{Y}$, it holds that $\gamma_t \in \{\zeta \geq 0 : \zeta \tilde{x}_t \in \mathcal{Y}\}$ and therefore $\zeta_t \geq \gamma_t$.

2: Since, $v_t = \zeta_t \tilde{x}_t$ and $\zeta_t \in [\gamma_t, 1]$, it follows from Lemma 2 that

$$\theta^\top (\tilde{x}_t - v_t) = \theta^\top \tilde{x}_t (1 - \zeta_t) \leq S(1 - \zeta_t) \leq S(1 - \gamma_t) \leq \frac{2S}{b} \beta_t \|x_t\|_{V_t^{-1}}.$$

3: First, we will show that $\zeta_* = \max\{\zeta \geq 0 : \zeta x_* \in \mathcal{Y}\} = 1$. If this were not the case, then either $\zeta_* < 1$ or $\zeta_* > 1$. The case $\zeta_* < 1$ would imply that x_* is not in $\mathcal{Y}$, while the case that $\zeta_* > 1$ would imply the existence of a point $x = \zeta x_* \in \mathcal{Y}$ with $\zeta > 1$ such that $\theta^\top x = \zeta(\theta^\top x_*) > \theta^\top x_*$ (where we use $\theta^\top x_* > 0$ from Assumption 2). Either case contradicts the definition of x_* and therefore cannot hold.

Now, we turn to the statement. If there exists $\alpha > 0$ such that $x_t = \alpha x_*$, then,

$$\zeta_t = \max\{\zeta \geq 0 : \zeta \tilde{x}_t \in \mathcal{Y}\} = \gamma_t \max\{\zeta' \geq 0 : \zeta' x_t \in \mathcal{Y}\} = \frac{\gamma_t}{\alpha} \max\{\tilde{\zeta} \geq 0 : \tilde{\zeta} x_* \in \mathcal{Y}\} = \frac{\gamma_t}{\alpha},$$

where we use the mapping $\zeta' = \frac{1}{\gamma_t}\zeta$ in the first equality and $\tilde{\zeta} = \alpha\zeta'$ in the second equality. Therefore, it follows that

$$v_t = \zeta_t \tilde{x}_t = \frac{\zeta_t}{\gamma_t} x_t = \frac{\alpha \zeta_t}{\gamma_t} x_* = x_*.$$

4: First, note that if there does not exist $\alpha > 0$ such that $x_t = \alpha x_*$, then there does not exist $\alpha' > 0$ such that $v_t = \alpha' x_*$ as $v_t = \frac{\zeta_t}{\gamma_t} x_t$. Then, since $v_t \in \mathcal{Y}$, it follows from the definition of Δ that,

$$\Delta = \inf_{x \in \mathcal{Y}:\ x \neq \alpha x_*\ \ \forall \alpha > 0} \theta^\top (x_* - x) \leq \theta^\top (x_* - v_t).$$

□

Then, we restate and prove Theorem 2.

Theorem 5 (Duplicate of Theorem 2)**.** *Let Assumptions 1, 2 and 3 hold. If $\Delta > 0$, then the number of wrong directions chosen by ROFUL (Algorithm 1) satisfies*

$$B_T \leq \frac{32 d \beta_T^2}{\nu^2 \Delta^2} \log\left(1 + \frac{T}{\lambda d}\right)$$

with probability at least $1 - \delta$.

Proof. In order to bound the number of times that the wrong action is chosen, we study the regret due to the wrong choice of direction,

$$\tilde{R}_T := \sum_{t=1}^{T} \theta^\top (x_* - v_t),$$

where $v_t = \zeta_t \tilde{x}_t$ with ζ_t defined in (14). We will denote the instantaneous regret due to the wrong choice of direction as $\tilde{r}_t = \theta^\top (x_* - v_t)$. It follows from Lemma 6 (#3 and #4) that $\tilde{r}_t = 0$ if there exists $\alpha > 0$ such that $x_t = \alpha x_*$ (i.e. x_t is in the correct direction) and $\tilde{r}_t \geq \Delta$ otherwise. Therefore,

$$\begin{aligned}
\tilde{R}_T = \sum_{t=1}^{T} \tilde{r}_t &= \sum_{t=1}^{T} \tilde{r}_t \mathbb{1}\{\nexists\ \alpha > 0 : x_t = \alpha x_*\} \\
&\geq \Delta \sum_{t=1}^{T} \mathbb{1}\{\nexists\ \alpha > 0 : x_t = \alpha x_*\} \\
&= \Delta B_T.
\end{aligned}$$

Since $\tilde{R}_T \geq \Delta B_T$ and $\Delta > 0$, an upper bound on $\tilde{R}_T$ implies an upper bound on B_T.

Then, we bound $\tilde{r}_t$ in the following,

$$\begin{aligned}
\tilde{r}_t &= \theta^\top x_* - \theta^\top v_t \\
&\leq \hat{\theta}_t^\top \tilde{x}_t + \beta_t \|\tilde{x}_t\|_{V_t^{-1}} - \theta^\top v_t && \text{(a)} \\
&= \theta^\top \tilde{x}_t + (\hat{\theta}_t - \theta)^\top \tilde{x}_t + \beta_t \|\tilde{x}_t\|_{V_t^{-1}} - \theta^\top v_t \\
&\leq \theta^\top \tilde{x}_t + 2\beta_t \|\tilde{x}_t\|_{V_t^{-1}} - \theta^\top v_t && \text{(b)}
\end{aligned}$$

$$
\begin{aligned}
&= \theta^\top(\tilde{x}_t - v_t) + 2\beta_t\|\tilde{x}_t\|_{V_t^{-1}} \\
&\leq \frac{2S}{b}\beta_t\|x_t\|_{V_t^{-1}} + 2\beta_t\|\tilde{x}_t\|_{V_t^{-1}} && \text{(c)} \\
&\leq \frac{2S}{b}\beta_t\|x_t\|_{V_t^{-1}} + \frac{2}{\nu}\beta_t\|x_t\|_{V_t^{-1}} && \text{(d)} \\
&= \frac{4S}{b}\beta_t\|x_t\|_{V_t^{-1}},
\end{aligned}
$$

where (a) follows from the fact that $\tilde{x}_t$ is optimistic (i.e. (12)), (b) is from the definition of the confidence set, (c) is from Lemma 6 (#2), and (d) is due to (13).

Since either $\tilde{r}_t \geq \Delta$ or $\tilde{r}_t = 0$, it holds that $\tilde{r}_t \leq (\tilde{r}_t)^2/\Delta$. Then, we have that

$$
\begin{aligned}
\tilde{R}_T &= \sum_{t=1}^{T} \tilde{r}_t \\
&\leq \sum_{t=1}^{T} \frac{(\tilde{r}_t)^2}{\Delta} \\
&\leq \frac{16S^2\beta_T^2}{b^2\Delta} \sum_{t=1}^{T} \|x_t\|^2_{V_t^{-1}} \\
&\leq \frac{32S^2\beta_T^2 d}{b^2\Delta} \log\left(1 + \frac{T}{\lambda d}\right)
\end{aligned}
$$

where the last inequality uses Lemma 5. Finally, we have that

$$
B_T \leq \frac{\tilde{R}_T}{\Delta} \leq \frac{32S^2\beta_T^2 d}{b^2\Delta^2} \log\left(1 + \frac{T}{\lambda d}\right).
$$

□

Now, we turn out attention to Corollary 1. To do so, we state PD-ROFUL more formally in Algorithm 2. Note that in the second phase of the algorithm, we reduce the problem to a 1-dimensional safe linear bandit problem which is defined formally in the following.

Definition 1 (Reduction to 1-dimensional problem)**.** *Given a direction $u_* \in \mathbb{S}$, the safe linear bandit problem (Section 2) reduces to a 1-dimensional setting. For each round t of this setting, the learner chooses $\xi_t \in \mathbb{R}_+$ and then $\xi_t u_*$ is played in the original setting (Section 2).*

Using this, we give psuedo-code for PD-ROFUL.

Corollary 2 (Duplicate of Corollary 1)**.** *Let Assumptions 1, 3 and 2 hold. With $\Delta > 0$, the regret of PD-ROFUL (Algorithm 2) satisfies*

$$
R_T \leq \frac{4S}{b}\beta_{2\bar{B}+1}\sqrt{2d(2\bar{B}+1)\log\left(1 + \frac{2\bar{B}+1}{\lambda d}\right)} + \frac{4S}{b}\tilde{\beta}_T\sqrt{2T\log\left(1 + \frac{T}{\lambda d}\right)}
$$

with probability at least $1 - 2\delta$, where $\tilde{\beta}_T$ is β_T with $d = 1$.

Proof. We condition on the confidence sets holding jointly for both the first and second phases, which occurs with probability at least $1 - 2\delta$.

First, we argue that the optimal direction is correctly identified, i.e. $u_* = x_*/\|x_*\|$. Intuitively, this holds because Theorem 2 says that wrong directions are chosen at most $\bar{B}$ times, so any single direction that is chosen more than this must be the optimal direction. Concretely, Theorem 2 implies that for every single wrong direction $u \in \mathbb{S}, u \neq x_*/\|x_*\|$, the actions chosen by ROFUL in the first $\tau - 1$ rounds will satisfy

$$
\sum_{t=1}^{\tau-1} \mathbb{1}\{u_t = u\} \leq \sum_{t=1}^{\tau-1} \mathbb{1}\{\nexists\ \alpha > 0 : x_t = \alpha x_*\} \leq B_T \leq \bar{B},
$$

Algorithm 2: Problem-Dependent ROFUL (PD-ROFUL)

Set $\mathcal{A} = \{\}$, $\bar{N} = 0$ and $t = 1$.
while $\bar{N} \leq \bar{B} = \frac{32S^2 d}{b^2\Delta^2}\beta_T^2 \log\left(1 + \frac{T}{\lambda d}\right)$ **do**
  ROFUL plays x_t and observes y_t, z_t.
  if $x_t \neq \mathbf{0}$ **and** $u_t = x_t/\|x_t\| \notin \mathcal{A}$ **then**
    $\mathcal{A} = \mathcal{A} \cup \{u_t\}$.
    $N_{u_t} = 1$.
  else if $x_t \neq \mathbf{0}$ **and** $u_t = x_t/\|x_t\| \in \mathcal{A}$ **then**
    $N_{u_t} = N_{u_t} + 1$
  Set $\bar{N} = \max_{u \in \mathcal{A}} N_u$.
  Set $t = t + 1$.
end
Set $u_* = \arg\max_{u \in \mathcal{A}} N_u$ and $\tau = t$.
For $t \in [\tau + 1, T]$, ROFUL is restarted and plays 1-dimensional setting (Definition 1) in direction u_* for remaining rounds.

where we use the notation $u_t = x_t/\|x_t\|$ for nonzero x_t as specified in Algorithm 2. Then, since u_* is specified in PD-ROFUL to be a direction that ROFUL plays more than $\bar{B}$ times in $\tau - 1$ rounds,

$$\sum_{t=1}^{\tau-1} \mathbb{1}\{u_t = u_*\} > \bar{B},$$

and therefore u_* cannot be a wrong direction or equivalently, $u_* = x_*/\|x_*\|$.

Then, we show the bound on the duration of the first phase $\tau - 1 \leq 2\bar{B} + 1$. The intuition is that u_* is played at most $\bar{B} + 1$ times and wrong directions are played at most $\bar{B}$ times, so the total duration of the first phase must be less than $2\bar{B} + 1$. Concretely, it follows from the fact that $u_* = x_*/\|x_*\|$,

$$\begin{aligned}\tau - 1 &= \sum_{t=1}^{\tau-1} \mathbb{1}\{\exists\ \alpha > 0 : x_t = \alpha x_*\} + \sum_{t=1}^{\tau-1} \mathbb{1}\{\nexists\ \alpha > 0 : x_t = \alpha x_*\} \\ &= \underbrace{\sum_{t=1}^{\tau-1} \mathbb{1}\{u_t = u_*\}}_{\leq \bar{B}+1} + \underbrace{\sum_{t=1}^{\tau-1} \mathbb{1}\{\nexists\ \alpha > 0 : x_t = \alpha x_*\}}_{\leq \bar{B}} \leq 2\bar{B} + 1,\end{aligned}$$

where the bound on the first term holds because the first round ends when the correct direction is played more than $\bar{B}$ times and the bound on the second term follows from Theorem 2.

Finally, we study the regret. To do so, we decompose the regret due to the first and second phases respectively,

$$R_T = \underbrace{\sum_{t=1}^{\tau-1} r_t}_{R_T^{\mathrm{I}}} + \underbrace{\sum_{t=\tau}^{T} r_t}_{R_T^{\mathrm{II}}}.$$

Then, we use Theorem 1 and the fact that $\tau - 1 \leq 2\bar{B} + 1$ to bound R_T^{I},

$$R_T^{\mathrm{I}} = \sum_{t=1}^{\tau-1} r_t \leq \frac{4S}{b}\beta_{\tau-1}\sqrt{2d(\tau-1)\log\left(1 + \frac{\tau - 1}{\lambda d}\right)} \leq \frac{4S}{b}\beta_{2\bar{B}+1}\sqrt{2d(2\bar{B}+1)\log\left(1 + \frac{2\bar{B}+1}{\lambda d}\right)},$$

For the remainder of the rounds, play of the algorithm is ROFUL with $d = 1$. Also, the duration of the second phase is less than the time horizon, i.e. $T - (\tau - 1) \leq T$. Therefore, R_T^{II} is less than that given in Theorem 1 with $d = 1$. □

Algorithm 3: Safe Phased Elimination (Safe-PE)

Input: $\mathcal{X}, S, b, d, \rho, \delta \in (0,1), \lambda \geq 1$

1 Set $\mathcal{A}_1 = [k]$, $\zeta_{i,1} = \min(b/S, \alpha_i)\ \forall i \in [k]$, $\mathcal{Y}_1^p = \{\zeta_{i,1} u_i\}_{i\in[k]}$, $J = \lceil \log_2(T+1) \rceil$, $t_j = 2^{j-1}$.

2 **for** $j = 1$ **to** J **do**

3     $V_{t_j} = \lambda I$.

4     $\tau_j = \min(t_{j+1} - 1, T)$.

5   **for** $t = t_j$ **to** τ_j **do**

6   Play $x_t \in \arg\max_{x \in \mathcal{Y}_j^p} \|x\|_{V_t^{-1}}$.

7         $V_{t+1} = V_t + x_t x_t^\top$.

8   **end**

9   Set $\hat{\theta}_j = \bar{V}_j^{-1} \sum_{s=t_j}^{\tau_j} x_s y_s$ and $\hat{a}_j = \bar{V}_j^{-1} \sum_{s=t_j}^{\tau_j} x_s z_s$, where $\bar{V}_j = V_{\tau_j+1}$.

10   Find $\hat{x}_j \in \arg\max_{x\in\mathcal{Y}_j^p} \left(\hat{\theta}_j^\top x - \beta \|x\|_{\bar{V}_j^{-1}} \right)$.

11   Update $\mathcal{A}_{j+1} = \left\{ i \in \mathcal{A}_j : \hat{\theta}_j^\top (\hat{x}_j - \zeta_{i,j} u_i) \leq \beta \|\hat{x}_j\|_{\bar{V}_j^{-1}} + \beta \zeta_{i,j} \|u_i\|_{\bar{V}_j^{-1}} + \frac{2S\beta\zeta_{i,j}\|u_i\|_{\bar{V}_{j-1}^{-1}}}{b} \right\}$.

12   Update $\mu_{i,j+1} := \max\left\{ \alpha \in [0, \alpha_i] : \alpha \left(u_i^\top \hat{a}_j + \beta \|u_i\|_{\bar{V}_j^{-1}} \right) \leq b \right\}$ and $\zeta_{i,j+1} = \max(\zeta_{i,j}, \mu_{i,j+1})$ for all $i \in \mathcal{A}_{j+1}$.

13   Update $\mathcal{Y}_{j+1}^p = \{\zeta_{i,j+1} u_i\}_{i \in \mathcal{A}_{j+1}}$.

14 **end**

C Details of Safe-PE Algorithm

In this section, we give the details of the Safe-PE algorithm (Algorithm 3) discussed in Section 4. This algorithm relies on the action set being a finite star convex set, which we formally assume in the following.

Assumption 5. *The action set satisfies*

$$\mathcal{X} := \bigcup_{i\in[k]} \{\alpha u_i : \alpha \in [0, \alpha_i]\},$$

where $u_1, ..., u_k \in \mathbb{S}$ are unit vectors and $\alpha_1, ..., \alpha_k \in \mathbb{R}_{++}$ are the maximum scalings for each unit vectors.

The Safe-PE algorithm builds on SpectralEliminator from Valko et al. (2014) and Kocák et al. (2020). It differs in that it eliminates directions in each phase rather than distinct actions and only plays actions from a verifiably safe set.

C.1 Operation of algorithm

The Safe-PE algorithm consists of phases $j = 1, 2, ...$, which are each of duration 2^{j-1}. Throughout its operation, the algorithm maintains a set of direction indexes $\mathcal{A}_j$ and safe actions $\mathcal{Y}_j^p$. The key parts of each phase are:

1. For 2^{j-1} rounds, chooses the action in $\mathcal{Y}_j^p$ with the largest confidence set width $\|\cdot\|_{V_t^{-1}}$ (line 6).
2. Eliminates directions from $\mathcal{A}_j$ with too low of estimated reward (line 11).
3. Updates $\mathcal{Y}_j^p$ with the maximum scaling of each direction that is verifiably safe (lines 12 and 13).

The algorithm relies on a confidence set to determine which directions should be eliminated and to ensure that the constraints are not violated. Different from the confidence set in Lemma 1, the radius of the following confidence set does not grow with d. We prove such a confidence set in the following lemma.

Lemma 7. *Let Assumptions 4 and 5 hold and fix some $\delta \in (0,1)$. Then, for all $x \in \mathcal{X}$ and all $j \in [J]$ it holds that $|x^\top(\hat{\theta}_j - \theta)| \leq \beta \|x\|_{\bar{V}_j^{-1}}$ and $|x^\top(\hat{a}_j - a)| \leq \beta \|x\|_{\bar{V}_j^{-1}}$ where $\beta = \rho\sqrt{2\log\left(\frac{4kJ}{\delta}\right)} + \sqrt{\lambda}S$ with probability at least $1 - \delta$.*

Proof. First, we find a confidence set that applies for a fixed $u \in \{u_1, ..., u_k\}$. To do so, we start as

$$
\begin{aligned}
u^\top(\hat{\theta}_j - \theta) &= u^\top \left(\bar{V}_j^{-1} \sum_{t=t_j}^{\tau_j} x_t y_t - \theta \right) \\
&= u^\top \left(\bar{V}_j^{-1} \sum_{t=t_j}^{\tau_j} x_t (x_t^\top \theta + \epsilon_t) - \theta \right) \\
&= \underbrace{u^\top \left(\bar{V}_j^{-1} \sum_{t=t_j}^{\tau_j} x_t x_t^\top - I \right) \theta}_{\text{Term I}} + \underbrace{u^\top \bar{V}_j^{-1} \sum_{t=t_j}^{\tau_j} x_t \epsilon_t}_{\text{Term II}}
\end{aligned}
$$

Using the notation $\tilde{V} = \sum_{t=t_j}^{\tau_j} x_s x_s^\top$, we study the first term,

$$
\begin{aligned}
|\text{Term I}| &= \left| u^\top \left(\bar{V}_j^{-1} \tilde{V} - I \right) \theta \right| \\
&= \left| u^\top \bar{V}_j^{-1} \left(\tilde{V} - \bar{V}_j \right) \theta \right| \\
&= \lambda \left| u^\top \bar{V}_j^{-1} \theta \right| && \text{(a)} \\
&\leq \lambda \|\theta\| \|u^\top \bar{V}_j^{-1}\| \\
&= \lambda \|\theta\| \sqrt{u^\top \bar{V}_j^{-1} \bar{V}_j^{-1} u} \\
&\leq \sqrt{\lambda} S \|u\|_{\bar{V}_j^{-1}}, && \text{(b)}
\end{aligned}
$$

where (a) is due to the fact that $\tilde{V} - \bar{V}_j = -\lambda I$, and (b) is due to the fact that $\bar{V}_j \succeq \lambda I$ and therefore $y^\top \bar{V}_j^{-1} y \leq \|y\|^2/\lambda$ for any $y \in \mathbb{R}^d$.

Now, we look at the second term, which can be written as

$$
\text{Term II} = \sum_{t=t_j}^{\tau_j} \left(u^\top \bar{V}_j^{-1} x_t \epsilon_t \right).
$$

Since the noise terms ϵ_t are independent and the actions x_t within each phase (i.e. $t \in [t_j, \tau_j]$) are deterministic given the history at the beginning of the phase, we can use standard concentration of subgaussian random variables (e.g. as in equation 20.2 of Lattimore and Szepesvári (2020)) to get that

$$
|\text{Term II}| > \rho \sqrt{2 \log\left(\frac{2}{\delta'}\right) \sum_{t=t_j}^{\tau_j} (u^\top \bar{V}_j^{-1} x_s)^2}
$$

with probability at least $1 - \delta'$ for some $\delta' \in (0, 1)$. Also, since $\tilde{V} \preceq \bar{V}_j$,

$$
\sum_{t=t_j}^{\tau_j} (u^\top \bar{V}_j^{-1} x_t)^2 = u^\top \bar{V}_j^{-1} \left(\sum_{t=t_j}^{\tau_j} x_t x_t^\top \right) \bar{V}_j^{-1} u = u^\top \bar{V}_j^{-1} \tilde{V} \bar{V}_j^{-1} u \leq u^\top \bar{V}_j^{-1} \bar{V}_j \bar{V}_j^{-1} u = \|u\|^2_{\bar{V}_j^{-1}}.
$$

Therefore, with probability at least $1 - \delta'$, it holds that

$$
|u^\top(\hat{\theta}_j - \theta)| \leq |\text{Term I}| + |\text{Term II}| \leq \left(\rho \sqrt{2 \log\left(\frac{2}{\delta'}\right)} + \sqrt{\lambda} S \right) \|u\|_{\bar{V}_j^{-1}}.
$$

Note that the same applies replacing θ with a. Then, by taking

$$
\beta = \rho \sqrt{2 \log\left(\frac{4kJ}{\delta}\right)} + \sqrt{\lambda} S,
$$

it holds that $|u^\top(\hat{\theta}_j - \theta)| \leq \beta \|u\|_{\bar{V}_j^{-1}}$ and $|u^\top(\hat{a}_j - a)| \leq \beta \|u\|_{\bar{V}_j^{-1}}$ for all $u \in \{u_1, ..., u_k\}$ and all $j \in [J]$ with probability at least $1-\delta$. Since all $x \in \mathcal{X}$ can be written as αu for some $\alpha \geq 0$ and $u \in \{u_1, ..., u_k\}$, it holds under the same conditions that

$$|x^\top(\hat{\theta}_j - \theta)| = \alpha |u^\top(\hat{\theta}_j - \theta)| \leq \alpha\beta \|u\|_{\bar{V}_j^{-1}} = \beta \|x\|_{\bar{V}_j^{-1}}.$$

□

We define the event that this confidence set holds as

$$\mathcal{E}_{\text{conf}} := \left\{ |x^\top(\hat{\theta}_j - \theta)| \leq \|x\|_{\bar{V}_j^{-1}}\beta, |x^\top(\hat{a}_j - a)| \leq \|x\|_{\bar{V}_j^{-1}}\beta \quad \forall x \in \mathcal{X} \; \forall j \in [J] \right\},$$

which occurs with probability at least $1-\delta$.

C.2 Regret analysis

Now, we will prove the regret bound for Safe-PE. In order to do so, we need some more notation. The true maximum scaling for each direction is denoted by

$$\bar{\zeta}_i := \max \left\{ \alpha \in [0, \alpha_i] : \alpha u_i^\top a \leq b \right\}. \tag{18}$$

Also, let $v_{i,j} := \zeta_{i,j} u_i$ and $\bar{v}_i := \bar{\zeta}_i u_i$. The index of the direction played at round t and the optimal direction are denoted by i_t and i_*, respectively. When i_t or i_* are used in a subscript, the shorthand t and $*$ are used. With this, we prove a bound on the safe scalings.

Lemma 8. *Let Assumptions 1, 2 and 5 hold and assume that $\mathcal{E}_{\text{conf}}$ holds. For all $i \in [k]$ and $j \geq 1$, it holds that*

$$\frac{\zeta_{i,j}}{\bar{\zeta}_i} \geq 1 - \frac{2}{b} \|v_{i,j}\|_{\bar{V}_{j-1}^{-1}} \beta$$

and furthermore that

$$\frac{\zeta_{i,j}}{\bar{\zeta}_i} \geq 1 - \frac{2}{b^2} S \|v_{i,j-1}\|_{\bar{V}_{j-1}^{-1}} \beta.$$

Proof. We condition on $\mathcal{E}_{\text{conf}}$ throughout the proof without further reference. We aim to find a $\gamma \geq 0$ such that $v_{i,j} = \gamma \bar{v}_i$. First, we show that $\mu \bar{v}_i$ is in $\mathcal{Y}_j^p$, where $\mu = \frac{b}{2\|\bar{v}_i\|_{\bar{V}_{j-1}^{-1}}\beta + b}$. This holds because,

$$\begin{aligned}
\hat{a}_{j-1}^\top(\mu \bar{v}_i) + \beta \|\mu \bar{v}_i\|_{\bar{V}_{j-1}^{-1}} &= \mu \left(\hat{a}_{j-1}^\top \bar{v}_i + \beta \|\bar{v}_i\|_{\bar{V}_{j-1}^{-1}} \right) \\
&= \mu \left(a^\top \bar{v}_i + (\hat{a}_{j-1}^\top - a)^\top \bar{v}_i + \beta \|\bar{v}_i\|_{\bar{V}_{j-1}^{-1}} \right) \\
&\leq \mu \left(a^\top \bar{v}_i + 2\beta \|\bar{v}_i\|_{\bar{V}_{j-1}^{-1}} \right) \\
&\leq \mu \left(b + 2\beta \|\bar{v}_i\|_{\bar{V}_{j-1}^{-1}} \right) = b
\end{aligned}$$

where the first inequality is from the definition of $\mathcal{E}_{\text{conf}}$ and the second inequality is due to the fact that $\bar{v}_i$ satisfies the constraints by definition. It follows that $\gamma \geq \frac{b}{2\|\bar{v}_i\|_{\bar{V}_{j-1}^{-1}}\beta + b}$. Then, using the fact that $\gamma \|\bar{v}_i\|_{\bar{V}_{j-1}^{-1}} = \|\gamma \bar{v}_i\|_{\bar{V}_{j-1}^{-1}} = \|v_{i,j}\|_{\bar{V}_{j-1}^{-1}}$, we can rearrange this to get that

$$\frac{\zeta_{i,j}}{\bar{\zeta}_i} = \gamma \geq 1 - \frac{2\|v_{i,j}\|_{\bar{V}_{j-1}^{-1}}\beta}{b}$$

This is the first claim.

Also, we know from the definition of $\zeta_{i,j+1}$ that

$$\alpha_i \geq \zeta_{i,j+1} \geq \zeta_{i,j} \geq \zeta_{i,0} = \min\left(\frac{b}{S}, \alpha_i\right),$$

and therefore

$$\frac{\zeta_{i,j}}{\zeta_{i,j-1}} \leq \frac{\alpha_i}{\zeta_{i,0}} = \frac{\alpha_i}{\min\left(\frac{b}{S}, \alpha_i\right)} = \frac{1}{\min\left(\frac{b}{S\alpha_i}, 1\right)} \leq \frac{1}{\min\left(\frac{b}{S}, 1\right)} = \frac{S}{b}, \tag{19}$$

where we use $\frac{b}{S} \leq 1$ from Assumption 2. It follows that

$$\frac{\zeta_{i,j}}{\bar{\zeta}_i} \geq 1 - \frac{2\|v_{i,j}\|_{\bar{V}_{j-1}^{-1}}\beta}{b} = 1 - \frac{2(\zeta_{i,j}/\zeta_{i,j-1})\|v_{i,j-1}\|_{\bar{V}_{j-1}^{-1}}\beta}{b} \geq 1 - \frac{2S\|v_{i,j-1}\|_{\bar{V}_{j-1}^{-1}}\beta}{b^2},$$

giving the second claim. □

Next, we show that the optimal action is never eliminated with high probability.

Lemma 9. *Let Assumptions 1, 2 and 5 hold and suppose that $\mathcal{E}_{\text{conf}}$ holds. It follows that i_* is in $\mathcal{A}_j$ for all $j \in [J]$.*

Proof. First note that, given Lemma 8, it holds that

$$\begin{aligned}
\theta^\top(x_* - v_{*,j}) &= \theta^\top u_*(\bar{\zeta}_* - \zeta_{*,j}) \\
&= \bar{\zeta}_*\theta^\top u_*(1 - \zeta_{*,j}/\bar{\zeta}_*) \\
&\leq \bar{\zeta}_* S(1 - \zeta_{*,j}/\bar{\zeta}_*) \\
&\leq \frac{2S\|v_{*,j}\|_{\bar{V}_{j-1}^{-1}}\beta}{b}
\end{aligned} \tag{20}$$

Then, conditioning on $\mathcal{E}_{\text{conf}}$, we have for all $j \in [J]$ that

$$\begin{aligned}
\hat{\theta}_j^\top(\hat{x}_j - v_{*,j}) &= \hat{x}_j^\top(\hat{\theta}_j - \theta) + \theta^\top\hat{x}_j - \theta^\top v_{*,j} + v_{*,j}^\top(\theta - \hat{\theta}_j) \\
&\leq \hat{x}_j^\top(\hat{\theta}_j - \theta) + \theta^\top(x_* - v_{*,j}) + v_{*,j}^\top(\theta - \hat{\theta}_j) \\
&\leq \beta\|\hat{x}_j\|_{\bar{V}_j^{-1}} + \frac{2S\|v_{*,j}\|_{\bar{V}_{j-1}^{-1}}\beta}{b} + \beta\|v_{*,j}\|_{\bar{V}_j^{-1}},
\end{aligned} \tag{21}$$

where the first inequality comes from the optimality of x_* (i.e. $\theta^\top\hat{x}_j \leq \theta^\top x_*$) and the second inequality applies the confidence set to the first and third terms and (20) to the second term. Note that (21) is exactly the condition for directions to be retained by the algorithm in line 11. Therefore, if i_* is in $\mathcal{A}_j$ for some j, then it will be in $\mathcal{A}_{j+1}$. Then, since $i_* \in \mathcal{A}_1$, it holds that $i_* \in \mathcal{A}_j$ for all $j \in [J]$ by induction. □

Next, we relate the actions in $\mathcal{Y}_j^p$ to the chosen actions.

Lemma 10. *For all $j \in [J-1]$, it holds that*

$$w_j := \max_{x \in \mathcal{Y}_j^p} \|x\|_{\bar{V}_j^{-1}} \leq \frac{1}{t_{j+1} - t_j} \sum_{t=t_j}^{t_{j+1}-1} \|x_t\|_{V_t^{-1}}. \tag{22}$$

Proof. This proof essentially follows from Lemma 39 in Kocák et al. (2020) but we give it for completeness. Note that for any $t \in [t_j, \tau_j]$, it holds that $\bar{V}_j = \sum_{s=t+1}^{\tau_j} x_s x_s^\top + V_t \succeq V_t \succ 0$ and therefore that $\|x\|_{\bar{V}_j^{-1}} \leq \|x\|_{V_t^{-1}}$ for any x. Also, since $j \leq J-1$, it holds that $\tau_j = t_{j+1} - 1$. It follows for any $x \in \mathcal{Y}_j^p$ that

$$\begin{aligned}
(t_{j+1} - t_j)\|x\|_{\bar{V}_j^{-1}} &\leq \sum_{t=t_j}^{t_{j+1}-1} \|x\|_{V_t^{-1}} \\
&\leq \sum_{t=t_j}^{t_{j+1}-1} \max_{x \in \mathcal{Y}_j^p} \|x\|_{V_t^{-1}} \\
&= \sum_{t=t_j}^{t_{j+1}-1} \|x_t\|_{V_t^{-1}}.
\end{aligned}$$

□

Lastly, we put everything together and prove the complete regret bound for the Safe-PE in Theorem 6, which shows that the regret is $\mathcal{O}(\frac{1}{b^2}\sqrt{dT\log(T)\log(k\log(T))})$.

Theorem 6 (Complete version of Theorem 3)**.** *Let Assumptions 1, 2, 4 and 5 hold. Then, the regret of Safe-PE (Algorithm 3) satisfies*

$$R_T \leq 6S + 5\beta\left(\frac{24S^2}{b^2} + 10\right)\sqrt{2dT\log\left(1 + \frac{T}{\lambda d}\right)}$$

with probability at least $1-\delta$*.*

Proof. Without further reference, we condition on $\mathcal{E}_{\text{conf}}$ throughout. We decompose the instantaneous regret for $t \in [t_j, \tau_j]$ as

$$r_t = \theta^\top x_* - \theta^\top x_t = \underbrace{\theta^\top x_* - \theta^\top v_{*,j-1}}_{\text{Term I}} + \underbrace{\theta^\top v_{*,j-1} - \theta^\top v_{t,j-1}}_{\text{Term II}} + \underbrace{\theta^\top v_{t,j-1} - \theta^\top x_t}_{\text{Term III}}.$$

We study each of the terms individually in the following.

Term I: It follows from Lemma 8 that

$$\text{Term I} = \theta^\top(x_* - v_{*,j-1}) = \bar{\zeta}_*\theta^\top u_*(1 - \zeta_{*,j-1}/\bar{\zeta}_*) \leq \frac{2S^2\beta}{b^2}\|v_{*,j-2}\|_{\bar{V}_{j-2}^{-1}}.$$

Term II: We further decompose Term II as

$$\begin{aligned}\text{Term II} &= \theta^\top v_{*,j-1} - \theta^\top v_{t,j-1} \\ &= \underbrace{\hat{\theta}_{j-1}^\top v_{*,j-1} - \hat{\theta}_{j-1}^\top v_{t,j-1} - \beta\|v_{*,j-1}\|_{\bar{V}_{j-1}^{-1}}}_{\text{Term II.A}} + \underbrace{(\theta - \hat{\theta}_{j-1})^\top v_{*,j-1} + (\hat{\theta}_{j-1} - \theta)^\top v_{t,j-1} + \beta\|v_{*,j-1}\|_{\bar{V}_{j-1}^{-1}}}_{\text{Term II.B}}\end{aligned}$$

Then, we have that

$$\begin{aligned}\text{Term II.A} &= \hat{\theta}_{j-1}^\top v_{*,j-1} - \hat{\theta}_{j-1}^\top v_{t,j-1} - \beta\|v_{*,j-1}\|_{\bar{V}_{j-1}^{-1}} \\ &\leq \hat{\theta}_{j-1}^\top \hat{x}_{j-1} - \hat{\theta}_{j-1}^\top v_{t,j-1} - \beta\|\hat{x}_{j-1}\|_{\bar{V}_{j-1}^{-1}} \\ &\leq 2\beta\|v_{t,j-1}\|_{\bar{V}_{j-1}^{-1}} + \frac{2S\beta}{b}\|v_{t,j-1}\|_{\bar{V}_{j-2}^{-1}} \\ &\leq 2\beta\|v_{t,j-1}\|_{\bar{V}_{j-1}^{-1}} + \frac{2S^2\beta}{b^2}\|v_{t,j-2}\|_{\bar{V}_{j-2}^{-1}}\end{aligned}$$

where the first inequality follows from the definition of $\hat{x}_j$ in line 10 of Algorithm 3 given that $v_{*,j-1} \in \mathcal{Y}_{j-1}^p$, the second inequality comes from the fact that $i_t \in \mathcal{A}_j$ and therefore satisfied the criteria in line 11, and the third inequality comes from (19).

Also, we have that

$$\begin{aligned}\text{Term II.B} &= (\theta - \hat{\theta}_{j-1})^\top v_{*,j-1} + (\theta - \hat{\theta}_{j-1})^\top v_{t,j-1} + \beta\|v_{*,j-1}\|_{\bar{V}_{j-1}^{-1}} \\ &\leq 2\beta\|v_{*,j-1}\|_{\bar{V}_{j-1}^{-1}} + \beta\|v_{t,j-1}\|_{\bar{V}_{j-1}^{-1}}\end{aligned}$$

where the inequality is from the definition of $\mathcal{E}_{\text{conf}}$.

Putting everything together, we have that

$$\text{Term II} \leq 3\beta\|v_{t,j-1}\|_{\bar{V}_{j-1}^{-1}} + 2\beta\|v_{*,j-1}\|_{\bar{V}_{j-1}^{-1}} + \frac{2S^2\beta}{b^2}\|v_{t,j-2}\|_{\bar{V}_{j-2}^{-1}}.$$

Term III: We have that

$$\text{Term III} = \theta^\top(v_{t,j-1} - x_t)$$

$$
\begin{aligned}
&= \theta^\top u_t(\zeta_{t,j-1} - \zeta_{t,j}) \\
&\leq S|\zeta_{t,j-1} - \zeta_{t,j}| \\
&= S(\zeta_{t,j} - \zeta_{t,j-1}) && \text{(a)} \\
&\leq S(\bar{\zeta}_t - \zeta_{t,j-1}) && \text{(b)} \\
&= S\bar{\zeta}_t(1 - \zeta_{t,j-1}/\bar{\zeta}_t) \\
&\leq \frac{2S^2\beta}{b^2} \|v_{t,j-2}\|_{\bar{V}_{j-2}^{-1}} && \text{(c)}
\end{aligned}
$$

where (a) is from the fact that $\zeta_{t,j} \geq \zeta_{t,j-1}$ since $\zeta_{i,j}$ is monotone in j by definition (see line 12), (b) is due to the fact that $\zeta_{t,j} \leq \bar{\zeta}_t$ give that $\mathcal{Y}_j^p \subseteq \mathcal{Y}$ (conditioned on $\mathcal{E}_{\text{conf}}$), and (c) is from Lemma 8.

Completing the proof: Using the bounds established for each of the terms, it holds that

$$
\begin{aligned}
r_t &= \text{Term I} + \text{Term II} + \text{Term III} \\
&\leq 3\beta\|v_{t,j-1}\|_{\bar{V}_{j-1}^{-1}} + 2\beta\|v_{*,j-1}\|_{\bar{V}_{j-1}^{-1}} + \frac{2S^2\beta}{b^2}\|v_{*,j-2}\|_{\bar{V}_{j-2}^{-1}} + \frac{4S^2\beta}{b^2}\|v_{t,j-2}\|_{\bar{V}_{j-2}^{-1}} \\
&\leq 5\beta w_{j-1} + \frac{6S^2\beta}{b^2} w_{j-2},
\end{aligned}
$$

where we use the fact $i_* \in \mathcal{A}_{j-1} \subseteq \mathcal{A}_{j-2}$ (due to Lemma 9) and $i_t \in \mathcal{A}_j \subseteq \mathcal{A}_{j-1} \subseteq \mathcal{A}_{j-2}$ which implies that $v_{*,j-1}, v_{t,j-1} \in \mathcal{Y}_{j-1}^p$ and $v_{*,j-2}, v_{t,j-2} \in \mathcal{Y}_{j-2}^p$. Therefore $\|v_{t,j-1}\|_{\bar{V}_{j-1}^{-1}} \leq w_{j-1}$, $\|v_{*,j-1}\|_{\bar{V}_{j-1}^{-1}} \leq w_{j-1}$, $\|v_{t,j-2}\|_{\bar{V}_{j-2}^{-1}} \leq w_{j-2}$ and $\|v_{*,j-2}\|_{\bar{V}_{j-2}^{-1}} \leq w_{j-2}$ (where w_j is defined in (22)).

Then, we study the regret within a single phase $j \geq 3$,

$$
\begin{aligned}
\sum_{t=t_j}^{\tau_j} r_t &\leq (\tau_j - t_j + 1)\left(\frac{6S^2\beta w_{j-2}}{b^2} + 5\beta w_{j-1}\right) \\
&\leq (t_{j+1} - t_j)\left(\frac{6S^2\beta w_{j-2}}{b^2} + 5\beta w_{j-1}\right) && \text{(a)} \\
&= \frac{24S^2\beta}{b^2}(t_{j-1} - t_{j-2})w_{j-2} + 10\beta(t_j - t_{j-1})w_{j-1} && \text{(b)} \\
&\leq \frac{24S^2\beta}{b^2}\sum_{t=t_{j-2}}^{t_{j-1}-1}\|x_t\|_{V_t^{-1}} + 10\beta\sum_{t=t_{j-1}}^{t_j-1}\|x_t\|_{V_t^{-1}} && \text{(c)} \\
&\leq \frac{24S^2\beta}{b^2}\sqrt{t_{j-2}\sum_{t=t_{j-2}}^{t_{j-1}-1}\|x_t\|^2_{V_t^{-1}}} + 10\beta\sqrt{t_{j-1}\sum_{t=t_{j-1}}^{t_j-1}\|x_t\|^2_{V_t^{-1}}} && \text{(d)} \\
&\leq \frac{24S^2\beta}{b^2}\sqrt{2dt_{j-2}\log\left(1 + \frac{t_{j-2}}{\lambda d}\right)} + 10\beta\sqrt{2dt_{j-1}\log\left(1 + \frac{t_{j-1}}{\lambda d}\right)}, && \text{(e)}
\end{aligned}
$$

where (a) is due to the fact that $\tau_j \leq t_{j+1} - 1$, (b) is due to the fact that $t_{j+1} - t_j = 2(t_j - t_{j-1})$, (c) is due to Lemma 10, (d) is Cauchy-Schwarz and (e) is from Lemma 5.

Finally, we can apply this to the total regret as

$$
\begin{aligned}
R_T &= \sum_{j=1}^{J}\sum_{t=t_j}^{\tau_j}\theta^\top(x_* - x_t) \\
&\leq 6S + \sum_{j=3}^{J}\sum_{t=t_j}^{\tau_j}\theta^\top(x_* - x_t) \\
&\leq 6S + \sum_{j=3}^{J}\left(\frac{24S^2\beta}{b^2}\sqrt{2dt_{j-2}\log\left(1 + \frac{t_{j-2}}{\lambda d}\right)} + 10\beta\sqrt{2dt_{j-1}\log\left(1 + \frac{t_{j-1}}{\lambda d}\right)}\right)
\end{aligned}
$$

$$\leq 6S + \beta \left(\frac{24S^2}{b^2} + 10 \right) \sqrt{2d \log \left(1 + \frac{T}{\lambda d} \right)} \sum_{j=3}^{J} \sqrt{t_j}$$
$$\leq 6S + 5\beta \left(\frac{24S^2}{b^2} + 10 \right) \sqrt{2dT \log \left(1 + \frac{T}{\lambda d} \right)},$$

where the last inequality uses $\sum_{j=1}^{J} \sqrt{t_j} = \sum_{j=1}^{J} (\sqrt{2})^j \leq 5\sqrt{T}$. $\square$

D Proofs for linked convex constraints

In this section, we prove the regret guarantees for the setting with linked convex constraints. First, we give some notation and specialize the assumptions from the original setting to this setting. We denote the vector formed from the ith row of A as a_i such that $A = [a_1 \; ... \; a_n]^\top$, and the ith element of z_t as $z_{t,i}$.

Assumption 6. *There exists positive reals S_A and S_θ such that $\|a_i\| \leq S_A$ for all $i \in [n]$ and $\|\theta\| \leq S_\theta$. Let $S := \max(S_A, S_\theta)$. Also, there exists positive real r such that $r\mathbb{B} \subseteq \mathcal{G}$. Lastly, it holds that $\nu := \frac{r}{\sqrt{n} S_A} \leq 1$.*[5]

In the following subsections will first study ROFUL in this setting and then GenOP and Safe-PE.

D.1 ROFUL under linked convex constraints

We first update the definitions of ROFUL to this setting, then will prove the regret bounds. We define the estimator of the vector a_i as

$$\hat{a}_{t,i} = V_t^{-1} \sum_{k=1}^{t-1} x_k z_{k,i}$$

and $\hat{A}_t = [\hat{a}_{t,1} \; ... \; \hat{a}_{t,n}]^\top$. We then state the specific structural assumption on the noise terms.

Assumption 7. *For all $t \in [T]$, it holds that $\mathbb{E}[\epsilon_t | x_1, \epsilon_1, ..., \epsilon_{t-1}, x_t] = 0$ and $\mathbb{E}[\exp(\lambda \epsilon_t) | x_1, \epsilon_1, ..., \epsilon_{t-1}, x_t] \leq \exp(\frac{\lambda^2 \rho^2}{2}), \forall \lambda \in \mathbb{R}$. The same holds replacing ϵ_t with $\eta_{t,i}$ for each $i \in [n]$.*

With this, we give a generalization of the confidence sets originally defined in Lemma 1.

Lemma 11 (Theorem 2 in Abbasi-Yadkori et al. (2011))**.** *Let Assumptions 1, 6 and 7 hold. Also, let*

$$\beta_t := \rho \sqrt{d \log \left(\frac{1 + (t-1)/\lambda}{\delta/(n+1)} \right)} + \sqrt{\lambda} S. \tag{23}$$

Then with probability at least $1 - \delta$, it holds that both $|x^\top (\hat{\theta}_t - \theta)| \leq \beta_t \|x\|_{V_t^{-1}}$ and $(\hat{A}_t - A)x \in \beta_t \|x\|_{V_t^{-1}} \mathbb{B}_\infty$ for all $x \in \mathbb{R}^d$ and all $t \geq 1$.

We use $\mathcal{E}_{\text{conf}}$ to refer to the event that the confidence sets in Lemma 11 hold for all rounds. The optimistic and pessimistic sets then become

$$\mathcal{Y}_t^o = \{x \in \mathcal{X} : \hat{A}_t x + \beta_t \|x\|_{V_t^{-1}} \mathbb{B}_\infty \cap \mathcal{G} \neq \emptyset\} \tag{24}$$

and

$$\mathcal{Y}_t^p = \{x \in \mathcal{X} : \hat{A}_t x + \beta_t \|x\|_{V_t^{-1}} \mathbb{B}_\infty \subseteq \mathcal{G}\}. \tag{25}$$

The main challenge in this setting is characterizing the scaling required to take any point in $\mathcal{Y}_t^o$ in to $\mathcal{Y}_t^p$, which we lower bound in the following lemma.

Lemma 12. *Let Assumption 1 hold. Also, let x be any point in $\mathcal{Y}_t^o$ and $\zeta = \max\{\mu \in [0,1] : \mu x \in \mathcal{Y}_t^p\}$. Then, for all t, it holds that*

$$\zeta \geq \frac{r}{r + 2\sqrt{n} \beta_t \|x\|_{V_t^{-1}}},$$

and, with $\bar{x} = \zeta x$, that

$$\zeta \geq 1 - \frac{2\sqrt{n}}{r} \beta_t \|\bar{x}\|_{V_t^{-1}}.$$

[5]If $\nu > 1$, then for all $x \in \mathcal{X}$ it holds that $\|Ax\| \leq \|A\|_F \|x\| < \sqrt{n} S \nu = r$ given that $\|x\| \leq 1 < \nu$ for all $x \in \mathcal{X}$.

Proof. From the definition of $\mathcal{Y}_t^o$, we can choose a $v \in \mathbb{B}_\infty$ such that

$$u := \hat{A}_t x + \beta_t \|x\|_{V_t^{-1}} v \in \mathcal{G}.$$

For $\alpha \in [0,1]$, we know that

$$\begin{aligned}
\hat{A}_t \alpha x + \beta_t \|\alpha x\|_{V_t^{-1}} \mathbb{B}_\infty &\subseteq \hat{A}_t \alpha x + \beta_t \|\alpha x\|_{V_t^{-1}} (2\mathbb{B}_\infty + v) \\
&= \hat{A}_t \alpha x + \beta_t \|\alpha x\|_{V_t^{-1}} v + 2\beta_t \|\alpha x\|_{V_t^{-1}} \mathbb{B}_\infty \\
&= \alpha \left(\hat{A}_t x + \beta_t \|x\|_{V_t^{-1}} v \right) + 2\beta_t \|\alpha x\|_{V_t^{-1}} \mathbb{B}_\infty \\
&= \alpha u + 2\beta_t \|\alpha x\|_{V_t^{-1}} \mathbb{B}_\infty.
\end{aligned}$$

From Fact 1 and the fact that u is in $\mathcal{G}$, we know that $\alpha u + (1-\alpha) r \mathbb{B} \subseteq \mathcal{G}$. We choose $\alpha = \frac{r}{r + 2\sqrt{n}\beta_t \|x\|_{V_t^{-1}}}$ such that $\alpha = \frac{r}{\sqrt{n} 2\beta_t \|x\|_{V_t^{-1}}}(1-\alpha)$ to get that

$$\begin{aligned}
\hat{A}_t \alpha x + \beta_t \|\alpha x\|_{V_t^{-1}} \mathbb{B}_\infty &\subseteq \alpha u + 2\beta_t \|\alpha x\|_{V_t^{-1}} \mathbb{B}_\infty \\
&\subseteq \alpha u + 2\sqrt{n}\beta_t \|\alpha x\|_{V_t^{-1}} \mathbb{B} \\
&= \alpha u + r(1-\alpha)\mathbb{B} \subseteq \mathcal{G}.
\end{aligned}$$

Since $\hat{A}_t \alpha x + \beta_t \|\alpha x\|_{V_t^{-1}} \mathbb{B}_\infty \subseteq \mathcal{G}$ and αx is in $\mathcal{X}$ due to the fact that it is star-convex, we know that $\alpha x \in \mathcal{Y}_t^p$. It follows that

$$\zeta = \max \{\mu \geq 0 : \mu x \in \mathcal{Y}_t^p\} \geq \alpha = \frac{r}{r + 2\sqrt{n}\beta_t \|x\|_{V_t^{-1}}}, \tag{26}$$

which proves the first inequality in the statement of the lemma. Then, given that $\zeta x = \bar{x}$ and $\zeta \geq 0$, it holds that

$$\zeta \|x\|_{V_t^{-1}} = \|\zeta x\|_{V_t^{-1}} = \|\bar{x}\|_{V_t^{-1}}.$$

With this, we can rearrage (26) to get that

$$\zeta r + 2\sqrt{n}\beta_t \|\bar{x}\|_{V_t^{-1}} \geq r \quad \iff \quad \zeta \geq 1 - \frac{2\sqrt{n}}{r}\beta_t \|\bar{x}\|_{V_t^{-1}},$$

which proves the second inequality in the statement of the lemma. □

The regret bound for ROFUL in this setting then follows from this.

Theorem 7. *Let Assumptions 1, 6 and 7 hold. Then, the regret of ROFUL in the setting with linked convex constraints satisfies*

$$R_T \leq \frac{2\sqrt{n}}{r} (S_\theta + S_A) \beta_T \sqrt{2dT \log\left(1 + \frac{T}{\lambda d}\right)}$$

with probability at least $1-\delta$.

Proof. We condition on $\mathcal{E}_{\text{conf}}$ throughout the proof without further explicit reference to it. From Lemma 12 and using the same reasoning as Lemma 2, we know that

$$\gamma_t \geq \max\left(1 - \frac{2\sqrt{n}}{r}\beta_t \|x_t\|_{V_t^{-1}}, \nu\right).$$

Then, we know that

$$\begin{aligned}
\theta^\top (\tilde{x}_t - x_t) &\leq S_\theta (1 - \gamma_t) \\
&\leq \frac{2\sqrt{n} S_\theta}{r} \beta_t \|x_t\|_{V_t^{-1}}
\end{aligned}$$

Also, it holds that

$$\begin{aligned}\theta^\top(x_* - \tilde{x}_t) &\leq \hat{\theta}_t^\top \tilde{x}_t + \beta_t \|\tilde{x}_t\|_{V_t^{-1}} - \theta^\top \tilde{x}_t \\ &\leq 2\beta_t \|\tilde{x}_t\|_{V_t^{-1}} \\ &\leq \frac{2}{\nu}\beta_t \|x_t\|_{V_t^{-1}}\end{aligned}$$

Then, the instantaneous regret satisfies

$$\begin{aligned}r_t &= \theta^\top(x_* - x_t) \\ &\leq \frac{2\sqrt{n}}{r}\left(S_\theta + S_A\right)\beta_t \|x_t\|_{V_t^{-1}}.\end{aligned}$$

The proof follows from the elliptic potential lemma (Lemma 5) as used in Theorem 1. □

D.2 GenOP under linked convex constraints

In this section, we prove regret guarantees of GenOP under linked convex constraints. Note that this also gives regret bounds of GenOP in the original setting (i.e. given in (5) in Section 3.3) by taking $n = 1$ and $r = b$. Before giving the regret guarantees, we first give a corollary to Lemma 12 that bounds the scaling required to take any point in $\mathcal{Y}$ in to $\mathcal{Y}_t^p$.

Corollary 3. *Assume the same as Lemma 12 and let $\mathcal{E}_{\text{conf}}$ hold. Let y be any point in $\mathcal{Y}$ and $\zeta_1 = \max\{\mu \in [0,1] : \mu y \in \mathcal{Y}_t^p\}$. Then, for all t, it holds that*

$$\zeta_1 \geq \frac{r}{r + 2\sqrt{n}\beta_t \|y\|_{V_t^{-1}}},$$

and, with $\bar{y} = \zeta_1 y$, that

$$\zeta_1 \geq 1 - \frac{2\sqrt{n}}{r}\beta_t \|\bar{y}\|_{V_t^{-1}}.$$

Proof. Conditioned on $\mathcal{E}_{\text{conf}}$, it holds that $\mathcal{Y} \subseteq \mathcal{Y}_t^o$. Therefore, y is in $\mathcal{Y}_t^o$ and we can apply Lemma 12 to get the statement of the corollary. □

With this, we prove the regret bound for GenOP in the following theorem.

Theorem 8. *Let Assumptions 1, 6 and 7 hold. Then, playing GenOP with $\kappa = 1 + \frac{2\sqrt{n}S_\theta}{r}$ in the setting with linked convex constraints satisfies*

$$R_T \leq 2\left(1 + \frac{\sqrt{n}S_\theta}{r}\right)\beta_T \sqrt{2dT \log\left(1 + \frac{T}{\lambda d}\right)}$$

with probability at least $1 - \delta$.

Proof. Since x_* is in $\mathcal{Y}$ and using Corollary 3, we know that αx_* is in $\mathcal{Y}_t^p$, where $\alpha = \left(\frac{r}{r + 2\sqrt{n}\beta_t \|x_*\|_{V_t^{-1}}}\right)$. Using this, we show that the upper confidence bound of the actions played by GenOP is in fact larger than the optimal reward.

Since x_t is chosen by a maximization over $\mathcal{Y}_t^p$, we can reason that

$$\begin{aligned}\hat{\theta}_t^\top x_t + \kappa\beta_t \|x_t\|_{V_t^{-1}} &\geq \alpha\left(\hat{\theta}_t^\top x_* + \kappa\beta_t \|x_*\|_{V_t^{-1}}\right) \\ &= \alpha\left(\theta^\top x_* + (\hat{\theta}_t - \theta)^\top x_* + \kappa\beta_t \|x_*\|_{V_t^{-1}}\right) \\ &\geq \alpha\left(\theta^\top x_* - \beta_t \|x_*\|_{V_t^{-1}} + \kappa\beta_t \|x_*\|_{V_t^{-1}}\right)\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\frac{2\sqrt{n}\beta_t \|x_*\|_{V_t^{-1}}}{r} + 1} \left(\theta^\top x_* + \frac{2\sqrt{n} S_\theta}{r} \beta_t \|x_*\|_{V_t^{-1}} \right) \\
&= \frac{\theta^\top x_* + \frac{2\sqrt{n} S_\theta}{r} \beta_t \|x_*\|_{V_t^{-1}}}{\frac{2\sqrt{n}}{r} \beta_t \|x_*\|_{V_t^{-1}} + 1} \\
&\geq \theta^\top x_*,
\end{aligned}
$$

where the last inequality holds according to the reasoning that

$$
\begin{aligned}
&\frac{\theta^\top x_* + \frac{2\sqrt{n} S_\theta}{r} \beta_t \|x_*\|_{V_t^{-1}}}{\frac{2\sqrt{n}}{r} \beta_t \|x_*\|_{V_t^{-1}} + 1} \geq \theta^\top x_* \\
&\iff \theta^\top x_* + \frac{2\sqrt{n} S_\theta}{r} \beta_t \|x_*\|_{V_t^{-1}} \geq \theta^\top x_* \left(\frac{2\sqrt{n}}{r} \beta_t \|x_*\|_{V_t^{-1}} + 1 \right) \\
&\iff \frac{2\sqrt{n} S_\theta}{r} \beta_t \|x_*\|_{V_t^{-1}} \geq \theta^\top x_* \frac{2\sqrt{n}}{r} \beta_t \|x_*\|_{V_t^{-1}} \\
&\iff S_\theta \geq \theta^\top x_*,
\end{aligned}
$$

which holds by Assumption 1. Therefore, we know that

$$
\begin{aligned}
r_t &= \theta^\top (x_* - x_t) \\
&\leq \hat{\theta}_t^\top x_t + \kappa \beta_t \|x_t\|_{V_t^{-1}} - \theta^\top x_t \\
&\leq (\hat{\theta}_t - \theta)^\top x_t + \kappa \beta_t \|x_t\|_{V_t^{-1}} \\
&\leq (1 + \kappa) \beta_t \|x_t\|_{V_t^{-1}}
\end{aligned}
$$

Applying Cauchy-Schwarz to the cumulative regret and then the elliptic potential lemma (Lemma 5) as used in Theorem 1 completes the proof. □

D.3 Safe-PE under linked convex constraints

In this section, we give regret bounds for Safe-PE under linked convex constraints. Let the estimator of each a_i in phase j be $\hat{a}_{j,i}$ and let $\hat{A}_j = [\hat{a}_{j,1} \ldots \hat{a}_{j,n}]^\top$. We then state the specific structural assumption on the noise terms.

Assumption 8. *The noise sequences $(\epsilon_t)_{t=1}^T$ and $(\eta_{t,i})_{t=1}^T$ are independent ρ-subgaussian random variables for all $i \in [n]$.*

With this, we can then define the confidence set for the parameters in this setting which follows immediately from Lemma 7.

Lemma 13. *Then, for all $x \in \mathcal{X}$ and all $j \in [J]$ it holds that $|x^\top (\hat{\theta}_j - \theta)| \leq \|x\|_{V_j^{-1}} \beta$ and $(\hat{A}_t - A)x \in \|x\|_{V_j^{-1}} \beta \mathbb{B}_\infty$ where $\beta = \rho \sqrt{2 \log \left(\frac{4nkJ}{\delta} \right)} + \sqrt{\lambda} S$ with probability at least $1 - \delta$.*

Then, the only change to the algorithm is the definition of the maximum safe scalings (i.e. line 7 in Algorithm 3), which is

$$
\mu_{i,j+1} := \max \left\{ \alpha \in [0, \alpha_i] : \alpha \left(u_i^\top \hat{a}_j + \|u_i\|_{V_j^{-1}} \beta_j \mathbb{B}_\infty \right) \subseteq \mathcal{G} \right\}.
$$

We then apply Corollary 3 to bound the scaling of each direction in the pessimistic set as proven in the following lemma. Recall the notation from Appendix C.

Lemma 14 (Lemma 8 for linked convex constraints)**.** *Let Assumptions 1, 2 and 5 hold. For all $i \in [k]$, it holds that $\zeta_{i,j}/\bar{\zeta}_i \geq 1 - \frac{2\sqrt{n}}{r} \|v_{i,j}\|_{V_{j-1}^{-1}} \beta$ for all $j \geq 1$. Furthermore, $\zeta_{i,j}/\bar{\zeta}_i \geq 1 - \frac{2nS}{r^2} \|v_{i,j-1}\|_{V_{j-1}^{-1}} \beta$.*

Proof. Note the similarity between $\mu_{i,j}$ and the definition of ζ_1 in Corollary 3. Therefore, we can follow the reasoning of Corollary 3 to get that $(\zeta_{i,j}/\bar{\zeta}_i) \geq 1 - \frac{2\sqrt{n}}{r} \beta \|v_{i,j}\|_{V_{j-1}^{-1}}$ and then following the proof of Lemma 8 gives the claim. □

With this, we can then give the regret bound for Safe-PE in the following theorem.

Theorem 9. *Let Assumptions 1, 2, 4 and 5 hold. Then, the regret of Safe-PE (Algorithm 3) satisfies*

$$R_T \leq 6S + 5\beta \left(\frac{24S^2 n}{r^2} + 10 \right) \sqrt{2dT \log \left(1 + \frac{T}{\lambda d} \right)}$$

with probability at least $1 - \delta$.

Proof. Following the proof of Theorem 3 using Lemma 14 yields the result. □

E Problem-dependent analysis of GenOP

In this section, we show that the problem dependent analysis approach in Theorem 2 and Corollary 1 applies to GenOP.

First, we give some useful facts in the following lemma, which is analogous to Lemma 6 in the problem-dependent analysis of ROFUL.

Lemma 15. *Let Assumptions 1 and 2 hold, and let* $\mathcal{E}_{\mathrm{conf}}$ *hold. Also, let*[6]

$$\zeta_t := \max\{\zeta \geq 0 : \zeta x_t \in \mathcal{Y}\}, \tag{27}$$

and $v_t = \zeta_t x_t$*. Then, it follows that:*

1. $1/\zeta_t \in \left[1 - \frac{2}{b} \beta_t \|x_t\|_{V_t^{-1}}, 1 \right]$
2. $\theta^\top (x_t - v_t) \leq \frac{2S}{b} \beta_t \|x_t\|_{V_t^{-1}}$
3. *If there exists* $\alpha > 0$ *such that* $x_t = \alpha x_*$*, then* $v_t = x_*$.
4. *If there does not exists* $\alpha > 0$ *such that* $x_t = \alpha x_*$*, then* $\theta^\top (x_* - v_t) \geq \Delta$.

Proof. We condition on $\mathcal{E}_{\mathrm{conf}}$ throughout the proof without further reference. We will first give some useful facts. In particular, it holds that,

$$\hat{\theta}_t^\top x_t + \kappa \beta_t \|x_t\|_{V_t^{-1}} \geq \theta^\top x_* > 0 \tag{28}$$

where the first inequality is due to optimism and the second is Assumption 2. It follows from (28) that $x_t \neq \mathbf{0}$ and therefore the set $\{\zeta \geq 0 : \zeta x_t \in \mathcal{Y}\}$ is compact. Also, note that $\{\zeta \geq 0 : \zeta x_t \in \mathcal{Y}\}$ contains 0 and is therefore nonempty, so ζ_t is well-defined. Next, we prove each item individually in the following.

1: First, we argue that $\zeta_t \geq 1$. This holds because $x_t \in \mathcal{Y}_t^p \subseteq \mathcal{Y}$ and therefore 1 is in $\{\zeta \geq 0 : \zeta x_t \in \mathcal{Y}\}$. It follows that $1/\zeta_t \leq 1$.

Then, we show that $1/\zeta_t \geq 1 - \frac{2}{b} \beta_t \|x_t\|_{V_t^{-1}}$. In order to do this, we first show that $1/\zeta_t \geq \max\{\mu \in [0, 1] : \mu v_t \in \mathcal{Y}_t^p\}$. Suppose, this was not the case, i.e. that there exists $\mu \in \{\mu \in [0, 1] : \mu v_t \in \mathcal{Y}_t^p\}$ such that $\mu > 1/\zeta_t$. Since, $\mu v_t \in \mathcal{Y}_t^p$, this would imply that,

$$\hat{\theta}_t^\top x_t + \kappa \beta_t \|x_t\|_{V_t^{-1}} = \max_{x \in \mathcal{Y}_t^p} \left(\hat{\theta}_t^\top x + \kappa \beta_t \|x\|_{V_t^{-1}} \right) \geq \hat{\theta}_t^\top (\mu v_t) + \kappa \beta_t \|\mu v_t\|_{V_t^{-1}}. \tag{29}$$

At the same time, given that $v_t = \zeta_t x_t$,

$$\begin{aligned} \hat{\theta}_t^\top x_t + \kappa \beta_t \|x_t\|_{V_t^{-1}} &= (1/\zeta_t) \zeta_t \left(\hat{\theta}_t^\top x_t + \kappa \beta_t \|x_t\|_{V_t^{-1}} \right) \\ &< \mu \zeta_t \left(\hat{\theta}_t^\top x_t + \kappa \beta_t \|x_t\|_{V_t^{-1}} \right) \\ &= \mu \left(\hat{\theta}_t^\top v_t + \kappa \beta_t \|v_t\|_{V_t^{-1}} \right) \\ &= \hat{\theta}_t^\top (\mu v_t) + \kappa \beta_t \|\mu v_t\|_{V_t^{-1}}, \end{aligned} \tag{30}$$

[6] Note that this definition of ζ_t differs from the one used for the problem-dependent analysis of ROFUL in Lemma 6.

where the inequality uses (28). Since (29) and (30) cannot simultaneously hold, it follows that $1/\zeta_t \geq \max\{\mu \in [0,1] : \mu v_t \in \mathcal{Y}_t^p\}$.

Finally, using Corollary 3 by taking $n = 1$ and $r = b$ (since the setting with linked convex constraints is a more general case), we have that

$$1/\zeta_t \geq \max\{\mu \in [0,1] : \mu v_t \in \mathcal{Y}_t^p\} \geq 1 - \frac{2}{b}\beta_t \|x_t\|_{V_t^{-1}}.$$

2: Since, $v_t = \zeta_t x_t$ and $1/\zeta_t \in \left[1 - \frac{2}{b}\beta_t\|x_t\|_{V_t^{-1}}, 1\right]$, it holds that

$$\theta^\top(x_t - v_t) = \theta^\top v_t(1/\zeta_t - 1) \leq S|1/\zeta_t - 1| = S(1 - 1/\zeta_t) \leq \frac{2S}{b}\beta_t\|x_t\|_{V_t^{-1}}.$$

3: From Lemma 6, we know that $\max\{\zeta \geq 0 : \zeta x_* \in \mathcal{Y}\} = 1$. Then, if there exists $\alpha > 0$ such that $x_t = \alpha x_*$,

$$\zeta_t = \max\{\zeta \geq 0 : \zeta x_t \in \mathcal{Y}\} = \frac{1}{\alpha}\max\{\tilde{\zeta} \geq 0 : \tilde{\zeta} x_* \in \mathcal{Y}\} = \frac{1}{\alpha},$$

where we use the mapping $\tilde{\zeta} = \alpha\zeta'$. Therefore, it follows that

$$v_t = \zeta_t x_t = \alpha\zeta_t x_* = x_*.$$

4: First, note that if there does not exist $\alpha > 0$ such that $x_t = \alpha x_*$, then there does not exist $\alpha' > 0$ such that $v_t = \alpha' x_*$ as $v_t = \zeta_t x_t$. Then, since $v_t \in \mathcal{Y}$, it follows from the definition of Δ that,

$$\Delta = \inf_{x \in \mathcal{Y}:\ x \neq \alpha x_*\ \forall \alpha > 0} \theta^\top(x_* - x) \leq \theta^\top(x_* - v_t).$$

□

Theorem 10. *Let Assumptions 1, 2 and 3 hold. If $\Delta > 0$, then the number of wrong directions chosen by GenOP with $\kappa = 1 + \frac{2S}{b}$ (defined by (4)) satisfies*

$$B_T \leq \frac{1}{\Delta^2} 8d\left(1 + \frac{2S}{b}\right)^2 \beta_T^2 \log\left(1 + \frac{T}{\lambda d}\right)$$

with probability at least $1 - \delta$.

Proof. We condition on $\mathcal{E}_{\text{conf}}$ defined in (8) without further mention. Consider the the instantaneous directional regret,

$$\begin{aligned}
\tilde{r}_t &= \theta^\top(x^* - v_t)\\
&\leq \hat{\theta}_t^\top x_t + \kappa\beta_t\|x_t\|_{V_t^{-1}} - \theta^\top v_t\\
&= \theta^\top x_t + (\theta - \hat{\theta}_t)^\top x_t + \kappa\beta_t\|x_t\|_{V_t^{-1}} - \theta^\top v_t\\
&\leq \theta^\top(x_t - v_t) + (\kappa+1)\beta_t\|x_t\|_{V_t^{-1}}\\
&\leq \frac{2S}{b}\beta_t\|x_t\|_{V_t^{-1}} + (\kappa+1)\beta_t\|x_t\|_{V_t^{-1}}\\
&\leq 2(1 + \frac{2S}{b})\beta_t\|x_t\|_{V_t^{-1}},
\end{aligned}$$

where the first inequality uses optimism (with $\kappa = 1 + \frac{2S}{b}$), the second inequality uses the definition of the confidence set, the third inequality uses Lemma 15 (#2). Then, from Lemma 15 (#3, #4), we know that either $\tilde{r}_t = 0$ if there exists $\alpha > 0$ such that $x_t = \alpha x_*$ or $\tilde{r}_t \geq \Delta$ otherwise. Then, using the bound $B_T \leq \tilde{R}_T/\Delta$ and the fact that $\tilde{r}_t \leq \tilde{r}_t^2/\Delta$, we have that

$$B_T \leq \frac{\tilde{R}_T}{\Delta} = \frac{1}{\Delta}\sum_{t=1}^T \tilde{r}_t \leq \frac{1}{\Delta^2}\sum_{t=1}^T (\tilde{r}_t)^2 = \frac{4}{\Delta^2}\left(1 + \frac{2S}{b}\right)^2 \beta_T^2 \sum_{t=1}^T \|x_t\|_{V_t^{-1}}^2 \leq \frac{1}{\Delta^2} 8d\left(1 + \frac{2S}{b}\right)^2 \beta_T^2 \log\left(1 + \frac{T}{\lambda d}\right),$$

where the last inequality comes from Lemma 5. □

It then immediately follows that it is possible to achieve regret that only depends on d in $\mathcal{O}(\text{polylog}(T))$ terms with the same reasoning as Corollary 1.

Corollary 4. *Let Assumptions 1, 2 and 3 hold. If $\Delta > 0$, consider the algorithm:*

1. *Play GenOP until any single direction has been played more than $\bar{B} := \frac{1}{\Delta^2} 8d \left(1 + \frac{2S}{b}\right)^2 \beta_T^2 \log\left(1 + \frac{T}{\lambda d}\right)$ times. Let this direction be denoted by u_*.*
2. *For the remaining rounds, play GenOP (after restarting) for the 1-dimensional safe linear bandit problem of choosing $\xi_t \in \mathbb{R}_+$ and then playing $\xi_t u_*$.*

Then, with probability at least $1 - 2\delta$,

$$R_T \leq \frac{8S}{b} \beta_{2\bar{B}} \sqrt{d\bar{B} \log\left(1 + \frac{2\bar{B}}{\lambda d}\right)} + \frac{4S}{b} \tilde{\beta}_T \sqrt{2T \log\left(1 + \frac{T}{\lambda d}\right)}$$

where $\tilde{\beta}_T$ is β_T with $d = 1$.

F Details on numerical experiments

In this section, we give the details of the numerical experiments that were not included in the body of the paper as well as details on the computing setup and additional results. The computing hardware specifications are given in Section F.1. Then, the details on the simulation results for the settings with linear contraints, linked convex constraints, and star convex multi-armed bandit are given in Sections F.2, F.3, and F.4, respectively.

F.1 Computing hardware

All simulations were run on a Lenovo ThinkPad T470 with an Intel Core i7 processor and 16 GB of memory.

F.2 Linear constraints

In the first setting (results in Figure 2a), we take $d = 2$ and $T = 5 \times 10^4$, and also take the action set to be a finite star convex set $\mathcal{X} = \bigcup_{i \in [10]} \{\alpha u_k : \alpha \in [0, 1]\}$. For each trial, we sample $\theta \sim \mathcal{U}(\mathbb{B}_\infty)$, $u_k \sim \mathcal{U}(\mathbb{S})$ for all $k \in [10]$ (where we resample $\{u_k\}_k$ until it holds that $\theta^\top x_* > 0$), $b \sim \mathcal{U}[0.25, 1]$, $a \sim \mathcal{U}(\mathbb{B}_\infty)$. The learner is only given the prior information on these parameters that $\|\theta\| \leq \sqrt{2}$ and $\|a\| \leq \sqrt{2}$. As such, the algorithm can take $S_a = S_\theta = \sqrt{2}$. The noise terms are sampled i.i.d as $\eta_t \sim \mathcal{N}(\sigma)$ and $\epsilon_t \sim \mathcal{N}(\sigma)$, where $\sigma = 0.1$. The learner is given σ. For the regularization parameter, we use $\lambda = 1$ for all algorithms tested.

The second setting (results in Figure 2b) is the same except that $T = 1 \times 10^5$ and $b \sim [0.05, 0.25]$.

To implement the algorithms in this setting, we enumerate over the directions in order to calculate the algorithms' updates. To show how this matches the specified update for GenOP, consider the update specified in (4),

$$\underset{x \in \mathcal{Y}_t^p}{\arg\max} \left(\hat{\theta}_t^\top x + \kappa \beta_t \|x\|_{V_t^{-1}}\right). \tag{31}$$

Since the pessimistic set can be defined as

$$\mathcal{Y}_t^p = \bigcup_{i \in [10]} \left\{\alpha u_i : \alpha \in [0, 1], \alpha \left(\hat{a}_t^\top u_i + \beta_t \|u_i\|_{V_t^{-1}}\right) \leq b\right\},$$

and the objective is convex, we know that for each of the line segments, the maximum objective value is attained at the origin or the maximum scaling in that direction. Therefore, we find a point in (31) by optimizing over these points, i.e.

$$x_t \in \underset{x \in \{\mathbf{0}, \zeta_1^p u_1, \ldots, \zeta_{10}^p u_{10}\}}{\arg\max} \left(\hat{\theta}_t^\top x + \kappa \beta_t \|x\|_{V_t^{-1}}\right),$$

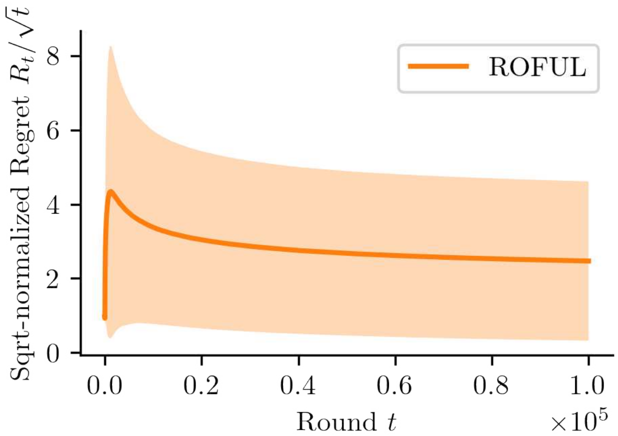


Figure 3: Simulation results for setting with linked convex constraints and box constraints.

where

$$\zeta_i^p = \begin{cases} \min(b/(\hat{a}_t^\top u_i + \beta_t \|u_i\|_{V_t^{-1}}), 1), & \text{if } \hat{a}_t^\top u_i + \beta_t \|u_i\|_{V_t^{-1}} > 0 \\ 1 & \text{else.} \end{cases}$$

The calculation used for ROFUL uses a similar idea. In particular, the optimistic action (line 4 of Algorithm 1) is calculated as

$$\tilde{x}_t \in \underset{x \in \{\mathbf{0}, \zeta_1^o u_1, \ldots, \zeta_{10}^o u_{10}\}}{\arg\max} \left(\hat{\theta}_t^\top x + \beta_t \|x\|_{V_t^{-1}} \right),$$

where

$$\zeta_i^o = \begin{cases} \min(b/(\hat{a}_t^\top u_i - \beta_t \|u_i\|_{V_t^{-1}}), 1), & \text{if } \hat{a}_t^\top u_i - \beta_t \|u_i\|_{V_t^{-1}} > 0 \\ 1 & \text{else.} \end{cases}$$

Then, μ_t from line 6 is calculated as

$$\mu_t = \begin{cases} \min(b/(\hat{a}_t^\top \tilde{x}_t + \beta_t \|\tilde{x}_t\|_{V_t^{-1}}), 1), & \text{if } \hat{a}_t^\top \tilde{x}_t + \beta_t \|\tilde{x}_t\|_{V_t^{-1}} > 0 \\ 1 & \text{else.} \end{cases} \tag{32}$$

F.3 Linked convex constraints

In this (results Figure 2b), $d = 2$, $n = 2$, $\mathcal{X} = \mathbb{B}$, and $T = 10^5$. We take $\mathcal{G} = b\mathbb{B}$, where $b \sim \mathcal{U}[0.25, 1]$ for each trial. The constraint matrix and reward vector are randomly sampled, where each row of A is sampled as $a_i \sim \mathcal{U}(\mathbb{B}_\infty)$ for all $i \in [n]$ and $\theta \sim \mathcal{U}(\mathbb{B}_\infty)$. The learner is only given the prior information on these parameters that $\|\theta\| \leq \sqrt{2}$ and $\|a_i\| \leq \sqrt{2}$ for all $i \in [n]$. As such, the algorithms can take $S = \sqrt{2}$ and $r_1 = 0.25$. The noise terms are sampled i.i.d as $\eta_t \sim \mathcal{N}(\sigma I)$ and $\epsilon_t \sim \mathcal{N}(\sigma)$, where $\sigma = 0.1$. The learner is given σ. For the regularization parameter, we use $\lambda = 1$ for both algorithms. We simulate this setting for 30 trials, where different realizations of the problem parameters are used for each trial. In Figure 2b, the mean of the regret at each round t normalized by square-root t is shown along with the plus-or-minus one standard deviation.

For this setting, we relax the optimistic and pessimistic sets such that they use the 2-norm ball instead of the infinity-ball. In particular, we use the sets

$$\mathcal{Y}_t^o = \{x \in \mathcal{X} : \hat{A}_t x + \sqrt{n}\beta_t \|x\|_{V_t^{-1}} \mathbb{B} \cap \mathcal{G} \neq \emptyset\}$$

and

$$\mathcal{Y}_t^p = \{x \in \mathcal{X} : \hat{A}_t x + \sqrt{n}\beta_t \|x\|_{V_t^{-1}} \mathbb{B} \subseteq \mathcal{G}\}.$$

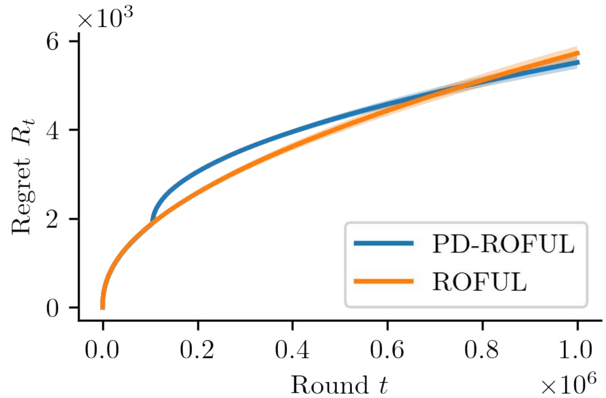


Figure 4: Simulation results for ROFUL and PD-ROFUL in problem with $\Delta > 0$ and known.

F.3.1 Additional Results

We also includes additional results (given in Figure 3), where everything is the same except that $\mathcal{G} = b\mathbb{B}_\infty$ and $\mathcal{X} = \bigcup_{k\in[10]}\{\alpha u_k : \alpha \in [0,1]\}$ where $u_k \sim \mathcal{U}(\mathbb{S})$ for all $k \in [10]$. In Figure 3, the mean of the regret at each round t normalized by square-root t is shown along with plus-or-minus one standard deviation.

F.4 Star convex multi-armed bandit

In this setting (results in Figure 2d), the action set only consists of the coordinate directions with scalings between 0 and 1. We set $\theta = a = [1\ 0\ ...\ 0]^\top$ and $b = 0.5$ and use i.i.d. Gaussian noise of standard deviation 0.1. In this case, we only give the learner the knowledge that $\|a\|, \|\theta\| \leq S = 2$ because if the learner was given the information that $\|a\| \leq 1$, it would be initially known that the optimal action satisfies the constraint. In this setting, we simulate ROFUL and Safe-PE for 3 trials for $d = 10$.

F.4.1 Additional results

We also simulate PD-ROFUL (Algorithm 2) and ROFUL in the same setting with, except with $b = 0.9$ and $S = 1.5$. We make this modification to make the initial phase of duration less than T so that there is a difference between ROFUL and PD-ROFUL. We simulate both algorithms for 5 trials with $d = 10$ and show the mean and standard deviation of the regret in Figure 4.