IEEE Open Journal of
Control Systems

Received 14 January 2025; accepted 12 March 2025. Date of publication 7 April 2025;
date of current version 30 April 2025. Recommended by Guest Editor Kyriakos G. Vamvoudakis.

Digital Object Identifier 10.1109/OJCSYS.2025.3558118

Optimistic Algorithms for Safe Linear Bandits Under General Constraints

SPENCER HUTCHINSON, ARGHAVAN ZIBAIE, RAMTIN PEDARSANI (Senior Member, IEEE), AND MAHNOOSH ALIZADEH

(Intersection of Machine Learning with Control)

Department of Electrical and Computer Engineering, University of California-Santa Barbara, Santa Barbara, CA 93106 USA

CORRESPONDING AUTHOR: Spencer Hutchinson (e-mail: shutchinson@ucsb.edu).

This work was supported by the National Science Foundation under Grant 2330154.

ABSTRACT The stochastic linear bandit problem has emerged as a fundamental building-block in machine learning and control, and a realistic model for many applications. By equipping this classical problem with safety constraints, the *safe linear bandit problem* further broadens its relevance to safety-critical applications. However, most existing algorithms for safe linear bandits only consider *linear constraints*, making them inadequate for many real-world applications, which often have non-linear constraints. To alleviate this limitation, we study the problem of safe linear bandits under general (non-linear) constraints. Under a novel constraint regularity condition that is weaker than convexity, we give two algorithms with $\tilde{\mathcal{O}}(d\sqrt{T})$ regret. We then give efficient implementations of these algorithms for several specific settings. Lastly, we give simulation results demonstrating the effectiveness of our algorithms in choosing dynamic pricing signals for a demand response problem under distribution power flow constraints.

INDEX TERMS Safe bandit, safe learning, stochastic linear bandit.

I. INTRODUCTION

The stochastic linear bandit problem is a sequential decision-making framework where, in each round, a learner chooses a vector action and then receives a stochastic reward that is linear with respect to the action [1], [15]. In choosing these actions, the learner aims to maximize her total payout over a number of rounds, despite the fact that the underlying reward functions are unknown. This problem has emerged as a fundamental building-block in both machine learning and control, with notable generalizations being linear reinforcement learning [19], [47] and adaptive control [2], [16]. Furthermore, the stochastic linear bandit problem enjoys a wide variety of real-world applications, including online advertising [28], clinical trials [37] and the smart grid [38].

Motivated by the safety-critical nature of many such applications, the stochastic linear bandit problem has recently been extended to incorporate safety constraints via the framework of *safe linear bandits*, e.g. [4], [30], [33]. This framework has garnered significant attention in recent years, and has inspired the development of other safe learning settings such as safe reinforcement learning [6], [26] and safe online convex optimization [13]. Despite the significance of the safe linear bandit problem, most prior work in this area is limited to *linear* constraints, which may be inadequate for many real-world settings. As a concrete example, bandit algorithms for active demand management in distribution networks (e.g. [38]) need to ensure satisfaction of power flow constraints, which are known to be highly non-linear [29] and linear approximations may be inadequate. In this paper, we aim to fill this gap by developing safe linear bandit algorithms that can handle *general* (non-linear) safety constraints.

In particular, we formulate the problem of safe linear bandits under general constraints, and give two algorithms for this setting: one that builds on the *scaled confidence-set* approach from [30], [33], and one that builds on the *scaled action* approach from [17]. By introducing a novel constraint regularity condition that is weaker than convexity, we show that both of these algorithms enjoy (near) optimal $\tilde{\mathcal{O}}(d\sqrt{T})$ regret. We then give efficient versions of these algorithms under several specific settings (see Table 1), and numerically demonstrate their efficacy in toy settings and a smart grid example.

A. RELATED WORK

In the following sections, we discuss related work broadly classified in to the areas of (1) bandits with round-wise constraints, (2) bandits with cumulative constraints and (3) safe learning and control generally.

1) BANDITS WITH ROUND-WISE CONSTRAINTS

The distinctive characteristic of bandits with round-wise constraints is that the constraints are enforced in each round (as opposed to cumulatively across all rounds). One such research direction considers constraints on the reward function without any additional feedback [4], [21], [31]. This differs from our problem in that we consider a separate constraint function for which the learner receives feedback. More closely related to our setting, another research direction considers an auxiliary linear constraint function $Ax_t \leq b$ with separate feedback $Ax_t + \epsilon_t$ [17], [30], [33]. In fact, we consider the same feedback $Ax_t + \epsilon_t$, except with a more general constraint $Ax_t \in \mathcal{G}_t$, where $\mathcal{G}_t$ is the union of (potentially infinite) convex sets and therefore the constraint is in general non-linear and non-convex. Non-linear constraints have also been considered by [41] of the form $f(x_t) \leq b$, where f is a β-smooth function. However, they require that the constraint is loose on the optimal action with a gap proportional to the smoothness constant, i.e. $f(x_*) \leq b - c\beta$ for constant c. This differs from our setting, which allows for the constraint to be tight on the optimal action. Lastly, we point out that our prior conference paper [17] gave algorithms for the special case where the constraint set $\mathcal{G}_t$ is convex, although it did not provide efficient versions of these algorithms. In this work, we extend the analysis approach in [17] to the setting where $\mathcal{G}_t$ is the union of convex sets, and give efficient implementations for several special cases (including the case of convex $\mathcal{G}_t$). One of the key difficulties in developing efficient implementations for this setting is that the standard "relaxed-confidence-set" approach for designing efficient bandit algorithms [15] is only applicable when the constraints are linear. To address this challenge, we introduce a technical approach that uses generalized inequalities (Section VI-A) and the properties of the norm ball (Section VI-B) to develop efficient algorithms for several types of non-linear constraints.

2) BANDITS WITH CUMULATIVE CONSTRAINTS

Cumulative constraints have been considered in a number of different bandit settings, including knapsacks [3], [8], [12], budget constraints [14], [45], and conservatism constraints [20], [46]. There has also been some efforts to give unifying frameworks to handle cumulative constraints generally [25], [27]. However, all of these works fundamentally differ from our setting because they do not guarantee constraint satisfaction in every round.

3) SAFE LEARNING AND CONTROL

Constraint satisfaction under uncertainty has also been considered in the context of control [11], [16], [22], [43], reinforcement learning [6], [10], [26], optimization [39], [40], and Gaussian process (GP) bandits [5], [35], [36]. The most directly relevant is the safe GP bandit setting, which handles non-linear constraint and cost functions by modeling them as a GP with a specified covariance kernel. This is a highly general model that captures many non-linear functions. However, for many popular kernels, existing approaches suffer exponentially in the problem dimension in both convergence rate and computation complexity, making them unsuitable for many high-dimensional settings (such as in the smart grid). Furthermore, existing regret guarantees in the safe GP setting [5] use the assumption that the constraint is loose on the optimal action, i.e. $h(x_*) < 0$ for constraint function h and optimal action x_*. This assumption is particularly restrictive in the case of linear costs, where the optimal action is necessarily in the boundary of the feasible set. In light of the above discussion, our setting serves as an alternative to GP for handling non-linear constraints, and has several advantages with respect to safe GP bandits: (a) linear dimension dependence in the regret, (b) efficient implementations for convex sets and several types of non-convex sets (see Table 1), (c) regret guarantees include the case where constraint is tight on the optimal action.

B. PAPER ORGANIZATION

We specify the problem of safe linear bandits with general constraints in Section II, and give the model assumptions in Section III. We then introduce a framework for studying this problem in Section IV, give suitable algorithms in Section V, and give efficient versions of these algorithms in Section VI. Lastly, we demonstrate the numerical efficacy of these methods in Section VII.

C. NOTATION

We use the following notations: $[n] := \{1, 2, \dots, n\}$ for natural number n, $\mathbb{R}^d_-$ is the non-positive orthant, $M^\top$ is transpose of matrix M, $\|\cdot\|$ is 2-norm, $\|x\|_A := \sqrt{x^\top Ax}$ for vector x and p.d. matrix A, $\mathbb{B}_p := \{x : \|x\|_p \leq 1\}$ for $p \in [1, \infty]$, $f : \mathcal{X} \rightrightarrows \mathcal{Y}$ is set-valued mapping from $\mathcal{X}$ to subsets of $\mathcal{Y}$, $\mathcal{A} \oplus \mathcal{B} := \{a + b : a \in \mathcal{A}, b \in \mathcal{B}\}$ and $\mathcal{A} \ominus \mathcal{B} := \{x : \mathcal{B} + x \subseteq \mathcal{A}\}$ for sets $\mathcal{A}, \mathcal{B}$.

II. PROBLEM SETUP

In each round $t \in [T]$, the learner receives the action set $\mathcal{X}_t \subseteq \mathbb{R}^d$ and constraint set $\mathcal{G}_t \subseteq \mathbb{R}^n$, chooses an action $x_t \in \mathcal{X}_t$, and then observes the reward feedback $y_t = \theta^\top x_t + \epsilon_t$ and the constraint feedback $z_t = Ax_t + \eta_t$. The parameters $\theta \in \mathbb{R}^d$ and $A \in \mathbb{R}^{n \times d}$ are unknown and $\epsilon_t \in \mathbb{R}$ and $\eta_t \in \mathbb{R}^n$ are random noise terms.

The learner needs to ensure constraint satisfaction $Ax_t \in \mathcal{G}_t$ for all $t \in [T]$ with high probability. At the same time, the learner aims to accumulate a large expected reward. To benchmark the performance of the learner in this task, we compare the actions chosen by the learner to the actions that have the highest expected reward in each round. In particular, we use the well-known notion of pseudo-regret,

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

where $x_t^\star$ are the optimal constraint-satisfying actions, i.e.

$$x_t^\star : = \arg\max_{x \in \mathcal{Y}_t} \theta^\top x, \tag{2}$$

$$\mathcal{Y}_t : = \{x \in \mathcal{X}_t : Ax \in \mathcal{G}_t\}. \tag{3}$$

For brevity, we will use the term *regret* in place of pseudo-regret going forward. In order to demonstrate that our proposed approaches are effective, we will aim to establish a high probability bound on the growth of the regret.

Remark 1: This setting can be extended to the case where the noisy constraint feedback needs to satisfy the constraint with high probability, i.e. $z_t = Ax_t + \eta_t \in \mathcal{G}_t$ w.h.p. We show this in Appendix A.

III. ASSUMPTIONS

We first assume that the action set and unknown parameters are bounded and that the noise is conditionally-subgaussian, all of which are standard in the stochastic linear bandit literature, e.g. [1], [15].

Assumption 1 (Bounded Actions and Parameters): We assume that the action set $\mathcal{X}_t$ is closed and bounded, and that the unknown parameters θ, A are bounded. That is, there exists positive reals D, S_θ, S_A such that[1] $\|\mathcal{X}_t\| \leq D$, $\|\theta\| \leq S_\theta$, and $\|a_i\| \leq S_A$ for all $i \in [n]$, where $a_i^\top$ denotes the ith row of A. Let $S := \max(S_\theta, S_A)$.

Assumption 2 (Conditionally-Subgaussian Noise): We assume that ϵ_t, η_t are conditionally-subgaussian, i.e. given variance proxies σ_r^2, σ_c^2, it holds that

$$\mathbb{E}[\epsilon_t | \mathcal{F}_t] = 0, \ \mathbb{E}[e^{\lambda \epsilon_t} | \mathcal{F}_t] \leq \exp\left(\frac{\lambda^2 \sigma_r^2}{2}\right) \forall \lambda \in \mathbb{R},$$

$$\mathbb{E}[\eta_{t,i} | \mathcal{F}_t] = 0, \ \mathbb{E}[e^{\lambda \eta_{t,i}} | \mathcal{F}_t] \leq \exp\left(\frac{\lambda^2 \sigma_c^2}{2}\right) \forall \lambda \in \mathbb{R} \ \forall i \in [n],$$

where $\mathcal{F}_t := \sigma(\epsilon_1, \eta_1, x_1, \ldots, \epsilon_{t-1}, \eta_{t-1}, x_t)$ is the filtration defined by all randomness up to the time that y_t and z_t are observed.

Then, in Assumption 3, we assume that the origin is a feasible action, and that the action set is always star-convex. Knowledge of a feasible point is necessary to ensure feasibility from the first round and it is often available in practice. Indeed, real-world systems often have a conservative action that has been determined previously and therefore can be used for this initially feasible action. Also note that the assumption of a star-convex action set is strictly weaker than convexity and allows us to interpolate between any action and the origin, which is critical to our approach. Lastly, we point out that our setting extends to the case where the known feasible point is not the origin as discussed in Remark 2. We do not explicitly consider this setting for ease of notation.

Assumption 3 (Known Feasible Action and Star-convexity): For all $t \in [T]$, it holds that $\mathbf{0} \in \mathcal{Y}_t$ and $\mathcal{X}_t$ is star-convex, i.e. $\alpha x \in \mathcal{X}_t$ for all $\alpha \in [0, 1]$ and $x \in \mathcal{X}_t$.

Remark 2 (Non-zero Feasible Action): If there are known (non-zero) feasible actions $\bar{x}_t \in \mathcal{Y}_t$ with known constraint values $\bar{z}_t = A\bar{x}_t$ and the action set is star convex about $\bar{x}_t$, we can reduce the problem to an equivalent one satisfying Assumption 3 by passing the algorithm the shifted past actions $x_t \leftarrow x_t - \bar{x}_t$ (for calculations relating to the constraint), the shifted constraint feedback $z_t \leftarrow z_t - \bar{z}_t$ and shifted constraint set $\mathcal{G}_t \leftarrow \mathcal{G}_t - \bar{z}_t$. The actions chosen by the algorithm are then shifted back as $x_t \leftarrow x_t + \bar{x}_t$ before being played.

Lastly, we assume in Assumption 4 that the constraint set is the union of (possibly infinite) convex sets that have the origin in the interior. Roughly speaking, this assumption ensures that the constraint set $\mathcal{G}_t$ is sufficiently "spacious" to allow for safe and efficient learning of the constraints. More precisely, it guarantees that by scaling down any point in the set (i.e. moving it towards the origin), we can create separation between this scaled point and the boundary of the set. This property will be critical to our algorithm and analysis, as we will use this "scaling" to create separation with the set boundary, and therefore ensure that a given point is safe despite uncertainty on the constraint function. Furthermore, it is worth pointing out that by Assumption 4, $\mathcal{G}_t$ is necessarily connected, but not necessarily convex.

Assumption 4 (Constraint is Union of Convex sets): For all $t \in [T]$, the constraint set $\mathcal{G}_t$ is closed and the union of (possibly infinite) convex sets with the origin in the interior. Precisely, there exists a set-valued mapping $\mathcal{D}_t : \mathcal{I} \rightrightarrows \mathbb{R}^d$ from an index set $\mathcal{I}$ such that $\mathcal{G}_t = \bigcup_{i \in \mathcal{I}} \mathcal{D}_t(i)$, where $\mathcal{D}_t(i)$ is convex for all $i \in \mathcal{I}$ and there exists $r_t(i) \in \mathbb{R}$ such that $r_t(i)\mathbb{B}_\infty \subseteq \mathcal{D}_t(i)$ for all $i \in \mathcal{I}$ with $\bar{r} := \inf_{i \in \mathcal{I}, t \in [T]} r_t(i) > 0$. We assume that $\nu := \frac{\bar{r}}{S_A} \leq D$.

Remark 3: Our assumption that $\nu = \frac{\bar{r}}{S_A} \leq D$ ensures that the problem is nontrivial. In fact, if $\nu > D$, then it is known that every $x \in \mathcal{X}_t$ satisfies the constraints[2] and the problem can be treated as a conventional stochastic linear bandit problem. Assuming $\nu \leq D$ will also simplify presentation of results.

IV. PRELIMINARIES

In this section, we give the main tools that we will use to study the specified problem. In particular, we discuss *confidence bounds* in Section IV-A, *pessimistic sets* in Section IV-B and *optimistic algorithms* in Section IV-C.

[1] We use the notation $\|\mathcal{X}_t\| \leq D$ to mean that every element in the image of $\mathcal{X}_t$ under $\|\cdot\|$ is less than D, i.e. $\max_{x \in \mathcal{X}_t} \|x\| \leq D$.

[2] Indeed, if $\bar{r} > S_A D$, then for all $x \in \mathcal{X}_t$, it holds that $\|Ax\|_\infty = \max_i |a_i^\top x| \leq S_A D < r \Rightarrow Ax \in \bar{r}\mathbb{B}_\infty \subseteq \mathcal{G}_t$.

A. CONFIDENCE BOUNDS

A central challenge of the specified problem is that the reward function $\theta^\top x$ and the constraint function Ax are unknown. In order to ensure both constraint satisfaction and low regret, it is paramount that we can tightly bound the true functions $\theta^\top x$ and Ax using past observations of the reward and constraint. To do so, we follow the standard approach in the stochastic linear bandit literature [1], [15], and consider high-probability *confidence bounds* of the form,

$$Ax \in \hat{A}_t x + \beta_t^A \|x\|_{V_t^{-1}} \mathbb{B}_\infty, \tag{4}$$

$$\theta^\top x \in \hat{\theta}_t^\top x + \beta_t^\theta \|x\|_{V_t^{-1}} [-1, 1], \tag{5}$$

for any $x \in \mathbb{R}^d$, where $\hat{\theta}_t$ and $\hat{A}_t$ are the regularized least-squares estimator,

$$\hat{\theta}_t := V_t^{-1} \sum_{s=1}^{t-1} y_s x_s, \qquad \hat{A}_t^\top := V_t^{-1} \sum_{s=1}^{t-1} z_s^\top x_s,$$

$V_t := \sum_{s=1}^{t-1} x_s x_s^\top + \lambda I$ is the Gram matrix (with $\lambda > 0$), and $\beta_t^A, \beta_t^\theta > 0$ are appropriately-chosen hyper-parameters. In fact, a good choice of $\beta_t^A, \beta_t^\theta$ is given in Lemma 1, which guarantees that (4) and (5) hold with high probability for all rounds.

Lemma 1 (Theorem 2 in [1]): Let Assumptions 1 and 2 hold. Fix $\delta \in (0, 1)$ and let[3]

$$\beta_t^\theta := \sigma_r \sqrt{d \log\left(\frac{1 + (t-1)D^2/\lambda}{\delta/(n+1)}\right)} + \sqrt{\lambda} S_\theta$$

$$\beta_t^A := \sigma_c \sqrt{d \log\left(\frac{1 + (t-1)D^2/\lambda}{\delta/(n+1)}\right)} + \sqrt{\lambda} S_A$$

With this choice of $\beta_t^\theta, \beta_t^A$, let $\mathcal{E}$ be the event that (5) and (4) both hold for all $t \in \mathbb{N}$. Then, $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$.

B. PESSIMISTIC SETS

Since constraint satisfaction is a central objective of our problem setting, we will next specify a set of actions that are guaranteed to satisfy the constraints. In particular, we use the confidence bounds in (4) and the fact that $A(\nu\mathbb{B}) \subseteq \mathcal{G}_t$ (see Remark 3) to define the set of actions that are guaranteed to satisfy the constraints, which we call the *pessimistic set*,

$$\mathcal{Y}_t^p := \{x \in \mathcal{X}_t : \hat{A}_t x + \beta_t^A \|x\|_{V_t^{-1}} \mathbb{B}_\infty \subseteq \mathcal{G}_t\} \cup (\mathcal{X}_t \cap \nu\mathbb{B}),$$

which ensures that $\mathcal{Y}_t^p \subseteq \mathcal{Y}_t$ for all $t \in [T]$ under the high probability event $\mathcal{E}$ (defined in Lemma 1). Using a pessimistic set to ensure constraint satisfaction is a standard technique in the safe linear bandit literature, e.g. [4], [30], [33]. However, these existing works are only applicable to the special case of our setting in which there are linear constraints, i.e. $\mathcal{G}_t = b + \mathbb{R}_-$. Lastly, note that without further assumptions on the problem setting or improved confidence bounds for least-squares estimators, the action set $\mathcal{Y}_t^p$ is about as large as is possible to ensure high probability constraint satisfaction.

C. OPTIMISTIC ALGORITHMS

In this section, we specify a class of algorithms that choose actions from the pessimistic set and are guaranteed to have low regret. To do so, we build on the paradigm of *optimism in the face of uncertainty*, which is known to be successful in the stochastic bandit literature, e.g. [1], [7], [23]. This concept is distilled in the notion of *optimistic algorithms* as defined in Definition 1. The key property of an optimistic algorithm (as we have defined it) is that it chooses actions in such a way that the the optimal reward $\theta^\top x_t^\star$ is upper bounded by the realized reward $\theta^\top x_t$ plus a multiple of the weighted norm at the played actions $\|x_t\|_{V_t^{-1}}$, as specified in (6). This property guarantees low regret, which we prove later in this section.

Definition 1 (Optimistic Algorithm): Let the relevant information in round t be denoted by $\mathcal{J}_t = (\mathcal{G}_t, \mathcal{X}_t, \hat{A}_t, \hat{\theta}_t, V_t, \beta_t^A, \beta_t^\theta)$. Then, an *optimistic algorithm*, denoted by $\mathcal{A}$, maps $\mathcal{J}_t$ to the pessimistic set, i.e. $x_t = \mathcal{A}(\mathcal{J}_t) \in \mathcal{Y}_t^p$, such that there exists $K_A, K_\theta \geq 0$ where for all $t \in [T]$,

$$\theta^\top x_t^\star \leq \theta^\top x_t + (K_A \beta_t^A + K_\theta \beta_t^\theta) \|x_t\|_{V_t^{-1}}, \tag{6}$$

under the event $\mathcal{E}$.

When the constraint is loose (i.e. $\nu > D$ as discussed in Remark 3) then the problem is unconstrained, and the classical LinUCB algorithm [1], [7] provides a valid optimistic algorithm. This is shown in Example 1. The key idea behind the LinUCB algorithm is that it chooses the action that maximizes the upper confidence bound on the reward (see (7)). Since the upper confidence bound is larger than the true reward (with high probability) and the optimal action is in the pessimistic set (when the constraint is loose), this algorithm gives an upper bound on the optimal reward that satisfies Definition 1.

Example 1 (Optimistic Algorithm for Unconstrained Case): In the case where $\nu > D$ and therefore the constraint is known to be loose, a valid optimistic algorithm is,

$$\mathcal{A}(\mathcal{J}_t) = \arg\max_{x \in \mathcal{X}_t} (\hat{\theta}_t^\top x + \beta_t^\theta \|x\|_{V_t^{-1}}), \tag{7}$$

where $K_\theta = 2$ and $K_A = 0$.

Proof: Since $\nu > D$, it holds that $x_t^\star \in \mathcal{Y}_t = \mathcal{X}_t = \mathcal{Y}_t^p$, and therefore with $x_t = \mathcal{A}(\mathcal{J}_t)$,

$$\begin{aligned}\theta^\top x_t^\star &\leq \max_{x \in \mathcal{Y}_t} (\hat{\theta}_t^\top x + \beta_t^\theta \|x\|_{V_t^{-1}}) \\ &= \max_{x \in \mathcal{X}_t} (\hat{\theta}_t^\top x + \beta_t^\theta \|x\|_{V_t^{-1}}) \\ &= \hat{\theta}_t^\top x_t + \beta_t^\theta \|x_t\|_{V_t^{-1}} \\ &\leq \theta^\top x_t + 2\beta_t^\theta \|x_t\|_{V_t^{-1}},\end{aligned}$$

where both inequalities use the confidence bounds from (5). ■

Constructing an optimistic algorithm is more difficult in the case of general constraints because there needs to be a way to upper bound the optimal reward, despite the fact that the

[3] We modify Theorem 2 in [1] by taking union bounds over the rows of A as well as θ (i.e. $n + 1$ union bounds).

Algorithm 1: ROFUL.

```
input $\mathcal{G}_t, \mathcal{X}_t, \hat{A}_t, \hat{\theta}_t, V_t, \beta_t^A, \beta_t^\theta$.
1:  Construct optimistic set:
      $\mathcal{Y}_t^o := \{x \in \mathcal{X}_t : \hat{A}_t x + \beta_t^A \|x\|_{V_t^{-1}} \mathbb{B}_\infty \cap \mathcal{G}_t \neq \emptyset\}$.
2:  Compute optimistic action:
      $\tilde{x}_t \in \arg\max_{x \in \mathcal{Y}_t^o} (\hat{\theta}_t^\top x + \beta_t^\theta \|x\|_{V_t^{-1}})$.
3:  Compute safe scaling:
      $\gamma_t = \max\{\gamma \in [0, 1] : \gamma \tilde{x}_t \in \mathcal{Y}_t^p\}$.
4:  **return** $x_t = \gamma_t \tilde{x}_t$.
```

optimal action may not yet be in the pessimistic set. The next section will be dedicated to developing optimistic algorithms for this problem. In any case, the following proposition shows that if there is an optimistic algorithm and K_A, K_θ do not depend on d or T, then we can guarantee near-optimal $\tilde{\mathcal{O}}(d\sqrt{T})$ regret.[4]

Proposition 1 (Regret of Optimistic Algorithms): Let $\mathcal{A}$ be an optimistic algorithm in the sense of Definition 1. Then, choosing $x_t = \mathcal{A}(\mathcal{J}_t)$ for all $t \in [T]$, and $\lambda \geq \max(1, D^2)$, ensures that,

$$R_T \leq (K_\theta \beta_T^\theta + K_A \beta_T^A) \sqrt{2dT \log\left(1 + \frac{T}{\lambda d}\right)},$$

under the event $\mathcal{E}$.

Proof: See Appendix B. ■

V. OPTIMISTIC ALGORITHMS

In this section, we present two different algorithms that satisfy Definition 1. Specifically, the ROFUL algorithm is given in Section V-A, and the OptPess algorithm is given in Section V-B. We provide some discussion on the difference between the two in Section V-C.

A. ROFUL

The ROFUL algorithm (Algorithm 1) operates by first constructing an *optimistic set* $\mathcal{Y}_t^o$, which is an outer approximation of the feasible set $\mathcal{Y}_t \subseteq \mathcal{Y}_t^o$. It then finds the intermediate iterate $\tilde{x}_t$ by maximizing an upper bound on the reward over the optimistic set, which provides an upper bound on the optimal reward, i.e. $\hat{\theta}_t^\top \tilde{x}_t + \beta_t^\theta \|\tilde{x}_t\|_{V_t^{-1}} \geq \theta^\top x_t^\star$. However, $\tilde{x}_t$ is not necessarily in the feasible set, so selecting it might violate the constraints. Instead, ROFUL scales $\tilde{x}_t$ into the pessimistic set, which ensures both that the chosen action is safe with high probability, and that the algorithm gains information *in the direction of the optimistic action*.

Remark 4: It may not be immediately apparent that the calculation of γ_t in line 3 of Algorithm 1 is well-defined. However, it follows from Assumption 3 that $\mathbf{0}$ is in $\mathcal{G}_t$ and $\mathcal{X}_t$, and therefore $\mathbf{0} \in \mathcal{Y}_t^p$. This ensures that the set $\{\mu \in [0, 1] : \mu \tilde{x}_t \in \mathcal{Y}_t^p\}$ is nonempty (and trivially, closed and bounded) so the maximum element exists. Also, note that $\tilde{x}_t \in \mathcal{X}_t$, so by the star-convexity of $\mathcal{X}_t$, it holds that $x_t = \gamma_t \tilde{x}_t \in \mathcal{X}_t$ since $\gamma_t \in [0, 1]$.

In the following theorem, we show that ROFUL does in fact satisfy the requirements of Definition 1.

Theorem 1: ROFUL (Algorithm 1) is a valid optimistic algorithm in the sense of Definition 1, where $K_A = \frac{2DS_A}{\bar{r}}$ and $K_\theta = \frac{2DS_\theta}{\bar{r}}$.

The central piece of this theorem is the following lemma, which bounds the scaling required to go from the optimistic set in to the pessimistic set. This ultimately gives a bound on γ_t and will also be used for the analysis of OptPess in the next section.

Lemma 2: Consider any $x \in \mathcal{Y}_t^o$ and $\mu = \max\{\mu \in [0, 1] : \mu x \in \mathcal{Y}_t^p\}$. Then, it holds that,

$$\mu \geq \max\left(\frac{\bar{r}}{\bar{r} + 2\beta_t^A \|x\|_{V_t^{-1}}}, \frac{\bar{r}}{DS_A}\right). \tag{8}$$

Furthermore, setting $y = \mu x$ gives that,

$$\mu \geq \max\left(1 - \frac{2}{\bar{r}} \beta_t^A \|y\|_{V_t^{-1}}, \frac{\bar{r}}{DS_A}\right). \tag{9}$$

Proof: First let,

$$\alpha := \frac{\bar{r}}{\bar{r} + 2\beta_t^A \|x\|_{V_t^{-1}}}$$

$$\iff \quad \alpha = \frac{\bar{r}}{2\beta_t^A \|x\|_{V_t^{-1}}}(1 - \alpha).$$

Then, we show that $\alpha x \in \mathcal{Y}_t^p$. From the definition of $\mathcal{Y}_t^o$, there exists $v \in \mathbb{B}_\infty$ such that,

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

Then, since $\mathcal{G}_t$ is the union of convex sets $\{\mathcal{D}_t(i)\}_i$, there exists $i \in \mathcal{I}$ such that $u \in \mathcal{D}_t(i)$. Therefore, it holds that

$$\begin{aligned}
&\hat{A}_t(\alpha x) + \beta_t^A \|\alpha x\|_{V_t^{-1}} \mathbb{B}_\infty \\
&\subseteq \hat{A}_t(\alpha x) + \beta_t^A \|\alpha x\|_{V_t^{-1}} (2\mathbb{B}_\infty + v) \\
&= \hat{A}_t(\alpha x) + \beta_t^A \|\alpha x\|_{V_t^{-1}} v + 2\beta_t^A \|\alpha x\|_{V_t^{-1}} \mathbb{B}_\infty \\
&= \alpha \left(\hat{A}_t x + \beta_t^A \|x\|_{V_t^{-1}} v\right) + 2\alpha \beta_t^A \|x\|_{V_t^{-1}} \mathbb{B}_\infty \\
&= \alpha u + (1 - \alpha) \bar{r} \mathbb{B}_\infty \\
&\subseteq \alpha \mathcal{D}_t(i) \oplus (1 - \alpha) \bar{r} \mathbb{B}_\infty \\
&\subseteq \alpha \mathcal{D}_t(i) \oplus (1 - \alpha) r(i) \mathbb{B}_\infty \\
&\subseteq \alpha \mathcal{D}_t(i) \oplus (1 - \alpha) \mathcal{D}_t(i) = \mathcal{D}_t(i) \subseteq \mathcal{G}_t,
\end{aligned}$$

where the last equality is from the convexity of $\mathcal{D}_t(i)$. Thus, we have shown that $\alpha x \in \mathcal{Y}_t^p$ (and $\alpha \in [0, 1]$ trivially), so it follows that

$$\mu = \max\{\mu \in [0, 1] : \mu x \in \mathcal{Y}_t^p\} \geq \alpha. \tag{10}$$

[4] To see that the bound in Proposition 1 is $\tilde{\mathcal{O}}(d\sqrt{T})$, note that $\beta_t^A = \tilde{\mathcal{O}}(\sqrt{d})$ and $\beta_t^\theta = \tilde{\mathcal{O}}(\sqrt{d})$ from their definitions in Lemma 1.

Also, it holds that $(\nu\mathbb{B})\cap\mathcal{X}_t\subseteq\mathcal{Y}_t^p$ and therefore $\min\left(\frac{\nu}{\|x\|},1\right)x\in\mathcal{Y}_t^p$ so,

$$\mu\geq\min\left(\frac{\nu}{\|x\|},1\right)\geq\min\left(\frac{\nu}{D},1\right)=\frac{\bar{r}}{DS_A},\tag{11}$$

where we use the fact that $\nu\leq D$ from Assumption 4. Combining (10) and (11) gives the inequality (8).

Then, noting that $y=\mu x$ and therefore $\|y\|_{V_t^{-1}}=\mu\|x\|_{V_t^{-1}}$, we can rearrange (10) to get that,

$$\begin{aligned}\bar{r}&\leq\mu\bar{r}+\mu2\beta_t^A\|x\|_{V_t^{-1}}=\mu\bar{r}+2\beta_t^A\|y\|_{V_t^{-1}}\\ \iff\ \mu&\geq1-\frac{2}{\bar{r}}\beta_t^A\|y\|_{V_t^{-1}}.\end{aligned}\tag{12}$$

Combining (11) and (12) gives the inequality (9). ■

With this, we then give the proof of Theorem 1.

Proof of Theorem 1: First, note that we can apply (9) in Lemma 2 with $x\leftarrow\tilde{x}_t$, $\mu\leftarrow\gamma_t$ and $y\leftarrow x_t$ to get that,

$$\gamma_t\geq\max\left(1-\frac{2}{\bar{r}}\beta_t^A\|x_t\|_{V_t^{-1}},\frac{\bar{r}}{DS_A}\right).\tag{13}$$

Then, consider the decomposition,

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

When $\mathcal{E}$ holds, we can bound Term I as,

$$\begin{aligned}\text{Term I}&=\theta^\top(x_t^\star-\tilde{x}_t)\\ &\leq\hat{\theta}_t^\top\tilde{x}_t+\beta_t^\theta\|\tilde{x}_t\|_{V_t^{-1}}-\theta^\top\tilde{x}_t\\ &=(\hat{\theta}_t-\theta)^\top\tilde{x}_t+\beta_t^\theta\|\tilde{x}_t\|_{V_t^{-1}}\\ &\leq2\beta_t^\theta\|\tilde{x}_t\|_{V_t^{-1}}\\ &\leq2\beta_t^\theta\frac{\|x_t\|_{V_t^{-1}}}{\gamma_t}\\ &\leq\frac{2DS_A}{\bar{r}}\beta_t^\theta\|x_t\|_{V_t^{-1}},\end{aligned}$$

where the last inequality applies $\gamma_t\geq\frac{\bar{r}}{DS_A}$ from (13).

As for Term II, it holds (almost surely) that,

$$\begin{aligned}\text{Term II}&=\theta^\top(\tilde{x}_t-x_t)\\ &=\theta^\top\tilde{x}_t(1-\gamma_t)\\ &\leq S_AD(1-\gamma_t)\\ &\leq\frac{2S_\theta D}{\bar{r}}\beta_T^A\|x_t\|_{V_t^{-1}},\end{aligned}$$

where we use $\gamma_t\geq1-\frac{2}{\bar{r}}\beta_t^A\|x_t\|_{V_t^{-1}}$ from (13).

Therefore, ROFUL satisfies Definition 1 with $K_A=\frac{2DS_A}{\bar{r}}$ and $K_\theta=\frac{2S_\theta D}{\bar{r}}$. ■

B. OPTPESS

The OptPess algorithm (Algorithm 2) performs a maximization directly on the pessimistic set. To provide an upper bound on the optimal reward, it incorporates both the uncertainty on the reward $\beta_t^\theta\|x\|_{V_t^{-1}}$ and the uncertainty on the constraint $\beta_t^A\|x\|_{V_t^{-1}}$ into this maximization. The idea of expanding the confidence bounds in the maximization was proposed by [33] in the UCB case and [30] in the Thompson Sampling case. Here, we extend the approach to non-linear constraints (i.e. $\mathcal{G}_t\neq\mathbb{R}_-$) and differentiated confidence radii for costs and constraints (i.e. $\beta_t^A\neq\beta_t^\theta$).

Algorithm 2: OptPess.

```
input $\mathcal{G}_t, \mathcal{X}_t, \hat{A}_t, \hat{\theta}_t, V_t, \beta_t^A, \beta_t^\theta$.
    Optimistic-pessimistic update:
        $x_t \in \arg\max_{x\in\mathcal{Y}_t^p}(\hat{\theta}_t^\top x + (\beta_t^\theta + \frac{2S_\theta D}{\bar{r}}\beta_t^A)\|x\|_{V_t^{-1}})$.
    return $x_t$.
```

In the following theorem, we show that OptPess is an optimistic algorithm in the sense of Definition 1.

Theorem 2: OptPess (Algorithm 2) is a valid optimistic algorithm in the sense of Definition 1, where $K_A=\frac{2DS_A}{\bar{r}}$ and $K_\theta=2$.

Proof: First note that since $x_t^\star\in\mathcal{Y}_t\subseteq\mathcal{Y}_t^o$ under $\mathcal{E}_{\text{conf}}$, we can apply Lemma 2 with $x\leftarrow x_t^\star$ to get that $\alpha x_t^\star\in\mathcal{Y}_t^p$ where,

$$\alpha=\frac{1}{1+\frac{2}{\bar{r}}\beta_t^A\|x_t^\star\|_{V_t^{-1}}}.$$

Therefore, from the maximization in Algorithm 2, it holds under $\mathcal{E}_{\text{conf}}$ that,

$$\begin{aligned}&\hat{\theta}_t^\top x_t+\left(\beta_t^\theta+\frac{2S_\theta D}{\bar{r}}\beta_t^A\right)\|x_t\|_{V_t^{-1}}\\ &\geq\hat{\theta}_t^\top(\alpha x_t^\star)+\left(\beta_t^\theta+\frac{2S_\theta D}{\bar{r}}\beta_t^A\right)\|\alpha x_t^\star\|_{V_t^{-1}}\\ &=\alpha\left(\hat{\theta}_t^\top x_t^\star+\left(\beta_t^\theta+\frac{2S_\theta D}{\bar{r}}\beta_t^A\right)\|x_t^\star\|_{V_t^{-1}}\right)\\ &\geq\alpha\left(\theta^\top x_t^\star+\frac{2S_\theta D}{\bar{r}}\beta_t^A\|x_t^\star\|_{V_t^{-1}}\right)\\ &\geq\alpha\theta^\top x_t^\star\left(1+\frac{2}{\bar{r}}\beta_t^A\|x_t^\star\|_{V_t^{-1}}\right)=\theta^\top x_t^\star,\end{aligned}$$

where the second inequality uses the confidence bounds (5), and the third inequality uses that $S_\theta D\geq\|\theta\|\|x_t^\star\|\geq\theta^\top x_t^\star$. Applying (5) again gives that,

$$\theta^\top x_t^\star\leq\theta^\top x_t+\left(2\beta_t^\theta+\frac{2S_\theta D}{\bar{r}}\beta_t^A\right)\|x_t\|_{V_t^{-1}}$$

which satisfies Definition 1 with $K_A=\frac{2S_\theta D}{\bar{r}}$ and $K_\theta=2$. ■

C. COMPARISON OF ROFUL AND OPTPESS

One of the key differences between the two algorithms is that ROFUL requires solving an optimization problem over the optimistic set $\mathcal{Y}_t^o$, while OptPess requires solving an optimization problem over the pessimistic set $\mathcal{Y}_t^p$. When the action set $\mathcal{X}_t$ and the constraint set $\mathcal{G}_t$ are convex, the pessimistic set $\mathcal{Y}_t^p$

TABLE 1. Summary of efficient implementations.

Section	Conditions on $\mathcal{X}_t$	Conditions on $\mathcal{G}_t$	Base Algorithm	Computation	Regret
VI.A	convex	union of N convex cones	ROFUL	$4Nd^2$ convex programs	$\tilde{\mathcal{O}}(d^{3/2}\sqrt{T})$
VI.B	convex	union of N norm balls (arbitrary norms)	ROFUL	$4Nd^2$ convex programs	$\tilde{\mathcal{O}}(d^{3/2}\sqrt{T})$
VI.C	convex	convex	OptPess	$2d$ convex programs with $2dn$ constraints	$\tilde{\mathcal{O}}(nd^{3/2}\sqrt{T})$
VI.D	K-finite star-convex	concave/convex	ROFUL/OptPess	K root-findings with n evaluations each step	$\tilde{\mathcal{O}}(nd\sqrt{T})$

can be written as a convex set (see Appendix C) because it can then be specified as the *intersection* of convex sets. As a result, OptPess can be implemented efficiently (with modification) in such settings, which we discuss further in the following section.

However, even when $\mathcal{X}_t$ and $\mathcal{G}_t$ are convex, the optimistic set $\mathcal{Y}_t^o$ is not generally convex because it would be the *union* of convex sets. Nonetheless, we find that the optimistic set is amenable to efficient implementations when $\mathcal{G}_t$ is a *non-convex* set that can be represented as the union of certain convex sets. This is because, unlike for the pessimistic set, we can interchange the unions that define the constraint set and the unions that define the optimistic set. As a result, ROFUL can be implemented efficiently (with modification) when the constraint set is the union of certain types of convex sets, which we discuss further in the next section.

It is also worth pointing out that [17] observed empirical differences in performance between ROFUL and OptPess in the case of linear constraints.[5] In particular, they found that ROFUL performs better than OptPess in certain settings where the constraint is looser. Lastly, we note that both algorithms require knowledge of the problem parameters $D, S_A, S_\theta, \sigma_r, \sigma_c, \bar{r}$.

VI. EFFICIENT IMPLEMENTATIONS

In this section, we study novel efficient implementations of the developed algorithms. Note that as specified in the last section, neither ROFUL and OptPess yield efficient implementations in general. As such, we give versions of these algorithms that can be implemented efficiently for the several special cases detailed in Table 1.

A. UNION OF CONVEX CONES

In this section, we consider the case where $\mathcal{X}_t$ is convex and $\mathcal{G}_t$ is the union of a finite number of convex cones, i.e.

$$\mathcal{G}_t = \bigcup_{i \in \mathcal{I}} \mathcal{D}_i, \quad \text{where} \quad \mathcal{D}_i = \mathcal{K}_i + b_i, \tag{14}$$

where $b_i \in \mathbb{R}^n$, $\mathcal{I}$ is a finite set, and $\mathcal{K}_i$ are proper cones.[6]

To handle this setting, we use ROFUL (Algorithm 1) with the following computation to find $\tilde{x}_t$ in line 2:

$$\max_{(i,j,p,\xi,\zeta) \in \mathcal{I} \times [d]^2 \times \{-1,1\}^2} f(i,j,p,\xi,\zeta),$$

$$\begin{aligned} f(i,j,p,\xi,\zeta) &:= \max_{x \in \mathcal{X}_t} \left(\hat{\theta}_t^\top + \zeta \sqrt{d} \beta_t^\theta v_{t,p}^{-1/2}\right) x \\ &\text{s.t.} \left(\hat{A}_t - \frac{\xi \sqrt{d} b_i \beta_t^A}{r_i} v_{t,j}^{-1/2}\right) x \in \mathcal{D}_i \end{aligned} \tag{15}$$

We use $v_{t,p}^{-1/2}$ to denote the pth row of $V_t^{-1/2}$. Then, we compute γ_t in line 3:

$$\gamma_t = \max\left(\frac{\bar{r}^2}{\bar{r}^2 + 2\bar{b}\sqrt{d}\beta_t^A \|V_t^{-1/2}\tilde{x}_t\|_\infty}, \frac{\bar{r}}{DS_A}\right). \tag{16}$$

The optimization problem (15) requires solving $4|\mathcal{I}|d^2$ convex programs. In the following theorem, we show that using ROFUL with (15) yields $\tilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ regret.

Theorem 3: Playing ROFUL (Algorithm 1) with the maximizer of (15) for $\tilde{x}_t$ in line 2 and (16) instead of line 3 is an optimistic algorithm (in the sense of Definition 1) with $K_A = \frac{2DS_A\bar{b}\sqrt{d}}{\bar{r}^2}$ and $K_\theta = \frac{2S_\theta\sqrt{d}}{\bar{r}}$, where $\bar{b} := \max_{i \in \mathcal{I}} \|b_i\|_\infty$.

In order to prove Theorem 3, we will first establish an equivalent form for the feasible set of (15) that will allow for analysis. To do so, we define the notation for the feasible set for x in (15) as,

$$\tilde{\mathcal{Y}}_t^o := \bigcup_{\substack{i \in \mathcal{I} \\ \xi \in \{-1,1\} \\ j \in [d]}} \left\{x \in \mathcal{X}_t : \hat{A}_t x - \frac{\xi \sqrt{d} b_i \beta_t^A}{r_i} v_{t,j}^{-1/2} x \in \mathcal{D}_i\right\}. \tag{17}$$

Then, in the following, we establish an equivalent form for $\tilde{\mathcal{Y}}_t^o$.

Lemma 3: The set $\tilde{\mathcal{Y}}_t^o$, defined in (17), can be written as,

$$\tilde{\mathcal{Y}}_t^o = \bigcup_{i \in \mathcal{I}} \left\{x \in \mathcal{X}_t : \left(\hat{A}_t x + \frac{\sqrt{d}}{r_i} \beta_t^A \|x\|_{V_t^{-1}}^\infty \mathcal{W}_i\right) \cap \mathcal{D}_i \neq \emptyset\right\}, \tag{18}$$

where $\|x\|_{V_t^{-1}}^\infty := \|V_t^{-1/2} x\|_\infty$ and,

$$\mathcal{W}_i := \operatorname{conv}(r_i \mathbb{B}_\infty \cup \{-b_i\} \cup \{b_i\}).$$

Before proving Lemma 3, we first give some properties of the set $\mathcal{W}_i$ that will be useful for analysis. We will make use of the notion of generalized inequality, denoted by $x \preceq_\mathcal{K} y$

[5] Note that [17] uses the name GenOP to refer to the algorithm that we call OptPess.

[6] A cone is called proper if it is convex, closed, has nonempty interior, and is pointed [9].

with proper cone $\mathcal{K}$ and defined by $y - x \in \mathcal{K}$. See Section 2.4 in [9] for more details on generalized inequalities.

Lemma 4 (Properties of $\mathcal{W}_i$): It holds that,

1) $\mathcal{W}_i$ is symmetric, i.e. $\mathcal{W}_i = -\mathcal{W}_i$,
2) $\mathcal{W}_i \subseteq \mathcal{K}_i + b_i$,
3) $-b_i$ is a maximum element of $\mathcal{W}_i$ with respect to $\preceq_{\mathcal{K}_i}$, i.e. $-b_i \succeq_{\mathcal{K}_i} z$ for all $z \in \mathcal{W}_i$.

Proof: Property 1) holds because,

$$\begin{aligned}-\mathcal{W}_i &= -\mathrm{conv}(r_i\mathbb{B}_\infty \cup \{-b_i\} \cup \{b_i\})\\ &= \mathrm{conv}(-r_i\mathbb{B}_\infty \cup \{b_i\} \cup \{-b_i\})\\ &= \mathrm{conv}(r_i\mathbb{B}_\infty \cup \{b_i\} \cup \{-b_i\}) = \mathcal{W}_i.\end{aligned}$$

Property 2) holds because,

$$\begin{aligned}&r_i\mathbb{B}_\infty \subseteq \mathcal{K}_i + b \qquad \text{(Assumption 4)},\\ &b_i = \mathbf{0} + b_i \in \mathcal{K}_i + b_i \qquad (\mathbf{0} \in \mathcal{K}_i),\\ &-b_i = 2(\mathbf{0} - b_i) + b_i \in \mathcal{K}_i + b_i \qquad (\mathbf{0} - b_i \in \mathcal{K}_i).\end{aligned}$$

Since $\mathcal{W}_i$ is the convex hull of these terms and $\mathcal{K}_i + b_i$ is convex, it holds that $\mathcal{W}_i \subseteq \mathcal{K}_i + b_i$.

Property 3) holds because $-b_i \in \mathcal{W}_i$ and,

$$\begin{aligned}\mathcal{W}_i &= \mathrm{conv}(r_i\mathbb{B}_\infty \cup \{-b_i\})\\ &\subseteq -\mathrm{conv}(r_i\mathbb{B}_\infty \cup \{b_i\})\\ &\subseteq -b_i - \mathcal{K}_i,\end{aligned}$$

and therefore, from the definition of maximum element (e.g. Definition 4.8 in [18]), $-b_i$ is a maximum element with respect to $\preceq_{\mathcal{K}_i}$. ■

With this established, we prove Lemma 3 in the following.

Proof of Lemma 3: First, note that, because $\mathcal{W}_i$ is symmetric and convex,

$$\|x\|^\infty_{V_t^{-1}}\mathcal{W}_i = \max_{\substack{\xi\in\{-1,1\}\\ j\in[d]}} (\xi v_{t,j}^{-1/2} x)\mathcal{W}_i = \bigcup_{\substack{\xi\in\{-1,1\}\\ j\in[d]}} \xi v_{t,j}^{-1/2} x\mathcal{W}_i. \tag{19}$$

Then, note that $-b_i$ is a maximum element of $\mathcal{W}_i$ (via Lemma 4). Therefore, there exists ξ_*, j_* such that $-\xi_* v_{t,j_*}^{-1/2} x b_i$ is the maximum element of $\|x\|^\infty_{V_t^{-1}}\mathcal{W}_i$, and, at the same time, it holds that $\xi v_{t,j}^{-1/2} x b_i \in \|x\|^\infty_{V_t^{-1}}\mathcal{W}_i$ for all ξ, j (due to (19)). It follows that for any $c \in \mathbb{R}^n$,

$$\begin{aligned}&\exists(\xi, j) \in \{-1, 1\} \times [d] \ : \ c \preceq_{\mathcal{K}_i} -\xi v_{t,j}^{-1/2} x b_i\\ &\iff \exists w \in \mathcal{W}_i : c \preceq_{\mathcal{K}_i} \|x\|^\infty_{V_t^{-1}} w.\end{aligned} \tag{20}$$

With this, we can show that the condition defining (17) is equivalent to the condition defining (18). In particular, choosing $c = \frac{r_i(b_i - \hat{A}_t x)}{\sqrt{d} b_i \beta_t^A}$ in the sense of (20), it holds that,

$$\begin{aligned}&\exists(\xi, j) \in \{-1, 1\} \times [d] : \hat{A}_t x - \frac{\xi\sqrt{d} b_i \beta_t^A}{r_i} v_{t,j}^{-1/2} x \in \mathcal{K}_i + b_i\\ &\iff \exists(\xi, j) \in \{-1, 1\} \times [d] : c \preceq_{\mathcal{K}_i} \xi v_{t,j}^{-1/2} x\\ &\iff \exists w \in \mathcal{W} : c \preceq_{\mathcal{K}_i} \|x\|^\infty_{V_t^{-1}} w\\ &\iff \left(\hat{A}_t x + \frac{\sqrt{d}}{r_i}\beta_t^A \|x\|^\infty_{V_t^{-1}}\mathcal{W}_i\right) \cap \mathcal{D}_i \neq \emptyset,\end{aligned}$$

which is the defining condition of (18). ■

Next, we show that $\tilde{\mathcal{Y}}_t^o$ can be "sandwiched" between two sets, which will facilitate the regret analysis.

Lemma 5: It holds that $\tilde{\mathcal{Y}}_t^o$ defined (18) satisfies the following,

$$\begin{aligned}&\mathcal{Y}_t^o \subseteq \tilde{\mathcal{Y}}_t^o\\ &\quad \subseteq \left\{x \in \mathcal{X}_t : \left(\hat{A}_t x + \frac{\bar{b}\sqrt{d}}{\bar{r}}\beta_t^A \|x\|^\infty_{V_t^{-1}}\mathbb{B}_\infty\right) \cap \mathcal{G}_t \neq \emptyset\right\},\end{aligned}$$

where $\bar{b} = \max_{i\in\mathcal{I}} \|b_i\|_\infty$ and $\bar{r} = \min_{i\in\mathcal{I}} r_i$.

Proof: First, note that by construction of $\mathcal{W}_i$ it holds that,

$$\mathbb{B}_\infty \subseteq \frac{1}{r_i}\mathcal{W}_i \subseteq \frac{\|b_i\|_\infty}{r_i}\mathbb{B}_\infty \subseteq \frac{\bar{b}}{\bar{r}}\mathbb{B}_\infty.$$

Also, it holds that $\|x\|_{V_t^{-1}} \leq \sqrt{d}\|x\|^\infty_{V_t^{-1}}$. It follows that,

$$\begin{aligned}\mathcal{Y}_t^o &= \{x \in \mathcal{X}_t : \hat{A}_t x + \beta_t^A \|x\|_{V_t^{-1}}\mathbb{B}_\infty \cap \mathcal{G}_t \neq \emptyset\}\\ &\subseteq \{x \in \mathcal{X}_t : \hat{A}_t x + \sqrt{d}\beta_t^A \|x\|^\infty_{V_t^{-1}}\mathbb{B}_\infty \cap \mathcal{G}_t \neq \emptyset\}\\ &= \bigcup_{i\in\mathcal{I}}\{x \in \mathcal{X}_t : \hat{A}_t x + \sqrt{d}\beta_t^A \|x\|^\infty_{V_t^{-1}}\mathbb{B}_\infty \cap \mathcal{D}_i \neq \emptyset\}\\ &\subseteq \bigcup_{i\in\mathcal{I}}\left\{x \in \mathcal{X}_t : \hat{A}_t x + \frac{\sqrt{d}}{r_i}\beta_t^A \|x\|^\infty_{V_t^{-1}}\mathcal{W}_i \cap \mathcal{D}_i \neq \emptyset\right\}\\ &= \tilde{\mathcal{Y}}_t^o\\ &\subseteq \bigcup_{i\in\mathcal{I}}\left\{x \in \mathcal{X}_t : \hat{A}_t x + \frac{\bar{b}\sqrt{d}}{\bar{r}}\beta_t^A \|x\|^\infty_{V_t^{-1}}\mathbb{B}_\infty \cap \mathcal{D}_i \neq \emptyset\right\}\\ &\subseteq \left\{x \in \mathcal{X}_t : \hat{A}_t x + \frac{\bar{b}\sqrt{d}}{\bar{r}}\beta_t^A \|x\|^\infty_{V_t^{-1}}\mathbb{B}_\infty \cap \mathcal{G}_t \neq \emptyset\right\},\end{aligned}$$

completing the proof. ■

With this established, we prove Theorem 3 in the following.

Proof of Theorem 3: First, note that (15) is equivalent to,

$$\begin{aligned}&\max_{\substack{x\in\tilde{\mathcal{Y}}_t^o\\ p\in[d]\\ \zeta\in\{-1,1\}}} \left(\hat{\theta}_t^\top + \zeta\sqrt{d}\beta_t^\theta v_{t,p}^{-1/2}\right) x\\ &= \max_{x\in\tilde{\mathcal{Y}}_t^o} \hat{\theta}_t^\top x + \sqrt{d}\beta_t^\theta \|x\|^\infty_{V_t^{-1}}\\ &\geq \max_{x\in\tilde{\mathcal{Y}}_t^o} \hat{\theta}_t^\top x + \beta_t^\theta \|x\|_{V_t^{-1}}\\ &\geq \max_{x\in\mathcal{Y}_t^o} \hat{\theta}_t^\top x + \beta_t^\theta \|x\|_{V_t^{-1}} \geq \theta^\top x_t^\star,\end{aligned} \tag{21}$$

where we use Lemma 5 and condition on $\mathcal{E}$. Therefore, it holds that $\hat{\theta}_t^\top \tilde{x}_t + \sqrt{d}\beta_t^\theta \|\tilde{x}_t\|^\infty_{V_t^{-1}} \geq \theta^\top x_t^\star$ conditioned on $\mathcal{E}$.

Then, also note that, with Lemma 2 and the substitution $\beta_t^A \leftarrow \frac{\bar{b}\sqrt{d}}{\bar{r}}\beta_t^A$, it holds that $x_t = \gamma_t \tilde{x}_t$ is in the set,

$$\left\{x \in \mathcal{X}_t : \left(\hat{A}_t x + \frac{\bar{b}\sqrt{d}}{\bar{r}}\beta_t^A \|x\|_{V_t^{-1}}^{\infty} \mathbb{B}_\infty\right) \subseteq \mathcal{G}_t\right\} \cup (\mathcal{X}_t \cap \nu\mathbb{B}),$$

which is a subset of $\mathcal{Y}_t^p$. Therefore, it holds that $x_t \in \mathcal{Y}_t^p$. The rest of Theorem 1 then applies with the substitution $\beta_t^\theta \leftarrow \sqrt{d}\beta_t^\theta$. This results in ROFUL being an optimistic algorithm (in the sense of Definition 1) with $K_A = \frac{2DS_A\bar{b}\sqrt{d}}{\bar{r}^2}$ and $K_\theta = \frac{2S_\theta\sqrt{d}}{\bar{r}}$. ■

B. UNION OF NORM BALLS

In this section, we consider the case where $\mathcal{X}_t$ is convex and $\mathcal{G}_t$ is the union of a finite number of norm balls, i.e.

$$\mathcal{G}_t = \bigcup_{i\in\mathcal{I}} \mathcal{D}_i \quad \text{where} \quad \mathcal{D}_i = \{z \in \mathbb{R}^n : \|z\|_{\mathcal{D}_i} \le 1\},$$

where $\|\cdot\|_{\mathcal{D}_i}$ is a norm for all $i \in \mathcal{I}$. Equivalently, we can say that $\mathcal{D}_i$ is a compact, convex, and symmetric set that contains $\mathbf{0}$ and has a nonempty interior.

For this setting, we use ROFUL with the following computation to find $\tilde{x}_t$ in line 2:

$$\max_{(i,j,p,\xi,\zeta)\in\mathcal{I}\times[d]^2\times\{-1,1\}^2} f(i,j,p,\xi,\zeta),$$

$$f(i,j,p,\xi,\zeta) := \max_{x\in\mathcal{X}_t} \left(\hat{\theta}_t^\top + \zeta\sqrt{d}\beta_t^\theta v_{t,p}^{-1/2}\right) x \tag{22}$$

$$\text{s.t. } \|\hat{A}_t x\|_{\mathcal{D}_i} - \frac{\xi\sqrt{d}\beta_t^A}{r_i} v_{t,j}^{-1/2} x \le 1$$

We use $v_{t,p}^{-1/2}$ to denote the pth row of $V_t^{-1/2}$. Then, we compute γ_t in line 3:

$$\gamma_t = \max\left(\frac{\bar{r}^2}{\bar{r}^2 + 2C\sqrt{d}\beta_t^A\|V_t^{-1/2}\tilde{x}_t\|_\infty}, \frac{\bar{r}}{DS_A}\right). \tag{23}$$

We use the notation $C := \max_{i\in\mathcal{I}} \max_{z\in\mathcal{D}_i} \|z\|_\infty$. Solving (22) requires solving $4|\mathcal{I}|d^2$ convex programs. Now, we state the regret guarantees obtained by using (22) with ROFUL.

Theorem 4: Playing ROFUL (Algorithm 1) with the maximizer of (22) for $\tilde{x}_t$ in line 2 and (23) in place of line 3 is an optimistic algorithm (in the sense of Definition 1) with $K_A = \frac{2DS_AC\sqrt{d}}{\bar{r}^2}$ and $K_\theta = \frac{2S_\theta\sqrt{d}}{\bar{r}}$.

Same as in the previous section, the key difficulty in proving Theorem 4 is to establish an equivalent form of the feasible set that allows for analysis. Reusing notation, we define the feasible set in (22) as,

$$\tilde{\mathcal{Y}}_t^o = \bigcup_{i\in\mathcal{I}} \bigcup_{\substack{\xi\in\{-1,1\}\\ j\in[d]}} \left\{x \in \mathcal{X}_t : \|\hat{A}_t x\|_{\mathcal{D}_i} - \frac{\xi\sqrt{d}\beta_t^A}{r_i} v_{t,j}^{-1/2} x \le 1\right\}, \tag{24}$$

where $v_{t,i}^{-1/2}$ is the ith row of $V_t^{-1/2}$.

We will then give an equivalent formulation of this set.

Lemma 6: It holds that $\tilde{\mathcal{Y}}_t^o$ defined in (25) is equal to,

$$\tilde{\mathcal{Y}}_t^o := \bigcup_{i\in\mathcal{I}} \left\{x \in \mathcal{X}_t : \hat{A}_t x + \frac{\sqrt{d}}{r_i}\beta_t^A \|x\|_{V_t^{-1}}^{\infty} \mathcal{D}_i \cap \mathcal{D}_i \neq \emptyset\right\}. \tag{25}$$

To prove this lemma, we first state the following fact, which gives us a closed-way to check the satisfaction of the norm constraint.

Fact 1: For arbitrary norm $\|\cdot\|_\dagger$ and $y \in \mathbb{R}^n$, it holds that

$$\min_{\|x\|_\dagger\le 1} \|x - y\|_\dagger = \max(\|y\|_\dagger - 1, 0)$$

Proof: We denote the minimizer as x_*. If $\|y\|_\dagger \le 1$, then choosing $x_* = y$ minimizes as the norm is non-negative. If $\|y\|_\dagger > 1$, then it holds for any feasible x that,

$$\|y\|_\dagger = \|y - x + x\|_\dagger \le \|y - x\|_\dagger + \|x\|_\dagger \le \|y - x\|_\dagger + 1.$$

This implies that the objective is lower bounded by $\|y\|_\dagger - 1$ for any feasible x which is attained by choosing $x_* = \frac{y}{\|y\|_\dagger}$. ■

With this in hand, we prove Lemma 6 in the following.

Proof of Lemma 6: We show that the condition that defines (25) is equivalent to the condition that defines (24). First, note that when $x = \mathbf{0}$ it is immediate, so we take $x \neq \mathbf{0}$ for the rest of the proof. In particular,

$$\begin{aligned}
&\hat{A}_t x + \frac{\sqrt{d}}{r_i}\beta_t^A \|x\|_{V_t^{-1}}^{\infty} \mathcal{D}_i \cap \mathcal{D}_i \neq \emptyset \\
\iff\; &\exists\, w \in \hat{A}_t x + \frac{\sqrt{d}}{r_i}\beta_t^A \|x\|_{V_t^{-1}}^{\infty} \mathcal{D}_i \;:\; \|w\|_{\mathcal{D}_i} \le 1 \\
\iff\; &\min_{w\in \hat{A}_t x + \frac{\sqrt{d}}{r_i}\beta_t^A \|x\|_{V_t^{-1}}^{\infty} \mathcal{D}_i} \|w\|_{\mathcal{D}_i} \le 1 \\
\iff\; &\min_{y\in\mathcal{D}_i} \left\| \hat{A}_t x + \frac{\sqrt{d}}{r_i}\beta_t^A \right\| x\|_{V_t^{-1}}^{\infty} y\|_{\mathcal{D}_i} \le 1 \\
\iff\; &\frac{\sqrt{d}}{r_i}\beta_t^A \|x\|_{V_t^{-1}}^{\infty} \min_{y\in\mathcal{D}_i} \left\| \frac{-\hat{A}_t x}{\frac{\sqrt{d}}{r_i}\beta_t^A \|x\|_{V_t^{-1}}^{\infty}} - y \right\|_{\mathcal{D}_i} \le 1 \\
\iff\; &\max\left(\|\hat{A}_t x\|_{\mathcal{D}_i} - \frac{\sqrt{d}}{r_i}\beta_t^A \|x\|_{V_t^{-1}}^{\infty}, 0\right) \le 1 \\
\iff\; &\|\hat{A}_t x\|_{\mathcal{D}_i} - \frac{\sqrt{d}}{r_i}\beta_t^A \|x\|_{V_t^{-1}}^{\infty} \le 1 \\
\iff\; &\min_{\substack{\xi\in\{-1,1\}\\ j\in[d]}} \left(\|\hat{A}_t x\|_{\mathcal{D}_i} - \frac{\xi\sqrt{d}\beta_t^A}{r_i} v_{t,j}^{-1/2} x\right) \le 1,
\end{aligned}$$

where we use Fact 1 in the third-to-last line. The condition in the last line is the same as in (24). ■

Finally, we prove Theorem 4 in the following.

Proof of Theorem 4: First note that, similar to Lemma 5, it holds for any $x \in \mathbb{R}^d$ that,

$$\begin{aligned} \|x\|_{V_t^{-1}} \mathbb{B}_\infty &\subseteq \sqrt{d} \|x\|^\infty_{V_t^{-1}} \mathbb{B}_\infty \\ &\subseteq \frac{\sqrt{d}}{r_i} \|x\|^\infty_{V_t^{-1}} \mathcal{D}_i \subseteq \frac{\sqrt{d}C}{\bar{r}} \|x\|^\infty_{V_t^{-1}} \mathbb{B}_\infty. \end{aligned}$$

And therefore,

$$\begin{aligned} \mathcal{Y}_t^o &= \bigcup_{i \in \mathcal{I}} \{x \in \mathcal{X}_t : (\hat{A}_t x + \beta_t^A \|x\|_{V_t^{-1}} \mathbb{B}_\infty) \cap \mathcal{D}_i \neq \emptyset\} \\ &\subseteq \bigcup_{i \in \mathcal{I}} \{x \in \mathcal{X}_t : \left(\hat{A}_t x + \beta_t^A \frac{\sqrt{d}}{r_i} \|x\|^\infty_{V_t^{-1}} \mathcal{D}_i \right) \cap \mathcal{D}_i \neq \emptyset\} \\ &= \tilde{\mathcal{Y}}_t^o \\ &\subseteq \bigcup_{i \in \mathcal{I}} \{x \in \mathcal{X}_t : \left(\hat{A}_t x + \beta_t^A \frac{\sqrt{d}C}{\bar{r}} \|x\|^\infty_{V_t^{-1}} \mathbb{B}_\infty \right) \cap \mathcal{D}_i \neq \emptyset\} \end{aligned}$$

Then, using similar reasoning as in (21), it holds that $\hat{\theta}_t^\top \tilde{x}_t + \sqrt{d} \beta_t^\theta \|\tilde{x}_t\|^\infty_{V_t^{-1}} \geq \theta^\top x_t^\star$ conditioned on $\mathcal{E}$.

Then, by applying Lemma 2 with the substitution $\beta_t^A \leftarrow \frac{\bar{b}\sqrt{d}C}{\bar{r}} \beta_t^A$, we know that $x_t = \gamma_t \tilde{x}_t$ is in the set,

$$\left\{ x \in \mathcal{X}_t : \left(\hat{A}_t x + \beta_t^A \frac{\sqrt{d}C}{\bar{r}} \|x\|^\infty_{V_t^{-1}} \mathbb{B}_\infty \right) \subseteq \mathcal{G}_t \right\} \cup (\nu \mathbb{B} \cap \mathcal{X}_t),$$

which is a subset of $\mathcal{Y}_t^p$. Therefore, $x_t \in \mathcal{Y}_t^p$. The rest of Theorem 1 then applies with the substitution $\beta_t^\theta \leftarrow \sqrt{d} \beta_t^\theta$. This results in ROFUL being an optimistic algorithm (in the sense of Definition 1) with $K_A = \frac{2DS_A C\sqrt{d}}{\bar{r}^2}$ and $K_\theta = \frac{2S_\theta \sqrt{d}}{\bar{r}}$.■

C. CONVEX SETS

In this section, we consider the case where both $\mathcal{X}_t$ and $\mathcal{G}_t$ are convex. To handle this setting, we use OptPess with the following calculation for x_t:

$$\begin{aligned} &\max_{(p,\zeta) \in [d] \times \{-1,1\}} f(p, \xi), \\ &f(p, \xi) := \max_{x \in \mathcal{X}_t} \left(\hat{\theta}_t^\top + \zeta \sqrt{d} \left(\beta_t^\theta + \frac{2nS_\theta D \beta_t^A}{\bar{r}} \right) v_{t,p}^{-1/2} \right) x \\ &\qquad \text{s.t.} \left(\mathrm{\hat{A}_t + \xi n \sqrt{d} \beta_t^A} \mathbf{e_k v_{t,j}^{-1/2}} \right) \mathrm{x} \in \mathcal{G}_\mathrm{t} \\ &\qquad \forall (k, j, \xi) \in [n] \times [d] \times \{-1, 1\} \qquad (26) \end{aligned}$$

We use $v_{t,p}^{-1/2}$ to denote the pth row of $V_t^{-1/2}$. Solving this requires $2d$ convex programs, where each program has $2dn$ constraints. We give the regret guarantees in the following.

Theorem 5: Playing the maximizer of (26) in each round is an optimistic algorithm (in the sense of Definition 1) with $K_A = \frac{2DS_A n\sqrt{d}}{\bar{r}}$ and $K_\theta = 2\sqrt{d}$ and therefore enjoys regret of $\tilde{\mathcal{O}}(nd^{3/2}\sqrt{T})$.

Same as in the previous section, we will establish an equivalent form of the feasible set in (26), which we denote as,

$$\tilde{\mathcal{Y}}_t^p = \bigcap_{\substack{k \in [n] \\ \xi \in \{-1,1\} \\ j \in [d]}} \left\{ x \in \mathcal{X}_t : \hat{A}_t x + \xi n \sqrt{d} \beta_t^A v_{t,j}^{-1/2} x \mathbf{e_k} \in \mathcal{G}_\mathbf{t} \right\},$$

where $v_{t,j}^{-1/2}$ is the jth row of $V_t^{1/2}$. Then, we state another form that allows for easier analysis.

Lemma 7: It holds that,

$$\tilde{\mathcal{Y}}_t^p = \{x \in \mathcal{X}_t : \hat{A}_t x + n\sqrt{d} \beta_t^A \|x\|^\infty_{V_t^{-1}} \mathbb{B}_1 \subseteq \mathcal{G}_t\}. \qquad (27)$$

Proof: It holds that,

$$\begin{aligned} &\hat{A}_t x + n\sqrt{d} \beta_t^A \|x\|^\infty_{V_t^{-1}} \mathbb{B}_1 \subseteq \mathcal{G}_t \\ &\quad \iff \hat{A}_t x + n\sqrt{d} \beta_t^A \max_{\xi \in \{-1,1\}, j \in [d]} (\xi v_{t,j}^{-1/2} x) \mathbb{B}_1 \subseteq \mathcal{G}_t \\ &\quad \iff \hat{A}_t x + n\sqrt{d} \beta_t^A \xi v_{t,j}^{-1/2} x \mathbb{B}_1 \subseteq \mathcal{G}_t, \quad \forall \xi, j \\ &\quad \iff \hat{A}_t x + n\sqrt{d} \beta_t^A \xi v_{t,j}^{-1/2} x \mathbf{e_k} \subseteq \mathcal{G}_\mathbf{t}, \quad \forall \xi, \mathbf{j}, \mathbf{k}, \end{aligned}$$

where the second line uses the fact that we can write $\|z\|_\infty = \max_{\xi \in \{-1,1\}, j \in [d]} \xi z_j$, and the last line holds because for convex set $\mathcal{D}$, it holds that $\mathbb{B}_1 \subseteq \mathcal{D} \iff \mathbf{e}_1, \ldots, \mathbf{e}_n \in \mathcal{D}$. To see that $\mathbb{B}_1 \subseteq \mathcal{D} \iff \mathbf{e}_1, \ldots, \mathbf{e}_n \in \mathcal{D}$, first note that $\Rightarrow$ is immediate, and then $\Longleftarrow$ holds because $\mathbb{B}_1 = \mathrm{conv}(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ and therefore any $b \in \mathbb{B}_1$ is convex combination of $\mathbf{e}_1, \ldots, \mathbf{e}_n$ and therefore must be in $\mathcal{D}$ given convexity. ■

With this, we then give the proof of Theorem 5.

Proof of Theorem 5: First, note that the $\tilde{\mathcal{Y}}_t^p \subseteq \mathcal{Y}_t^p$ as it holds for all $x \in \tilde{\mathcal{Y}}_t^p$ that,

$$\begin{aligned} \hat{A}_t x + \beta_t^A \|x\|_{V_t^{-1}} \mathbb{B}_\infty &\subseteq \hat{A}_t x + \sqrt{d} \beta_t^A \|V_t^{-1/2} x\|_\infty \mathbb{B}_\infty \\ &\subseteq \hat{A}_t x + n\sqrt{d} \beta_t^A \|V_t^{-1/2} x\|_\infty \mathbb{B}_1 \subseteq \mathcal{G}_t. \end{aligned}$$

It follows that all chosen actions satisfy $x_t \in \tilde{\mathcal{Y}}_t^p \subseteq \mathcal{Y}_t$.

Also, it holds that,

$$\begin{aligned} &\max_{(p,\zeta) \in [d] \times \{-1,1\}} f(p, \xi) \\ &\quad = \max_{\substack{x \in \tilde{\mathcal{Y}}_t^p \\ p \in [d] \\ \zeta \in \{-1,1\}}} \left(\hat{\theta}_t^\top + \zeta \sqrt{d} \left(\beta_t^\theta + \frac{2nS_\theta D \beta_t^A}{\bar{r}} \right) v_{t,p}^{-1/2} \right) x \\ &\quad = \max_{x \in \tilde{\mathcal{Y}}_t^p} \hat{\theta}_t^\top x + \sqrt{d} \left(\beta_t^\theta + \frac{2nS_\theta D \beta_t^A}{\bar{r}} \right) \|x\|^\infty_{V_t^{-1}}. \end{aligned}$$

Then, note that the steps of Theorem 2 apply with the substitution $\beta_t^\theta \leftarrow \sqrt{d} \beta_t^\theta$ and $\beta_t^A \leftarrow n\sqrt{d} \beta_t^A$. This immediately gives that $K_A = \frac{2DS_A n\sqrt{d}}{\bar{r}}$ and $K_\theta = 2\sqrt{d}$. ■

D. FINITE STAR-CONVEX ACTION SET

In this section, we consider the case where $\mathcal{X}_t$ is a finite star-convex set, i.e.,

$$\mathcal{X}_t = \bigcup_{k \in [K]} \{\nu u_k : \nu \in [0, \nu_k]\},$$

where $u_1, \ldots, u_K \in \mathbb{R}^d$ and $\nu_1, \ldots, \nu_K \in \mathbb{R}_+$. We assume that the constraint set can be represented as a functional inequality, i.e. $\mathcal{G}_t = \{x \in \mathbb{R}^n : g_t(x) \leq 0\}$ where $g_t : \mathbb{R}^n \to \mathbb{R}$ is continuous. Furthermore, we assume that $g_t(\mathbf{0}) < 0$ and will specify that g_t is either convex or concave.

When g_t is concave, we use ROFUL. In particular, to calculate $\tilde{x}_t$ in ROFUL (line 2), we use,

$$\max_{k\in[K]} \tilde{\mu}_k(\hat{\theta}_t^\top u_k + \beta_t^\theta \|u_k\|_{V_t^{-1}}), \tag{28}$$

where

$$\tilde{\mu}_k = \max\Bigg\{\mu \in [0, \nu_k] : \\ \min_{w\in\mathbf{e}_1,\ldots,\mathbf{e}_n} g_t(\mu(\hat{A}_t u_k + n\beta_t^A \|u_k\|_{V_t^{-1}} w)) \leq 0\Bigg\}. \tag{29}$$

For each k, $\tilde{\mu}_k$ can be solved by root-finding, where each step of the root-finding requires n evaluations of g_t. Note that root-finding algorithms such as bisection are guaranteed to converge linearly given that $g_t(\mathbf{0}) < 0$.

Proposition 2: When g_t is concave, playing ROFUL (Algorithm 1) with (28) in place of line 3 is an optimistic algorithm (in the sense of Definition 1) with $K_A = \frac{2DS_A n}{\bar{r}}$ and $K_\theta = \frac{2DS_\theta}{\bar{r}}$.

Proof: First, note that a vertex of a polytope is always a minimizer of a concave function, and therefore we can rewrite (29) with,

$$\begin{aligned} &\min_{w\in\mathbf{e}_1,\ldots,\mathbf{e}_n} g_t(\mu(\hat{A}_t u_k + n\beta_t^A \|u_k\|_{V_t^{-1}} w)) \leq 0 \\ \iff &\min_{w\in\mathbb{B}_1} g_t(\mu(\hat{A}_t u_k + n\beta_t^A \|u_k\|_{V_t^{-1}} w)) \leq 0 \\ \iff &\mu(\hat{A}_t u_k + n\beta_t^A \|u_k\|_{V_t^{-1}} \mathbb{B}_1) \cap \mathcal{G}_t \neq \emptyset. \end{aligned}$$

Therefore, Theorem 1 applies with $\beta_A \leftarrow n\beta_A$. ■

When g_t is convex, we use OptPess (Algorithm 2), where the action x_t in line 1 is calculated with:

$$\max_{k\in[K]} \mu_k\left(\hat{\theta}_t^\top u_k + \left(\beta_t^\theta + \frac{2S_\theta D}{\bar{r}}\beta_t^A\right)\|u_k\|_{V_t^{-1}}\right), \tag{30}$$

where

$$\mu_k = \max\Bigg\{\mu \in [0, \nu_k] : \\ \max_{w\in\mathbf{e}_1,\ldots,\mathbf{e}_n} g_t(\mu(\hat{A}_t u_k + n\beta_t^A \|u_k\|_{V_t^{-1}} w)) \leq 0\Bigg\}.$$

For each k, μ_k can be calculated with root-finding, where each step requires n evaluations of g_t.

Proposition 3: When g_t is convex, playing OptPess (Algorithm 2) with (30) in place of line 1 is an optimistic algorithm (in the sense of Definition 1) with $K_A = \frac{2DS_A n}{\bar{r}}$ and $K_\theta = 2$.

Proof: Follows from the fact that g_t is convex, and therefore a vertex is a maximum. ■

VII. NUMERICAL EXPERIMENTS

In this section, we give numerical experiments in a series of toy settings and also in a smart grid demand response setting.

A. TOY SETTINGS

We consider toy settings where the constraint set is either (a) *union of cones* (b) *union of norm balls* or (c) *convex*. For all settings, we take $n = 2$, $d = 2$, $\mathcal{X} = 2\mathbb{B}$, $A = I$, $\eta_t \sim \mathcal{N}(\mathbf{0}, 0.1I)$ and $\epsilon_t \sim \mathcal{N}(0, 0.1)$. Therefore, the feasible set is $\mathcal{Y} = \{x \in \mathcal{X} : Ax \in \mathcal{G}\} = \mathcal{X} \cap \mathcal{G}$. The constraint in each setting is:

a) *Union of cones:* $\mathcal{G} = (\mathbb{R}_- + \mathbf{1}) \cup (-\mathbb{R}_- - \mathbf{1})$

b) *Union of norm balls:* $\mathcal{G} = (1.18\mathbb{B}_\infty) \cup (1.59\mathbb{B}_1)$

c) *Convex:* $\mathcal{G} = \left\{x \in \mathbb{R}^n : \begin{bmatrix} 0 & 1 \\ \sqrt{3} & -1 \\ -\sqrt{3} & -1 \end{bmatrix} x \leq \begin{bmatrix} 1.2 \\ 1.2 \\ 1.2 \end{bmatrix}\right\}$

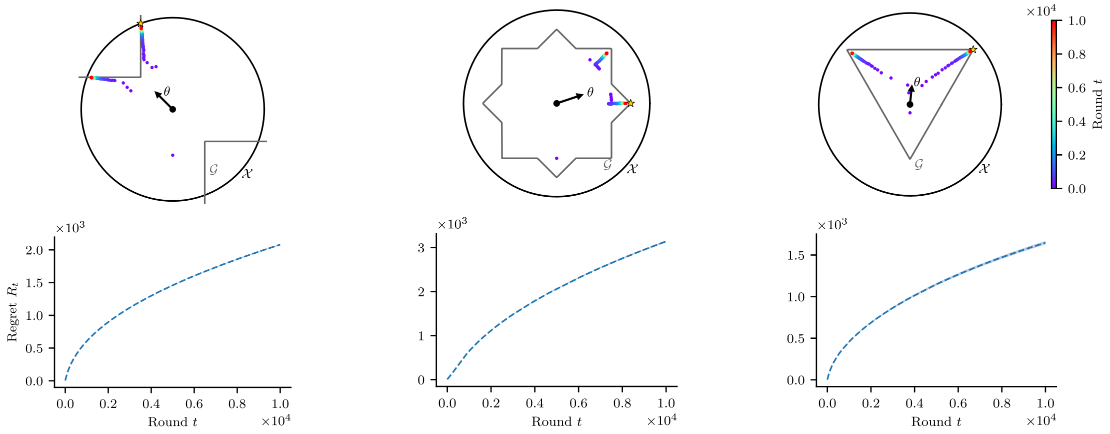
In each of these settings, we implement the applicable algorithm from Section VI as summarized in Table 1.

The results are shown in Fig. 1. The top row of Fig. 1 shows the played actions, where the color indicates the round in which the action was played, and ⋆ shows the optimal action. We downsample the played actions so that every twentieth action is shown. The bottom row of Fig. 1 shows the regret at each round, where the dashed line is the mean over 10 trials with different realizations of the noise $\{\epsilon_t, \eta_t\}_{t\in[T]}$, and the shaded region indicates the standard deviation.

B. APPLICATION TO DEMAND RESPONSE IN THE SMART GRID

In this section, we show how our algorithms can be used for demand response in the smart grid. Demand response is a class of mechanisms where an aggregator (or other electricity-supplying entity) modifies the electricity demand of customers via interventions. In particular, we consider day-ahead real-time pricing, which is a popular demand response mechanism. In each *day* (t) of day-ahead real-time pricing, the aggregator chooses a vecctor of *prices* (p_t) to be applied to each user in the following day. We consider a realistic model where the aggregator does not know how users will respond to prices a priori and instead has to learn users' price response in *real-time* by choosing prices and observing the resulting demand. In particular, we model the user's demand as a stochastic and parametrically-linear function of price, i.e. $z_t = A\phi(p_t) + \eta_t$ where $p_t \in \mathcal{P}_t$ is the vector of prices and $\phi : \mathcal{P}_t \to \mathbb{R}^d$ is a known feature map. Therefore, without loss of generality, we work in the space of actions $x_t = \phi(p_t) \in \mathcal{X}_t = \phi(\mathcal{P}_t)$ as we can always map these back to prices. In choosing these prices, the aggregator aims to maximize an *objective* (e.g. customer satisfaction or operating costs), which we model as a stochastic linear function $y_t = \theta^\top x_t + \epsilon_t$. At the same time, the aggregator aims to ensure that the customers' demand satisfies the *grid constraints* ($\mathcal{G}$) due to the nodal voltage limits and line power limits.

As a proof-of-concept, we demonstrate our approach in a demand response setting with the 22-bus distribution system



(a) union of cones

(b) union of norm balls

(c) convex

FIGURE 1. The played actions (top row) and the resulting regret (bottom row) for each type of constraint: (a) union of cones, (b) union of norm balls and (c) convex.

from [34], which we access via the MATPOWER software [48]. Since power flow constraints are known to be non-convex and highly irregular [29], we approximate the constraints by using a convex subset of the true feasible set, which is known as a *convex restriction* [24]. Convex restrictions provide a tractable way to approximate power flow constraints, while also ensuring that the true constraints are satisfied. Many different convex restrictions of power flow constraints have been proposed in the literature, e.g. [24], [32], [44]. We use the convex restriction from [44] and assume that the power factor of each node is fixed as the nominal power factor given in the test distribution system. Therefore, the constraint set can be written as,

$$\mathcal{G} = \{z \in \mathbb{R}^n : \|M\text{diag}(z)\|_\infty \leq 1/4\},$$

where $M \in \mathbb{R}^{n\times n}$ is defined as,

$$M := |\mathbf{W}^{-1}\mathbf{Y}_{LL}^{-1}\overline{\mathbf{W}}^{-1}\text{diag}(\mathbf{1} - \rho j)|.$$

Here $|\cdot|$ denotes the element-wise complex magnitude, and $\overline{\cdot}$ denotes the complex conjugate. The notation ρ is the vector of nominal power factors at each of the nodes, $\mathbf{Y}_{LL}$ is the admittance matrix of the buses, and $\mathbf{W}$ is the diagonal matrix of zero-load voltages. See [44] for the exact equations for $\mathbf{Y}_{LL}$ and $\mathbf{W}$.

In our simulations, we consider $d = 3$ features, and take $\mathcal{X} = 2\mathbb{B}$, $A = (0.13)\mathbf{1}$, $\eta_t \sim \mathcal{N}(\mathbf{0}, 0.01I)$ and $\epsilon_t \sim \mathcal{N}(0, 0.01)$. We sample $\theta \sim \mathcal{U}(\mathbb{S} \cap \mathbb{R}_+)$ in each trial. We use the ROFUL algorithm for constraint sets that are the union of norm balls (see Section VI-B). Since the constraint set is a norm (i.e. the union of $N = 1$ norm balls), we calculate γ with the exact equation (line 3 in Algorithm 1) as a root-finding problem instead of the lower bound specified in (23). The resulting regret is shown in Fig. 2, where the line indicates the average of 10 trials and the shading indicates the standard deviation.

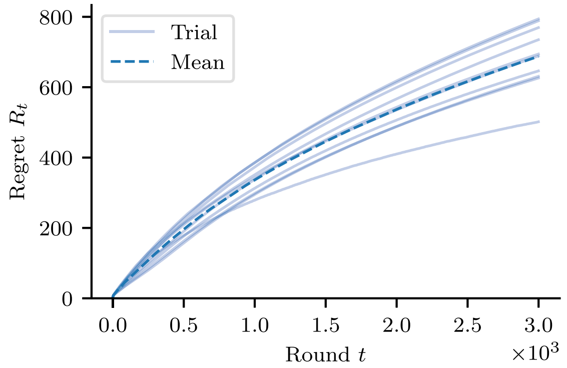


FIGURE 2. Regret of ROFUL in demand response pricing problem.

VIII. CONCLUSION

In this work, we study the problem of stochastic linear bandits under general safety constraints. We present two different algorithms for this setting, and give efficient versions of these algorithm for several special cases. Furthermore, we give simulation results that demonstrate how our algorithms can be used to efficiently and safely choose dynamic pricing signals in the smart grid.

An interesting direction for future work is to identify additional special cases that enjoy efficient implementations beyond those in Table 1. This would expand the possible applications of our algorithms. Another direction for future work is to perform a more detailed study on the effectiveness of our algorithms for dynamic pricing in the smart grid. For example, we could consider a larger distribution network and different power flow approximations, beyond the ones we consider in this paper.

APPENDIX

A. CHANCE-CONSTRAINED FORMULATION

In this section, we show how our problem setting (as defined in Section II) can be extended to handle chance-constraints on the noisy feedback $z_t = Ax_t + \eta_t$, rather than just constraints on the expected feedback Ax_t. In particular, we specify that the constraint needs to hold for all possible distributions on η_t, where the set of possible distributions is specified by the conditionally-subgaussian assumption (Assumption 2). After formulating the problem with chance-constraints, we show that our original setting (Section II) subsumes this model, and therefore that our previous results are applicable.

We consider a setting where the learner needs to ensure high-probability satisfaction of the noisy constraint term $z_t = Ax_t + \eta_t$ under all possible distributions for η_t, i.e.

$$\inf_{P_t \in \mathcal{P}} P_t(Ax_t + \eta_t \in \bar{\mathcal{G}}_t) \geq 1 - \delta', \qquad \forall t \in [T], \tag{31}$$

where $\bar{\mathcal{G}}_t$ is the constraint set, P_t is the conditional distribution of η_t given $\mathcal{F}_t$ and $\mathcal{P}$ is the set of possible conditional distributions. Since η_t is assumed to be conditionally-subgaussian (Assumption 2), we can specify $\mathcal{P}$ as the set of distributions on $\mathbb{R}^n$ such that for all $P \in \mathcal{P}$,[7]

$$\begin{gathered} \mathbb{E}_{\eta \sim P}[\eta] = \mathbf{0}, \\ P\left(|\eta_i| \leq \sigma\sqrt{\log(2/\delta)}, \forall i \in [n]\right) \geq 1 - \delta, \end{gathered} \tag{32}$$

for any $\delta \in (0, 1)$. Therefore, for a given δ', we can equivalently state (31) as,

$$\begin{aligned} Ax_t \in \mathcal{G}_t &:= \left\{ z \in \mathbb{R}^d : z + \sigma\sqrt{\log(2/\delta')}\mathbb{B}_\infty \subseteq \bar{\mathcal{G}}_t \right\} \\ &= \bar{\mathcal{G}}_t \ominus (\sigma\sqrt{\log(2/\delta')}\mathbb{B}_\infty), \end{aligned}$$

which matches the constraint satisfaction requirement in Section II, provided that the stated assumptions on $\mathcal{G}_t$ are satisfied.

In order to measure the performance of the learner in this setting, we compare her performance to a "bechmark learner" that knows the parameters θ, A, but does not know the realization of the noise terms $\eta_1, \ldots, \eta_T$ or $\epsilon_1, \ldots, \epsilon_T$. Therefore, in each round, the benchmark learner plays the action $x_t^\star$ which solves the chance-constrained optimization problem,

$$\begin{aligned} x_t^\star := \arg\max\nolimits_{x \in \mathcal{X}_t} &\ \theta^\top x \\ \text{s.t.} &\ \inf_{\mathrm{P_t} \in \mathcal{P}} \mathbb{P}_{\eta \sim \mathrm{P_t}}(\mathrm{Ax_t} + \eta \in \bar{\mathcal{G}}_\mathrm{t}) \geq 1 - \delta' \\ = \arg\max\nolimits_{x \in \mathcal{X}_t, Ax \in \mathcal{G}_t} &\theta^\top x. \end{aligned}$$

We can then define the (psuedo) regret as the cumulative difference in reward between the benchmark learners' and the learners' actions,

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

which matches the definition in (1).

[7] The σ in (32) does not necessarily match the σ_c in Assumption 2. However, the existence of each constant implies the existence of the other [42, Proposition 2.5.2].

B. PROOF OF PROPOSITION 1

Before stating the proof, we first give the so-called elliptic potential lemma.

Lemma 8 (Elliptic Potential Lemma, Lemma 11 [1]): Let Assumption 1 hold. If $\lambda \geq \max(1, D^2)$, then it holds for all $T \in \mathbb{N}$ that,

$$\sum_{t=1}^T \|x_t\|^2_{V_t^{-1}} \leq 2d \log\left(1 + \frac{T}{\lambda d}\right).$$

Using this, we then state the proof of Proposition 1 in the following.

Proof of Proposition 1: Under event $\mathcal{E}$, it holds that,

$$\begin{aligned} R_T &= \sum_{t=1}^T \theta^\top (x_t^\star - x_t) \\ &\leq \sum_{t=1}^T (K_A \beta_t^A + K_\theta \beta_t^\theta)\|x_t\|_{V_t^{-1}} && \text{(a)} \\ &\leq (K_A \beta_T^A + K_\theta \beta_T^\theta) \sum_{t=1}^T \|x_t\|_{V_t^{-1}} && \text{(b)} \\ &\leq (K_A \beta_T^A + K_\theta \beta_T^\theta) \sqrt{T \sum_{t=1}^T \|x_t\|^2_{V_t^{-1}}} && \text{(c)} \\ &\leq (K_A \beta_T^A + K_\theta \beta_T^\theta) \sqrt{2dT \log\left(1 + \frac{T}{\lambda d}\right)}, && \text{(d)} \end{aligned}$$

where (a) uses the UCB property in Definition 1, (b) uses the fact that $\beta_t^A, \beta_t^\theta$ are increasing in t, (c) uses Cauchy-Schwarz, and (d) uses Lemma 8. ■

C. CONVEX VERSION OF PESSIMISTIC SET

In this section, we show that the pessimistic set can be formulated in a convex way when used with the OptPess. This is shown in the following proposition.

Proposition 4: If $\mathcal{X}_t$ and $\mathcal{G}_t$ are convex, then the following set is convex,

$$\bar{\mathcal{Y}}_t^p := \{x \in \mathcal{X}_t : \hat{A}_t x + \beta_t^A \|x\|_{V_t^{-1}} \mathbb{B}_\infty \subseteq \mathcal{G}_t\}. \tag{33}$$

Furthermore, playing OptPess with $\mathcal{Y}_t^p \leftarrow \bar{\mathcal{Y}}_t^p$ results in the same guarantees as Theorem 2.

Proof: First, we show that $\bar{\mathcal{Y}}_t^p$ is convex. In particular, consider $z = \nu x + (1 - \nu)y$ for any $x, y \in \bar{\mathcal{Y}}_t^p$ and $\nu \in [0, 1]$. First, note that $z \in \mathcal{X}_t$ by convexity. Then, it holds that,

$$\begin{aligned} &\hat{A}_t z + \beta_t^A \|z\|_{V_t^{-1}} \mathbb{B}_\infty \\ &\subseteq \hat{A}_t z + \beta_t^A (\nu\|x\|_{V_t^{-1}} + (1 - \nu)\|y\|_{V_t^{-1}})\mathbb{B}_\infty \\ &= \hat{A}_t z + \nu\beta_t^A \|x\|_{V_t^{-1}} \mathbb{B}_\infty \oplus (1 - \nu)\|y\|_{V_t^{-1}} \mathbb{B}_\infty \\ &= \nu(\hat{A}_t x + \beta_t^A \|x\|_{V_t^{-1}} \mathbb{B}_\infty) \oplus (1 - \nu)(\hat{A}_t y + \|y\|_{V_t^{-1}} \mathbb{B}_\infty) \\ &\subseteq \nu\mathcal{G}_t \oplus (1 - \nu)\mathcal{G}_t = \mathcal{G}_t, \end{aligned}$$

Therefore, $z \in \bar{\mathcal{Y}}_t^p$ and thus $\bar{\mathcal{Y}}_t^p$ is convex.

Then, we show that the regret guarantees of OptPess are unchanged with $\mathcal{Y}_t^p \leftarrow \bar{\mathcal{Y}}_t^p$. This holds because the proof of

Theorem 2 only uses the fact that $\mathcal{Y}_t^p \supseteq \bar{\mathcal{Y}}_t^p$, and therefore the regret guarantees apply. ■

REFERENCES

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2011, vol. 24, pp. 2312–2320.

[2] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of linear quadratic systems," in *Proc. 24th Annu. Conf. Learn. Theory*, 2011, pp. 1–26.

[3] S. Agrawal, N. R. Devanur, and L. Li, "An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives," in *Proc. Conf. Learn. Theory*, 2016, pp. 4–18.

[4] S. Amani, M. Alizadeh, and C. Thrampoulidis, "Linear stochastic bandits under safety constraints," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 9256–9266.

[5] S. Amani, M. Alizadeh, and C. Thrampoulidis, "Regret bound for safe Gaussian process bandit optimization," in *Proc. 2nd Conf. Learn. Dyn. Control*, 2020, pp. 158–159.

[6] S. Amani, C. Thrampoulidis, and L. Yang, "Safe reinforcement learning with linear function approximation," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 243–253.

[7] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, pp. 235–256, 2002.

[8] A. Badanidiyuru, J. Langford, and A. Slivkins, "Resourceful contextual bandits," in *Proc. Conf. Learn. Theory*, 2014, pp. 1109–1134.

[9] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[10] A. Castellano, H. Min, J. A. Bazerque, and E. Mallada, "Learning to act safely with limited exposure and almost sure certainty," *IEEE Trans. Autom. Control*, vol. 68, no. 5, pp. 2979–2994, May 2023.

[11] M. Castroviejo-Fernandez and I. Kolmanovsky, "Safe constraint learning for reference governor implementation in constrained linear systems," *IEEE Contr. Syst. Lett.*, vol. 8, pp. 3117–3122, 2024.

[12] S. Cayci, A. Eryilmaz, and R. Srikant, "Budget-constrained bandits over general cost and reward distributions," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 4388–4398.

[13] S. Chaudhary and D. Kalathil, "Safe online convex optimization with unknown linear safety constraints," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, pp. 6175–6182.

[14] R. Combes, C. Jiang, and R. Srikant, "Bandits with budgets: Regret lower bounds and optimal algorithms," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 43, no. 1, pp. 245–257, 2015.

[15] D. Varsha, T. P Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," in *Proc. 21st Annu. Conf. Learn. Theory*, 2008, pp. 355–366.

[16] S. Dean, S. Tu, N. Matni, and B. Recht, "Safely learning to control the constrained linear quadratic regulator," in *Proc. Amer. Control Conf.*, 2019, pp. 5582–5588.

[17] S. Hutchinson, B. Turan, and M. Alizadeh, "Directional optimism for safe linear bandits," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2024, pp. 658–666.

[18] J. Jahn et al., *Vector Optimization*. Berlin, Germany: Springer, 2009.

[19] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan, "Provably efficient reinforcement learning with linear function approximation," in *Proc. Conf. Learn. Theory*, 2020, pp. 2137–2143.

[20] A. Kazerouni, M. Ghavamzadeh, Y. A. Yadkori, and B. Van Roy, "Conservative contextual linear bandits," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 3913–3922.

[21] K. Khezeli and E. Bitar, "Safe linear stochastic bandits," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 10202–10209.

[22] T. Koller, F. Berkenkamp, M. Turchetta, and A. Krause, "Learning-based model predictive control for safe exploration," in *Proc. IEEE Conf. Decis. Control*, 2018, pp. 6059–6066.

[23] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.

[24] D. Lee, H. D Nguyen, K. Dvijotham, and K. Turitsyn, "Convex restriction of power flow feasibility sets," *IEEE Trans. Control Netw. Syst.*, vol. 6, no. 3, pp. 1235–1245, Sep. 2019.

[25] Q. Liu, W. Xu, S. Wang, and Z. Fang, "Combinatorial bandits with linear constraints: Beyond knapsacks and fairness," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 2997–3010.

[26] T. Liu, R. Zhou, D. Kalathil, P. Kumar, and C. Tian, "Learning policies with zero or bounded constraint violation for constrained MDPs," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 17183–17193.

[27] X. Liu, B. Li, P. Shi, and L. Ying, "An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 24075–24086.

[28] H. Brendan McMahan et al., "Ad click prediction: A view from the trenches," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2013, pp. 1222–1230.

[29] D. K. Molzahn et al., "A survey of relaxations and approximations of the power flow equations," *Found. Trends Elect. Energy Syst.*, vol. 4, no. 1–2, pp. 1–221, 2019.

[30] A. Moradipari, S. Amani, M. Alizadeh, and C. Thrampoulidis, "Safe linear thompson sampling with side information," *IEEE Trans. Signal Process.*, vol. 69, pp. 3755–3767, 2021.

[31] A. Moradipari, C. Thrampoulidis, and M. Alizadeh, "Stage-wise conservative linear bandits," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 11191–11201.

[32] H. D. Nguyen, K. Dvijotham, and K. Turitsyn, "Constructing convex inner approximations of steady-state security regions," *IEEE Trans. Power Syst.*, vol. 34, no. 1, pp. 257–267, Jan. 2019.

[33] A. Pacchiano, M. Ghavamzadeh, P. Bartlett, and H. Jiang, "Stochastic bandits with linear constraints," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 2827–2835.

[34] M. R. Raju, K. V. S. Ramachandra Murthy, and K. Ravindra, "Direct search algorithm for capacitive compensation in radial distribution systems," *Int. J. Elect. Power Energy Syst.*, vol. 42, no. 1, pp. 24–30, 2012.

[35] Y. Sui, A. Gotovos, J. Burdick, and A. Krause, "Safe exploration for optimization with Gaussian processes," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 997–1005.

[36] Y. Sui, V. Zhuang, J. Burdick, and Y. Yue, "Stagewise safe Bayesian optimization with Gaussian processes," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4781–4789.

[37] A. Tewari and S. A. Murphy, "From ads to interventions: Contextual bandits in mobile health," in *Mobile Health: Sensors, Analytic Methods, and Applications*. Berlin, Germany: Springer, 2017, pp. 495–517.

[38] N. Tucker, A. Moradipari, and M. Alizadeh, "Constrained thompson sampling for real-time electricity pricing with grid reliability constraints," *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 4971–4983, Nov. 2020.

[39] I. Usmanova, A. Krause, and M. Kamgarpour, "Safe convex learning under uncertain constraints," in *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, 2019, pp. 2106–2114.

[40] I. Usmanova, A. Krause, and M. Kamgarpour, "Safe non-smooth blackbox optimization with application to policy search," in *Proc. Conf. Learn. Dyn. Control*, 2020, pp. 980–989.

[41] K. N. Varma, S. Lale, and A. Anandkumar, "Stochastic linear bandits with unknown safety constraints and local feedback," in *Proc. ICML Workshop New Front. Learn., Control, Dyn. Syst.*, 2023. [Online]. Available: https://openreview.net/forum?id=xFXaZXLhpK

[42] R. Vershynin, *High-Dimensional Probability: An Introduction With Applications in Data Science*, vol. 47. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[43] K. P. Wabersich, L. Hewing, A. Carron, and M. N. Zeilinger, "Probabilistic model predictive safety certification for learning-based control," *IEEE Trans. Autom. Control*, vol. 67, no. 1, pp. 176–188, Jan. 2022.

[44] C. Wang, A. Bernstein, J.-Y. Le Boudec, and M. Paolone, "Existence and uniqueness of load-flow solutions in three-phase distribution networks," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 3319–3320, Jul. 2017.

[45] H. Wu, R. Srikant, X. Liu, and C. Jiang, "Algorithms with logarithmic or sublinear regret for constrained contextual bandits," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2015, vol. 28, pp. 433–441.

[46] Y. Wu, R. Shariff, T. Lattimore, and C. Szepesvári, "Conservative bandits," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1254–1262.

[47] L. Yang and M. Wang, "Sample-optimal parametric Q-learning using linearly additive features," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6995–7004.

[48] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, "MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 12–19, Feb. 2011.