2024 European Control Conference (ECC)
June 25-28, 2024. Stockholm, Sweden

Convex Methods for Constrained Linear Bandits

Amirhossein Afsharrad[1], Ahmadreza Moradipari[2], Sanjay Lall[1]

Abstract— **Recently, bandit optimization has received significant attention in real-world safety-critical systems that involve repeated interactions with humans. While there exist various algorithms with performance guarantees in the literature, practical implementation of the algorithms has not received as much attention. This work presents a comprehensive study on the computational aspects of safe bandit algorithms, specifically safe linear bandits, by introducing a framework that leverages convex programming tools to create computationally efficient policies. In particular, we first characterize the properties of the optimal policy for safe linear bandit problem and then propose an end-to-end pipeline of safe linear bandit algorithms that only involves solving convex problems. We also numerically evaluate the performance of our proposed methods.**

I. Introduction

Recently, bandit optimization has received significant attention in real-world cyber-physical systems that involve repeated interactions with humans. In such cases, a learner repeatedly interacts with an unknown environment. During each interaction, it selects an action from a given action set and observes its corresponding reward. The learner's goal is to maximize the accumulated reward. However, these systems are bound by safety constraints that must be respected during these interactions. Consequently, traditional bandit algorithms may not be directly applicable in these contexts. Indeed, proper and nontrivial modifications are necessary to enable the use of bandit algorithms in safety-critical systems. To achieve this, new research directions have emerged, focusing on designing constraint bandit algorithms with provable guarantees. In these settings, the environment is subject to a set of unknown operational constraints. Depending on the nature of these constraints, various constrained stochastic bandit settings have been formulated and analyzed. In our work, we concentrate on the linear stochastic bandit problem that is constrained by a set of unknown linear constraints.

A linear bandit (LB) is a variant of the multi-armed bandit (MAB) problem in which each action is associated with a feature vector x and the expected reward of playing each action is equal to the inner product of its feature vector and an unknown parameter vector θ_*. Two efficient approaches have been developed for LB: linear UCB (LUCB) [1]–[3] and linear Thompson sampling [4], [5]. A diverse body of related works on linear stochastic bandits has considered the effect of safety constraints that need to be respected during all the rounds of the algorithm. An algorithm is called stage-wise safe if the safety constraint is not violated with high probability over all rounds. Such algorithms have been proposed for for linear UCB [6] and for linear Thompson sampling [7], [8]. In the more relaxed setting, where the algorithm is allowed to violate the safety constraint for some limited rounds, [9] has proposed safe algorithms with a provable upper bound on the total number constraint violations. Our setting is inspired by the work of [10], where the agent's objective is to produce a series of policies that yield the highest expected cumulative reward, all the while maintaining that the expected cost of the policy constructed in each round stays below a specified threshold.

In this work, we investigate the computational aspects of safe linear bandit algorithms. Various methods have been developed as shown in [7], [10]–[14], which produce policies with precise performance guarantees. In this paper, we utilize convex programming tools to build a framework using these algorithms, allowing for explicit computation of policies. We aim to address two main challenges. First, standard methods require solving a non-convex optimization problem at each time step of the bandit algorithms. This poses a computational challenge, as finding a globally efficient solution for this class of problems can become NP-hard in certain cases, as noted in [15]. Second, standard algorithms necessitate optimization over a set of probability distributions. While straightforward for convex decision sets, the complexity is dependent on the form of the decision set and can pose challenges for some non-convex decision sets. Our primary contribution is an end-to-end pipeline of algorithms for constrained bandits with performance guarantees, which only involve solving convex optimization problems. This ensures computational efficiency as all the algorithms can be efficiently implemented using only a convex solver. In order to address the second aforementioned challenge, we focus on decision sets that are a union of convex sets, each described by convex inequalities.

Acknowledging the significance of our work in the broader context of real-world applications, particularly in cyber-physical systems, is crucial. Cyber-physical systems, which integrate computational algorithms with physical processes, necessitate a careful balancing of computational decisions against physical limitations. Notable applications underscore the relevance of our research, including personalized medicine, where the aim is to customize treatments to optimize patient recovery while managing constraints like dosage limits and side effects. Similarly, in cloud computing, the challenge lies in efficient resource allocation among tasks, ensuring optimal performance without resource monopolization. These examples highlight the essential role of advanced

[1] Department of Electrical Engineering, Stanford University, {afsharrad, lall}@stanford.edu. [2] Department of Electrical and Computer Engineering, University of California Santa Barbara, ahmadreza_moradipari@ucsb.edu. This work was supported by NSF under ECCS CPS project number 2125511.

algorithms in addressing constrained linear bandit problems, with implications for enhancing operational efficiency, safety, and reliability across critical sectors. Though our focus here is on the computational tools for safe linear bandits, the foundational framework we propose is poised for application across a spectrum of safety-critical systems, hinting at a vast landscape of potential future research directions and implementations.

The rest of the paper is organized as follows: Section II presents some preliminary material. In Section III we state the formal version of the problem we are addressing. In Section IV-A, we provide characteristics of an optimal policy, offering insight into what one might expect from such a policy, and propose a method to compute such a policy. Section IV-B introduces a general computationally efficient algorithm with performance guarantees to address the constrained bandit problem. In Section IV-C we propose a novel problem-dependent approach that improves the performance bound of the previous section and can achieve optimal performance for specific classes of problems. Section V presents experiments that illustrate the performance of our methods.

II. Preliminaries

Before introducing the main problem formulation and our results, we introduce a set of definitions and lemmas in this section.

Notations. For a positive integer n, the set $\{1, 2, \ldots, n\}$ is denoted by $[n]$.

For a vector $x \in \mathbb{R}^d$ and positive definite matrix $\Sigma \in \mathbb{R}^{d\times d}$ we define $\|x\|_{\Sigma,p} = \|\Sigma^{1/2}x\|_p$. In particular, in the case of $p = 2$, we have $\|x\|_{\Sigma,2} = \sqrt{x^\top \Sigma x}$.

Lemma 1 (Caratheodory's theorem)**.** *Every point in the convex hull of a set $S \subset \mathbb{R}^d$ can be expressed as a convex combination of at most $d + 1$ points from S.*

Lemma 2 (Linear program basic feasible solution)**.** *The linear program*

$$\begin{aligned} \text{maximize} \quad & c^\top x \\ \text{subject to} \quad & Ax = b, \quad x \geq 0 \end{aligned}$$

has a solution with at most p non-zero entries, where $A \in \mathbb{R}^{p\times q}$ is a fat full-rank matrix. This solution is called a basic feasible solution.

Lemma 3 (Convex hull of the union of convex sets [16])**.** *Consider the problem*

$$\begin{aligned} \text{minimize} \quad & f_0(z) \\ \text{subject to} \quad & z \in \textbf{conv}\left(\bigcup_{i=1}^{k} \mathcal{D}^i\right) \end{aligned} \tag{1}$$

where $\mathcal{D}^i = \{x : f_{ij}(x) \leq 0,\ j = 1, \cdots, k_i\}$ and each function $f_{ij} : \mathbb{R}^d \to \mathbb{R}$ is convex.

An approach to solving this problem is to solve the convex program

$$\begin{aligned} \text{minimize} \quad & f_0(z) \\ \text{subject to} \quad & \alpha_i f_{ij}(x_i/\alpha_i) \leq 0, \quad i \in [k], j \in [k_i] \\ & \mathbf{1}^\top \alpha = 1 \\ & \alpha \geq 0 \\ & z = x_1 + \cdots + x_k \end{aligned} \tag{2}$$

over the variables $z, x_1, \cdots, x_k \in \mathbb{R}^d$ and $\alpha_1, \cdots, \alpha_k \in \mathbb{R}$. If $(z^\star, x_1^\star, \cdots, x_k^\star, \alpha_1^\star, \cdots, \alpha_k^\star)$ is an optimal solution of (2), then z^* is an optimal solution of (1).

III. Problem Formulation

Initial setup. We consider the linear bandit with linear constraints characterized by the reward parameter $\theta_* \in \mathbb{R}^d$ and the cost parameter $\Gamma_* \in \mathbb{R}^{m\times d}$. In each round t, the agent is given a decision set $\mathcal{D}_t \subset \mathbb{R}^d$ from which it has to choose an action x_t. We assume that $\mathcal{D}_t$ is the union of n_t convex sets $\mathcal{D}_t^1, \cdots, \mathcal{D}_t^{n_t}$, each of which being described via convex inequalities, *i.e.*,

$$\mathcal{D}_t = \bigcup_{i=1}^{n_t} \mathcal{D}_t^i, \quad \mathcal{D}_t^i = \left\{x : f_t^{ij}(x) \leq 0,\ j = 1, \cdots, k_t^i\right\}. \tag{3}$$

Upon taking action $x_t \in \mathcal{D}_t$, the agent observes a reward signal $r_t = \theta_*^\top x_t + \eta_t^r$ and a cost signal vector $c_t = \Gamma_* x_t + \eta_t^c$, where $\eta_t^r \in \mathbb{R}$ and $\eta_t^c \in \mathbb{R}^m$ are random variables of reward and cost noise, satisfying conditions that will be specified later. The agent selects its action $x_t \in \mathcal{D}_t$ in each round t according to its policy $\pi_t \in \Delta_{\mathcal{D}_t}$ at that round, *i.e.*, $x_t \sim \pi_t$.

Objective. The objective of the agent is to generate a sequence of policies $\{\pi_t\}_{t=1}^T$ maximizing the *expected cumulative reward* over T rounds. This should be achieved while satisfying the *linear constraints*

$$\mathbb{E}_{x\sim\pi_t}(\Gamma_* x) \leq \tau, \quad \forall t \in [T], \tag{4}$$

where the ith row of Γ_* is represented by μ_{*i}. The vector $\tau \in \mathbb{R}^m$ is termed the *constraint threshold vector* and is known to the agent. Additionally, the vector inequality in (4) is interpreted element-wise.

Consequently, the policy π_t that the agent chooses in each round $t \in [T]$ must reside within the set of feasible policies defined over the action set $\mathcal{D}_t$, *i.e.*,

$$\Pi_t = \left\{\pi \in \Delta_{\mathcal{D}_t} : \mathop{\mathbb{E}}_{x\sim\pi}(\Gamma_* x) \leq \tau\right\}. \tag{5}$$

Optimizing for the maximum expected cumulative reward over T rounds can be rephrased as minimizing the constrained pseudo-regret across T rounds

$$\mathcal{R}_\Pi(\theta_*, T) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{x\sim\pi_t^*}\left(\theta_*^\top x\right) - \mathop{\mathbb{E}}_{x\sim\pi_t}\left(\theta_*^\top x\right), \tag{6}$$

where $\pi_t, \pi_t^* \in \Pi_t$ for all $t \in [T]$. Here, π_t^* signifies the *optimal feasible policy* during round t, defined as

$$\pi_t^* = \max_{\pi\in\Pi_t} \mathop{\mathbb{E}}_{x\sim\pi_t}\left[\theta_*^\top x\right]. \tag{7}$$

It is worth emphasizing that π_t^* refers to the optimal *omniscient* feasible policy, one that is achievable by an agent that is informed of the hidden parameters θ_* and Γ_*. This should be distinctly recognized from the best achievable policy by an agent observing only noisy rewards and costs.

Assumptions. We operate under the following assumptions in our setting, which are standard in the linear bandit literature.

Assumption 1. The constraint parameter matrix $\Gamma_* \in \mathbb{R}^{m\times d}$ is fat and full-rank, *i.e.*, $m < d$ and $\mathbf{rank}(\Gamma_*) = m$.

Assumption 2. For all $t \in T$, the reward and cost noise random variables η_t^r, η_t^c are conditionally R-sub-Gaussian,

$$\mathbb{E}\left[\eta_t^r|\mathcal{F}_{t-1}\right]=0, \quad \mathbb{E}\left[\exp\left(\alpha\eta_t^r\right)|\mathcal{F}_{t-1}\right] \le \exp\left(\alpha^2R^2/2\right),$$
$$\mathbb{E}\left[\eta_{t,i}^c|\mathcal{F}_{t-1}\right]=0, \quad \mathbb{E}\left[\exp\left(\alpha\eta_{t,i}^c\right)|\mathcal{F}_{t-1}\right] \le \exp\left(\alpha^2R^2/2\right)$$

for any $\alpha \in \mathbb{R}, i \in [m]$, where $\mathcal{F}_t$ is the filtration that includes all events $(x_{1:t+1}, \eta_{1:t}^r, \eta_{1:t}^c)$ up to round t.

Assumption 3. There is a known constant $S > 0$, such that $\|\theta_*\| \le S$ and $\|\mu_{i*}\| \le S^2$ for all $i \in [m]$.

Assumption 4. The decision set $\mathcal{D}_t$ is bounded. Specifically, $\max_{t\in[T]} \max_{x\in\mathcal{D}_t} \|x\| \le L$.

Assumption 5. For all $t \in [T]$ and $x \in \mathcal{D}_t$, the mean rewards and costs are bounded, *i.e.*, $\theta_*^\top x \in [0,1]$ and $\mu_{i*}^\top x \in [0,1]$ for $i \in [m]$.

Assumption 6. There exists a universally safe action $x_0 \in \mathcal{D}_t$ for all $t \in [T]$ associated with the cost vector $c_0 \in \mathbb{R}^m$. This means that $\Gamma_* x_0 = c_0 < \tau$. For the sake of clarity, we assume that $c_0 = 0$ and that its value is known. Extending this to the cases where $c_0 \neq 0$ is known, or c_0 is unknown, is straightforward. See [17] for further details on these scenarios.

Summary. To summarize, the problem data includes the reward vector θ_*, the constraint matrix Γ_*, the constraint threshold vector τ, the problem horizon T, the observation noise sub-Gaussian parameter R, the reward and cost upper bound parameter S, the known safe action x_0, and the decision sets $\mathcal{D}_1, \ldots, \mathcal{D}_T$, where each $\mathcal{D}_t$ is characterized by a set of integers $n_t, k_t^1, \ldots, k_t^{n_t}$ and a set of convex functions $f_t^{i,j}$ with $i \in [n_t], j \in [k_t^i]$.

Note that we are working within the specified class of decision sets, *i.e.*, sets in the form of a union of convex sets each described by convex inequalities, exclusively for computational purposes. Nevertheless, it is important to highlight that our theoretical results and theorems remain valid for any arbitrary choice of decision sets.

IV. Main Results

A. The optimal feasible policy

At each time step t, the optimal feasible policy π_t^* is obtained by solving the following optimization problem:

$$\begin{aligned} \underset{\pi\in\Delta_{\mathcal{D}_t}}{\text{maximize}} \quad & \underset{x\sim\pi}{\mathbb{E}}\left(\theta_*^\top x\right) \\ \text{subject to} \quad & \underset{x\sim\pi}{\mathbb{E}}\left(\Gamma_* x\right) \le \tau \end{aligned} \tag{8}$$

While the reward and cost parameters θ_* and Γ_* are unknown in the bandit setting, it is valuable to understand the structure of the optimal feasible policy π_t^* even when these parameters are known. Specifically, the optimization in (8) considers probability distributions over the decision set $\mathcal{D}_t$, and since $\mathcal{D}_t$ can be any arbitrary set, characterizing the optimal feasible policy can be a complex task. The subsequent theorem, an extension of Lemma 5 in [17], provides a characterization of the optimal feasible policy π_t^*.

Theorem 1. *There exists an optimal feasible policy π_t^* that solves (8) with finite support of at most $m + 1$ elements.*

Proof: First, observe that while (8) is an optimization over all choices of distributions $\pi \in \mathcal{D}_t$, the only component of π that plays a role in the optimization is $\mathbb{E}_{x\sim\pi}(x)$. Thus, letting $z = \mathbb{E}_{x\sim\pi}(x)$, solving (8) is equivalent to first solving

$$\begin{aligned} \underset{z}{\text{maximize}} \quad & \theta_*^\top z \\ \text{subject to} \quad & \Gamma_* z \le \tau \\ & z \in \mathbf{conv}(\mathcal{D}_t) \end{aligned} \tag{9}$$

to find a solution z^*, and then find a distribution $\pi_t^* \in \Delta_{\mathcal{D}_t}$ such that $\mathbb{E}_{x\sim\pi_t^*}(x) = z^*$. Note that the constraint $z \in \mathbf{conv}(\mathcal{D}_t)$ has to be included in the new optimization problem since if $z^* \notin \mathbf{conv}(\mathcal{D}_t)$, then there is no distribution $\pi \in \Delta_{\mathcal{D}_t}$ whose expected value is z^*.

Now, let z^* be the solution of (9). Since $z \in \mathbf{conv}(\mathcal{D}_t)$, we know that z is given by a convex combination of a finite number of elements in $\mathcal{D}_t$. Moreover, according to Caratheodory's theorem presented in Lemma 1, one such convex combination exists with at most $d + 1$ points. Thus, a set of points $z_1, \cdots, z_{d+1} \in \mathcal{D}_t$ and a set of non-negative scalars $\alpha_1, \cdots, \alpha_{d+1}$ exist such that $z^* = \sum_{i=1}^{d+1} \alpha_i z_i = Z\alpha$ and $\sum_{i=1}^{d+1} \alpha_i = 1$, where $Z \in \mathbb{R}^{d\times(d+1)}$ is a matrix whose ith column is z_i and $\alpha \in \mathbb{R}^{d+1}$ is a vector whose i entry is α_i. Next, we form the following optimization problem:

$$\begin{aligned} \underset{\beta\in\mathbb{R}^{d+1}}{\text{maximize}} \quad & \theta^\top Z\beta \\ \text{subject to} \quad & \Gamma Z\beta \le \tau \\ & \mathbf{1}^\top \beta = 1, \quad \beta \ge 0 \end{aligned} \tag{10}$$

Note that if β is a solution of (10), then $z_\beta = Z\beta$ is a solution of (9). The final step would be to show that a specific solution β^* for (10) exists with at most $m+1$ non-zero entries. This step is taken via Lemma 2, according to which (10) has a basic feasible solution that has no more than $m+1$ non-zero elements. Note that (10) can be converted to the form given by Lemma 2 by adding slack variables. Now, letting β^* be a basic feasible solution of (10), the optimal feasible policy π_t^* with a support of at most $m + 1$ elements is given by

$$\underset{x\sim\pi_t^*}{\mathbb{P}}(x = z) = \begin{cases} \beta_i^* & z = z_i \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where z_i is the ith column of Z. This completes the proof.

The proof of Theorem 1 provides a straightforward algorithm to compute the optimal feasible policy π_t^* given θ_* and Γ_*. Algorithm 1 provides the steps to achieve this goal.

Algorithm 1 Computation of the optimal feasible policy

```
Input: $\theta_* \in \mathbb{R}^d, \Gamma_* \in \mathbb{R}^{m \times d}, \tau \in \mathbb{R}^d_+$
1: Solve (9) and find $z^* \in \mathbf{conv}(\mathcal{D}_t)$
2: Find $Z = \begin{bmatrix} z_1 \cdots z_{d+1} \end{bmatrix} \in \mathbb{R}^{d \times (d+1)}$ and $\alpha \in \mathbb{R}^d_+$ such that $z^* = Z\alpha$ and $\mathbf{1}^\top \alpha = 1$
3: Find $\beta^*$, a basic feasible solution of (10)
4: return $\pi_t^*$ according to (11)
```

With the decision set described in (3), lines 1 and 2 of the algorithm can be implemented simultaneously using the result of Lemma 3. According to Lemma 3, this can be done by solving the convex optimization problem

$$\begin{array}{ll} \text{minimize} & \theta^\top z \\ \text{subject to} & \Gamma z \leq \tau \\ & \alpha_i f_{ij}(x_i/\alpha_i) \leq 0, \quad i \in [k], j \in [k_i] \\ & \mathbf{1}^\top \alpha = 1, \quad \alpha \geq 0 \\ & z = x_1 + \cdots + x_k \end{array} \tag{12}$$

and finding the optimal $z^* = \sum_{i=1}^k \alpha_i z_i$, where $z_i = x_i/\alpha_i$ and x_i, α_i are solutions of (12). Note that in this case, instead of expressing z^* in terms of at most $d+1$ points, it is expressed in terms of k points. Based on how d and k compare, this can be a computational advantage or disadvantage. However, it does not affect the overall flow of Algorithm 1 as all the steps can be implemented and the only difference is that $d+1$ gets substituted by k.

While we have addressed the implementation issue in line 2 of Algorithm 1 for a special case, we do not have a general computationally efficient method to implement it without further knowledge of the set $\mathcal{D}_t$ and its representation.

Finally, line 3 of Algorithm 1 can be implemented using the Simplex method, and line 4 is constructed based on the output of line 3, which concludes our full algorithmic pipeline to calculate the optimal feasible policy π_t^*.

B. Computationally-tractable algorithms with performance guarantees for linear bandits with linear constraints

In the literature, there are numerous formulations of linearly-constrained linear bandits [6], [7], [10], [12], [18], [19]. Many associated algorithms [7], [10], [18], [19] follow similar strategies. Specifically, they establish confidence regions for both reward and cost parameters. These algorithms strive to optimistically maximize the reward, while taking a pessimistic stance regarding the cost. This means they account for the worst-case scenario that the cost parameter corresponds to the least favorable value within the confidence region.

In this section, we discuss the Optimistic-Pessimistic Linear Bandit (OPLB) Algorithm introduced by [17], which serves as our foundational algorithm. We elucidate its workings, identify computational barriers, and tackle these challenges by introducing computationally-tractable algorithms backed by performance guarantees. It is worth noting that, although our solutions are tailored to a specific formulation of the linearly constrained linear bandit problem, they can be readily extended to other formulations, given that they all encounter the same computational challenge.

Consider a linear bandit with linear constraints as described in III. For simplicity we assume $m = 1$. Consequently, the constraint matrix $\Gamma_* \in \mathbb{R}^{m \times d}$ simplifies to a row vector, which we denote by $\mu_*^\top \in \mathbb{R}^d$. This implies that only one linear constraint, $\mu_*^\top x \leq \tau$, is present. Extending this to the general case with m constraints is straightforward.

At each round $t \in [T]$, given the past actions $\{x_i\}_{i=1}^{t-1}$, observed rewards $\{r_i\}_{i=1}^{t-1}$, and cost signals $\{c_i\}_{i=1}^{t-1}$, we construct the Gram matrix

$$\Sigma_t = \lambda I + \sum_{i=1}^{t-1} x_i x_i^\top. \tag{13}$$

Then we compute the ℓ_2-regularized least squares estimates of θ_* and μ_* using the regularization parameter λ. These are given by

$$\widehat{\theta}_t = \Sigma_t^{-1} \sum_{i=1}^{t-1} r_i x_i, \qquad \widehat{\mu}_t = \Sigma_t^{-1} \sum_{i=1}^{t-1} c_i x_i. \tag{14}$$

As suggested by OPLB, we construct the confidence sets

$$\begin{aligned} \mathcal{C}^\theta_{t,\ell_2} &= \left\{ \theta \in \mathbb{R}^d : \left\| \theta - \widehat{\theta}_t \right\|_{\Sigma_t,2} \leq \rho\beta_t \right\}, \\ \mathcal{C}^\mu_{t,\ell_2} &= \left\{ \mu \in \mathbb{R}^d : \|\mu - \widehat{\mu}_t\|_{\Sigma_t,2} \leq \beta_t \right\}, \end{aligned} \tag{15}$$

where $\rho = 1 + \frac{2}{\tau - c_0}$, $\beta_t = R\sqrt{d \log \frac{1+(t-1)L^2/\lambda}{\delta}} + \sqrt{\lambda} S$, and $\|.\|_{\Sigma_t,2}$ is defined in Sectin II.

According to the principal theorem presented in [20], with probability at least $1-\delta$, the unidentified parameters θ_* and μ_* are contained within the sets $\mathcal{C}^\theta_{t,\ell_2}$ and $\mathcal{C}^\mu_{t,\ell_2}$, respectively.

The final step of OPLB is to solve the problem

$$\begin{array}{ll} \underset{\pi \in \Delta_{\mathcal{D}_t}, \theta \in \mathbb{R}^d}{\text{maximize}} & \underset{x \sim \pi}{\mathbb{E}} \left(\theta^\top x \right) \\ \text{subject to} & \theta \in \mathcal{C}^\theta_{t,\ell_2} \\ & \pi \in \Pi_t, \end{array} \tag{16}$$

where

$$\Pi_t = \{ \pi \in \Delta_{\mathcal{D}_t} : \underset{x \sim \pi}{\mathbb{E}} \left(\mu^\top x \right) \leq \tau, \ \forall \mu \in \mathcal{C}^\mu_{t,\ell_2} \} \tag{17}$$

is the pessimistic set of safe policies.

Proposition 1. *The optimization problem* (16) *is equivalent to*

$$\begin{array}{ll} \underset{z \in \mathbb{R}^d}{\text{maximize}} & \rho\beta_t \sqrt{z^\top \Sigma_t z} + \widehat{\theta}_t^\top z \\ \text{subject to} & \beta_t \sqrt{z^\top \Sigma_t z} + \widehat{\mu}_t^\top z \leq \tau \\ & z \in \mathbf{conv}(\mathcal{D}_t). \end{array} \tag{18}$$

Proof: First, we define $z = \mathbb{E}_{x \sim \pi}(x)$. Instead of tackling an optimization problem over a set of probability distributions, we aim to find the expected value. This step needs the condition $z \in \mathbf{conv}(\mathcal{D}_t)$. This reasoning follows the same lines as the proof of Theorem 1. The remainder of the proof, which explains the specific forms of the objective function and the constraint, directly stems from Proposition 1 in [17].

Once equation (18) is solved and the optimal solution $z^* = \sum_{i=1}^{d+1} \alpha_i z_i$ is identified as a convex combination of elements from $\mathcal{D}_t$, the optimal feasible policy π_t^* is expressed by

$$\mathop{\mathbb{P}}_{x\sim\pi_t^*}(x = z) = \begin{cases} \alpha_i & \text{if } z = z_i \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

The following theorem, a central result from [17], offers a regret bound on the algorithm's performance.

Theorem 2 (Theorem 2 of [17])**.** *Assuming the conditions presented in the problem formulation of Section III are satisfied, the regret of OPLB, with a probability greater than $1 - 2\delta$, is bounded by*

$$\begin{aligned}\mathcal{R}_\Pi(\theta_*, T) \leq {} & \frac{2L(\rho+1)\beta_T}{\sqrt{\lambda}}\sqrt{2T\log(1/\delta)} \\ & + (\rho+1)\beta_T\sqrt{2Td\log(1+TL^2/\lambda)}.\end{aligned} \tag{20}$$

While the outlined approach offers a comprehensive pipeline to tackle the constrained bandit problem, a primary obstacle arises from the computational complexity of solving the main optimization problem (16) or its equivalent (18). As noted in [15], the unconstrained variant of this problem, with a decision set that is represented as a polytope defined by the intersection of halfspaces, is NP-hard. This implies that searching for a universally applicable computational technique, irrespective of the decision set's nature, may be futile. Furthermore, as elaborated in Section IV-A, optimizing over probability distributions (or equivalently, with the constraint $z \in \textbf{conv}(\mathcal{D}_t)$) introduces its own set of challenges.

To navigate the first challenge, we propose a modified OPLB that, while computationally feasible, yields a more relaxed regret bound. This modification ensures a universally efficient algorithm. Later in Section IV-C, we present an alternative technique for addressing the original problem (18), which is suitable for specific cases but not universally. To tackle the second challenge, analogous to Section IV-A, we utilize the technique introduced in Lemma 3.

To make the OPLB more computationally efficient, we modify the confidence sets. Instead of using the confidence set C^θ_{t,ℓ_2} presented in (15), we switch to a confidence set using the ℓ_1 norm and an adjusted radius. Specifically, we define the confidence set as

$$\mathcal{C}^\theta_{t,\ell_1} = \left\{\theta \in \mathbb{R}^d : \left\|\theta - \widehat{\theta}_t\right\|_{\Sigma_t,1} \leq \rho\sqrt{d}\beta_t\right\}, \tag{21}$$

where ρ and β_t retain their previous definitions and $\|.\|_{\Sigma,1}$ is detailed in Section II. Note that, as will be shown in a subsequent lemma, an ℓ_1 confidence set for μ_* is unnecessary. Instead, we can continue using the $\mathcal{C}^\mu_{t,\ell_2}$ as previously defined.

Lemma 4. *For any $t \in [T]$ and any $\delta > 0$, the following holds:*

$$\mathbb{P}\left(\theta_* \in \mathcal{C}^\theta_{t,\ell_1}\right) \geq 1 - \delta. \tag{22}$$

Proof: For any vector $x \in \mathbb{R}^d$, we have that $\|x\|_1 \leq \sqrt{d}\|x\|_2$. This yields

$$\left\|\Sigma^{1/2}\left(\theta - \widehat{\theta}\right)\right\|_1 \leq \sqrt{d}\left\|\Sigma^{1/2}\left(\theta - \widehat{\theta}\right)\right\|_2.$$

Given that the right-hand side is bounded by $\sqrt{d}\rho\beta_t$ for any $\theta \in \mathcal{C}^\theta_{t,\ell_2}$, it follows that $\mathcal{C}^\theta_{t,\ell_2} \subseteq \mathcal{C}^\theta_{t,\ell_1}$. By the main theorem of [20], we know that $\theta_* \in \mathcal{C}^\theta_{t,\ell_2}$ with a probability of at least $1 - \delta$, which concludes the proof.

In the modified version of OPLB that incorporates the ℓ_1 confidence region, we address a new problem given by

$$\begin{aligned} \underset{z\in\mathbb{R}^d,\theta\in\mathbb{R}^d}{\text{maximize}} \quad & \theta^\top z \\ \text{subject to} \quad & \theta \in \mathcal{C}^\theta_{t,\ell_1} \\ & z \in S_t \\ & z \in \textbf{conv}(\mathcal{D}_t), \end{aligned} \tag{23}$$

where $S_t = \{z \in \mathbb{R}^d : \mu^\top z \leq \tau, \forall \mu \in \mathcal{C}^\mu_{t,\ell_2}\}$. Once this problem is solved and the optimal solution $z^* = \sum_{i=1}^{d+1} \alpha_i z_i$ is identified as a convex combination of elements from $\mathcal{D}_t$, the optimal feasible policy π_t^* is given by (19).

Proposition 2. *The optimization problem (23) can be decomposed and solved by addressing $2d$ individual convex optimization problems.*

Proof: We can express (23) in the following format:

$$\begin{aligned} \underset{\theta\in\mathbb{R}^d}{\text{maximize}} \quad & f(\theta) \\ \text{subject to} \quad & \theta \in \mathcal{C}^\theta_{t,\ell_1}, \end{aligned} \tag{24}$$

where the function f is defined as:

$$\begin{aligned} f(\theta) = \quad \max_{z\in\mathbb{R}^d} \quad & \theta^\top z \\ \text{s.t.} \quad & z \in S_t \\ & z \in \textbf{conv}(\mathcal{D}_t). \end{aligned} \tag{25}$$

Given that f is convex and the region $\mathcal{C}^\theta_{t,\ell_1}$ forms a polytope in $\mathbb{R}^d$, we realize that the solutions of (24) occur at the vertices of the polytope, and it suffices to evaluate f at the $2d$ vertices to solve this problem. Each evaluation corresponds to solving a convex optimization problem as shown in (25), which completes the proof.

Proposition 2 demonstrates that the modified OPLB can be efficiently solved. The subsequent step is to ascertain a guarantee for the regret bound. The theorem below provides this guarantee.

Theorem 3 (Modified OPLB regret bound)**.** *Given that the conditions outlined in Section III are met, the regret of the modified OPLB employing the ℓ_1 confidence region for the reward parameter θ_*, with a probability exceeding $1 - 2\delta$, can be upper-bounded as*

$$\begin{aligned}\mathcal{R}_\Pi(\theta_*, T) \leq {} & \frac{2L(\rho+1)\beta_T}{\sqrt{\lambda}}\sqrt{2Td\log(1/\delta)} \\ & + (\rho+1)\beta_T d\sqrt{2T\log(1+TL^2/\lambda)}.\end{aligned} \tag{26}$$

Proof: By examining the proof of Theorem 2, it becomes apparent that the regret bound depends on the confidence region radius of the reward parameter θ_*, namely $\rho\beta_T$, without specifically relying on the value of $\rho\beta_T$. Further inspection reveals that the confidence region radius of the cost parameter μ_* has no bearing on the bound. In the modified OPLB approach, the initial radius is scaled by a factor of $\sqrt{d}$, while the latter remains unchanged. Hence, in the expression (20), substituting β_t with $\sqrt{d}\beta_t$ results in the updated bound presented in (26), which completes the proof.

With Theorem 3, we now possess a comprehensive framework for tackling the constrained bandit problem using algorithms that are computationally efficient. It's important to highlight that a key step in this process is the evaluation of the function f as defined in (25). Although this is a convex optimization problem, one cannot overlook that its two constraints, in their most general form, may introduce complications unless they are further simplified.

The primary constraint, $z \in S_t$, can be replaced by the more direct constraint $\beta_t\sqrt{z^\top \Sigma_t z} + \widehat{\mu}^\top z \leq \tau$, following the guidelines of Proposition 1. The latter constraint, $z \in \mathbf{conv}\,(\mathcal{D}_t)$, while complicated in general, is navigated for the class of decision sets studied in this work using the technique introduced in Lemma 3. Consequently, the task of evaluating the function f from (25) simplifies to solving

$$\begin{aligned} f(\theta) = \max_{z\in\mathbb{R}^d} \quad & \theta^\top z \\ \text{s.t.} \quad & \beta_t\sqrt{z^\top \Sigma_t z} + \widehat{\mu}_t^\top z \leq \tau \\ & \alpha_i f_{ij}(x_i/\alpha_i) \leq 0, \quad i \in [k], j \in [k_i] \\ & \mathbf{1}^\top \alpha = 1, \quad \alpha \geq 0, \\ & z = x_1 + \cdots + x_k. \end{aligned} \tag{27}$$

The procedure is concisely summarized in Algorithm 2.

Algorithm 2 Modified OPLB

```
Input: $T \in \mathbb{N}$, $\delta \in \mathbb{R}_+$, $\gamma \in \mathbb{R}_+$, $\tau \in \mathbb{R}$
for $t = 1$ to $T$ do
    Observe $r_t, c_t$ and compute $\widehat{\theta}_t$ and $\widehat{\mu}_t$ using (14)
    $z^* \leftarrow$ Solve (23) using Proposition 2 and (27)
    Construct $\pi_t$ using $z^*$ according to (19)
    Play the action $x_t \sim \pi_t$
end for
```

C. The upper bound maximization method

Recall (18) which presents the original problem with ℓ_2 confidence sets that we initially sought to solve. Since solving this problem is challenging, our first approach was to present an ℓ_1 relaxation to this problem, as discussed in Section IV-B. In this section we introduce a problem-dependent method that has the potential to exactly solve (18). The following theorem provides the tools that we need for this method.

Theorem 4. *Let $g_1, g_2 : \mathbb{R}^d \to \mathbb{R}$ be arbitrary functions and let $C \subseteq \mathbb{R}^d$ be an arbitrary set. Consider the optimization problems*

$$\begin{aligned} \underset{z\in\mathbb{R}^d}{\text{maximize}} \quad & g_1(z) \\ \text{subject to} \quad & g_1(z) \leq g_2(z) \\ & z \in C \end{aligned} \tag{28}$$

and

$$\begin{aligned} \underset{z\in\mathbb{R}^d}{\text{maximize}} \quad & g_2(z) \\ \text{subject to} \quad & g_1(z) \leq g_2(z) \\ & z \in C. \end{aligned} \tag{29}$$

If z^* is an optimal solution for (29) and $g_1(z^*) = g_2(z^*)$, then z^* is also an optimal solution for (28).

Proof: Suppose $\tilde{z}$ is an optimal solution for (28) and z^* is not. Then, $g_1(\tilde{z}) > g_1(z^*) = g_2(z^*) \geq g_2(\tilde{z})$. The first inequality stems from the optimality of $\tilde{z}$ and the non-optimality of z^* for (28), the equality follows directly from the assumption of the theorem, and the last inequality arises because z^* maximizes g_2. This leads to $g_1(\tilde{z}) > g_2(\tilde{z})$, a violation of the constraint $g_1(z) \leq g_2(z)$, thus a contradiction. This concludes that z^* is an optimal solution for (28).

This theorem allows us to solve (29) instead of (28). If the condition $g_1(z^*) = g_2(z^*)$ holds, then we have an optimal solution for (28) as well. This may be quite useful if (29) is more tractable than (28).

We now apply the result of Theorem 4 to (18). For clarity, we restate this problem as follows:

$$\begin{aligned} \underset{z\in\mathbb{R}^d}{\text{maximize}} \quad & \rho\beta_t\sqrt{z^\top \Sigma_t z} + \hat{\theta}_t^\top z \\ \text{subject to} \quad & \beta_t\sqrt{z^\top \Sigma_t z} + \hat{\mu}_t^\top z \leq \tau \\ & z \in \mathbf{conv}(\mathcal{D}_t). \end{aligned} \tag{30}$$

Setting $g_1(z) = \rho\beta_t\sqrt{z^\top \Sigma_t z} + \hat{\theta}_t^\top z$, $g_2(z) = \rho\tau + (\hat{\theta}_t - \rho\hat{\mu}_t)^\top z$, and $C = \mathbf{conv}(\mathcal{D}_t)$, (30) becomes a particular instance of (28). Consequently, the counterpart of (29) in our setting is

$$\begin{aligned} \underset{z\in\mathbb{R}^d}{\text{maximize}} \quad & \rho\tau + \left(\hat{\theta}_t - \rho\hat{\mu}_t\right)^\top z \\ \text{subject to} \quad & \beta_t\sqrt{z^\top \Sigma_t z} + \hat{\mu}_t^\top z \leq \tau \\ & z \in \mathbf{conv}(\mathcal{D}_t), \end{aligned} \tag{31}$$

which is a convex optimization problem. This provides a potentially more efficient approach to solve the original OPLB problem with ℓ_2 confidence sets and yield exact solutions. It involves solving (31), a convex optimization problem amenable to efficient computation. Upon solving this problem, one must check whether the first constraint is active. If it is, then the obtained solution also solves (30). If not, the process shifts back to addressing the ℓ_1 version of the problem, as outlined in (23) or (27). To handle the second constraint in (31), we utilize the technique proposed in Lemma 3. Algorithm 3 summarizes the entire methodology. We refer to this technique as the *Upper Bound Maximization*

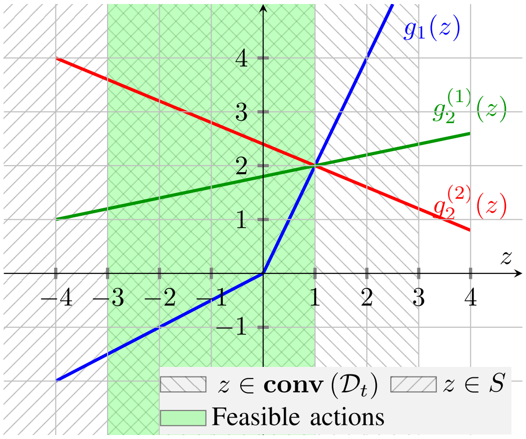

Fig. 1 Example of the upper bound maximization test

(UBM) method, as it entails maximizing an upper bound on the objective function rather than the objective function itself.

Algorithm 3 Enhanced OPLB with UBM

```
    Input: $T \in \mathbb{N}$, $\delta \in \mathbb{R}_+$, $\gamma \in \mathbb{R}_+$, $\tau \in \mathbb{R}$
for $t = 1$ to $T$ do
    Observe $r_t, c_t$, and compute $\hat{\theta}_t$, $\hat{\mu}_t$ using (14)
    $z^* \leftarrow$ Solve (31)
    if $\beta_t\sqrt{z^{*\top}\Sigma_t z^*} + \hat{\mu}_t^\top z^* < \tau$ then
        $z^* \leftarrow$ Solve (23) using Proposition 2 and (27)
    end if
    Construct $\pi_t$ using $z^*$ according to (19)
    Play the action $x_t \sim \pi_t$
end for
```

Each iteration of Algorithm 3 involves solving either the ℓ_2 or the ℓ_1 confidence set problem. Thus, the ultimate regret bound will be no worse than that provided by Theorem 3 but may approach the bound of Theorem 2, depending on the frequency at which the first constraint becomes active in (31).

Example. Figure 1 illustrates a one-dimensional example of our setup. The decision set $\mathcal{D}_t$ is defined such that $\mathbf{conv}\,(\mathcal{D}_t) = \{z : |z| \leq 3\}$. Two distinct upper bound functions, $g_2^{(1)}(z)$ and $g_2^{(2)}(z)$, are introduced, each corresponding to a different value of $\widehat{\mu}_t$. The set S represents the points where the safety constraint $g_1(z) \leq g_2(z)$, as described in (30) and (31), is satisfied.

The implications of Theorem 4 are observable in Figure 1, where the conditions under which UBM is effective become apparent. Specifically, when the upper bound is described by $g_2^{(1)}(z)$, maximizing this function also optimizes the original objective $g_1(z)$, with the constraint $g_1(z) \leq g_2(z)$ becoming active at the optimum. Conversely, when the upper bound is $g_2^{(2)}(z)$, UBM does not lead to an optimal solution, as maximizing $g_2^{(2)}(z)$ does not make the constraint active, rendering the approach ineffective in this case.

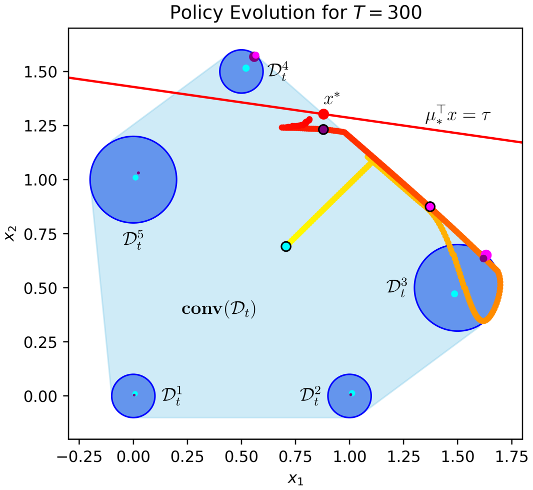

Fig. 2 Mean policy trajectory with $\mathcal{D}_t$ as a union of convex sets

V. Experiments

In this section, we present empirical evaluations of the proposed algorithms through two distinct experiments.

A. *Enhanced OPLB policy evaluation with non-convex decision sets*

The first experiment considers a two-dimensional scenario with a non-convex decision set represented by a union of five disks in $\mathbb{R}^2$, all subject to a single linear constraint. Figure 2 illustrates the policies chosen by the algorithm at each time step over a total of $T = 300$ rounds. The trajectory depicting the mean value of the policy is shown in Figure 2, which transitions from yellow to red as time progresses. Notably, at three specific time steps—$t = 0$, $t = 40$, and $t = 100$—the mean policy values are highlighted in cyan, magenta, and purple respectively, each delineated with a black border. Corresponding to each of these mean values, five points are plotted, representing the five potential actions, one of which is to be randomly selected according to a specific probability for the policy to be effective. The radius of each point is proportional to its probability weight in the policy's construction, with all weights summing up to one. Furthermore, the constraint boundary, defined by $x^\top\mu = \tau$, is represented as a line within the figure, and the mean value of the optimal policy is denoted as x^*.

Observations from the figure reveal that initially, the trajectory of the points moves along the boundary of the convex hull of the decision set and away from the optimal policy. However, as time progresses, the trajectory redirects towards the optimal policy and ultimately converges to the optimal solution. Furthermore, the mean value of the policy always remains within the safe region, indicating that the pessimism in action selection has been effective, ensuring that the algorithm does not violate the safety constraint at any point.

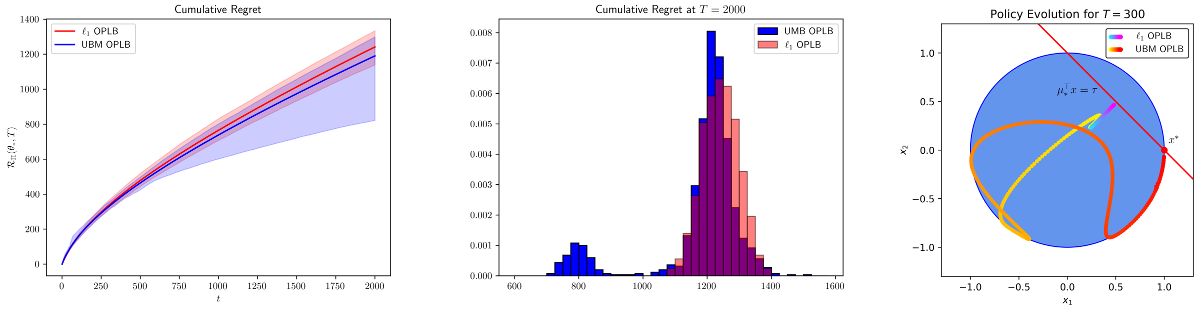

Fig. 3 Left: Cumulative regret of ℓ_1 OPLB vs. OPLB with UBM. Middle: Histogram of cumulative regret at $T = 2000$. Right: Mean policy trajectory of ℓ_1 OPLB vs. UBM OPLB

B. Cumulative regret comparison

In the second experiment, we compare the cumulative regrets of Algorithms 2 and 3, namely the ℓ_1 OPLB and UBM OPLB. Figure 3 (left) presents the cumulative regret of both algorithms given the parameters $\theta_* = [3, 2.5]^\top$, $\mu_* = [0.5, 0.5]^\top$, and $\tau = 0.5$, with the decision set being the unit disk. The results indicate a marginally better cumulative regret for UBM OPLB. This plot reveals an interesting phenomenon: asymmetric confidence bands around the UBM OPLB's regret, with a lower confidence band that is notably further below the mean compared to the upper band. Further investigation into this observation is conducted by examining Figure 3 (middle), which displays a histogram of the cumulative regrets for both algorithms at time $t = 2000$ over $N = 1000$ simulations. The histogram suggests that, although UBM OPLB's performance is largely in line with that of ℓ_1 OPLB, it exhibits a secondary mode where the cumulative regret is substantially lower. This accounts for the observed lower confidence band in the first plot. In certain cases, UBM OPLB significantly outperforms ℓ_1 OPLB. For a closer look at this behavior, we examine the mean policy trajectories of ℓ_1 OPLB and UBM OPLB under the aforementioned superior performance. Figure 3 (right) delineates these trajectories with the evolution from yellow to red and cyan to magenta, respectively, for a span of $T = 300$ steps. Clearly, ℓ_1 OPLB does not approach the optimal policy as closely as UBM OPLB, resulting in greater regret, whereas UBM OPLB tends toward the optimal policy, exhibiting minimal regret. Although this phenomenon is problem-specific and not universally observed, it presents an intriguing aspect for further research.

References

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

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

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

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

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

[6] S. Amani, M. Alizadeh, and C. Thrampoulidis, "Linear stochastic bandits under safety constraints," in *Advances in Neural Information Processing Systems*, 2019, pp. 9252–9262.

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

[8] A. Moradipari, M. Alizadeh, and C. Thrampoulidis, "Linear thompson sampling under unknown linear constraints," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3392–3396.

[9] T. Chen, A. Gangrade, and V. Saligrama, "Strategies for safe multi-armed bandits with logarithmic regret and risk," in *International Conference on Machine Learning*. PMLR, 2022, pp. 3123–3148.

[10] A. Pacchiano, M. Ghavamzadeh, P. Bartlett, and H. Jiang, "Stochastic bandits with linear constraints," in *International conference on artificial intelligence and statistics*. PMLR, 2021, pp. 2827–2835.

[11] K. N. Varma, S. Lale, and A. Anandkumar, "Stochastic linear bandits with unknown safety constraints and local feedback," in *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023.

[12] K. Khezeli and E. Bitar, "Safe linear stochastic bandits," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 06, 2020, pp. 10 202–10 209.

[13] S. Hutchinson, B. Turan, and M. Alizadeh, "The impact of the geometric properties of the constraint set in safe optimization with bandit feedback," in *Learning for Dynamics and Control Conference*. PMLR, 2023, pp. 497–508.

[14] T. Chen, A. Gangrade, and V. Saligrama, "A doubly optimistic strategy for safe linear bandits," *arXiv preprint arXiv:2209.13694*, 2022.

[15] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," in *Annual Conference Computational Learning Theory*, 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:9134969

[16] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[17] A. Pacchiano, M. Ghavamzadeh, P. Bartlett, and H. Jiang, "Stochastic bandits with linear constraints," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Banerjee and K. Fukumizu, Eds., vol. 130. PMLR, 13–15 Apr 2021, pp. 2827–2835. [Online]. Available: https://proceedings.mlr.press/v130/pacchiano21a.html

[18] A. Moradipari, C. Thrampoulidis, and M. Alizadeh, "Stage-wise conservative linear bandits," *Advances in neural information processing systems*, vol. 33, pp. 11 191–11 201, 2020.

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

[20] Y. Abbasi-yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds., vol. 24. Curran Associates, Inc., 2011.