The Impact of the Geometric Properties of the Constraint Set in Safe Optimization with Bandit Feedback

Spencer Hutchinson SHUTCHINSON@UCSB.EDU
University of California, Santa Barbara

Berkay Turan BTURAN@UCSB.EDU
University of California, Santa Barbara

Mahnoosh Alizadeh ALIZADEH@UCSB.EDU
University of California, Santa Barbara

Editors: N. Matni, M. Morari, G. J. Pappas

Abstract

We consider a safe optimization problem with bandit feedback in which an agent sequentially chooses actions and observes responses from the environment, with the goal of maximizing an arbitrary function of the response while respecting stage-wise constraints. We propose an algorithm for this problem, and study how the geometric properties of the constraint set impact the regret of the algorithm. In order to do so, we introduce the notion of the *sharpness* of a particular constraint set, which characterizes the difficulty of performing learning within the constraint set in an uncertain setting. This concept of sharpness allows us to identify the class of constraint sets for which the proposed algorithm is guaranteed to enjoy sublinear regret. Simulation results for this algorithm support the sublinear regret bound and provide empirical evidence that the sharpness of the constraint set impacts the performance of the algorithm.

Keywords: Safe Learning, Bandits, Optimization

1. Introduction

As contemporary learning and control paradigms expand into domains with stringent safety requirements, the need for control mechanisms that can provide such safety guarantees has grown significantly. This has resulted in a plethora of novel safe learning problems in the literature through the lens of model predictive control Koller et al. (2018); Hewing et al. (2020); Chen et al. (2022), reinforcement learning Junges et al. (2016); Garcıa and Fernández (2015), optimization Usmanova et al. (2019); Fereydounian et al. (2020), bandits Sui et al. (2015), and many others. Such problems are well suited for applications in which the control algorithm interacts with humans, which introduces uncertainties that need to be considered to ensure safety (e.g., clinical trials Sui and Burdick (2017), robotic systems Berkenkamp et al. (2021), and resource allocation in societal infrastructure through pricing Hutchinson et al. (2022)).

In this work, we are interested in a sequential decision making problem, where the decisions must be within an arbitrary and unknown compact safety set. We consider a safe optimization framework with bandit feedback, where the reward and the constraint set are known non-linear functions of the matrix multiplication of the action with an unknown parameter. Compared to the existing literature, this problem is uniquely challenging because (1) both the decisions and the feedback from the environment are vectors, (2) the reward is an arbitrary function of the decision vector,

and (3) the safety constraint on the decision vector is an arbitrary compact set. These challenges are however warranted, given that problems of this form appear in many real-world applications. For example, power flow constraints are nonlinear and nonconvex in general (Molzahn et al. (2017)) and often solved with (nonlinear) convex relaxations (e.g. Bai et al. (2008); Farivar and Low (2013)).

We handle the challenge general safety sets present by introducing a geometric property of a set, which we call *sharpness*, that is related to how difficult it is to perform learning within a particular safety set. This allows us to characterize the performance of our learning algorithm, measured in terms of regret, as a function of the sharpness of the safety set. Accordingly, we identify the class of safety sets (that includes all convex sets) for which we can establish a sublinear regret bound.

Related work: Sequential decision making under uncertainty with safety constraints has been an increasingly popular area of research among scholars. In particular, there have been various works that study optimization problems with uncertain constraint functions. Depending on the specific problem setting, the constraint function is assumed to be linear Usmanova et al. (2019); Chaudhary and Kalathil (2022); Fereydounian et al. (2020), have a Gaussian process prior Sui et al. (2015, 2018); Berkenkamp et al. (2021) or be generally Lipschitz Usmanova et al. (2020). In all of these works, the constraint is specified as requiring that the output of the (unknown) function is in the nonpositive orthant (i.e. $g(x) \leq 0$), whereas we model the constraint as requiring that the output of the unknown function is in some arbitrary set (i.e. $g(x) \in \mathcal{E}$ for some arbitrary $\mathcal{E}$). This model warrants a unique analysis approach where we study how the geometry of this arbitrary constraint set impacts the performance of our algorithm.

Uncertain constraints have also been studied in the stochastic linear bandit setting. In the conventional stochastic linear bandit setting (without uncertain constraints), an agent chooses an action vector at each time step and then receives a reward that is linear in expectation with respect to the action, with the aim of maximizing her cumulative reward (see e.g. Dani et al. (2008); Abbasi-Yadkori et al. (2011)). One type of stochastic linear bandit problem with uncertain constraints is so-called conservative linear bandits Kazerouni et al. (2017); Moradipari et al. (2020), where the expected reward at each round needs to be close to a baseline reward. Others consider a setting where there is an auxiliary constraint function. Specifically, in Amani et al. (2019) the constraint function depends on the (linearly transformed) reward parameter, while in Pacchiano et al. (2021); Moradipari et al. (2021); Wang et al. (2022) the constraint function is unrelated to the reward parameter and the learner receives noisy feedback of it. Similar to stochastic linear bandits, we consider a problem where the expected response from the environment is a linear function of the action. However, we take the response from the environment to be a vector rather than a scalar, consider the reward to be an arbitrary function of the expected response, and require the expected response from the environment to be within an arbitrary set. The main novelty of this problem is the arbitrary constraint set, which necessitates new analysis techniques that might find broader applicability.

Notation: For a vector $v \in \mathbb{R}^d$ and positive definite matrix $A \in \mathbb{R}^{d \times d}$, we denote the weighted 2-norm as $\|v\|_A = \sqrt{v^\top A v}$. The minimum and maximum eigenvalues of a square matrix M are denoted $\lambda_{min}(M)$ and $\lambda_{max}(M)$, respectively. We denote the closed and open ball with radius r and norm $\|\cdot\|$ as $\bar{\mathcal{B}}_{\|\cdot\|}(r)$ and $\mathcal{B}_{\|\cdot\|}(r)$, which are centered at the origin. The condition number of a matrix M is denoted as $\kappa(M)$. For a set $\mathcal{D}$ and matrix M, we use the notation $M\mathcal{D} = \{Mx : x \in \mathcal{D}\}$. The set $\{1, 2, ..., n\}$ is denoted $[n]$. We use $\tilde{\mathcal{O}}$ to refer to big-O notation that ignores logarithmic factors. We use the notation $\mathbf{e}_i$ to refer to a vector with 1 as the ith position and 0 everywhere else.

2. Problem Setup

We study a sequential decision-making problem where, in each round, an *agent* chooses an *action* and then the environment chooses a *response* according to the action taken. Similar to stochastic linear bandits, we assume that the response from the environment is an unknown linear function of the action and that the agent observes the output of this linear function plus some noise. However, our problem differs from stochastic linear bandits in that the response is multi-dimensional and the agent's goal is to accumulate *reward*, which is an arbitrary known function of the response. In our problem, the agent also needs to ensure that the response from the environment lies within a *safety set* every round.

The details of the problem are described as follows. In each round $t \in [T]$, an agent chooses an action x_t from the compact action set $\mathcal{A} \subset \mathbb{R}^d$ and then observes the noisy response $y_t := \Theta_* x_t + \epsilon_t$. The matrix $\Theta_* \in \mathbb{R}^{n \times d}$ is an unknown parameter that is full rank, $\epsilon_t \in \mathbb{R}^n$ is random noise, and we have that $n \leq d$. Upon choosing an action, the agent earns the reward $f(\Theta_* x_t)$, where the reward function $f : \mathbb{R}^n \to \mathbb{R}$ is known. The agent needs to ensure that when it chooses actions, the response $\Theta_* x_t$ lies within the known compact safety set $\mathcal{E} \subset \mathbb{R}^n$ that has nonempty interior, or equivalently, that x_t is in the *unknown* feasible action set $\mathcal{X} := \{x \in \mathcal{A} : \Theta_* x \in \mathcal{E}\}$.

With the actions that it chooses, the agent aims to maximize the cumulative reward that it achieves while ensuring that the safety constraint is satisfied for all rounds. Therefore, the performance of the agent can be measured with the cumulative regret, $R_T := \sum_{t=1}^{T} \left(f(\Theta_* x_*) - f(\Theta_* x_t)\right)$ where $x_* \in \arg\max_{x \in \mathcal{X}} f(\Theta_* x)$.

In the following two assumptions, we assume that the unknown parameter and feasible actions are bounded, and that the noise is subgaussian. These assumptions are standard in related literature (e.g. Abbasi-Yadkori et al. (2011); Pacchiano et al. (2021)).

Assumption 1 *For all x in $\mathcal{A}$, there exists a constant L such that $\|x\|_2 \leq L$. Additionally, there exists constant S such that $\|\theta_*^i\|_2 \leq S$ for all $i \in [n]$, where θ_*^i is the ith row of Θ_*. The constants S and L are known to the agent.*

Assumption 2 *For all $t \in [T]$, the noise ϵ_t is element-wise conditionally R-subgaussian, such that given the history $\mathcal{F}_t = \sigma(x_1, x_2, ..., x_{t+1}, \epsilon_1, \epsilon_2, ..., \epsilon_t)$ and denoting the ith element of ϵ_t as ϵ_t^i, it holds for all $i \in [n]$ that $\mathbb{E}[\epsilon_t^i | \mathcal{F}_{t-1}] = 0$ and $\mathbb{E}[e^{\lambda \epsilon_t^i} | \mathcal{F}_{t-1}] \leq \exp(\frac{\lambda^2 R^2}{2}), \forall \lambda \in \mathbb{R}$. The constant R is known to the agent.*

We additionally assume that the reward function is Lipschitz.

Assumption 3 *f is M-Lipschitz on $\mathcal{E}$ such that $|f(x) - f(y)| \leq M\|x - y\|_2$ for all x, y in $\mathcal{E}$.*

Lastly, we make an assumption which ensures that the knowledge provided to the agent by Assumption 1 is enough to choose initial actions that are safe. Since it is known that Θ_* is in $\mathcal{C}^0 = \{[\theta^1\ \theta^2\ ...\ \theta^n]^\top \in \mathbb{R}^{n \times d} : \|\theta^i\|_2 \leq S, \forall i \in [n]\}$ due to Assumption 1, then it is also known that $\mathcal{G}^0 := \{x \in \mathcal{A} : \Theta x \in \mathcal{E}, \forall \Theta \in \mathcal{C}^0\}$ is a subset of $\mathcal{X}$. Therefore, we ensure that the agent can initially choose safe actions by assuming that the interior of $\mathcal{G}^0$ is nonempty.

Assumption 4 *The initial feasible set $\mathcal{G}^0$ has a nonempty interior.*

We provide an algorithm for the stated problem in the next section.

Algorithm 1

Input: $\mathcal{A}, \mathcal{E}, f, L, S$

// Pure Exploration

for $t = 1$ **to** T' **do**

  Choose x_t by randomly sampling $\mathcal{G}^0$, and observe response y_t.

end

Construct $\mathcal{C}_{T'}$ and $\mathcal{G}_{T'}$ with (1) and (2) respectively.

// Exploration-Exploitation

for $t = T' + 1$ **to** T **do**

  Choose some $(x_t, \tilde{\Theta}_t) \in \arg\max_{(x,\Theta)\in\mathcal{G}_{t-1}\times\mathcal{C}_{t-1}} f(\Theta x)$, and observe response y_t.

  Update $\mathcal{C}_t$ and $\mathcal{G}_t$ with (1) and (2) respectively.

end

3. Proposed Algorithm

We propose an algorithm to address the stated problem that operates by first performing pure exploration for an appropriate duration T', as specified in the analysis, and then performing exploration-exploitation for the remaining rounds. The algorithm is given in Algorithm 1.

The pure exploration phase proceeds by randomly sampling actions from $\mathcal{G}^0$ such that $\lambda_- := \lambda_{\min}\left(\mathbb{E}\left[x_t x_t^\top\right]\right) > 0$ for $t \in [T']$. Such a scheme is possible given that $\mathcal{G}^0$ has a nonempty interior, although we leave the specific choice of sampling scheme as a design decision.[1]

Each round in the exploration-exploitation phase, $t \in (T', T]$, consists of first identifying the set of actions which will ensure safety given the current knowledge, and then choosing the optimistic action within this safe action set. In order to both identify which actions are safe and to choose actions optimistically, we use confidence sets in which the parameter Θ_* lies with high probability. Let θ_*^i be the ith row of Θ_*, such that $\Theta_* = [\theta_*^1\ \theta_*^2\ ...\ \theta_*^n]^\top$, and let y_t^i be the ith element of y_t, such that $y_t = [y_t^1\ y_t^2\ ...\ y_t^n]^\top$. Then the regularized least-squares estimator of each θ_*^i is given by $\hat{\theta}_t^i = [V_t]^{-1}\sum_{s=1}^t x_s y_s^i$ at round t, where the gram matrix is $V_t = \nu I + \sum_{s=1}^t x_s\left[x_s\right]^\top$. Using the regularized least-squares estimator for each row of Θ_*, we define the confidence set in the following theorem from Abbasi-Yadkori et al. (2011).

Theorem 1 *(Theorem 2 in Abbasi-Yadkori et al. (2011)) Let Assumptions 1 and 2 hold. Then if x_t is in $\mathcal{A}$ for all t, we have with probability at least $1-\delta$ that Θ_* lies in the set*

$$\mathcal{C}_t = \left\{[\theta^1\ \theta^2\ ...\ \theta^n]^\top \in \mathbb{R}^{n\times d} : \left\|\theta^i - \hat{\theta}_t^i\right\|_{V_t} \le \sqrt{\beta_t}, \forall i \in [n]\right\} \tag{1}$$

for all $t \ge 0$, where

$$\sqrt{\beta_t} = R\sqrt{d\log\left(\frac{1+tL^2/\nu}{\delta/n}\right)} + \sqrt{\nu}S.$$

1. We give an example of a sampling scheme with $\lambda_- > 0$. Since $\mathcal{G}^0$ has a nonempty interior, there exists $v \in \mathbb{R}^d$ and $r > 0$ such that the open ball $v + \mathcal{B}_2(r)$ is a subset of $\mathcal{G}^0$. It follows that the closed ball $v + \bar{\mathcal{B}}_2(r/2)$ is a subset of $\mathcal{G}^0$. Therefore, one possible sampling scheme is to uniformly sample u_t from the unit sphere i.i.d. such that $\mathbb{E}[u_t u_t^\top] = \frac{1}{d}I$, and then play $x_t = v + \frac{r}{2}u_t$. Therefore, $\mathbb{E}[x_t x_t^\top] = \mathbb{E}[vv^\top] + \frac{r}{2}\mathbb{E}[vu_t^\top + u_t v^\top] + \frac{r^2}{4}\mathbb{E}[u_t u_t^\top] = vv^\top + \frac{r^2}{4d}I$ given that v is fixed, and it follows that $\lambda_- = \frac{r^2}{4d} > 0$.

Using this set, we define a *conservative inner approximation* of the feasible action set ($\mathcal{X}$) as

$$\mathcal{G}_t := \{x \in \mathcal{A} : \Theta x \in \mathcal{E}, \forall \Theta \in \mathcal{C}_t\}. \tag{2}$$

Note that the sets $\mathcal{C}_t$ and $\mathcal{G}_t$ are updated at the end of each round such that the agent has access to $\mathcal{C}_{t-1}$ and $\mathcal{G}_{t-1}$ in round t. From the definition of $\mathcal{G}_t$, we can see that for any $\Theta \in \mathcal{C}_t$ and $x \in \mathcal{G}_t$, it is guaranteed that Θx is in the safety set $\mathcal{E}$. Theorem 1 states that Θ_* is in $\mathcal{C}_t$ for all rounds with high probability, so if the algorithm chooses x_t from $\mathcal{G}_{t-1}$, the responses from the environment $\{\Theta_* x_t\}_{\forall t \in (T', T]}$ will all be in $\mathcal{E}$ with high probability, and as such, they would ensure safety.

In order to choose these actions from the conservative action sets $\{\mathcal{G}_{t-1}\}_{\forall t \in (T', T]}$ such that the regret is favorable, the algorithm behaves *optimistically*. That is, the algorithm chooses the best action in $\mathcal{G}_{t-1}$ assuming that the true parameter Θ_* is as favorable as possible given available information. Since the agent knows that Θ_* is highly likely to be in the confidence set $\mathcal{C}_{t-1}$ in round t, the algorithm behaves optimistically by finding an action in $\mathcal{G}_{t-1}$ and parameter in $\mathcal{C}_{t-1}$ that maximize the possible reward. Accordingly, the algorithm chooses the action as

$$(x_t, \tilde{\Theta}_t) \in \underset{(x,\Theta) \in \mathcal{G}_{t-1} \times \mathcal{C}_{t-1}}{\arg\max} f(\Theta x), \tag{3}$$

for every round $t \in (T', T]$.

It is important to recognize that, because $\mathcal{G}_{t-1}$ is a conservative inner approximation of $\mathcal{X}$, the optimal action x_* may not be in $\mathcal{G}_{t-1}$. Hence, how well $\mathcal{G}_{t-1}$ approximates $\mathcal{X}$ has an impact on how far x_t is from the optimal action x_*, and hence how large the gap is between $f(\Theta_* x_*)$ and $f(\Theta_* x_t)$ (given the Lipschitz assumption on f). The tightness with which $\mathcal{G}_t$ approximates $\mathcal{X}$ is evidently impacted by the size of $\mathcal{C}_t$, but as we show in the following section, the geometric properties of the safety constraint $\mathcal{E}$ and the action set $\mathcal{A}$ play a significant role as well.

4. Regret Analysis

In this section, we prove an upper bound on the cumulative regret of Algorithm 1. A key aspect of the problem that impacts the regret is the geometric properties of the safety set and the action set. As of now, we have not made any assumptions on these set. However, we will show that the geometric properties of these sets will determine whether we can prove that the algorithm has sublinear regret. To aid in this analysis, we introduce a geometric property of sets that we refer to as *sharpness*, which plays a key role in the regret bound of the proposed algorithm.

4.1. Geometric Properties of Safety Sets

When the agent chooses an action, there is uncertainty as to what the response will be, necessitating the use of a conservative inner approximation of the set of safe actions. Choosing actions from this inner approximation maintains safety because it ensures that every reasonably possible response to the chosen action (i.e. every Θx_t for $\Theta \in \mathcal{C}_{t-1}$) satisfies the safety constraint. In essence, this ensures that some region around the expected response lies within the safety constraint. One can imagine that this region will not "fit" well in to any "sharp" corners that the safety set may have, and hence the inner approximation will be looser for safety sets with "sharp" corners, resulting in less favorable regret. We formalize this notion of *sharpness* in the following series of definitions. The proofs from this section are given in Appendix A in the full online version of this paper Hutchinson et al. (2023).

In order to study the impact that a safety set's geometry has on the tightness of the conservative inner approximation, we first present a more general type of inner approximation that we call the *shrunk version of a set*. Similar to how the conservative inner approximation ensures that the set of all reasonably possible responses are within the safety constraint, the shrunk version of a set ensures that a closed ball at each point is within the original set. This is formally defined in the following definition, which uses the closed ball, $\bar{\mathcal{B}}_{\|\cdot\|}(r) := \{x \in \mathbb{R}^m : \|x\| \leq r\}$, where r is the radius and the particular norm is $\|\cdot\|$.

Definition 2 *For a compact set $\mathcal{D} \subset \mathbb{R}^m$, a norm $\|\cdot\|$, and a nonnegative scalar Δ, we define the* shrunk version *of $\mathcal{D}$ as $\mathcal{D}_\Delta^{\|\cdot\|} := \{x \in \mathcal{D} : x + v \in \mathcal{D}, \forall v \in \bar{\mathcal{B}}_{\|\cdot\|}(\Delta)\}$.*[2]

Given the above definition of the shrunk version of a set, one can consider the maximum shrinkage that a set can withstand while still being nonempty. We introduce the *maximum shrinkage of a set* in the following definition.

Definition 3 *For a compact set $\mathcal{D} \subset \mathbb{R}^m$ and a norm $\|\cdot\|$, we define the* maximum shrinkage *of $\mathcal{D}$, as $H_\mathcal{D}^{\|\cdot\|} := \sup\{\Delta : \mathcal{D}_\Delta^{\|\cdot\|} \neq \emptyset\}$.*

We can now formally define the *sharpness of a set* as the maximum distance from any point in a set to the nearest point in the shrunk version of that set.

Definition 4 *For a compact set $\mathcal{D} \subset \mathbb{R}^m$ and norm $\|\cdot\|$, we define the* sharpness *of $\mathcal{D}$ as*

$$\text{Sharp}_\mathcal{D}^{\|\cdot\|}(\Delta) := \sup_{x \in \mathcal{D}} \inf_{y \in \mathcal{D}_\Delta^{\|\cdot\|}} \|y - x\|_2 ,$$

for all non-negative Δ such that $\mathcal{D}_\Delta^{\|\cdot\|}$ is nonempty.[3]

Sharpness is applicable to the analysis of safe learning algorithms because it upper bounds how far an optimal point within the safe set (e.g. some set $\mathcal{D}$) is from a conservative inner approximation of that safe set (e.g. $\mathcal{D}_\Delta^{\|\cdot\|}$ or a superset of $\mathcal{D}_\Delta^{\|\cdot\|}$). To demonstrate how the geometry of a set impacts its sharpness, the sharpness of several different sets in $\mathbb{R}^2$ is plotted in Figure 1. One can see that sets with "sharper" corners have greater sharpness for the same value of Δ. Also note that we use $\mathcal{D}_\Delta^p$, $H_\mathcal{D}^p$ and $\text{Sharp}_\mathcal{D}^p(\Delta)$ to refer to the shrunk set, maximum shrinkage and sharpness of some set $\mathcal{D}$ with respect to the p-norm.

We now show some simple properties related to when the shrunk version of a set is nonempty and therefore when the sharpness is defined. First, we have that the shrunk version of a compact set is nonempty for some positive shrinkage precisely when the set has a nonempty interior.

Proposition 5 *For a compact set $\mathcal{D} \subset \mathbb{R}^m$, there exists a $\Delta > 0$ such that $\mathcal{D}_\Delta^{\|\cdot\|}$ is nonempty if and only if $\mathcal{D}$ has a nonempty interior.*

2. We can equivalently define $\mathcal{D}_\Delta^{\|\cdot\|}$ using Minkowski subtraction. The Minkowski subtraction of sets $A, B \subseteq \mathbb{R}^m$ is defined as $A \ominus B := \{a - b : a \in A, b \in B\}$, or equivalently, $A \ominus B = \{x \in \mathbb{R}^m : x + B \subseteq A\}$ (Schneider (2014)). Therefore, we can write that $\mathcal{D}_\Delta^{\|\cdot\|} = \mathcal{D} \ominus \mathcal{B}_{\|\cdot\|}(\Delta)$ for $\Delta \geq 0$.
3. Sharpness can also be define with the Hausdorff metric between sets (see Schneider (2014) Sec. 1.8) such that $\text{Sharp}_\mathcal{D}^{\|\cdot\|}(\Delta) = d_H(\mathcal{D}, \mathcal{D}_\Delta^{\|\cdot\|})$.

Next, we show that the shrunk version of a compact set with nonempty interior is nonempty for all shrinkage less than or equal to the maximum shrinkage of the set. This indicates that sharpness is defined on the closed interval from zero to the maximum shrinkage.

Proposition 6 *For a compact set $\mathcal{D} \subset \mathbb{R}^m$ with nonempty interior, we have that $\mathcal{D}_\Delta^{\|\cdot\|}$ is nonempty for all $\Delta \in [0, H_\mathcal{D}^{\|\cdot\|}]$.*

For the remainder of this section, we will study the sharpness of different types of compact sets with nonempty interiors. The first type of set that we study is the polytope, which is the convex hull of a finite set of points, or equivalently, the bounded intersection of a finite number of closed half-spaces. Polytopes capture a wide variety of constraints in the real world and are frequently used for safety sets in safe learning (e.g., Chaudhary and Kalathil (2022); Fereydounian et al. (2020); Usmanova et al. (2019)). We use the polyhedron representation of a polytope, $\mathcal{D} = \{x \in \mathbb{R}^m : Ax \leq b\}$ with $A \in \mathbb{R}^{p \times m}$ and $b \in \mathbb{R}^p$, where there are no redundant constraints. We define $a_j \in \mathbb{R}^m$ as the jth row of A such that $A = [a_1 \ a_2 \ ... \ a_p]^\top$ and $b_j \in \mathbb{R}$ as the jth element of b such that $b = [b_1 \ b_2 \ ... \ b_p]^\top$. We also use $\mathcal{I}_A$ to refer to the collection of all sets of m indices such that for each $\{i_1, i_2, ..., i_m\} \in \mathcal{I}_A$ the vectors $a_{i_1}, a_{i_2}, ..., a_{i_m}$ are linearly independent. For each $\ell \in \mathcal{I}_A$ where $\ell = \{i_1, i_2, ..., i_m\}$, we write $A^\ell = [a_{i_1} \ a_{i_2} \ ... \ a_{i_m}]^\top$ and denote its condition number by $\kappa(A^\ell)$. Using this notation, the following proposition shows that the sharpness of a polytope is bounded by a function that is linear in shrinkage.

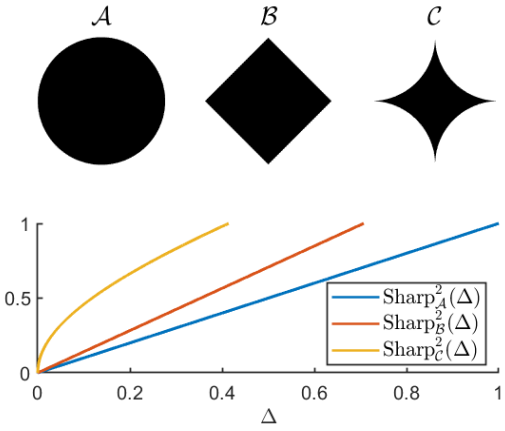

Figure 1: The 2-norm sharpness of three different sets in $\mathbb{R}^2$.

Proposition 7 *For a polytope $\mathcal{D} = \{x \in \mathbb{R}^m : Ax \leq b\}$ with nonempty interior, we have that $\text{Sharp}_\mathcal{D}^{\|\cdot\|}(\Delta) \leq \sqrt{m} C_{\|\cdot\|} K_\mathcal{D} \Delta$, where $K_\mathcal{D} := \max_{\ell \in \mathcal{I}_A} \kappa(A^\ell)$ and $C_{\|\cdot\|} := \max_{\|y\|=1} \|y\|_2$.*

The sharpness bound in Proposition 7 is proportional to the constant $K_\mathcal{D}$, which is the maximum condition number of all sets of m linearly independent constraints. Since there are m linearly independent constraints that are active at each vertex, $K_\mathcal{D}$ upper bounds the condition number of the active constraints at each vertex. This is an intuitive measure of the sharpness of a polytope, given that the condition number of the constraints indicate how close to parallel they are. Also, note that the term $C_{\|\cdot\|}$ in Proposition 7 may depend on the dimension. For example, when the infinity-norm is used, $C_{\|\cdot\|_\infty} = \sqrt{m}$, and when the 1-norm is used, $C_{\|\cdot\|_1} = 1$.

Using the sharpness bound that we developed for polytopes, we can study more general sets. The key intuition that we use to study more general sets is that we can define subsets of the original set which, with appropriate construction, bound the original set in terms of sharpness. In particular, we construct polytopic subsets of the original set in order to provide sharpness bounds that are linear with respect to shrinkage. Being able establish linear bounds on the sharpness is important because it allows us to establish sublinear regret bounds on the proposed algorithm, which we discuss in

(*a*) (*b*)

Figure 2: Examples of sets in $\mathbb{R}^2$ which are *polytope-sharp* and hence can be bounded linearly with respect to shrinkage via Proposition 10 (a), and sets that are not *polytope-sharp* (b).

the next section. In order to develop a bound that uses polytopic subsets, we define the families of polytopes that can be used to bound the sharpness of a given set.

Definition 8 *For a point x in the compact set $\mathcal{D} \subset \mathbb{R}^m$, we define $F_{\mathcal{D}}(x)$ as the family of polytopes with nonempty interior that contain x and are subsets of $\mathcal{D}$.*

From this, we define the class of sets for which we can use polytopic subsets to bound the sharpness.

Definition 9 *A compact set $\mathcal{D} \subset \mathbb{R}^m$ is referred to as* polytope-sharp *if $F_{\mathcal{D}}(x)$ is nonempty for all $x \in \mathcal{D}$.*

We can see that the class of polytope-sharp sets are those for which a collection of polytopes can be constructed to contain each point in the set while still being subsets of the original set. Examples of sets that meet this criterion and sets that do not meet this criterion are illustrated in Figure 2. With this definition, we can then present the following proposition, which provides a linear sharpness bound on sets that are polytope-sharp.

Proposition 10 *Let $\mathcal{D} \subset \mathbb{R}^m$ be a compact set that is polytope-sharp, and choose some arbitrary $\mathcal{F}_x \in F_{\mathcal{D}}(x)$ for each $x \in \mathcal{D}$. Then, we have that*

$$\text{Sharp}_{\mathcal{D}}^{\|\cdot\|}(\Delta) \leq \sqrt{m}\bar{C}_{\|\cdot\|}\Gamma_{\mathcal{D}}\Delta,$$

where, $\bar{C}_{\|\cdot\|} = \max(C_{\|\cdot\|}, 1)$ and

$$\Gamma_{\mathcal{D}} := \max\left\{\bar{K}_{\mathcal{F}}, \frac{r_{\mathcal{D}}}{\bar{H}_{\mathcal{F}}^{\|\cdot\|}}\right\},$$

with $\bar{K}_{\mathcal{F}} := \max_{x\in\mathcal{D}} K_{\mathcal{F}_x}$, $\bar{H}_{\mathcal{F}}^{\|\cdot\|} := \min_{x\in\mathcal{D}} H_{\mathcal{F}_x}^{\|\cdot\|}$ and $r_{\mathcal{D}} := \max_{x,y\in\mathcal{D}} \|x - y\|_2$.

In addition to providing a linear sharpness bound on polytope-sharp sets, Proposition 10 also indicates that, when the polytopes are small, i.e. $\bar{H}_{\mathcal{F}}^{\|\cdot\|}$ is small, or the polytopes are sharp, i.e. $\bar{K}_{\mathcal{F}}$ is large, then the sharpness bound is larger and therefore less favorable. From Proposition 10, we can also immediately show that every compact, convex set with nonempty interior is linearly sharp.

Corollary 11 *If a compact set $\mathcal{D} \subset \mathbb{R}^m$ with nonempty interior is convex, then it is polytope-sharp and it holds that* $\text{Sharp}_{\mathcal{D}}^{\|\cdot\|}(\Delta) \leq \sqrt{m}\bar{C}_{\|\cdot\|}\Gamma_{\mathcal{D}}\Delta$.

It is important to note that although all compact convex sets are polytope-sharp, there are also nonconvex sets that are polytope-sharp.

4.2. Regret Bound

We will now use the work in the previous section to establish a sublinear bound on the cumulative regret of Algorithm 1. In order to do so, we define the set of feasible responses as $\mathcal{Y} := \Theta_* \mathcal{A} \cap \mathcal{E}$, where we use the notation $\Theta_* \mathcal{A} = \{\Theta_* x : x \in \mathcal{A}\}$. The set $\mathcal{Y}$ reflects the set of responses that are possible given the action set $\mathcal{A}$ and the safety set $\mathcal{E}$. The sharpness of $\mathcal{Y}$ is used in the regret bound for Algorithm 1, as shown in Theorem 12. Although one might expect that the sharpness of $\mathcal{E}$ would be in the regret bound (instead of the sharpness of $\mathcal{Y}$), the set $\mathcal{A}$ can impact the distance from the optimal action x^* to the set $\mathcal{G}_{t-1}$ and hence it is insufficient to solely use the sharpness of $\mathcal{E}$ in the regret analysis. Therefore, we use the sharpness of $\mathcal{Y}$ to capture both the sharpness of $\mathcal{E}$ and any unfavorable effects due to the specific $\mathcal{A}$ in a particular problem setting.

Theorem 12 *Let Assumptions 1–4 hold. With probability at least* $1 - 2\delta$*, we have that the regret of Algorithm 1 is bounded as*

$$R_T \leq 2M\sqrt{n}LST' + M(T - T')\text{Sharp}_{\mathcal{Y}}^{\infty}\left(\frac{2\sqrt{2\beta_T}L}{\sqrt{2\nu + \lambda_- T'}}\right) + M\max(H_{\mathcal{Y}}^{\infty}, 1)\sqrt{n8\beta_T(T - T')d\log\left(\frac{1 + TL^2}{d\nu}\right)}.$$

for any $T' \geq \max(t_\delta, t_h)$ *where* $t_\delta := \frac{8L^2}{\lambda_-}\log(\frac{d}{\delta})$ *and* $t_h := \frac{8\beta_T L^2}{\lambda_-(H_{\mathcal{Y}}^{\infty})^2} - \frac{2\nu}{\lambda_-}$.

Corollary 13 *Assume the same as Theorem 12. If* $\mathcal{Y}$ *is polytope-sharp, then the regret of Algorithm 1 satisfies*

$$R_T \leq 2M\sqrt{n}LST' + \frac{2n\sqrt{2\beta_T}\Gamma_{\mathcal{Y}}LM(T - T')}{\sqrt{2\nu + \lambda_- T'}} + M\max(H_{\mathcal{Y}}^{\infty}, 1)\sqrt{n8\beta_T(T - T')d\log\left(\frac{1 + TL^2}{d\nu}\right)}$$

with probability at least $1-2\delta$ *when* $T' \geq \max(t_\delta, t_h)$*. In particular, choosing* $T' = \max(T^{2/3}, t_\delta, t_h)$ *ensures that* $R_T = \tilde{\mathcal{O}}\left(T^{2/3}\right)$.

We can see that the regret bound depends on the sharpness of $\mathcal{Y}$, and as shown in Corollary 13, is $\tilde{\mathcal{O}}\left(T^{2/3}\right)$ when $\mathcal{Y}$ is polytope-sharp. Note that the agent needs to know the maximum shrinkage of $\mathcal{Y}$, or a lower bound of it, in order to appropriately choose T'. If there is a known subset of $\mathcal{Y}$ or it is known that $\mathcal{E}$ is a subset of $\Theta_* \mathcal{A}$ then the agent can calculate a lower bound on the maximum shrinkage of $\mathcal{Y}$. Otherwise, there might be application specific information that provides a conservative estimate of the maximum shrinkage of $\mathcal{Y}$.

The complete proof of Theorem 12 is given in Appendix B of the full online version Hutchinson et al. (2023). This proof utilizes a decomposition of the instantaneous regret given by

$$r_t := f(\Theta_* x_*) - f(\Theta_* x_t) = \underbrace{f(\Theta_* x_*) - f(\tilde{\Theta}_t x_t)}_{\text{Term I}} + \underbrace{f(\tilde{\Theta}_t x_t) - f(\Theta_* x_t)}_{\text{Term II}}. \tag{4}$$

Term I captures the suboptimality of the optimistic pair $(x_t, \tilde{\Theta}_t)$ from (3), while Term II captures the shrinkage of the confidence set $\mathcal{C}_t$. The pair $(x_t, \tilde{\Theta}_t)$ may be suboptimal due to the fact that $\mathcal{G}_t$ is a strict subset of $\mathcal{X}$, which is necessary to ensure safety. Although the analysis of Term II can be handled with conventional bandit analysis, the analysis of Term I requires novel techniques, including sharpness, as we discuss in the following paragraph.

The bound on Term I is given in the following lemma.

Lemma 14 *Let Assumptions 1–4 hold. For* $t \in (T', T]$, Term I *is bounded as*

$$\text{Term I} := f(\Theta_* x_*) - f(\tilde{\Theta}_t x_t) \leq M \text{Sharp}_{\mathcal{Y}}^{\infty}\left(\frac{2\sqrt{2\beta_T}L}{\sqrt{2\nu + \lambda_- T'}}\right)$$

when $T' \geq \max(t_\delta, t_h)$ *with probability at least* $1 - 2\delta$.

The proof of Lemma 14 is given in Appendix B.1 of the full online version Hutchinson et al. (2023) and considers a shrunk version of $\mathcal{Y}$ such that every possible y in the shrunk version of $\mathcal{Y}$ can be given by Θx with some $\Theta \in \mathcal{C}_t$ and some $x \in \mathcal{G}_t$. This implies that $f(\tilde{\Theta} x_t)$ is greater than or equal to $f(y)$ for every y in the shrunk version of $\mathcal{Y}$ and hence we can bound Term I with the difference between the optimal reward ($f(y_*)$, where $y_* = \Theta_* x_*$) and the reward from some y in the shrunk version of $\mathcal{Y}$. With the Lipschitz assumption on f, this can be bounded with the difference between y_* and some y in the shrunk version $\mathcal{Y}$. By choosing y to be the point in the shrunk version of $\mathcal{Y}$ that is closest to y_*, we can ultimately bound the regret with the sharpness of $\mathcal{Y}$ as given in Lemma 14.

5. Numerical Experiments

We simulated the results of the algorithm with three different polytopic safety sets of different sharpness in a problem setting where $n = 3$ and $d = 3$. The cumulative sum of the instantaneous regret in the exploration-exploitation phase of the algorithm is shown in Figure 3 for each polytopic safety set. The solid line is an average of six trials and the shaded region indicates the 95% confidence interval over the trials. For each safety set, the plot shows its K constant, as defined in Proposition 7. Each simulation has a different realization of the noise $\{\epsilon_t\}_{t\in[T]}$. Otherwise, the problem and algorithm parameters are the same for every simulation, and all the polytopic safety sets have the same maximum shrinkage. Also, note that the action set $\mathcal{A}$ is chosen to be non-restrictive, such that $\mathcal{Y} = \mathcal{E}$. Figure 3 provides some empirical support for the sublinear regret bound in Theorem 12 and also indicates that the sharpness of the safe set impacts the regret of the algorithm. The details of the simulation are given in Appendix C of the full online version Hutchinson et al. (2023).

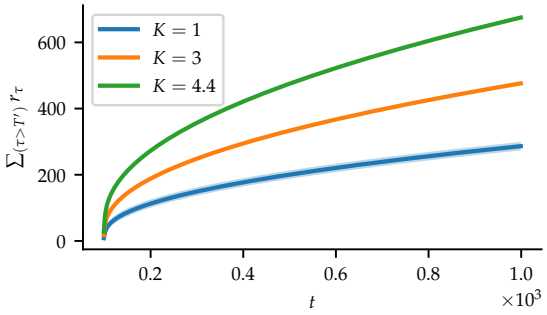

Figure 3: The cumulative sum of the instantaneous regret in the exploration-exploitation phase of the algorithm with polytopic constraint sets that have different K constants (as defined in Proposition 7).

Acknowledgments

This work was supported by NSF grant #1847096.

References

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

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

Xiaoqing Bai, Hua Wei, Katsuki Fujisawa, and Yong Wang. Semidefinite programming for optimal power flow problems. *International Journal of Electrical Power & Energy Systems*, 30(6-7): 383–392, 2008.

Felix Berkenkamp, Andreas Krause, and Angela P Schoellig. Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics. *Machine Learning*, pages 1–35, 2021.

Sapana Chaudhary and Dileep Kalathil. Safe online convex optimization with unknown linear safety constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6175–6182, 2022.

Shaoru Chen, Ning-Yuan Li, Victor M Preciado, and Nikolai Matni. Robust model predictive control of time-delay systems through system level synthesis. *arXiv preprint arXiv:2209.11841*, 2022.

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

Masoud Farivar and Steven H Low. Branch flow model: Relaxations and convexification—part i. *IEEE Transactions on Power Systems*, 28(3):2554–2564, 2013.

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

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Lukas Hewing, Kim P Wabersich, Marcel Menner, and Melanie N Zeilinger. Learning-based model predictive control: Toward safe learning in control. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:269–296, 2020.

Spencer Hutchinson, Berkay Turan, and Mahnoosh Alizadeh. A safe pricing mechanism for distributed resource allocation with bandit feedback. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 5092–5098. IEEE, 2022.

Spencer Hutchinson, Berkay Turan, and Mahnoosh Alizadeh. The impact of the geometric properties of the constraint set in safe optimization with bandit feedback. *arXiv preprint arXiv:2305.00889*, 2023.

Sebastian Junges, Nils Jansen, Christian Dehnert, Ufuk Topcu, and Joost-Pieter Katoen. Safety-constrained reinforcement learning for mdps. In *International conference on tools and algorithms for the construction and analysis of systems*, pages 130–146. Springer, 2016.

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

Torsten Koller, Felix Berkenkamp, Matteo Turchetta, and Andreas Krause. Learning-based model predictive control for safe exploration. In *2018 IEEE conference on decision and control (CDC)*, pages 6059–6066. IEEE, 2018.

Daniel K Molzahn, Florian Dörfler, Henrik Sandberg, Steven H Low, Sambuddha Chakrabarti, Ross Baldick, and Javad Lavaei. A survey of distributed optimization and control algorithms for electric power systems. *IEEE Transactions on Smart Grid*, 8(6):2941–2962, 2017.

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

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

Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 2827–2835. PMLR, 2021.

Rolf Schneider. *Convex bodies: the Brunn–Minkowski theory*. Number 151. Cambridge university press, 2014.

Yanan Sui and Joel W Burdick. Correlational dueling bandits with application to clinical treatment in large decision spaces. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2793–2799, 2017.

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

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

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

Ilnura Usmanova, Andreas Krause, and Maryam Kamgarpour. Safe non-smooth black-box optimization with application to policy search. In *Learning for Dynamics and Control*, pages 980–989. PMLR, 2020.

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