Safe Linear Bandits over Unknown Polytopes

Aditya Gangrade GANGRADE@BU.EDU
Boston Univesrity & University of Michigan

Tianrui Chen TRCHEN@BU.EDU
Boston University

Venkatesh Saligrama SRV@BU.EDU
Boston University

Editors: Shipra Agrawal and Aaron Roth

Abstract

The safe linear bandit problem (SLB) is an online approach to linear programming with unknown objective and unknown *roundwise* constraints, under stochastic bandit feedback of rewards and safety risks of actions. We study the tradeoffs between efficacy and smooth safety costs of SLBs over polytopes, and the role of aggressive doubly-optimistic play in avoiding the strong assumptions made by extant pessimistic-optimistic approaches.

We first elucidate an inherent hardness in SLBs due the lack of knowledge of constraints: there exist 'easy' instances, for which suboptimal extreme points have large 'gaps', but on which SLB methods must still incur $\Omega(\sqrt{T})$ regret or safety violations, due to an inability to resolve unknown optima to arbitrary precision. We then analyse a natural doubly-optimistic strategy for the safe linear bandit problem, DOSS, which uses optimistic estimates of both reward and safety risks to select actions, and show that despite the lack of knowledge of constraints or feasible points, DOSS simultaneously obtains tight instance-dependent $O(\log^2 T)$ bounds on efficacy regret, and $\widetilde{O}(\sqrt{T})$ bounds on safety violations, thus attaining near Pareto-optimality. Further, when safety is demanded to a finite precision, violations improve to $O(\log^2 T)$. These results rely on a novel dual analysis of linear bandits: we argue that DOSS proceeds by activating noisy versions of at least d constraints in each round, which allows us to separately analyse rounds where a 'poor' set of constraints is activated, and rounds where 'good' sets of constraints are activated. The costs in the former are controlled to $O(\log^2 T)$ by developing new dual notions of gaps, based on global sensitivity analyses of linear programs, that quantify the suboptimality of each such set of constraints. The latter costs are controlled to $O(1)$ by explicitly analysing the solutions of optimistic play.

Keywords: Safety; Linear Bandits; Optimism in Online Learning.

1. Introduction

The **Safe Linear Bandit (SLB) problem:** Consider a linear program $\max \theta^\top x : Ax \leq \alpha$ where the feasible set $\mathcal{S} := \{Ax \leq \alpha\}$ is known to be a nonempty bounded polytope in $\mathbb{R}^d$, but neither the objective $\theta \in \mathbb{R}^d$, nor the constraint matrix $A \in \mathbb{R}^{m\times d}$ are completely known a priori, and no action known a priori to be safe (i.e., feasible) is available. Instead, a learner sequentially picks actions x_t, with the goal of choosing x_t that are effective and safe *in each round*. Learning is enabled through stochastic bandit feedback in the form of a reward signal $R_t = \langle \theta, x_t\rangle + w_t^R$ and a risk signal $S_t : \mathbb{E}[S_t|x_t] = Ax_t + w_t^S$ where (w_t^R, w_t^S) is a noise process.

Ideally, we would explore only in $\mathcal{S}$, but since we do not know it (or any safe point) to start with, some safety violation must necessarily occur over the course of learning. It is natural in many

applications to penalise such violation 'softly'. With this view, we measure the performance of the learner over T rounds through the *efficacy regret*, $\mathscr{E}_T$, and the *net safety violation* $\mathscr{S}_T$ defined as

$$\mathscr{E}_T := \sum_{t \le T} \langle \theta, x^* - x_t \rangle)_+ \quad \textit{and} \quad \mathscr{S}_T := \sum_{t \le T} (\max_i \langle a^i, x_t \rangle - \alpha^i)_+, \tag{1}$$

where $(z)_+ := \max(z, 0)$, and x^* is the constrained optimum. These same ℓ_1 metrics were proposed in the finite-armed setting by Efroni et al. (2020) and Chen et al. (2022). The main structural property that makes $\mathscr{E}_T, \mathscr{S}_T$ pertinent in the roundwise scenario is that they accumulate only the *positive parts* of the roundwise inefficiency or safety violation. Indeed, since $\mathscr{E}_T$ sums over $(\langle \theta, x^* - x_t \rangle)_+$, playing any x_t with better reward than x^* leads to no decrease in it, and instead it increases $\mathscr{S}_T$ since such an x_t must be infeasible. Conversely, since $\mathscr{S}_T$ sums the largest roundwise violations, playing a safe but under-effective x_t increases $\mathscr{E}_T$ but does not reduce $\mathscr{S}_T$. Thus, the only way to make both $\mathscr{E}_T$ and $\mathscr{S}_T$ small is to ensure that most x_ts are near-safe *and* near-optimal. We note that the choice of the linear penalty on violations above is just out of convenience: any penalty of the form $f(\max_i(\langle a^i, x_t \rangle - \alpha^i)_+)$, where f smoothly decays to 0 near 0^+, is amenable to our analysis (§G).

Motivating Examples. The interplay of unknown rewards and constraints is a common feature of application domains of bandits. In drug trials, one needs to balance the efficacy of a treatment regimen with its risk of various side-effects (i.e., the probabilities that it induces harmful side-effects); in crowdsourcing, one must balance the cost of completing tasks with the quality of the resulting work; and recommmender systems must balance the click-rate of suggestions with their effects on engagement (such as watch-time or revisiting rates). In such cases, we must enforce the constraint in each round, e.g., completing one task well does not license us to be sloppy on the next. Further, it is nontrivial to find a feasible starting point, since, e.g., this requires knowing worker quality distributions a priori, or knowing which compounds balance the side-effects of active compounds a priori. Nevertheless, soft enforcement is meaningful, e.g., if the risk of a side-effect is slightly over α, this only leads to a slight increase in overall numbers of adverse effects realised; and a slight reduction in the mean watch-time is an acceptable price for learning. Thus strong control on $\mathscr{S}_T$ ensures that in the long run, the system performs arbitrarily close to safety.

Soft Roundwise Enforcement over Polytopes. We focus on understanding what performance can be achieved while ensuring that $\mathscr{S}_T = o(T)$. At the first glance, one expects control of the form $\max(\mathscr{E}_T, \mathscr{S}_T) = \widetilde{O}(\sqrt{T})$, which indeed follows from standard techniques (§4). However, this question is most interesting in a refined sense: since we are work over a *polytopal* domain,[1] prior work on linear bandits tells us that if $\mathcal{S}$ were known, one can derive *instance-dependent* bounds of $O(\log^2 T)$ on $\mathscr{E}_T, \mathscr{S}_T$ (e.g. Abbasi-Yadkori et al., 2011). This paper is concerned with the question

> Over polytopal domains, is it possible to attain instance-dependent polylogarithmic bounds on $\mathscr{E}_T$ and $\mathscr{S}_T$ without knowing $\mathcal{S}$ in advance?

Our Contributions approach this by studying the efficacy-safety tradeoffs for SLBs, and by analysing a natural doubly-optimistic method for the same. Concretely, we show that

- **Simultaneous logarithmic control on $\mathscr{E}_T$ and $\mathscr{S}_T$ is impossible.** We show that for any SLB algorithm, there exists an instance *with large 'gaps'* on which the algorithm incurs $\max(\mathscr{E}_T, \mathscr{S}_T) = \Omega(\sqrt{T})$. The key property of these instances is the large, i.e., $\Omega(1)$ gap, and due to this gap each

1. While obvious, let us explicitly note here that the problem over polytopal domains is of significant importance, since this corresponds to the ubiquitous questions of linear programming.

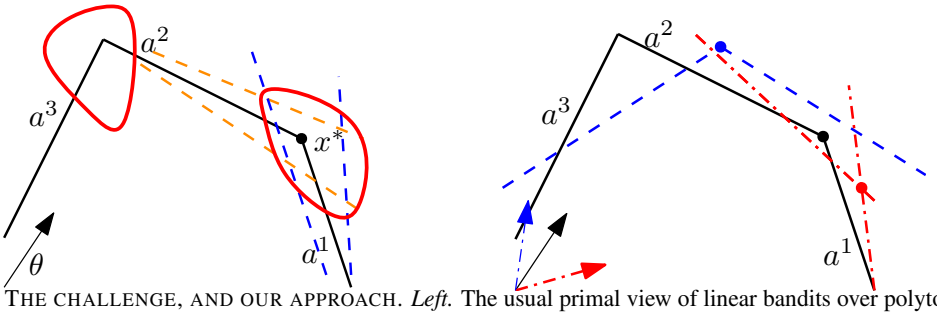

Figure 1: THE CHALLENGE, AND OUR APPROACH. *Left.* The usual primal view of linear bandits over polytopes breaks down, since noisy estimates of the unknown A induce a continuum of potential locations for extreme points (red blobs). *Right* Taking a dual linear programming view, we can identify extreme points as arising by saturating d independent constraints. We generalise this view by showing that DOSS plays by saturating noisy versions of d constraints. Poor play can arise from picking the wrong set of constraints (blue), or using a poor estimate for the right set of constraints (red).

instance could be solved $\max(\mathscr{E}_T, \mathscr{S}_T) = O(\log^2 T)$ if the feasible set $\mathcal{S}$ were known (§5). However, a polynomial lower bound arises since the lack of knowledge of $\mathcal{S}$ induces a '*precision barrier*,' that is, the fact that no method can *locate* effective and safe actions to precision better than $t^{-1/2}$ after t rounds of play. This same barrier also renders the standard primal approach of analysing polytopal linear bandits via their extreme points ineffective for SLBs(Fig. 1, left). We further note that the constructed instances are simple enough to embed into any nontrivial set of SLB instances, making the result generic rather than specific to the particular situation we study.

- **Nevertheless, doubly-optimistic (DO) methods can attain $\mathscr{E}_T = O(\log^2 T)$ and $\mathscr{S}_T = \widetilde{O}(\sqrt{T})$.** Specifically, we show that these bounds are attained by the DO method DOSS (§3.2), which generalises the finite-armed approach of Efroni et al. (2020) and Chen et al. (2022), and has been studied for aggregate enforcement (see below) by Agrawal and Devanur (2014). DOSS builds an 'optimistic' estimate $\widetilde{\mathcal{S}}_t$ of $\mathcal{S}$, and selects actions optimistically over the same. Since these bounds match our lower bounds up to polylog-factors, DOSS is near-Pareto-optimal for SLBs.
- **The aforementioned precision barrier is the sole obstruction to logarithmic bounds.** We argue that in important special cases, DOSS, with either no or mild changes, attains $\max(\mathscr{E}_T, \mathscr{S}_T) = O(\log^2 T)$. The key property of such settings is an innate way to avoid having to identify good primal actions to arbitrary precision, illustrating that this is the key obstruction in SLBs.

Technical novelty of the paper lies in the analysis of DOSS. Since the primal approach for obtaining polylog regret in linear bandits fails, we instead approach the problem through a novel dual analysis, that exploits the fact that extreme points of polytopes can be dually viewed as points saturating d constraints (Fig. 1, right). We show that this view generalises, i.e., DOSS picks actions by saturating a noisy version of d constraints. This allows us to break the analysis into two threads
a)*a combinatorial identification problem of whether the 'right' set of constraints is saturated, and*
b) whether effective points are played when the 'optimal' sets of constraints are saturated.
The efficacy loss due to the former is controlled to $O(\log^2 T)$ by developing a novel notion of 'dual gap' associated with each 'poor' set of constraints, which arise via a global LP sensitivity analysis approach. The second issue is handled via a careful analysis of optimistic play to argue that under mild nondegeneracy assumptions, it cannot play ineffective actions when saturating the 'optimal' set of constraints, which controls the efficacy loss due to such play to $O(1)$.

Related Work We briefly describe the two main lines of work on constrained bandits (also see §A).
Hard Roundwise Enforcement. Instead of the soft sense we study, one can demand roundwise enforcement in a *hard* sense, requiring that with high probability (whp), the constraints always be

met, i.e., whp, $\mathscr{S}_t = 0$. Since this is clearly not possible without knowing a safe point to start with, methods along these lines usually assume a priori knowledge of a point x^{s} in the *interior* of $\mathcal{S}$, i.e., with positive safety margin $M^{\mathsf{s}} := -\max_i(\langle a^i, x^{\mathsf{s}}\rangle - \alpha^i)$. Given the knowledge of $(x^{\mathsf{s}}, M^{\mathsf{s}})$, recent lines of work (Amani et al., 2019; Moradipari et al., 2021; Pacchiano et al., 2021; Afsharrad et al., 2023; Hutchinson et al., 2023; Varma et al., 2023; Pacchiano et al., 2024) have proposed various 'pessimistic-optimisic' (PO) methods for the SLB problem,[2] which operate by exploring in the vicinity of x^{s}, and build pessimistic estimates of $\mathcal{S}$, over which they act optimistically. While such methods attain the strong safety guarantee of $\mathscr{S}_T = 0$ whp, the associated costs are significant: (i) the knowledge of $(x^{\mathsf{s}}, M^{\mathsf{s}})$ is nontrivial to obtain, and the costs of obtaining the same are not accounted for in this literature,[3] and (ii) the resulting efficacy bounds, $\mathscr{E}_T = O(d\sqrt{T}/M^{\mathsf{s}})$, quantitatively depend on this safety margin.[4]

Aggregate Enforcement. Instead of roundwise metrics, *aggregate constraint enforcement* aims to control $\mathscr{R}_T = \sum\langle\theta, x^* - x_t\rangle$ and $\mathscr{A}_T = \sum_{t\le T}\max_i(\langle a^i, x_t\rangle - \alpha_i)$ (e.g. Badanidiyuru et al., 2013, 2014; Agrawal and Devanur, 2014, 2016; Agrawal et al., 2016). The key difference from the roundwise setting is that there is no nonlinearity in the roundwise penalties in $\mathscr{R}_T, \mathscr{A}_T$. This small change drastically affects allowable behaviour for such problems, e.g., we can ensure $\mathscr{A}_T = o(T)$ while alternating between playing 'very unsafe' and 'very safe' actions, since the negative costs of the latter cancel the positive costs of the former, but this would instead incur $\mathscr{S}_T = \Omega(T)$. Of course, $\mathscr{A}_T$ is an inappropriate metric for safety contexts, e.g., treating one patient unsafely cannot be balanced by assigning a placebo to the next. We note that while the analysis of Agrawal and Devanur (2014) *can* be extended to show $(\mathscr{E}_T, \mathscr{S}_T) = \widetilde{O}(\sqrt{T})$, we go much beyond this basic observation through our the finer grained upper bounds of $(\log^2 T, \widetilde{O}(\sqrt{T}))$, as well as our instance-wise obstructions, which are both novel. We also note that most of the literature on aggregate enforcement explicitly assumes that $x = 0$ is safe, and that the entries of A are positive, which we do not need. Aggregate enforcement remains an active area of research, e.g., 'Conservative bandits' (e.g. Wu et al., 2016) enforce properties of the form $\mathscr{A}_t = O(\sqrt{t})$ for most t, and Liu et al. (2021) show that given a Slater parameter, one can enforce $\mathscr{A}_t \le 0$ for all t large enough. We also note that most work on constrained MDPs is of this flavour (e.g. Vaswani et al., 2022, and references therein).

2. Problem Setup

For naturals $a \le b$, let $[a : b] := \{a, \dots, b\}$. $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ denote the inner-product and ℓ_2-norm in $\mathbb{R}^d$ respectively, and for a matrix $V \succ 0$, $\|z\|_V := \sqrt{\langle z, Vz\rangle}$. For a $p \times q$ matrix M, and a set $\mathsf{S} \subset [1 : p]$, $M(\mathsf{S})$ denotes the $|\mathsf{S}| \times q$ submatrix of M preserving rows indexed in S.

Setting. An instance of polytopal SLB problem is parameterised by an a polytopal region $\mathcal{X} = \{Bx \le \beta\} \subset \mathbb{R}^d$, a known constraint level vector $\alpha \in \mathbb{R}^U$, and latent objective $\theta \in \mathbb{R}^d$ and constraint matrix $A \in \mathbb{R}^{U\times d}$, which define the principal LP of relevance. Here, the constraints $\{Bx \le \beta\}$

2. and also safe MDPs (e.g. Turchetta et al., 2016; Wachi and Sui, 2020; Bernasconi et al., 2022; Vaswani et al., 2022)
3. Note that the need for a safety margin may make even seemingly simple settings challenging. E.g., if x is the amount of different drugs assigned to a treatment, one may think that the 'no-treatment' drug cocktail $x = 0$ is always 'safe', and can serve as x^{s}. However, in treatment regimens, it is common that any dose of compound 1 must be accompanied by a proportional dose of compound 2 to manage the side-effects induced by compound 1, i.e, the constraint may be of the form $\langle(a_1, -a_2), x\rangle \le 0$, in which case $x = 0$ has no safety margin, and so is unusable for PO methods.
4. We also include a simulation study in §8 that indicates that the safety violations of DOSS are considerably better behaved than the efficacy costs of the PO method SAFE-LTS (Moradipari et al., 2021).

should be thought of as arising from pre-determined hard limits on x.[5] For notational succinctness, we will embed these constraints into (A, α) by extending A to lie in $\mathbb{R}^{m\times d}$ for $m = U + K$, and setting the last K rows of A to B, and similarly augment α to include β. We shall often need the notation $\mathbf{1}_U = (1, \cdots, 1, 0, \cdots, 0)$, with U ones, which indicates the unknown constraints. With this notation, the principal LP of interest is $\max_{x\in\mathcal{X}}\langle\theta, x\rangle : Ax \leq \alpha$.

Play. The problem proceeds in rounds, indexed by t. For each t, we choose an $x_t \in \mathcal{X}$, and receive reward feedback r_t and safety feedback $\{s_t^i\}_{i\in[1:U]}$ that satisfy $r_t = \langle\theta, x_t\rangle + w_t^R$ and $s_t^i = \langle a^i, x_t\rangle + w_t^{S,i}$, where the various w_ts are each subGaussian noise processes, which need not be independent across i. The information set of the learner at time t is $\mathscr{H}_{t-1} := \{(x_\tau, r_\tau, \{s_\tau^i\}_{i\in[1:U]})_{\tau<t}\}$, and x_t must be adapted to the filtration induced by $\mathscr{H}_{t-1}$.

Metrics. We will control the *Efficacy Regret* and *Net Safety Violation* (1). We reiterate that these have pertinence to the SLB setting because they penalise only the positive parts of roundwise costs.

Assumptions. We conclude by noting standard assumptions due to Abbasi-Yadkori et al. (2011).

1. Boundedness: $\|\theta\| \leq 1, \|a^i\| \leq 1$ for all i, and $\mathcal{X} \subset \{\|x\| \leq 1\}$ is a bounded polytope.
2. SubGaussian Noise: $\forall t, w_t := (w_t^R, \{w_t^{S,i}\}_{i\in[1:U]})$ is conditionally centred and 1-subGaussian given $\mathcal{F}_t := \sigma(\mathcal{H}_{t-1}, x_t)$, i.e., $\forall t, \mathbb{E}[w_t|\mathcal{F}_t] = 0, \forall\lambda, \mathbb{E}[\exp(\lambda^\top w_t)|\mathcal{F}_t] \leq \exp(\|\lambda\|^2/2)$.

All subsequent results should be taken to hold only under the above. See §B.1 for more details.

3. A Doubly Optimistic Algorithm for Safe Linear Bandits

As previously discussed, our main method of interest is the natural approach of playing optimistically from an optimistic *permissible set* (Agrawal and Devanur, 2014; Efroni et al., 2020; Chen et al., 2022). We summarise the method, and establish key notation that is used throughout.

3.1. Confidence Sets and Noise Scales

We take the standard approach (Abbasi-Yadkori et al., 2011). Let the matrix $X_{1:t} = [x_1, \ldots, x_t]^\top$ and the vectors $R_{1:t} = [r_1, \ldots, r_t]^\top, S_{1:t}^i = [s_1^i, \ldots, s_t^i]^\top$ arise by stacking the actions and feedback. The 1-regularised least squares (RLS) estimate of θ, a^i using $\mathscr{H}_{t-1}$ is

$$\hat{\theta}_t = (X_{1:t}^\top X_{1:t} + \lambda I)^{-1} X_{1:t}^\top R_{1:t}, \quad \hat{a}_t^i = (X_{1:t}^\top X_{1:t} + \lambda I)^{-1} X_{1:t}^\top S_{1:t}^i..$$

Of course, if $i \in [U + 1 : m]$, then we do not need to estimate i, and we shall just set $\hat{a}_t^i = a^i$ instead. We will collate the $\hat{a}_t^i$s into a matrix $\hat{A}_t$ row-wise. Let us define the signal strength as $V_t := \sum_{s\leq t} x_s x_s^\top + I$, and for $\delta \in (0, 1)$, the m-confidence radius as $\sqrt{\omega_t(\delta)} = 1+\sqrt{\frac{1}{2}\log\frac{(U+1)\sqrt{\det V_{t-1}}}{\delta}}$. The main results are based on the following two concepts, which we explicitly delineate.

Definition 1 *For any time t, the* RLS confidence sets *are*

$\mathcal{C}_t^\theta(\delta) := \{\tilde{\theta} : \|\tilde{\theta} - \hat{\theta}_t\|_{V_{t-1}} \leq \sqrt{\omega_t(\delta)}\}$ *and* $\boldsymbol{\mathcal{C}}_t(\delta) := \{\tilde{A} : \forall$ *rows* $i, \|\tilde{a}^i - \hat{a}_t^i\|_{V_{t-1}} \leq \sqrt{\omega_t(\delta)}\mathbf{1}_U\}$,

and the local noise scale *is* $\rho_t(x; \delta) := 2\sqrt{\omega_t(\delta)}\|x\|_{V_{t-1}^{-1}}$.

The key properties we need are due to Abbasi-Yadkori et al. (2011), and are summarised below, and proved in §B.2. We will often drop the dependence of $\mathcal{C}_t^i(\delta), \boldsymbol{\mathcal{C}}_t(\delta)$, and $\rho_t(x; \delta)$ on δ.

Lemma 2 *The confidence sets are consistent, i.e.,* $\mathbb{P}\left(\forall t, \theta \in \mathcal{C}_t^\theta(\delta), A \in \boldsymbol{\mathcal{C}}_t(\delta)\right) \geq 1 - \delta$. *Further, under consistency, the noise scale $\rho_t(x; \delta)$ at any $x \in \mathcal{X}$ satisfies* $\forall x \in \mathcal{X}$,

$$\forall\tilde{A} \in \boldsymbol{\mathcal{C}}_t(\delta), |(\tilde{A} - A)x| \leq \rho_t(x; \delta)\mathbf{1}_U, \quad \textit{and} \quad \forall\tilde{\theta} \in \mathcal{C}_t^\theta(\delta), |\langle\tilde{\theta} - \theta, x\rangle| \leq \rho_t(x; \delta).$$

5. e.g., known box constraints in crowdsourcing account for maximum worker capacity, and nonnegativity of work.

Finally, for any sequence $\{x_t\}$, $\sum_{s\le t}\rho_s(x_s)^2 = O(d^2\log^2 t)$ *and* $\sum_{s\le t}\rho_s(x_s) = \widetilde{O}(\sqrt{d^2 t})$.

3.2. Doubly-Optimistic Safe Selection

We describe the method, DOSS (Algorithm 1). The key construction herein is the optimistic *permissible set* of points x that are safe according to at least one choice of constraints in $\boldsymbol{\mathcal{C}}_t$:

$$\widetilde{\mathcal{S}}_t(\delta) := \{x : \exists \tilde{A} \in \boldsymbol{\mathcal{C}}_t(\delta) \text{ s.t. } \tilde{A}x \le \alpha\}. \tag{2}$$

The set $\widetilde{\mathcal{S}}_t$ consists of all actions that may *plausibly* be safe given $\mathscr{H}_t$. The arm x_t is selected optimistically from $\widetilde{\mathcal{S}}_t$ as

$$(\tilde{\theta}_t, x_t) \in \arg\max\{\langle\tilde{\theta}, x\rangle : \tilde{\theta} \in \mathcal{C}_t^\theta(\delta), x \in \widetilde{\mathcal{S}}_t(\delta)\} \tag{3}$$

The optimistic construction of the permissible set is the main distinction between the DO and PO approaches (§1), which instead work with the pessimistic set $\Pi_t := \{x : \forall \tilde{A} \in \boldsymbol{\mathcal{C}}_t, \tilde{A}x \le \alpha\} \subset \mathcal{S}$ whp. Instead, $\widetilde{\mathcal{S}}_t(\delta) \supset \mathcal{S}$ whp. Of course, since the known constraints in A are enforced, $\widetilde{\mathcal{S}}_t(\delta) \subset \mathcal{X}$.

Algorithm 1 Doubly-Optimistic Safe Selection (DOSS) (δ)

Input: $\delta \in (0,1)$
for $t = 1, 2, \cdots$ **do**
  Construct $\widetilde{\mathcal{S}}_t(\delta)$ as in (2).
  Optimize (3) and play x_t.
  Observe r_{t,x_t}, $\{s^i_{t,x_t}\}$
  Update $X, R, \{S^i\}, V, C$
end for

4. Warm Up: Polynomial Bounds on Regret and Safety Cost, and Going Beyond

An immediate application of the approach of Abbasi-Yadkori et al. (2011) yields the following basic result, establishing that DOSS is a reasonable procedure.

Theorem 3 *The actions* $\{x_t\}$ *of* DOSS(δ) *yield, whp,* $\mathscr{E}_T = \widetilde{O}(\sqrt{d^2 T})$, *and* $\mathscr{S}_T = \widetilde{O}(\sqrt{d^2 T})$.

Proof Sketch. By Lemma 2, $\forall t, \theta \in \mathcal{C}_t^\theta, A \in \boldsymbol{\mathcal{C}}_t$ whp, and so $x^* \in \widetilde{\mathcal{S}}_t(\delta)$ whp. Thus, (3) ensures $\langle\tilde{\theta}, x_t\rangle \ge \langle\theta, x^*\rangle$. But, by the noise-scale characterisation in Lemma 2, $\langle\tilde{\theta}, x_t\rangle \le \langle\theta, x_t\rangle + \rho_t(x_t)$, and so $\langle\theta, x^* - x_t\rangle \le \rho_t(x_t)$. On the other hand, since $x_t \in \widetilde{\mathcal{S}}_t$, there exists some $\tilde{A} \in \widetilde{\mathcal{S}}_t : \tilde{A}x_t \le \alpha$. But again $\alpha \ge \tilde{A}x_t \ge Ax_t - \rho_t(x_t)\mathbf{1}_U$, and so $\max_i(\langle a^i, x\rangle - \alpha^i)_+ \le \rho_t(x_t)$ Consequently, $\mathscr{E}_T \le \sum_{t\le T}\rho_t(x_t)$, *and* $\mathscr{S}_T \le \sum_{t\le T}\rho_t(x_t)$, and the bound follows from Lemma 2.

Polytopes to Break Through $\sqrt{T}$? The above result holds in fact holds over any convex domain without change. However, our domain of interest is linear programming, i.e., $\mathcal{S}$ and $\mathcal{X}$ are polytopes, and thus is much more structured. Indeed, for linear bandits with *known* $\mathcal{S}$, optimistic play yields instance-dependent *logarithmic* regret bounds for large T (Abbasi-Yadkori et al., 2011). Such results rely on the observation that if $\mathcal{S}$ is known, any action that an optimistic method takes lies in the *finite* set of extreme points of $\mathcal{S}$. Therefore, $\exists\Delta > 0$ such that for any suboptimal x_t, $\langle\theta, x^* - x_t\rangle \ge \Delta$, and which directly leads to regret bounds of $O(\log^2(T)/\Delta)$.[6]

This raises the natural question: *can we also attain logarithmic bounds on* $(\mathscr{E}_T, \mathscr{S}_T)$ *when some of the constraints are unknown?* Answering this will occupy us for the remainder of this paper.

5. Impossibility of Simultaneous Logarithmic Bounds on Both Efficacy and Safety

The question we raised in §4 needs a little care to formulate: since we do not know $\mathcal{S}$, it is unreasonable to expect bounds that scale only with the optimality gap of actions, since unsafe points

6. The key trick is that $\mathscr{R}_T \le \sum\rho_t(x_t)\mathbf{1}\{\rho_t(x_t) \ge \Delta\} \le \sum\rho_t(x_t)^2/\Delta$.

outside of $\mathcal{S}$ must also be eliminated. We can account for this by also considering the spurious extreme points induced by the bounding polytope $\mathcal{X}$, and consider

$$\mathcal{E} := \{\text{extreme points of } \mathcal{S}\} \cup \{\text{extreme points of } \mathcal{X}\}.$$

Now, $\mathcal{E}$ is again a finite set, and for any $x \in \mathcal{E} \setminus \{x^*\}$, either x is feasible but suboptimal, in which case $\langle \theta, x^* - x \rangle > 0$ or it is infeasible, in which case $\max_i(\langle a^i, x\rangle - \alpha^i) > 0$. Let us say that an instance is Δ-well separated if the smallest such lower bound is at least Δ. Then note that if we knew $\mathcal{E}$, then it is easy to obtain $O(\Delta^{-1}\log^2 T)$ bounds using the technique described in §4. The refined question of interest is: *can we always attain logarithmic efficacy regret and safety violations for well-separated SLB instances?* Surprisingly, the answer to this is negative, as we show in §F.1.

Theorem 4 *For every SLB algorithm, there exists a* $1/8$*-well-separated instance on which the algorithm must incur* $\max(\mathbb{E}[\mathscr{E}_T], \mathbb{E}[\mathscr{S}_T]) = \Omega(\sqrt{T})$.

Proof Sketch. The obstruction is illustrated in Figure 2. We study the 1D problem $\max x$ under the known constraints $0 \le x \le 1$, reward parameter $\theta = 1$, and the unknown constraint $ax \le 1/4$. Consider the case $a \in \{(1\pm\kappa)/2\}$ for $\kappa \le 1/4$. For these instances, $\mathcal{E} = \{0, 1, 1/2(1\pm\kappa)\}$, and the last point is optimal. Further, 0 is at least $(2(1 \pm \kappa))^{-1} \ge 2/5$-inefficient, and 1 violates the constraint by $(1\pm2\kappa)/4 \ge \frac{1}{8}$, and so either instance is $1/8$-well-separated.

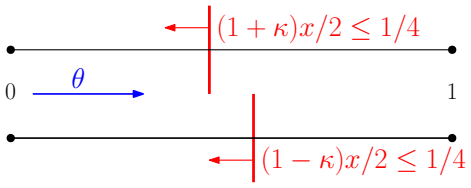

Figure 2: An obstruction to logarithmic bounds in safe linear bandits.

But, no matter the x_ts, we cannot estimate a to error better than $1/\sqrt{t}$, and so we cannot eliminate either of $1\pm\kappa/2$ if $t < \frac{1}{\kappa^2}$. Now, if the truth were $a = (1-\kappa)/2$, playing $x_t < 2/1-\kappa^2$ incurs inefficacy $\ge 2\kappa$, and conversely if $a = (1+\kappa)/2$, playing $x_t \ge 2/1-\kappa^2$ violates safety by 2κ. Thus, at least one of $\mathscr{E}_T^{(1+\kappa)/2}$ and $\mathscr{S}_T^{(1-\kappa)/2}$ must be $\Omega(\kappa \cdot \min(T, \kappa^{-2}))$. The bound follows by choosing $\kappa = 1/\sqrt{T}$. □

Impossibility of *instance-dependent* simultaneous logarithmic bounds. We highlight that the above lower bound scales as $\sqrt{T}$ despite *constant order separation* in the instance. This stands in sharp contrast to existing minimax lower bounds for standard bandits (e.g. Shamir, 2015), which set $\Delta \sim T^{-1/2}$ to show $\Omega(\sqrt{T})$ bounds. The barrier to logarithmic control in SLBs is more fundamental, and comes from an inability to refine the precise location of the optimal point, rather than because there are suboptimal points in the noiseless problem that have small gaps. In other words, the issue is one of *precision* rather than one of hardness in the underlying LP, and this makes it impossible to be both very efficient and very safe on all instances. We further observe that the construction is extremely simple, and thus can embed into essentially any class of instances (e.g., by revealing a line that the optimum lies on), and so this issue is pervasive, rather than limited to specific hard cases.

Nevertheless, the result does not preclude that *one* of $\mathscr{E}_T, \mathscr{S}_T$ is small. In fact, although they need the extra information $(x^{\mathsf{s}}, M^{\mathsf{s}})$, we can view PO schemes as saturating this bound, since they achieve $\mathscr{E}_T = \widetilde{O}(\sqrt{T}), \mathscr{S}_T = 0$. We shall show in the subsequent that the DO method DOSS saturates the bound as well, attaining $\mathscr{E}_T = O(\log^2 T), \mathscr{S}_T = \widetilde{O}(\sqrt{T})$, *without this extra information*.

A dual view, and our approach. From an analytic point of view, the failure to improve on $\sqrt{T}$ bounds can be seen as a breaking down of the assertion that *in polytopal domains, optimistic methods play on the finite set of extreme points of the polytope*. Indeed, in the SLB scenario, the polytope is not known, and these extreme points are effectively smeared out into sets of diameter $\Omega(t^{-1/2})$ due to estimation errors in $\hat{A}_t$. Thus, the primal approach to analysing polytopes breaks down.

As described in §1, our resolution to this issue lies in the dual view of extreme points of a polytope as points that activate exactly d independent constraints. Due to this, we can view optimism with known $\mathcal{S}$ as activating some d constraints of $\{Ax \leq \alpha\}$. This view generalises: we show that there exists some $\tilde{A} \in \boldsymbol{\mathcal{C}}_t$ such that under DOSS, x_t activates at least d constraints of $\mathcal{S}_{\tilde{A}} := \{\tilde{A}x \leq \alpha\}$. Naturally, such a set of constraints is a 'poor' choice if saturating these constraints for $\{Ax \leq \alpha\}$ yields poor or infeasible points. The key idea is that if I is 'poor', then the only way DOSS would prefer to activate noisy versions of the constraints in I is if the noise-scale $\rho_t(x_t)$ is large.

This sets up a two-step attack to control $\mathscr{E}_T$. First, we use the dual argument above to study a 'combinatorial identification' question of whether DOSS finds the 'right' set of constraints to saturate. This is addressed by developing new dual notions of gaps for sets of constraints, which arise by an approach reminiscent of the global sensitivity analysis of LPs (Bertsimas and Tsitsiklis, 1997, Ch.5), and is the subject of §6. Secondly, even if the 'right' set of constraints are activated, DOSS may play ineffectively due to noisy estimation of $\tilde{A}$. Standard arguments (such as §4) only yield a $\sqrt{T}$ control on this. Instead, we show that due to the optimism of (3), if $t \geq d$ then activating any 'optimal' set of constraints yields $x_t : \langle \theta, x_t - x^* \rangle > 0$, which controls efficacy loss due to such play to $O(1)$. This argument is elementary, but involved, and entails a careful analysis of the structure of (3) when optimal constraints are activated, as developed in §7, and §E.1.

6. Structural Behaviour of DOSS, and Noise-Scale Lower Bounds

We proceed to formally define basic index sets, as well as the gaps associated with these index sets, which lead lead to the key consequence that DOSS does not play 'suboptimal' index sets too often.

6.1. Basic Index Sets

We begin by formalising 'sets of constraints', and 'activation' as mentioned in §5.

Definition 5 *An* index set I *is a subset of* $[1:m]$. *Such a set is* I *is called a* basic index set (BIS) *if* $|I| = d$. *The set of points that* activate an index set I *is defined as* $\mathcal{X}^I := \{x \in \mathcal{S} : A(I)x = \alpha(I)\}$.

Notice that we demand that activating points are feasible, i.e., $\mathcal{X}^I \subset \mathcal{S}$. The set $\mathcal{X}^I$ may be empty, or a singleton, or an affine segment. We shall find the following linear-algebraic terminology useful.

Definition 6 *A BIS* I *is called* (i) feasible *if* $\mathcal{X}^I \neq \emptyset$ *and* infeasible *otherwise;* (ii) *suboptimal if* $x^* \notin \mathcal{X}^I$ *and* optimal *otherwise;* (iii) full rank *if the row vectors of* $A(I)$ *span* $\mathbb{R}^d$.

Example 7 *To illustrate these definitions, consider the LP*

$$\max x_1 + 2x_2 : \underbrace{x_2 \leq 1/2}_{\textit{unknown}}, \quad \underbrace{x_1 \geq x_2, x_1 \leq 1, x_2 \geq 0}_{\textit{known}}.$$

Foregoing normalisation for clarity, we have $m = 4, U = 1$ *and the parameters* $\theta = (1,2), a^1 = (0,1), a^2 = (-1,1), a^3 = (1,0), a^4 = (0,-1), \alpha = (0.5,0,1,0)$. *There are* $\binom{4}{2} = 6$ *basic index sets,*

$$I_1 = \{1,2\}, I_2 = \{1,3\}, I_3 = \{1,4\},$$
$$I_4 = \{2,3\}, I_5 = \{2,4\}, I_6 = \{3,4\}.$$

Of these, I_2 *is optimal, and the rest suboptimal, with* $x^* = (1, 1/2)$; I_3 *is rank-deficient while the rest are full-rank;* I_3 *and* I_4 *are infeasible, while the rest are feasible.*

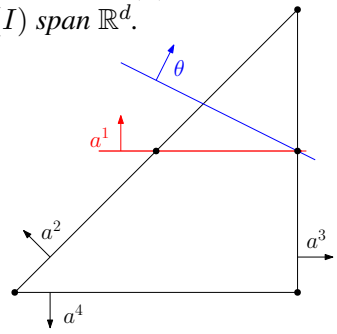

Figure 3: Illustration of Ex. 7. The black lines represent the known constraints, the red line is the unknown constraint, and the blue line is the locus of optimality.

Noisy Activation. For SLBs, instead of the true constraint matrix A, DOSS must work with noisy estimates of it, the $\tilde{A}$s. We extend the notion of BIS activation to handle this fuzziness in constraints.

Definition 8 *The set of points that* noisily activates *a BIS I at time t is*

$$\widetilde{\mathcal{X}}_t^I := \{x \in \widetilde{\mathcal{S}}_t : \exists \tilde{A} \in \boldsymbol{\mathcal{C}}_t, \tilde{A}(I)x = \alpha(I)\}.$$

Note that $\widetilde{\mathcal{X}}_t^I \subset \widetilde{\mathcal{S}}_t \subset \mathcal{X}$. The main structural result is the following observation.

Proposition 9 *The actions of* DOSS *must noisily activate at least one BIS, i.e.* $\forall t, \exists I_t : x_t \in \widetilde{\mathcal{X}}_t^{I_t}$.

If x_t noisily activates the BIS I at time t, we shall say that I is *played* at time t. Note that more than one BIS may be played at a time (since x_t can lie in the intersection of many $\widetilde{\mathcal{X}}_t^I$s).

6.2. Gaps of Suboptimal BISs

We argue that *if* DOSS *noisily activates a suboptimal BIS at t, then the noise scale $\rho_t(x_t;\delta)$ must be large.* To show this, we develop two *gaps* for suboptimal BISs: the *feasibility gap* and the *efficacy gap*, which respectively exploit the permissibility and optimism of x_t. Our results will lower bound $\rho_t(x_t;\delta)$ by the *larger* of these gaps when suboptimal BISs are played. The overall constructions are essentially via a reduction to global linear programming sensitivity analysis. This is necessary: since we do not know the constraints in A or θ, perturbations in this matrix (as represented by $\tilde{A}$) may, and indeed do, cause the optimal x^* to appear suboptimal.

The basic structure we use is the following localisation of x_ts played by DOSS, proved in §D.1 as a simple consequence of Lemma 2. From here onwards, we shall just write ρ_t instead of $\rho_t(x_t;\delta)$.

Lemma 10 *For $\zeta \in [0,\infty)$, define the* activation polytope *of I at scale ζ as*

$$\mathcal{T}(\zeta; I) := \{x : Ax \le \alpha + \zeta \mathbf{1}_U, A(I)x \ge \alpha(I) - \zeta \mathbf{1}_U(I)\}.$$

If the confidence sets are consistent, and if the action of DOSS *at time t, x_t, noisily activates the BIS I, then $x_t \in \mathcal{T}(\rho_t; I)$, and further, $\langle \theta, x^* - x_t \rangle \le \rho_t$.*

6.2.1. INTUITIVE ILLUSTRATION OF GAPS

To expose the key components that allow DOSS to control the play of suboptimal BISs, we will first consider the feasible, full-rank, and suboptimal BIS $I_1 = \{1,2\}$ in Ex. 7. Due to the full-rank, the constraints of I are activated by a unique point, x^I. Since I is suboptimal, there is a positive 'efficacy separation' between x^I and x^*, denoted $\gamma(I) := \langle \theta, x^* - x^I \rangle$. In our example, $x^{I_1} = (1/2, 1/2)$, and $\gamma(I_1) = 1/2$.

Efficacy Gap. Under noisy activation of I, the point x_t may depart from x^I, but it cannot go too far. Indeed, by Lemma 10, x_t must lie in the activation polytope $\mathcal{T}(\rho_t; I)$, which is a skewed ℓ_∞-box of scale ρ_t containing x^I. In our example, $\mathcal{T}(\rho_t; I_1) = \{x : x_1 = x_2, x_2 \in 1/2 \pm \rho_t\}$. This localisation constrains how large $\langle \theta, x_t \rangle$ can be. Indeed, there exists a constant $\mathfrak{s}(I)$, which we call the spread of I, such that

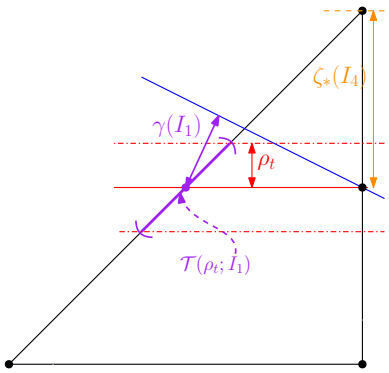

Figure 4: Illustration of gaps in Ex. 7. x^{I_1} is the purple dot, and the activation polytope $\mathcal{T}(\rho_t; I_1)$ is shown in purple, along with the separation $\gamma(I_1)$. The spread $\mathfrak{s}(I_1)$ is the inner product of the direction in which $\mathcal{T}$ varies and θ. For I_4, the feasibility gap $\zeta_*(I)$ is illustrated geometrically in orange.

$\max_{x \in \mathcal{T}(\zeta;I)} \langle \theta, x - x^I \rangle \leq \zeta \mathfrak{s}(I)$. In effect, $\mathfrak{s}(I)$ is a measure of how well the geometry induced by I near x^I aligns with θ, e.g., for $I_1, \mathfrak{s}(I_1)$ is the inner product between θ and $(1,1)$, the direction along which $\mathcal{T}$ varies.

Thus, $\langle \theta, x_t \rangle \leq \langle \theta, x^I \rangle + \rho_t \mathfrak{s}(I)$. But, since $\langle \theta, x^I - x^* \rangle = -\gamma(I)$, this implies $\langle \theta, x_t \rangle \leq \langle \theta, x^* \rangle - \gamma(I) + \rho_t \mathfrak{s}(I)$. This lies in tension with Lemma 10, which states that $\langle \theta, x_t \rangle \geq \langle \theta, x^* \rangle - \rho_t$. Resolving this tension yields the lower bound $\rho_t \geq \eta_*(I) := \gamma(I)/(1 + \mathfrak{s}(I))$. We call the constant $\eta_*(I)$ the *efficacy gap* of I. For Ex. 7, $\eta_*(I_1) = 1/8$.

Safety Gap. It is also possible that x_t noisily activates an infeasible BIS, such as $I_4 = \{2, 3\}$ in Ex. 7. In this case, a conflict arises between the inequalities defining the activation polytope $\mathcal{T}(\zeta; I)$: if I is infeasible, then $\mathcal{T}(0; I) = \mathcal{X}^I = \{Ax \leq \alpha, A(I)x \geq \alpha(I)\} = \emptyset$, and by right-continuity $\mathcal{T}(\zeta; I)$ is empty for small ζ. Let us define $\zeta_*(I)$ to be the smallest scale at which $\mathcal{T}(\zeta; I)$ is nonempty. Since $x_t \in \mathcal{T}(\rho_t; I)$, it follows that if x_t activates a BIS I, then $\rho_t \geq \zeta_*(I)$. We call $\zeta_*(I)$ the *safety gap* of I. In Ex. 7, $\mathcal{T}(\zeta; I_4) = \{x : x_1 = 1, x_1 = x_2, x_2 \leq 1/2 + \zeta\}$, and so $\zeta_*(I_4) = 1/2$.

Summary. The above illustrates two basic tensions in selecting suboptimal BISs. If a BIS I is infeasible, then activating it requires that ρ_t dominates its safety gap, and if I is feasible but suboptimal, then activation requires that ρ_t exceeds its efficacy gap. We formalise this concept below.

6.2.2. FORMAL DEFINITIONS OF THE GAPS

We give a unified treatment of the safety and efficacy gaps by analysing a parameterised LP with feasible set determined by the local structure induced by a BIS I, as encapsulated in Lemma 10.

Definition 11 *For a BIS I, and $\zeta \geq 0$, the* optimistic LP *at scale ζ induced by I is defined as* $P(\zeta; I) := \sup\{\langle \theta, x \rangle : x \in \mathcal{T}(\zeta; I)\}$, *with the convention that* $\sup \emptyset = -\infty$.

Since by Lemma 10, x_t lies in $\mathcal{T}(\rho_t; I)$ if it noisily activates I, this yields $\langle \theta, x_t \rangle \leq P(\rho_t; I)$. So, the behaviour of $P(\zeta; I)$ with ζ let us capture the tensions we illustrated in the previous section.

Definition 12 *We define the* ***feasibility gap*** *of a BIS I as*

$$\zeta_*(I) := \inf\{\zeta \geq 0 : P(\zeta; I) > -\infty\}.$$

We define the efficacy separation *of I as $\gamma(I) := \langle \theta, x^* \rangle - P(\zeta_*(I); I)$, and the* spread *of I as* $\mathfrak{s}(I) := \inf\{C : \forall \zeta \geq \zeta_*(I), P(\zeta; I) \leq P(\zeta_*(I)) + C(\zeta - \zeta_*(I))$, *which yield the* ***efficacy gap*** *of I,*

$$\eta_*(I) = \frac{\gamma(I) + \zeta_*(I)\mathfrak{s}(I)}{1 + \mathfrak{s}(I)}.$$

The definitions above concretise the quantities described in §6.2.1. The key consequence of these definitions is the following 'noise-scale lower bound on activating poor BISs,' shown in §D.2.

Lemma 13 *For any suboptimal BIS I, $\max(\zeta_*(I), \eta_*(I)) > 0$. Further, under consistency of the confidence sets, if x_t activates a suboptimal BIS I, then $\rho_t(x_t; \delta) \geq \max(\zeta_*(I), \eta_*(I))$.*

Note here that the noise-scale needed to play I is driven by the *larger* of the efficacy and safety gap at I. This is natural: these quantities measure the 'extent' of infeasibility or inefficacy of I, and thus the larger one determines the rate at which evidence of the suboptimality of I is accumulated.

6.3. Gap of the Problem, and Controlling the Play of Suboptimal BISs

In light of Lemma 13, the following is natural.

Definition 14 *The **gap of an SLB instance** is defined as* $\Gamma := \min_I \max(\zeta_*(I), \eta_*(I))$.

The main result of this section shows that Γ^{-2} bounds how often suboptimal BISs are played.

Theorem 15 *Let* $\{x_t\}$ *denote the actions of* DOSS(δ) *on an SLB instance. Then, with probability at least* $1-\delta$, *if at any time* t, x_t *noisily activates a suboptimal BIS, then* $\rho_t(x_t;\delta) > \Gamma$. *Further, the total number of times suboptimal BISs are played is bounded as*

$$\sum_t \mathbb{1}\{\exists \textit{suboptimal BIS } I : x_t \in \widetilde{\mathcal{X}}_t^I\} = O\left(\Gamma^{-2}\left(d^2 \log^2 T + d\log(T)\log(U/\delta)\right)\right).$$

This result, shown in §D.3, implies that most of the time, DOSS plays actions such that the noisy constraints they activate are precisely those that x^* saturates. In other words, while the method may not be able to locate x^* itself with precision better than $O(1/\sqrt{t})$, it can identify the binding constraints, and, most of the time, the actions of DOSS focus on activating these constraints.

7. Controlling Efficacy Regret and Total Safety Violation

We now come to the main results of the paper. The previous section tells us that suboptimal BISs cannot be played too often, effectively controlling a 'dual' type of regret. We proceed to translate these results into bounds on the 'primal' quantities $\mathscr{E}_T$ and $\mathscr{S}_T$. This requires us to account for the times when only optimal BISs (I such that $x^* \in I$) are played. We can control the behaviour of such times under the following weak nondegeneracy condition at the optimum.

Assumption 16 *Every optimal BIS (i.e.,* $I : x^* \in \mathcal{X}^I$*) is full-rank. Further, the noise* w_t^S *is generic in the sense that the probability that it lies in any subspace of less than* d *dimensions is zero.*

Note that the condition does not require the uniqueness of the optimum. Instead, nondegeneracy is demanded in the sense that any size d subset of all the constraints that x^* saturates constitutes a full rank BIS. The effect of this is to mainly exclude pathologies, such as the case in $\mathbb{R}^2$ where two identical constraints are placed on the system, and both pass through the optimum (i.e., $(a^i, \alpha^i) = c(a^j, \alpha^j)$ for some pair i, j). Notice that in standard linear programmming, such constraints would be eliminated during pre-processing, which we cannot do since we do not know all of the constraints. Nevertheless, since the constraints represent limitations on different safety scores, it is unlikely in practice that these would be linearly dependent. Further, note that Assumption 16 allows x^* to be degenerate in the sense that it may lie on many more than d constraints. Of course, the genericity of noise is a standard condition, and can be met by adding an arbitrarily small continuous noise to the feedback. The main utility of this assumption is the following result, which is argued in §E.1.

Lemma 17 *Under assumption 16, if the confidence sets are consistent,* $t \geq d+1$, *and the action* x_t *of* DOSS(δ) *is that* x_t *only noisily activates the optimal BIS, then* $\langle \theta, x_t \rangle \geq \langle \theta, x^* \rangle$.

In other words, when only the optimal BISs are played, the action x_t cannot be ineffective! The proof relies on using the optimal BIS I to construct a 'localised' program that the solutions $(\tilde{\theta}_t, x_t)$ and witness $\tilde{A}_t$ of (3) must also optimise. The assumption is used to make this part effective, and in general the same holds if $\theta \in$ row-span$(A(I))$. The final statement then follows through an elementary, but involved, analysis of structure of optimal solutions of this localised program.

Coupling the above with Theorem 15 yields our main result, shown in §E.2

Theorem 18 *Under assumption 16, w.p. $\geq 1-\delta$, the actions of* DOSS*(δ) yield*

$$\mathscr{E}_T = O\left(\Gamma^{-1}(d^2 \log^2 T + d \log T \log(U/\delta))\right), \text{ and } \mathscr{S}_T = \widetilde{O}\left(\sqrt{d^2 T(\log^2 T + \log T \log(U/\delta))}\right).$$

In light of Theorem 4, we see that up to polylog factors, DOSS saturates the lower bound, with a bias towards minimising the efficacy regret. While the gain in efficacy performance over PO methods is evident, we again stress the advantage in terms of the lack of prior knowledge of a safe ball in $\mathcal{X}$. We further note that the costs scale logarithmically with the number of unknown constraints, U.

Tightness of Dependence on Γ. Exploiting a subtle reduction of safe Multi-Armed Bandits problems to SLB problems, we show in §F.2 that the inverse dependence on Γ is necessary.

Theorem 19 *Fix a $c \in (0,1)$. For any $\Gamma \leq 1/16$, and any method that ensures that in every SLB instance, $\max(\mathscr{E}_T, \mathscr{S}_T) = O(T^{1-c})$, there exists an instance of the SLB problem with gap at least Γ, such that $\liminf \frac{\max(\mathbb{E}[\mathscr{E}_T], \mathbb{E}[\mathscr{S}_T])}{\log T} \geq c/108 \cdot \Gamma^{-1}$.*

7.1. Improved Safety Performance Under Tolerance

While Theorem 18 is tight in terms of $\mathscr{S}_T$, given that it achieves polylogarithmic $\mathscr{E}_T$, the polynomial dependence can nevertheless be considered prohibitive. To improve upon this, we study three concrete scenarios in which this dependence may be improved. At the core, each of these cases relaxes the SLB problem so that the precision barrier discussed in §5 does not arise, thus illustrating that this condition is the sole obstruction to polylogarithmic control on $\mathscr{S}_T$.

Finite Precision Slack in Constraint Levels As a first pass, we may allow for a finite amount of violation of constraints without any penalty, e.g., through the ε-precision metric $\mathscr{S}_T^{\varepsilon} := \sum_{t \leq T} \max_i (\langle a^i, x \rangle - \alpha^i)_+ \mathbb{1}\{\exists i : \langle a^i, x \rangle - \alpha^i > \varepsilon\}$. Such a relaxation is quite pertinent in scenarios such as drug trials or engineering design applications (where ε can be set to a small factor of α^i) or if the α^i are estimated values[7] (where ε can be the error level in these estimates). In this context, we show in §E.3.1 that

Theorem 20 *With probability at least $1-\delta$,* DOSS*(δ) ensures that* simultaneously for every $\varepsilon > 0$

$$\mathscr{E}_T = O\left(\Gamma^{-1} d^2 \log^2 T\right) \quad \text{and} \quad \mathscr{S}_T = O\left(\varepsilon^{-1} d^2 \log^2 T\right).$$

The main point of interest in the result above is that it holds *simultaneously* for every value of ε. Indeed, DOSS does not need ε as a parameter, and it only arises in the analysis. This means that the method adapts to the precision requirements of the domain at hand. Note further that setting $\varepsilon = T^{-c}$ for $c > 1/2$ yields $\sum_{t \leq T} (\max_i \langle a^i, x_t \rangle - \alpha^i - T^{-c})_+ = \widetilde{O}(T^{1-c})$, i.e., as $T \nearrow \infty$, DOSS rapidly converges towards feasibility, and gains over Theorem 18 are realised with decaying precision slack.

Finite Precision in Constraint Parameters Rather than treating the precision in the constraint levels, it may be possible that the constraint parameters are restricted to a finite grid. Generically, such a structure arises in settings modeled as integer programs (up to a unit factor), and particular examples include drug discovery (e.g. Radhakrishnan and Tidor, 2008), where constraints indicate requirements that a compound binds to certain receptors, and so are naturally binary. We can formalise this by specifying a finite set P which describes the 'grid' that A must lie in. Naturally, we can modify DOSS to exploit this by restricting the construction of $\widetilde{\mathcal{S}}_t(\delta)$ in (2) to $\tilde{A} \in \mathcal{C}_t^{\mathsf{P}} = \mathcal{C}_t \cap \mathsf{P}$. We argue in §E.3.2 that this change implicitly introduces a finite set of possible actions when only optimal BISs are activated, which in turn yields the following result.

7. this is quite common: process and measurement variations mean that an exact threshold for the quality of components necessary to ensure safe behaviour is not known, and must usually be fixed empirically.

Theorem 21 *If the constraint parameters lie in a finite precision set, then there exists a constant* $\pi > 0$ *such that w.p.* $\geq 1-\delta$*, the actions of* DOSS*(*δ*) satisfy* $\max(\mathscr{E}_T, \mathscr{S}_T) = O(\min(\Gamma, \pi)^{-1} d^2 \log^2 T)$.

Finite Action Spaces Finally, if we instead consider the commonly studied case of only having a finite number of possible actions (Abbasi-Yadkori et al., 2011; Dani et al., 2008; Agrawal and Devanur, 2016), then the issues of primal precision do not arise, since we do not need to exactly know the constraints in order to exactly locate any action. If we simply define $\Delta = \min_{\mathcal{X}} \max(\langle \theta, x^* - x\rangle, \max_i(\langle a^i, x\rangle - \alpha^i)_+)$, then merely employing the techniques described in §4 yields (see§E.3.3)

Proposition 22 *Over finite actions spaces, with probability at least* $1-\delta$*, the actions of* DOSS*(*δ*) ensure that* $\max(\mathscr{E}_T, \mathscr{S}_T) = O(\Delta^{-1} d^2 \log^2 T)$.

8. Simulations

We verify the theoretical study above with simulations over Example 7, and study the relative performance of DOSS and the optimistic-pessimistic method Safe-LTS Moradipari et al. (2021). These implementations are based on the following relaxation of Algorithm 1.

Computationally Feasible Relaxation A well-known barrier to implementing Algorithm 1 is that even if all constraints were known, the program (3) is non-convex (Dani et al., 2008). In our case, this is further complicated by the fact that the set $\widetilde{\mathcal{S}}_t$ needs to be determined. Following Dani et al. (2008), we approach these issues by constructing *box confidence sets*, i.e.,

$$\mathcal{C}_{t,1} := \{\tilde{A} : \forall i, \|(\tilde{a}^i - \hat{a}^i) V_t^{1/2}\|_1 \leq \sqrt{d\beta_t}\}.$$

Since $\|\cdot\|_2 \leq \|\cdot\|_1 \leq \sqrt{d}\|\cdot\|_2, \mathcal{C}_{t,1} \subset \mathcal{C}_t$. Further, due to the same equivalence, the ℓ_2-based analysis persists, up to a blowup of $\sqrt{d}$ in ρ_t, and thus running DOSS with $\mathcal{C}_{t,1}$ worsens our bounds from $(d^2 \log^2 T, \sqrt{d^2 T})$ to $(d^3 \log^2 T, \sqrt{d^3 T})$.

The main advantage of $\mathcal{C}_{t,1}$ lies in the fact that the box-confidence sets are polytopes. Due to this, the $\tilde{A}_t$ that are active for the optimistic action x_t must lie at the extreme points of these sets. Since each set has only $2d$ extreme points, this allows us to determine x_t by solving $(2d)^{U+1}$ convex programs, which is computationally feasible so long as U is small. Of course, this complexity remains painfully slow as U grows. Finding versions of DOSS that are computationally practical for a large number of unknown constraints remains an interesting open problem.

Setting. We implement DOSS on with the L_∞ relaxation above on the instance of Example 7 over the horizon $T = 10^4$, and with the parameters $\lambda = 2, \delta = 1/(4T) = 2.5 \times 10^{-5}$. The noise in observations is independent and Gaussian, with variance 0.1. Notice that for this instance, $\Gamma = 1/8$.

Behaviour of DOSS. Our main observation is that DOSS *is very effective, and has well-controlled violations.* Figure 5 shows the efficacy regret $\mathscr{E}_t$ and both the arbitrary precision safety violations $\mathscr{S}_t$ and the finite precision safety violations $\mathscr{S}_t^\varepsilon$ for the value $\varepsilon = 0.05 = 2\Gamma/5$. The simulations validate our main claims of strong efficacy regret control, and well-behaved growth of safety violations. Indeed, observe that the efficacy regret is essentially zero over most of the runs (with rare runs rising to $\mathscr{E}_{10^4} \approx 100$). This property arises since DOSS very rarely plays suboptimal BISs (see the following discussion and Figure 6), and when it plays the optimal BIS, it plays a 'over-efficient' but unsafe point. Further, the extent of the lack of safety of the actions chosen by DOSS is well-controlled, as seen in the behaviour of $\mathscr{S}_T$. The finite precision regret shows even stronger control, with growth essentially halted at $t \approx 5000$, validating the analysis underlying Theorem 18.

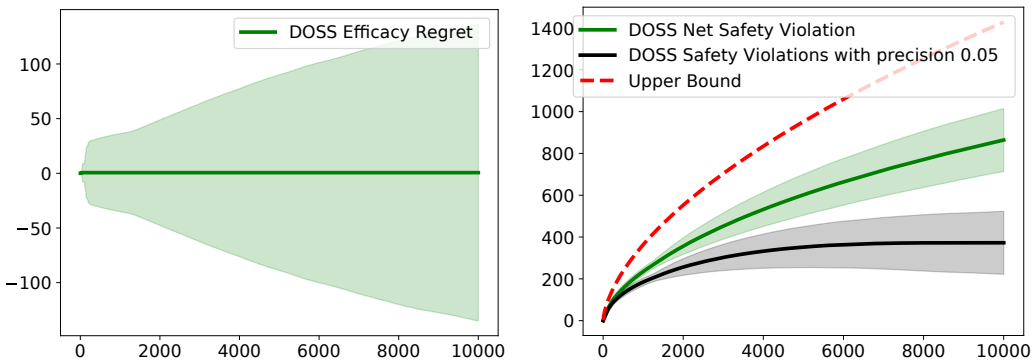

Figure 5: Efficacy Regret and Safety Violation of DOSS . We plot averages and one standard deviation confidence regions over 30 runs for $\mathscr{E}_T$ (left) and both $\mathscr{S}_t$ and $\mathscr{S}_t^{0.05}$ (right). We also plot the upper bounds we show in the latter to contextualise the observations. Observe that the efficacy regret is marginal: the mean is essentially 0, and the variance limited. Further, observe that the growth of the net safety violation $\mathscr{S}_t$ is well-controlled, and lies far below the bounds of §7. Further, the finite precision violations show a strong flattening, as is expected from Theorem 18.

DOSS rarely activates suboptimal index sets. In Figure 6, we plot the number of times that DOSS noisily activates a suboptimal BIS, i.e., any index set other than $I_2 = \{1, 3\}$. The main observation is that this occurs very rarely: indeed, over the horizon of 10^4, most runs do not activate suboptimal BISs more than 100 times. This is far below the upper bound of Theorem 15.

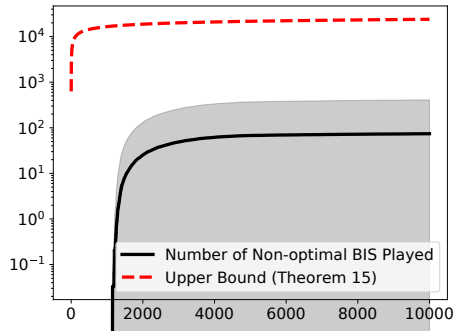

Figure 6: Suboptimal BIS activation by DOSS in the instance of Example 7. Observe that such activation is very rare, typically far less than 1% of the times, and the growth is essentially flat.

DOSS compares favourably to PO methods In §H, we report the behaviour of the PO method *safe-LTS* (Moradipari et al., 2021) on the same instance. Our key observation is that the efficacy regret of safe-LTS increases much faster than the safety-violations of DOSS, indicating that the safe-LTS type methods expands their safe sets towards optimality slowly, while DOSS contracts towards safety much faster. Indeed while the safety-violation of DOSS at $T = 10^4$ is about 800, the efficacy violation of safe-LTS at the same time scale is 3000, indicating that it is at least 0.25-separated from the boundary of $\mathcal{S}$ even at $T = 10^4$.

9. Discussion

The SLB problem is inherently challenging due to the roundwise enforcement of constraints. Our works offers new, and refined insights into both the hardness of the problem through our instance-dependent superlogarithmic lower bound, and to the effectiveness of doubly-optimistic methods for the same through our strong control on $\mathscr{E}_T$. In the process, we developed a new dual viewpoint of the SLB problem, by developing gaps for *sets* of constraints, which we believe is a conceptually important tool for such problems. Of course, a number of interesting questions remain open, e.g., are there computationally efficient ways to implement doubly-optimistic strategies for large U; or if one can design methods that attain the strong safety guarantees of PO methods, but without making the strong assumptions of prior knowledge of safe points. We believe that tackling these challenges is key to the effective use of bandit feedback in practical scenarios.

Acknowledgments

We acknowledge support by the Air Force Research Laboratory grant FA8650-22-C1039, Army Research Office grant W911NF2110246, and the National Science Foundation grants CCF-2007350 and CCF-1955981.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

Amirhossein Afsharrad, Ahmadreza Moradipari, and Sanjay Lall. Convex methods for constrained linear bandits. *arXiv preprint arXiv:2311.04338*, 2023.

Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 29:3450–3458, 2016.

Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 989–1006, 2014.

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

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. *arXiv preprint arXiv:1908.05814*, 2019.

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

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

Martino Bernasconi, Federico Cacciamani, Matteo Castiglioni, Alberto Marchesi, Nicola Gatti, and Francesco Trovò. Safe learning in tree-form sequential decision making: Handling hard and soft constraints. In *International Conference on Machine Learning*, pages 1854–1873. PMLR, 2022.

Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena scientific Belmont, MA, 1997.

Romain Camilleri, Andrew Wagenmaker, Jamie Morgenstern, Lalit Jain, and Kevin Jamieson. Active learning with safety constraints. *arXiv preprint arXiv:2206.11183*, 2022.

Emil Carlsson, Debabrota Basu, Fredrik D Johansson, and Devdatt Dubhashi. Pure exploration in bandits with linear constraints. *arXiv preprint arXiv:2306.12774*, 2023.

Tianrui Chen, Aditya Gangrade, and Venkatesh Saligrama. Strategies for safe multi-armed bandits with logarithmic regret and risk. In *Proceedings of the 39th International Conference on Machine Learning*, pages 3123–3148, 2022.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, 2008.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

Spencer B Gales, Sunder Sethuraman, and Kwang-Sung Jun. Norm-agnostic linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 73–91. PMLR, 2022.

Spencer Hutchinson, Berkay Turan, and Mahnoosh Alizadeh. The impact of the geometric properties of the constraint set in safe optimization with bandit feedback. In *Learning for Dynamics and Control Conference*, pages 497–508. PMLR, 2023.

Julian Katz-Samuels and Clayton Scott. Top feasible arm identification. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1593–1601. PMLR, 2019.

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

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

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

Aldo Pacchiano, Mohammad Ghavamzadeh, and Peter Bartlett. Contextual bandits with stage-wise constraints. *arXiv preprint arXiv:2401.08016*, 2024.

Mala L Radhakrishnan and Bruce Tidor. Optimal drug cocktail design: methods for targeting molecular ensembles and insights from theoretical model systems. *Journal of chemical information and modeling*, 48(5):1055–1073, 2008.

Ohad Shamir. On the complexity of bandit linear optimization. In *Conference on Learning Theory*, pages 1523–1551. PMLR, 2015.

Han Shao, Xiaotian Yu, Irwin King, and Michael R Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. *Advances in Neural Information Processing Systems*, 31, 2018.

Matteo Turchetta, Felix Berkenkamp, and Andreas Krause. Safe exploration in finite markov decision processes with gaussian processes. *Advances in Neural Information Processing Systems*, 29, 2016.

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

Sharan Vaswani, Lin Yang, and Csaba Szepesvari. Near-optimal sample complexity bounds for constrained MDPs. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

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

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

Appendix A. Related Work on Pure Exploration.

While we study the regret formulation, work on constrained bandits has naturally also appeared in the pure exploration setting. The typical such paper aims to recover arms that are both nearly-safe and nearly-optimal, in a PAC sense. Katz-Samuels and Scott (2019) study this quesiton for finite-armed bandits, and Wang et al. (2022) extend this study under a structured multi-armed bandit setting where each arm has a continuous parameter that must be selected, and monotonically affects reward and safety of the arm. Most pertinently, Camilleri et al. (2022); Carlsson et al. (2023) study best feasible arm identifaction in the linear bandit setting with the same structure as us, although they assume that the set of possible actions is finite and known a priori. It is interesting to note that even in the identification setting, where safety is not enforced during learning, methods that can identify good arms quickly can only give guarantees of safety up to a given precision. This complements our observations in the regret setting.

Appendix B. On the Assumptions, and Background on Online Linear Regression

We give an expanded discussion of the standard assumptions made in §2, and discuss a standard result from online linear regression controlling $\sum \|x_t\|_{V_{t-1}^{-1}}$ that is key to our analysis.

B.1. A closer look at the assumptions

The assumptions made in the main text are slightly simplified version of standard assumptions from the literature on linear bandits.

Boundedness. The boundedness assumption has two parts: firstly that the underlying parameters are bounded, i.e., $\|\theta\|, \|a^i\| \leq 1$ and secondly we assume that the domain is bounded, i.e., $\|x\| \leq 1$ for all $x \in \mathcal{X} = \{Bx \leq \beta\}$.

The bounded domain assumption is used chiefly to ensure that the underlying optimisation problem of interest has finite value. Quantitatively, this may be replaced with a generic bound $\|x\| \leq L$ instead without appreciably changing the study. The principal way this affects DOSS is via the choice of the regulariser: instead of setting $V_t = (I + \sum x_s x_s^\top)$, this requires us to set $V_t = \lambda I + \sum x_s x_s^\top$ for some $\lambda > L^2$. Concretely, the validity of of appropriate modification of Lemma 2 to handle general regularisation requires using

$$\sqrt{\omega_t(\delta;\lambda)} = \sqrt{\frac{1}{2}\log\left(\frac{(U+1)\det(V_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \lambda^{1/2}$$

for a λ that $\lambda \geq \max_t \|x_t\|^2$, which may be ensured by setting $\lambda \geq L^2$. The main paper simplifies this notational clutter by just setting $\lambda = 1$ and assuming $\|x\| \leq 1$. A second aspect that is affected by the quantity L is that the upper bound of Lemma 23 would read $\log(1 + TL^2/\lambda d)$ instead of $\log(1 + T/d)$, which mildly affects some logarithmic terms in the regret bounds (and in fact no bound reported in the main text needs modification if we set $\lambda \geq L^2$ and assume that $L \leq d$).

The assumption of bounded parameters is largely without loss of generality - indeed, if we had a bound $\|\theta\|, \max_i \|a^i\| \leq S$ instead, the only change required is that the confidence set radius ω_t would need to be set as

$$\sqrt{\omega_t(\delta;\lambda,S)} = \sqrt{\omega_t(\delta;\lambda)} + (S-1)\sqrt{\lambda},$$

i.e., only the additive $\sqrt{\lambda}$ term in $\sqrt{\omega_t}$ above would need adjustment. We note that in general, the norm bounds on the various a^i and θ need not agree, and it is in fact possible to adapt to their norms without prior knowledge of the same, by setting distinct ω_t^is for each a^i, and using the techniques of the recent work of Gales et al. (2022).

SubGaussianity. While the subGaussianity condition can also be relaxed (for instance, linear bandits with heavy tailed noise have been studied (Shao et al., 2018)), it yields significant technical convenience whilst remaining quite a generic setting. In the assumption, we concretely assume that the noise is conditionally 1-subGaussian. This may be relaxed to conditionally R-subGaussian. This too can be handled with a small change in ω_t to

$$\sqrt{\omega_t(\delta;\lambda,R)} = R\sqrt{\frac{1}{2}\log\left(\frac{(U+1)\det(V_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \lambda^{1/2}.$$

This change is somewhat stronger than the corresponding change induced by altering $\|\theta\|$ and $\|a^i\|$, since the scaling is now applied to the first term of ω_t, which grows with t unlike the constant $\sqrt{\lambda}$ penalty.

Overall Confidence Radius with General Parameters. To sum up, under the generic conditions $\|x\| \le L, \|\theta\| \le S, \|a^i\| \le S$, and R-subGaussianity of $\{\gamma_t^i\}$, the entirety of our following analysis will go through, but with the blown up confidence radii

$$\sqrt{\omega_t(\delta;\lambda,L,S,R)} = R\sqrt{\frac{1}{2}\log\left(\frac{(U+1)\det(V_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + S\lambda^{1/2},$$

and under the condition $\lambda \ge L^2$. This results in roughly an increase in the regret bounds of a factor of at most $\max(R,S)$, along with a potential increase in the logarithmic terms to $\log(1+TL^2/\delta)$ instead of $\log(1+T/\delta)$. For the remainder of our analysis, we shall stick to the default parameters $R = S = L = \lambda = 1$.

B.2. Quantitative Bounds from the Theory of Online Linear Regression

We conclude the preliminaries with the following generic statement, which holds due to a couple of applications of the matrix-determinant lemma. The result is standard - see the discussions of Abbasi-Yadkori et al. (2011, Lemma 11) for historical discussions.

Lemma 23 *Let $\{x_t\}$ be the actions of* DOSS. *Suppose that for all t, $\|x_t\| \le 1$, and let $\lambda \ge 1$. Then for any T,*

$$\sum_{t=1}^{T} \|x_t\|^2_{V_{t-1}^{-1}} \le \frac{3}{2}\log\left(\frac{\det(V_T)}{\det(\lambda I)}\right) \le \frac{3}{2} d \log\left(1 + \frac{T}{\lambda d}\right).$$

Proof of Lemma 23. First notice that since $V_t = V_{t-1} + x_t x_t^\top$, by the matrix-determinant lemma,

$$\det(V_t) = \det(V_{t-1})\det(I + V_{t-1}^{-1/2} x_t x_t^T (V_{t-1}^{-1/2})^\top = \det(V_{t-1})(1 + \|x_t\|^2_{V_{t-1}^{-1}}),$$

and induction yields

$$\det(V_T) = \det(\lambda I)\prod(1 + \|x_t\|^2_{V_{t-1}^{-1}}).$$

where we have used that $V_0 = \lambda I$.

Now, notice that since $V_{t-1} \succ \lambda I$ for each t, it follows that $\|x_t\|^2_{V_{t-1}^{-1}} \leq \|x_t\|^2/\lambda \leq 1$. But for $z \in [0,1], z \leq \frac{3}{2}\log(1+z)$, which implies that

$$\sum \|x_t\|^2_{V_{t-1}^{-1}} \leq \frac{3}{2}\sum \log(1 + \|x_t\|^2_{V_{t-1}^{-1}}) = \frac{3}{2}\log\frac{\det(V_T)}{\det(\lambda I)}.$$

Finally, note that since V_T is positive definite, by an application of the AM-GM inequality, $\det(V_T) \leq (\mathrm{trace}(V_T)/d)^d$, and further, $\mathrm{trace}(V_T) = d\lambda + \sum_t \|x_t\|_2^2 \leq d\lambda + T$. Further observing that $\det(\lambda I) = \lambda^d$, we conclude that

$$\log\frac{\det(V_T)}{\det(V)} \leq d\log\frac{(d\lambda + T)/d}{\lambda} = d\log\left(1 + \frac{T}{d\lambda}\right).$$

■

An immediate consequence of the above is the following pair of observations which we shall use frequently.

Lemma 24 *Let $\{x_t\}$ be the actions of* DOSS *run with the parameters λ, δ. For every $T > 0$,*

$$\sum_{t\leq T} \rho_t(x_t;\delta)^2 \leq 3d^2\log^2\left(1 + \frac{T}{\lambda d}\right) + 6d\log\left(1 + \frac{T}{\lambda d}\right)\left(\log\frac{U+1}{\delta} + 2\lambda\right), \tag{4}$$

$$\sum_{t\leq T} \rho_t(x_t;\delta) \leq d\sqrt{3T}\log\left(1 + \frac{\log T}{d\lambda}\right) + \sqrt{3dT\log\left(1 + \frac{T}{\lambda d}\right)}\left(\sqrt{2\lambda} + \sqrt{\log\frac{U+1}{\delta}}\right). \tag{5}$$

These bounds supply the core bounds needed to convert the control we develop on ρ_t in §6.3 and §7 into control on $\mathscr{E}_T$ and $\mathscr{S}_T$. Observe that the main terms in the above results do not show dependence on the failure probability parameter δ.

Proof of Lemma 24. Recall that $\rho_t(x_t;\delta) = 2\sqrt{\omega_t(\delta)}\cdot\|x_t\|_{V_{t-1}^{-1}}$. Further observe that ω_t is an increasing function of t. Immediately by Lemma 23,

$$\sum \rho_t^2 \leq 4\omega_T(\delta)\sum \|x_t\|^2_{V_{t-1}^{-1}} \leq 6d\omega_T(\delta)\log\left(1 + \frac{T}{d\lambda}\right).$$

Further, once again applying Lemma 23, and noting that $(\sqrt{u} + \sqrt{v})^2 \leq 2u + 2v$,

$$\sqrt{\omega_T(\delta)} = \sqrt{\lambda} + \sqrt{\frac{1}{2}\log\frac{(U+1)}{\delta} + \frac{1}{4}\log\frac{\det(V_T)}{\det(\lambda I)}}$$
$$\implies \omega_T(\delta) \leq 2\lambda + \log\frac{(U+1)}{\delta} + \frac{d}{2}\log\left(1 + \frac{T}{\lambda d}\right).$$

Multiplying these two bounds controls $\sum \rho_t^2$.

Further, by the Cauchy-Schwarz inequality,

$$\sum_{t=1}^T \rho_t \leq \sqrt{T}\cdot\sqrt{\sum_{t=1}^T \rho_t^2}.$$

The bound (5) follows upon applying the bound on $\sum \rho_t^2$ above, and then using the trivial relation $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$. ■

Finally, let us argue that the quantity $\rho_t(x_t;\delta)$ indeed controls the noise scale of the problem by showing Lemma 2.

Proof of Lemma 2. We refer to the proof of Abbasi-Yadkori et al. (2011, Thm. 2) for the consistency, and observe only that the factor $(U+1)/\delta$ enters our confidence radius $\sqrt{\omega_t(d)}$ by hitting their analysis with the union bound to ensure concentration over the unknown objective and over the U unknown constraints simultaneously. Of course, the factor of $\mathbf{1}_U$ arises in the definition of $\boldsymbol{\mathcal{C}}$ since each known constraint is already 'estimated' exactly by setting $\hat{a}_t^i = a^i$ for $i \in [U+1 : U+K]$.

To show that the noise scale limits the deviations $\langle \tilde{\theta} - \theta, x \rangle$, observe that under the assumption of consistency, $\theta \in \mathcal{C}_t^\theta$. Therefore

$$|\langle \tilde{\theta} - \theta, x \rangle| \leq |\langle \tilde{\theta} - \hat{\theta}, x \rangle| + |\langle \theta - \hat{\theta}, x \rangle|.$$

By exploiting the positive definiteness of V_{t-1} and the Cauchy-Schwarz inequality, we can further observe that

$$|\langle \theta - \hat{\theta}, x \rangle| = |\langle (\theta - \hat{\theta}) V_{t-1}^{1/2}, V_{t-1}^{-1/2} x \rangle| \leq \|\theta - \hat{\theta}\|_{V_{t-1}} \cdot \|x\|_{V_{t-1}^{-1}}.$$

Running the same calculation of $\tilde{\theta}$ and adding the bounds, we conclude that

$$|\langle \tilde{\theta} - \theta, x \rangle| \leq (\|\theta - \hat{\theta}\|_{V_{t-1}} + \|\tilde{\theta} - \hat{\theta}\|_{V_{t-1}}) \|x\|_{V_{t-1}^{-1}}$$

But both $\theta, \tilde{\theta} \in \mathcal{C}_t^\theta$, which by definitions means that their V_{t-1}-norm distance from $\hat{\theta}$ is bounded by $\sqrt{\omega_t(\delta)}$. The claim is immediate upon recalling that $\rho_t(x;\delta) := 2\sqrt{\omega_t(\delta)} \|x\|_{V_{t-1}^{-1}}$.

Of course, the same argument applies to every $\tilde{a}^i$, and thus to $\tilde{A}$. Again, for known constraints, the radius of the confidence set is 0, so $\tilde{a}^i = \hat{a}^i = a^i$, and hence the factor of $\mathbf{1}_U$ in $|(\tilde{A} - A)x| \leq \rho_t(x;\delta)\mathbf{1}_U$.

Finally, the bounds on $\sum \rho_t(x_t;\delta)$ and $\sum \rho_t(x_t;\delta)^2$ follow directly from Lemma 24. ■

Appendix C. Appendix on the Structural Behaviour of DOSS

This section is devoted to showing the key structural properties of the behaviour of DOSS that we discussed in §6. In particular, we show the main result of §6.1, namely that any point that DOSS plays must noisily activate some BIS. To this end, we first characterise the behaviour of DOSS relative to polytopes contained in the permissible set. Before stating the same, recall that an extreme point of a polytope (and indeed a closed convex set), is any point that is not contained on a line joining two other points in the polytope. Further, each extreme point of a polytope in $\mathbb{R}^d$ must satisfy at least d constraints with equality. For a polytope $\mathcal{P}$, we will denote its extreme points as $\mathcal{E}_\mathcal{P}$.

Lemma 25 *Suppose that $\mathcal{P}$ is a polytope such that $\mathcal{P} \subset \widetilde{\mathcal{S}}_t$. If* DOSS *plays in $\mathcal{P}$, then x_t must be an extreme point of $\mathcal{P}$, i.e., $x_t \in \mathcal{P} \implies x_t \in \mathcal{E}_\mathcal{P}$.*

Let us first argue that the Proposition 9 follows from the above Lemma.

Proof of Proposition 9. For a choice of $\tilde{A} \in \boldsymbol{\mathcal{C}}_t$, define the polytope

$$\mathcal{P}(\tilde{A}) = \{x : \tilde{A}x \leq \alpha\}.$$

Now, observe that

$$\widetilde{\mathcal{S}}_t = \bigcup_{\{\tilde{A} \in \boldsymbol{\mathcal{C}}_t\}} \{x : \tilde{A}x \leq \alpha\} = \bigcup_{\tilde{A} \in \boldsymbol{\mathcal{C}}_t} \mathcal{P}(\tilde{A}),$$

i.e., $\widetilde{\mathcal{S}}_t$ can be decomposed as a union of polytopes. But then the selected point x_t must lie in one of these polytopes, say $\mathcal{P}^*$.

Now, we have that $\mathcal{P}^* \subset \widetilde{\mathcal{S}}_t$, and $x_t \in \mathcal{P}^*$, and so by Lemma 25, x_t must be an extreme point of $\mathcal{P}^*$. But this implies that there are at least d linearly independent constraints amidst the $\tilde{A}x \leq \alpha$ that x_t activates, i.e., that there exists some $I \subset [1:m]$ such that $|I| = d$ and $\tilde{A}(I)x = \alpha(I)$. By definition, then $x_t \in \widetilde{X}_t^I$, showing the claim. ■

It remains to show the preceding Lemma. Before proceeding, let us comment that the statement above is intuitively obvious, but appears to be somewhat cumbersome to prove (as the argument below suggests, although nothing says that a cleaner proof could not be found). Of course, this statement extends also to the OFUL algorithm, and to our knowledge this has not been directly argued previously: instead, when working on polytopal domains, typically it is directly stated that it suffices to play on the extreme points of the polytope.

Proof of Lemma 25. Suppose that $x_t \in \mathcal{P}$. Then, due to the optimistic choice, there also exists some $\tilde{\theta}_t \in \mathcal{C}_t^0$ such that

$$(\tilde{\theta}_t, x_t) \in \operatorname*{arg\,max}_{\tilde{\theta} \in \mathcal{C}_t^0, x \in \mathcal{P}} \langle \tilde{\theta}, x \rangle.$$

Notice also that x_t is a solution of the linear program $\max_{x \in \mathcal{P}} \langle \tilde{\theta}_t, x \rangle$, and so lies on the boundary of $\mathcal{P}$. Similarly, $\tilde{\theta}_t$ lies on the boundary of $\mathcal{C}_t^0$. We need to argue that x_t must in fact be an extreme point of $\mathcal{P}$, i.e., it does not lie in the interior of some face of dimension ≥ 1 of $\mathcal{P}$.

For this, first suppose for the sake of contradiction that x_t lies in the interior of some 1-dimensional face of $\mathcal{P}$, say $\mathcal{F}$. Let u be the direction of variation of $\mathcal{F}$. Then it must hold that $\langle \tilde{\theta}_t, u \rangle = 0$, else $\langle \tilde{\theta}_t, x_t + \varepsilon u \rangle$ would exceed $\langle \tilde{\theta}_t, x_t \rangle$ for some small choice of ε. Now, let us rotate the domain so that u is directed along one coordinate axis, and project onto the 2D subspace spanned by the (orthogonal) directions u and $\tilde{\theta}_t$. Next, rescale the vectors so that both u and $\tilde{\theta}_t$ have norm 1, and finally translate the polytope so that the uth component of x_t is 0. Notice that the projection of an ellipsoid is an ellipsoid, and so doing the same transformations to $\mathcal{C}_{t-1}^{\theta}$ produces a 2-dimensional convex confidence ellipsoid D.

Let us relabel the axes of the resulting system as u_1 and u_2. In the resulting coordinate system, $\tilde{\theta} = (0, 1)$, and $\mathcal{F}$ is a line segment of the form $\{u_1 \in [p, q], u_2 = r\}$, where $p < 0 < q, r = \langle \tilde{\theta}_t, x_t \rangle / \|\tilde{\theta}_t\|$ and $x_t = (0, r)$. Observe that $\tilde{\theta}_t$ must lie on the boundary of D. We shall argue that there is some other $z \in \mathcal{F}$ and some other $\phi \in D$ such that $\langle z, \phi \rangle > r$, which violates the assumption.

We first take the case of $r > 0$. Observe that if any point of D has u_2 coordinate greater than 1, then we immediately have a contradiction, since then for such a point ϕ, $\langle \phi, x_t \rangle > \langle \tilde{\theta}_t, x_t \rangle$. But, since $\tilde{\theta}_t = (0, 1) \in D$, it follows that the ellipse D is tangent to $u_2 = 1$. But this means that for small ε, D must contain points $\phi_\varepsilon = (\varepsilon, 1 - f(\varepsilon))$ where $0 \leq f(\varepsilon) = O(\varepsilon^2)$. But this implies a contradiction - indeed, take $\varepsilon > 0$, and consider $z_\varepsilon = (\varepsilon^{1/2}, r)$. Then $z_\varepsilon \in \mathcal{F}$ for small enough ε, and

$\langle z_\varepsilon, \phi_\varepsilon \rangle - r = \varepsilon^{3/2} - rf(\varepsilon)$. Since $f(\varepsilon) = O(\varepsilon^2)$, this is positive for small enough ε, demonstrating a contradiction.

If $r < 0$, the same argument can be run mutatis mutandis - now D must lie above the line $u_2 = 1$, but still be tangent to it, and we can develop points of the form $(\varepsilon, 1 + f(\varepsilon))$ for $0 \leq f = O(\varepsilon^2)$ in D, and the analogous inner product $\langle z_\varepsilon, \phi_\varepsilon \rangle - r = \varepsilon + rf(\varepsilon)$ which is again positive for small enough ε.

Finally, we have the case $r = 0$, wherein x_t lies at the origin. But in this case any point in D of non-zero u_1 coordinate serves as a contradiction (since either $(p, 0)$ or $(0, q)$ will yield a positive inner product).

Together, the above paragraphs imply that x_t cannot lie on the interior of an edge of $\mathcal{P}$. But this argument generalises to the interior of any non-trivial face. Indeed, since $\tilde{\theta}_t$ must be orthogonal to the affine subspace formed by this face, we can argue that there must be a point in the interior of a 1-D face (that forms a boundary of the larger face) that must also attain the optimal value for $\langle \tilde{\theta}, x \rangle$, and then run the above argument for this point. It follows that x_t cannot lie in the interior of any non-trivial face of $\mathcal{P}$. ■

The above argument is not restricted to confidence ellipsoids of the form of §3.1, but extends to any $\mathcal{C}_t$ with a smooth and convex boundary. Indeed, this further extends to convex $\mathcal{C}_t$ with continuous boundaries, barring the case where $\tilde{\theta}_t$ is itself the extreme point of a polytope (with large 'curvature' at $\tilde{\theta}_t$). In such a case the property that $f(\varepsilon) = O(\varepsilon^2)$ does not hold, and a more global argument may be needed. One attack may pass through the use of continuous noise, in which case the confidence sets would almost surely not produce any extreme points that are orthogonal to the faces of a polytope (since such directions lie in a union of a finite number of dimension $d - 1$ affine subspaces, which in turn is Lebesgue null), and so we may almost surely avoid this disadvantageous case.

Let us also note the following interesting observation that can also be inferred using Lemma 25, and further characterises the behaviour of doubly-optimistic play.

Proposition 26 *Suppose that all confidence sets are valid. Then there exists at least one BIS I that x_t noisily activates, and such that $A(I)x_t \geq \alpha(I)$.*

In other words, for at least one BIS, the action x_t not only noisily activates it, it further either activates it or violates all of the true constraints of this BIS. Notice that if the BIS shown to exist above has at least one unknown constraint, then this basically means that DOSS must violate safety (since meeting this with equality for the unknown constraint would be rare).

Proof of Proposition 26. Fix x_t. We call $\tilde{A} \in \boldsymbol{\mathcal{C}}_t$ a witness for x_t if $\tilde{A}x_t \leq \alpha$, i.e., if $\tilde{A}$ witnesses the presence of x_t in $\widetilde{\mathcal{S}}_t$. Since x_t is the optimistic optimum over the entirety of $\widetilde{\mathcal{S}}_t$, it follows that for every witness $\tilde{A}$ of x_t, it holds that $x_t \in \arg\max_x \max_{\tilde{\theta} \in \mathcal{C}_t^\theta} \langle \tilde{\theta}, x \rangle : \tilde{A}x \leq \alpha$.

Now, let I_0 be all of the constraints that x_t noisily activates, and let $I_{\geq} := \{i \in [1 : m] : \langle a^i, x_t \rangle \geq \alpha^i\}$. We claim that $|I_{\geq}| \geq d$, which suffices to show the claim.

For the sake of contradiction, assume that $|I_{\geq}| \leq d - 1$. For each $i \in I_0 \setminus I_{\geq}$, we have $\langle a^i, x_t \rangle < \alpha^i$. Let us form the matrix $\tilde{A}_<$ formed by taking each of the ith rows in $\tilde{A}$ for which $i \in I_0 \setminus I_{\geq}$, and replacing the $\tilde{a}^i$ in the row by $\tilde{a}^i_< = a^i$. This matrix remains a witness, since the resulting $\tilde{A}_<$ lies in $\boldsymbol{\mathcal{C}}_t$ (as we have replaced rows by the rows of A, each of which lie in the corresponding confidence sets for individual rows), and by definition for each replaced row, $\langle \tilde{a}^i_<, x_t \rangle < \alpha^i$, since each such i lies in $I_0 \setminus I_{\geq}$.

Then x_t lies in the interior of the polytope $\mathcal{P}_< := \{x : \tilde{A}_< x \leq \alpha\}$, since by construction it activates at most $|I_\geq| \leq d-1$ constraints of this matrix. But since $\tilde{A}_< \in \boldsymbol{\mathcal{C}}_t$, it holds that $\mathcal{P}_< \subset \widetilde{\mathcal{S}}_t$, and thus the algorithm plays in the intrior of a polytope contained in the permissible set, contradicting Lemma 25. Therefore, our hypothesis is untenable, and $|I_\geq| \geq d$. ■

Appendix D. Controlling the Play of Suboptimal BISs

We now show the noise scale lower bound, and the subsequent control on the play of suboptimal BISs as discussed in §6.

D.1. Localising Actions when a BIS is Activated

We show Lemma 10 as a simple consequence of consistency and optimism.

Proof of Lemma 10. Suppose that the confidence sets are consistent, and that x_t noisily activates the BIS I. Since x_t is the action of DOSS, it is also permissible. Together, these two properties imply that there exists some $\tilde{A} \in \boldsymbol{\mathcal{C}}_t$ such that

$$\tilde{A}x_t \leq \alpha$$
$$\tilde{A}(I)x_t = \alpha(I)$$

But, since $\boldsymbol{\mathcal{C}}_t$ is consistent, Lemma 2 yields

$$Ax_t - \rho_t \mathbf{1}_U \leq \tilde{A}x_t \leq Ax_t + \rho_t \mathbf{1}_U.$$

The claim follows directly from this, since

$$\alpha \geq \tilde{A}x_t \geq Ax_t - \rho_t \mathbf{1}_U \implies Ax_t \leq \alpha + \rho_t \mathbf{1}_U,$$

and

$$\alpha(I) = \tilde{A}(I)x \leq A(I)x_t + \rho_t \mathbf{1}_U(I) \implies A(I)x_t \geq \alpha(I) - \rho_t \mathbf{1}_U(I).$$

Further, due to the optimistic selection of x_t, it is a maximiser amongst the permissible set of $\max_{\tilde{\theta} \in \mathcal{C}_t^\theta} \langle \tilde{\theta}, x \rangle$. But under consistency, $\theta \in \mathcal{C}_t^\theta$, and $x^* \in \widetilde{\mathcal{S}}_t$. Thus, it follows that if $\tilde{\theta}$ is the optimal choice in the above program, then

$$\langle \tilde{\theta}, x_t \rangle \geq \langle \theta, x^* \rangle.$$

But, again using consistency and Lemma 2, it holds that $\langle \tilde{\theta}, x_t \rangle \leq \langle \theta, x_t \rangle + \rho_t$, from which the claim is forthcoming. ■

D.2. Proof of Noise Scale Lower Bound and the Positivity of the Gaps of Suboptimal BISs

The argument underlying the proof of the noise-scale lower bound is essentially encapsulated in §6.2.1, but refined through the use of the LP $P(\zeta; I)$. The bulk of the following proofgoes into showing that the gap we define is meaningful, i.e., that if I is a suboptimal BIS, then $\max(\zeta_*(I), \eta_*(I)) > 0$. This essentially boils down to showing that $\mathfrak{s}(I)$ is finite for feasible BISs.

Proof of Lemma 13. We will first show that under consistency of the confidence sets, playing a suboptimal BIS I implies that $\rho_t(x_t;\delta) \geq \max(\zeta_*(I), \eta_*(I))$. Observe that under the assumption of consistency,

$$\langle \theta, x_t \rangle \leq P(\rho_t; I),$$

since $x_t \in \mathcal{T}(\rho_t; I)$ by Lemma 10. Further, by the final line of Lemma 10, $\langle \theta, x_t \rangle \geq \langle \theta, x^* \rangle - \rho_t$.

Since $x_t \in \mathcal{T}(\rho_t; I)$, this set is nonempty, and therefore by definition $\rho_t \geq \zeta_*(I)$. Note that if $\zeta_*(I) = \infty$, we can conclude already, since this means that $\rho_t(x_t;\delta) \geq \infty \geq \max(\zeta_*(I), \eta_*(I))$. If $\zeta_*(I) < \infty$, then by the definition of the spread $\mathfrak{s}(I)$, and the efficacy separation $\gamma(I)$, we have

$$P(\rho_t; I) \leq P(\zeta_*(I); I) + \mathfrak{s}(I)(\rho_t - \zeta_*(I)) = \langle \theta, x^* \rangle - \gamma(I) + \mathfrak{s}(I)(\rho_t - \zeta_*(I)).$$

But then we conclude that

$$-\rho_t \leq -\gamma(I) + \mathfrak{s}(I)(\rho_t - \zeta^*(I)) \iff \rho_t(\mathfrak{s}(I)+1) \geq \gamma(I) + \mathfrak{s}(I)\zeta_*(I) \iff \rho_t \geq \eta_*(I).$$

Thus, the claim follows.

We now proceed to argue that for any suboptimal BIS I, at least one of $\zeta_*(I)$ and $\eta_*(I)$ is positive. Fix the BIS I. Note that if $\zeta_*(I) = \infty$, then there is nothing to show. So, suppose $\zeta_*(I) < \infty$. By expanding out the definition of $\mathcal{T}(\zeta; I)$, the program P is

$$\begin{aligned} P(\zeta; I) = \max_x \langle \theta, x \rangle \\ \text{s.t. } Ax &\leq \alpha + \zeta \mathbf{1}_U \\ -A(I)x &\leq -\alpha(I) + \zeta \mathbf{1}_U(I). \end{aligned}$$

We recall that this is a linear program, which is of course evident in the above. Since $\zeta_*(I) < \infty$, the above program is feasible for $\zeta \geq \zeta_*(I)$. Further, since $\mathcal{X} \supset \mathcal{T}(\zeta; I)$ is a bounded polytope, the program is finite. Thus strong duality applies to the above program.

Let us introduce dual variables (λ, μ) respectively for the two blocks of constraints. By standard techniques, the dual program is

$$\begin{aligned} D(\zeta; I) = \min_{\lambda,\mu} \langle \lambda, \alpha + \zeta \mathbf{1}_U \rangle + \langle \mu, -\alpha(I) + \zeta \mathbf{1}_U(I) \rangle \\ \text{s.t. } A^\top \lambda - A(I)^\top \mu &= \theta, \\ \lambda \geq 0, \mu &\geq 0. \end{aligned}$$

For succinctness, let us write

$$\begin{aligned} f(\lambda, \mu) &= \langle \lambda, \mathbf{1}_U \rangle + \langle \mu, \mathbf{1}_U(I) \rangle \\ g(\lambda, \mu) &= \langle \lambda, \alpha + \zeta_*(I)\mathbf{1}_U \rangle + \langle \mu, -\alpha(I) + \zeta_*(I)\mathbf{1}_U(I) \rangle, \\ h(\lambda, \mu) &:= A^\top \lambda - A(I)^\top \mu - \theta. \end{aligned}$$

Further, let $\boldsymbol{\lambda} = (\lambda, \mu)$. We can succinctly write the dual as

$$D(\zeta; I) = \min_{\boldsymbol{\lambda}} (\zeta - \zeta_*(I)) f(\boldsymbol{\lambda}) + g(\boldsymbol{\lambda}) : h(\boldsymbol{\lambda}) = 0, \lambda \geq 0, \mu \geq 0.$$

Note that since the primal is bounded and feasible for $\zeta \geq \zeta_*(I)$, so is the dual, and by strong duality $D(\zeta_*(I); I) = P(\zeta_*(I); I)$. But

$$D(\zeta_*(I); I) = \min_{\boldsymbol{\lambda}} g(\boldsymbol{\lambda}) : h(\boldsymbol{\lambda}) = 0, \lambda \geq 0, \mu \geq 0.$$

It follows that the set

$$\mathcal{F} := \{\boldsymbol{\lambda} : g(\boldsymbol{\lambda}) \leq P(\zeta_*(I); I), h(\boldsymbol{\lambda}) = 0, \lambda \geq 0, \mu \geq 0\}$$

is nonempty. Observe that ζ does not appear anywhere in the definition of $\mathcal{F}$.

Let us define the two programs

$$\begin{aligned} D'(\zeta; I) &:= \min_{\boldsymbol{\lambda}} (\zeta - \zeta_*(I)) f(\boldsymbol{\lambda}) + g(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \mathcal{F}, \\ E(I) &:= \min_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \mathcal{F} \end{aligned}$$

Note that both of the above programs are feasible. As a feasible minimisation program we also have that $E(I) < \infty$. Further, since introducing extra constraints cannot decrease the value of a minimisation program, we note that $D(\zeta; I) \leq D'(\zeta; I)$. But observe that since the constraints of $D'(\zeta; I)$ include the requirement that $g(\boldsymbol{\lambda}) \leq P(\zeta_*(I); I)$, we have for every $\zeta \geq \zeta_*(I)$ that

$$\begin{aligned} D'(\zeta; I) &\leq P(\zeta_*(I); I) + \min\{(\zeta - \zeta_*(I)) f(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \mathcal{F}\} \\ &= P(\zeta_*(I); I) + (\zeta - \zeta_*(I)) \cdot \min\{f(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \mathcal{F}\} \\ &= P(\zeta_*(I); I) + (\zeta - \zeta_*(I)) E(I). \end{aligned}$$

But, then by strong duality,

$$P(\zeta; I) = D(\zeta; I) \leq P(\zeta_*(I); I) + (\zeta - \zeta_*(I)) E(I),$$

and we conclude that $\mathfrak{s}(I) \leq \max(0, E(I)) < \infty$.

Now, since $\mathfrak{s}(I)$ is finite, in order to show that $\max(\zeta_*(I), \eta_*(I)) > 0$, it suffices to argue that for any suboptimal BIS, $\max(\zeta_*(I), \gamma(I)) > 0$. But observe that if $\zeta_*(I) = 0$, then $\lim_{\zeta \searrow 0} P(\zeta; I) > -\infty$, and due to the right-continuity of P, this implies that $P(0; I) > -\infty \implies \mathcal{X}^I \neq \emptyset$, in other words, I is a feasible BIS. But if a BIS I is both feasible and suboptimal, then for every $x \in I$, it must hold that $\langle \theta, x \rangle < \langle \theta, x^* \rangle$, since otherwise I would be optimal. But, since $\mathcal{X}^I = \mathcal{T}(0; I)$ is a compact set, this means that $P(\zeta_*(I); I) = P(0; I) < \langle \theta, x^* \rangle \iff \gamma(I) > 0$. ■

D.3. Bounding the Play of Suboptimal BISs

With the above ingredients in place, we show the main result of §6.3.

Proof of Theorem 15. Let us again abbreviate $\rho_t(x_t; \delta)$ as ρ_t. By Lemma 13, if a suboptimal BIS I is played, then $\rho_t \geq \max(\eta_*(I), \zeta_*(I))$. But then any time a suboptimal BIS is played, $\rho_t \geq \min\{\max(\eta_*(I), \zeta_*(I)) : I \text{ is a suboptimal BISs}\}$, i.e., $\rho_t \geq \Gamma$.

Now observe that

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{1}\{\exists\text{suboptimal BIS } I : x_t \in \widetilde{\mathcal{X}}_t^I\} &\leq \sum_{t=1}^{T} \mathbb{1}\{\rho_t \geq \Gamma\} \\ &\leq \sum_{t=1}^{T} \frac{\rho_t^2}{\Gamma^2} \mathbb{1}\{\rho_t \geq \Gamma\} \\ &\leq \Gamma^{-2} \sum_{t \leq T} \rho_t^2,\end{aligned}$$

where the second inequality is using that if $\rho_t \geq \Gamma$, then $\rho_t/\Gamma \geq 1$. Applying Lemma 24 immediately bounds the above as $O\left(\frac{d^2 \log^2 T + d \log(T) \log(U/\delta)}{\Gamma^2}\right)$. ■

Appendix E. Proofs of Bounds on Efficacy Regret and Safety Violations

We proceed to discuss the proofs of the results of §7.

E.1. The Efficacy of the Actions of DOSS when Activating Optimal BISs

Our first order of business is to argue that playing only optimal BISs leads to actions x_t that are 'over-efficient', i.e., satisfy $\langle \theta, x_t \rangle \geq \langle \theta, x^* \rangle$. The following basic result is useful in our argument.

Lemma 27 *For a BIS I, define $K_I = I \cap [U+1 : m]$ to be the indices of the known constraints in I. Under the genericity of noise assumption, for any BIS I such that $A(K_I)$ is full row rank, for any $t \geq d$, it holds almost surely that $\hat{A}_t(I)$ is full rank.*

Proof of Lemma 27. Notice that since for any i, the noise in the feedback S_t^i is generic, it does not concentrate in any low-dimensional subspace of $\mathbb{R}^d$. This in turn means that the probability that any $\hat{a}_t^i$ lies in a low-dimensional subspace of $\mathbb{R}^d$ is exactly zero. The claim follows immediately: since $|I \setminus K_I| \leq d$, each $\hat{a}_t^i$ with probability one does not lie in the span of $\{\hat{a}^j\}_{j \in I \setminus \{i\}}$, and since by assumption the $A(K_I)$ is full rank. ■

With this in hand, we argue Lemma 17 by exploiting the weak-nondegeneracy condition of Assumption 16.

Proof of Lemma 17. We need to show that if *all* of the BISs x_t noisily activates are optimal, then $\langle \theta, x_t \rangle \geq \langle \theta, x^* \rangle$, which comprises the bulk of this proof. To this end, let us fix one such BIS, I.

By Assumption 16, we know that $\{x^*\} = \mathcal{X}^I$, and that I is full-rank. Notice that as a result, we may write

$$\langle \theta, x^* \rangle = \max \langle \theta, x \rangle : A(I)x = \alpha(I).$$

Indeed, due to the fact that I is full rank, the latter equality constraints already enforce that x^* is the sole feasible point. Further, by strong duality, there exists a choice of vectors μ such that

$$\mu^\top A(I) = \theta^\top.$$

Due to the optimistic selection rule, and the fact that x_t noisily saturates I, it must hold that x_t is a solution to

$$\max_{\tilde{\theta}\in\mathcal{C}_t^\theta, \tilde{A}\in\boldsymbol{\mathcal{C}}_t} \max_x \langle\tilde{\theta}, x\rangle : \tilde{A}(I)x = \alpha(I), \tilde{A}x \leq \alpha,$$

where the maximisation over $\tilde{A}$ is equivalent to optimisitic selection over $\widetilde{\mathcal{S}}_t = \{x : \exists \tilde{A} : \tilde{A}x \leq \alpha\}$, and the equality constraint arises since x_t noisily activates I. Now observe that in the optimisation above, we may restrict attention to $\tilde{A}$ such that $\tilde{A}(I)$ is full rank. Indeed, if this optimal choice were rank-deficient, then since the feasible set remains a polytope, there must exist some other constraints amongst the $\tilde{A}$ besides those in I that are activated by x_t (since otherwise we would be playing on the interior of a polytope, and thus violating Lemma 25). By dropping some linearly dependent rows, this would yield a different index set I' that x_t activates, and which is not rank-deficient. By the hypothesis, this index set must also be optimal, and we can run the argument for I' instead. But then note that x_t is exactly characterised by the equality conditions imposed by noisily activating the BIS I, which means that x_t is the optimiser of

$$\max_{\substack{\tilde{\theta}\in\mathcal{C}_t^\theta, \tilde{A}\in\boldsymbol{\mathcal{C}}_t, \\ \tilde{M}(I,\tilde{A}) \text{ is full-rank}}} \max_x \langle\tilde{\theta}, x\rangle : \tilde{A}(I)x = \alpha(I).$$

Now, let us write $\tilde{A} = A + \delta A, \tilde{\theta} = \theta + \delta\theta, x = x^* + \delta x$. Further denote the optima as $\delta\theta_t, \delta A_t, \delta x_t$. With this notation, our goal is to show that $\langle\theta, \delta x_t\rangle \geq 0$. To this end, observe that since the program above has the constraint $\tilde{A}(I)x = \alpha(I) = A(I)x^*$ we find that

$$\tilde{A}(I)x = A(I)x^* + \delta A(I)x + A(I)\delta x = \alpha(\mathsf{U}) \iff A(I)\delta x = -\delta A(I)x,$$

which imply that

$$\begin{aligned} &\langle\theta, \delta x\rangle = \langle A^\top\mu, \delta x\rangle = \langle\mu, A\delta x\rangle = -\langle\mu, \delta A x\rangle \\ \iff &\langle\theta, \delta x\rangle = \sum_{i\in I} -\mu^i\langle\delta A^i, x\rangle \end{aligned} \tag{6}$$

Thus, we can rewrite the program as

$$\max_x \max_{\delta\theta, \delta A} \langle\theta, x^*\rangle + \langle\delta\theta, x\rangle - \langle\mu, \delta A(I)x\rangle : \tilde{A}(I)x = \alpha(I).$$

Now, recall that the confidence sets are constructed around the RLS estimates $\hat{a}_t^i$ and $\hat{\theta}_t$, i.e.,

$$\mathcal{C}_t^\theta = \{\tilde{\theta} : \{\tilde{\theta} - \hat{\theta}_t\|_{V_{t-1}} \leq \omega_t\}, \mathcal{C}_t^i = \{\tilde{a} : \|\tilde{a} - \hat{a}_t^i\|_{V_{t-1}} \leq \omega_t \mathbf{1}_U^i\}.$$

To clearly express the choice of $\delta\theta, \delta A$, we define

$$\begin{aligned} &\Delta\theta_t = \hat{\theta}_t - \theta, \Delta a_t^i = \hat{a}_t^i - a^i, \Delta A = \hat{A}_t - A \\ &\partial\theta = \tilde{\theta} - \hat{\theta}_t \quad \partial a^i = \tilde{a}^i - \hat{a}_t^i, \partial A = \tilde{A} - \hat{A}_t. \end{aligned}$$

Observe then that

$$\delta\theta = \Delta\theta_t + \partial\theta; \delta a^i = \Delta a_t^i + \partial a^i.$$

Further, the decision variables of the program are only the $\partial\theta$ and ∂a^is, which lie in the set $\|\partial\theta\|_{V_{t-1}} \leq \omega_t$ and $\|\partial a^i\|_{V_{t-1}} \leq \omega_t \mathbf{1}^i_U$. Let us denote $U_I = I \cap [1:U]$ and $K_I = I \cap [U+1:m]$, and observe that $\Delta a^i = \partial a^i = 0$ for $i \in K_I$. Incorporating this structure, we can write the program as

$$\begin{aligned}\langle \theta, x^* \rangle + \max_x \max_{\partial\theta, \partial A} \langle \Delta\theta_t, x\rangle + - \sum_{i\in U_I} \mu^i \langle \Delta a^i_t, x\rangle + \langle \partial\theta, x\rangle - \sum_{i\in U_I} \mu_i \langle \partial a^i, x\rangle.\\ \text{s.t.} \quad \langle a^i + \Delta a^i_t, x\rangle + \langle \partial a^i, x\rangle = \alpha^i \quad \forall i \in I,\\ \|\partial\theta\|_{V_{t-1}} \leq \omega_t\\ \|\partial a^i\|_{V_{t-1}} \leq \omega_t \mathbf{1}^i_U \quad \forall i \in I.\end{aligned}$$

But now observe that the optimal choice of $\partial\theta$ in the above is exactly $\omega_t / \|x\|_{V_{t-1}^{-1}} V_{t-1}^{-1} x$. Indeed, recall that $\|u\|_{V_{t-1}} = \sqrt{u^\top V_{t-1} u} = \|V_{t-1}^{1/2} u\|$, and similarly $\|u\|_{V_{t-1}^{-1}} = \|V_{t-1}^{-1/2} u\|$. By the Cauchy-Schwarz inequality, $\langle \partial\theta, x\rangle = \langle V_{t-1}^{1/2}\partial\theta, V_{t-1}^{-1/2} x\rangle \leq \|\partial\theta\|_{V_{t-1}} \|x\|_{V_{t-1}^{-1}}$, and this is extremised when $V_{t-1}^{1/2}\partial\theta \propto V_{t-1}^{-1/2} x \iff \partial\theta \propto V_{t-1}^{-1} x$. Further, if for a scalar φ, $\partial\theta = \varphi \cdot V_{t-1}^{-1} x$, then

$$\partial\theta^\top V_{t-1} \partial\theta = \varphi^2 x^\top V_{t-1}^{-1} V_{t-1} V_{t-1}^{-1} x = \varphi^2 x^\top V_{t-1}^{-1} x,$$

or equivalently, $\|\varphi \cdot V_{t-1}^{-1} x\|_{V_{t-1}} = |\varphi| \|x\|_{V_{t-1}^{-1}}$, which means that to obey $\|\partial\theta\|_{V_{t-1}} \leq \omega_t$, we must set $\partial\theta = \pm \frac{\omega_t}{\|x\|_{V_{t-1}^{-1}}} V_{t-1}^{-1} x$, and of these the $+$ solution gives a positive value, and so is optimal.

Further notice that the optimal choice of ∂a^i for $i \in U_I$ must similarly be aligned with $V_{t-1}^{-1} x$. Indeed, write $V_{t-1}^{1/2}\partial a^i = \omega_t \sigma^i V_{t-1}^{-1/2} x + \psi^i$, where σ^i is a scalar, and ψ^i is a vector such that $\langle \psi^i, V_{t-1}^{-1/2} x\rangle = 0$. Then observe that due to the orthogonality,

$$\|\partial a^i\|^2_{V_{t-1}} = \langle V_{t-1}^{1/2}\partial a^i, V_{t-1}^{1/2}\partial a^i\rangle = \langle \omega_t\sigma^i V_{t-1}^{-1/2} x + \psi^i, \omega_t \sigma^i V_{t-1}^{-1/2} x + \psi^i\rangle = \omega_t^2 (\sigma^i)^2 \|x\|^2_{V_{t-1}^{-1}} + \|\psi^i\|^2,$$

and so the the constraint on ∂a^i becomes $(\omega_t\sigma^i)^2 \|x\|^{-1}_{V_{t-1}} + \|\psi^i\|^2 \leq \omega_t^2$. But ψ^i affects neither the first constraint on $\langle a^i + \Delta a^i_t + \partial a^i, x\rangle$, nor the objective, since

$$\langle \partial a^i, x\rangle = \langle V_{t-1}^{1/2}\partial a^i, V_{t-1}^{-1/2} x\rangle = \langle \omega_t \sigma^i V_{t-1}^{-1/2} x, V_{t-1}^{-1/2} x\rangle + \langle \psi^i, V_{t-1}^{-1/2} x\rangle = \sigma^i \|x\|^2_{V_{t-1}^{-1}}.$$

This means that dumping any energy into ψ^i affects neither the constraints nor the objective, so we can safety set it to zero in the following (in fact, as we shall see below, it must be zero since σ^i must saturate). This allows us to considerably simplify the above program: by introducing the real valued variables σ^i for $i \in I$, and noting that $\partial a^i = 0$ for $i \in K_I$ can be achieved by demanding $(\sigma^i)^2 \|x\|_{V_{t-1}^{-1}} \leq 0 = \mathbf{1}^i_U$ for $i \in K_I$, we may rewrite the program above as

$$\begin{aligned}\langle \theta, x^* \rangle + \max_x \max_{\{\sigma^i\}} \langle \Delta\theta_t, x\rangle - \sum_{i\in U_I} \mu^i \langle \Delta a^i_t, x\rangle + \omega_t \|x\|_{V_{t-1}^{-1}} - \sum_{i\in U_I} \mu^i \sigma^i \omega_t \|x\|^2_{V_{t-1}^{-1}}.\\ \text{s.t.} \quad \langle a^i + \Delta a^i_t, x\rangle = \alpha^i - \omega_t \sigma^i \|x\|^2_{V_{t-1}^{-1}} \quad \forall i \in I,\\ \langle a^i, x\rangle = \alpha^i \quad \forall i \in K_I\\ (\sigma^i)^2 \|x\|^2_{V_{t-1}^{-1}} \leq \mathbf{1}^i_U \quad \forall i \in I.\end{aligned}$$

Finally, observe that the first pair of constraints can be succinctly written in terms of $\hat{A}_t(\mathsf{U})$, giving us the following restatement, where σ is the vector formed by stacking the $\sigma^i s$.

$$\begin{aligned}\langle \theta, x^* \rangle + \max_x \max_\sigma \langle \Delta\theta_t, x\rangle - \sum_{i\in I} \mu^i \langle \Delta a_t^i, x\rangle + \omega_t \|x\|_{V_{t-1}^{-1}} - \langle \mu, \sigma\rangle \omega_t \|x\|^2_{V_{t-1}^{-1}}.\\ \text{s.t.} \quad \hat{A}_t(I)x = \alpha(I) - \omega_t \|x\|^2_{V_{t-1}^{-1}}\sigma\\ (\sigma^i)^2 \|x\|^2_{V_{t-1}^{-1}} \le \mathbf{1}^i_U \quad \forall i \in I.\end{aligned}$$

But notice that $A(K_I)$ is full row rank by assumption, and thus applying Lemma 27, with probability one, $\hat{A}_t(I)$ is full-rank. But this means that every value of σ that meets the final constraint is feasible for the above program, since we can find an appropriate x by inverting $\hat{A}_t(I)$. Of course, then the optimal choice of σ^i is then $-\mathbf{1}^i_U \text{sign}(\mu^i)/\|x\|_{V_{t-1}^{-1}}$, telling us that for each $i \in U_i$, the optimal ∂a^i at time t is

$$\partial a_t^i = -\mathbf{1}^i_U \text{sign}(\mu^i)\omega_t V_{t-1}^{-1} x / \|x\|_{V_{t-1}^{-1}} \implies \mu^i \langle \partial a_t^i, x\rangle = \mathbf{1}^i_U \omega_t |\mu^i| \cdot \|x\|_{V_{t-1}^{-1}}.$$

Now, finally, we observe that for each $i \in U_I$, and every x, $\omega_t |\mu^i| \|x\|_{V_{t-1}^{-1}} - \mu^i \langle \Delta a_t^i, x\rangle \ge 0$. Indeed, for $i \in K_I$ this is trivial since both $\Delta a_t^i, \partial a_t^i$ are 0 for such i. For $i \in U_I$, since the confidence sets are consistent, we know that $a^i \in \mathcal{C}_t^i \iff \|\Delta a_t^i\|_{V_{t-1}} \le \omega_t$. But then

$$|\mu^i \langle \Delta a_t^i, x\rangle| = |\mu^i| |\langle V_{t-1}^{1/2} \Delta a_t^i, V_{t-1}^{-1/2} x\rangle| \le |\mu^i| \|\Delta a_t^i\|_{V_{t-1}} \|x\|_{V_{t-1}^{-1}} \le |\mu^i| \omega_t \|x\|_{V_{t-1}^{-1}}.$$

But now we are in business. Indeed, using (6), we finally have

$$\begin{aligned}\langle \theta, \delta x_t\rangle &= \sum_{i\in I} -\mu^i \langle \delta a_t^i, x_t\rangle = \sum_{i\in I} -\mu^i \langle \partial a_t^i, x_t\rangle - \mu^i \langle \Delta a_t^i, x_t\rangle\\ &= \sum_{i\in I} |\mu^i| \omega_t \|x_t\|_{V_{t-1}^{-1}} - \mu^i \langle \Delta a_t^i, x_t\rangle \ge 0,\end{aligned}$$

and we are done. ■

The role of the non-degeneracy condition Assumption 16 in the above is fairly weak: all we really need is that x_t noisily activates some index set such that the true θ can be expressed via a linear combination of the true constraint vectors of the index set. In the absence of this, the proof does not quite work as stated, since it may be the case that some constraints that are needed to express θ are not noisily activated by x_t (although such constraints are activated by x^*). This removes the equality of the various programs we wrote, and would only leave us with a lower bound (in terms of some of these active at x^* but not noisily active at x_t constraints, along with the ones above), and it is unclear if x_t must also optimise this lower bound.

Nevertheless, we believe that this requirement is an artefact of our proof strategy: in general, optimistic play, when it leaks out of the safe set, has a tremendous freedom to activate any noisy constraints, and the conspiring of a choice of $\delta\theta$ and δA that makes the point suboptimal is severely constrained due to the presence of a large number of over-efficient actions in the vicinity of the safe set. Exactly nailing down an argument that cleanly expresses this intuition is an open problem.

E.2. Proof of the Main Theorem

With all the pieces in place, we proceed to argue our main claim.

Proof of Theorem 18. With probability at least $1-\delta$, all the confidence sets are consistent. We assume that this indeed occurs, and argue the claim under this event.

We first split the time horizon into two groups depending on whether x_t noisily activates suboptimal BISs or not by defining

$$\mathfrak{T}_1 := \{t \in [d+1:T] : \exists \text{ a suboptimal BIS } I \text{ such that } x_t \in \widetilde{\mathcal{X}}_t^I\}.$$

Notice that for $t \in [d+1:T] \setminus \mathfrak{T}_1$, x_t only activates optimal BISs.

Now, by Lemma 13, for all $t \in \mathfrak{T}_1$, $\rho_t(x_t;\delta) \geq \Gamma$, and further by the Lemma 17, for every $t \in [d+1:T] \setminus \mathfrak{T}_1$, it holds that $\langle \theta, x^* - x_t\rangle \leq 0$. Finally, we observe that it must hold that for all times

$$\langle \theta, x_t\rangle \geq \langle \theta, x^*\rangle - \rho_t(x_t;\delta).$$

Indeed, due to consistency, both θ and x^* are feasible choices for the actions of DOSS. Thus, if some $\tilde{\theta}, x_t$ are chosen instead, then $\langle \tilde{\theta}, x_t\rangle \geq \langle \theta, x^*\rangle$. But by Lemma 2, under consistency, $\langle \theta, x_t\rangle \geq \langle \tilde{\theta}, x_t\rangle - \rho_t(x_t;\delta)$, giving the above claim.

We thus have the efficacy control

$$\begin{aligned}
\mathscr{E}_T &= \sum_t \langle \theta, x^* - x_t\rangle_+ = \sum_{t\leq d} \langle \theta, x^* - x_t\rangle_+ + \sum_{t\in\mathfrak{T}_1} \langle \theta, x^* - x_t\rangle_+ + \sum_{t\notin\mathfrak{T}_1} \langle \theta, x^* - x_t\rangle_+ \\
&\leq d + \sum_{t\in\mathfrak{T}_1} \rho_t(x_t;\delta) + 0 \\
&\leq d + \sum_t \rho_t(x_t;\delta)\mathbb{1}\{\rho_t(x_t;\delta) \geq \Gamma\} \\
&\leq d + \sum_t \rho_t(x_t;\delta)\cdot\frac{\rho_t(x_t;\delta)}{\Gamma} \\
&= d + \frac{1}{\Gamma}\sum_t \rho_t(x_t;\delta)^2,
\end{aligned}$$

whence the claimed bound follows upon using Lemma 24. As in §5, we have used the trick that $\mathbb{1}\{u \geq v\} \leq u/v$ for positive v.

To control the safety behaviour, we observe that due to the property that $x_t \in \widetilde{\mathcal{S}}_t(\delta)$, there must exist some witness $\tilde{A}_t \in \boldsymbol{\mathcal{C}}_t(\delta)$ such that $\tilde{A}_t x_t \leq \alpha$. But, again by Lemma 2 that under consistency, for every i,

$$\langle \tilde{a}_t^i, x_t\rangle \geq \langle a^i, x_t\rangle - \rho_t(x_t;\delta),$$

which implies that

$$\max_i \langle \tilde{a}^i, x_t\rangle - \alpha^i \leq \rho_t(x_t;\delta).$$

But then

$$\mathscr{S}_T = \sum_{t\leq T} \max_i (\langle \tilde{a}^i, x_t\rangle - \alpha^i)_+ \leq \sum_{t\leq T} \rho_t(x_t;\delta),$$

and the claim is immediate from Lemma 24. Above, we have used the elementary fact that if $u \leq v$ and $v > 0$, then $(u)_+ \leq v$.

We note that the upper bound in Theorem 3 is also immediate from the above argument. The control on $\mathscr{S}_T$ can be repeated verbatim, while to control $\mathscr{E}_T$, we note that we began by showing that $\langle \theta, x^* - x_t \rangle \leq \rho_t(x_t; \delta)$, so the conclusion of the control on $\mathscr{S}_T$ above can be repeated verbatim. ■

E.3. Proofs of Polylogarithmic Safety Violation Claims from §7

Finally, we show the proof of the subsidiary observation from §7.

E.3.1. FINITE PRECISION IN CONSTRAINT LEVELS

The argument relies on the following observation.

Lemma 28 *Under consistency, for every $\varepsilon > 0, t$ if* DOSS(δ) *plays an action x_t such that $\max_i(\langle a^i, x_t \rangle - \alpha^i)_+ \geq \varepsilon$ then $\rho_t(x_t; \delta) \geq \varepsilon$.*

Proof. As in the proof of Theorem 18, if the algorithm plays x_t, then

$$\exists \tilde{A} \in \boldsymbol{C}_t(\delta) : \tilde{A} x_t \leq \alpha.$$

But, under consistency, by Lemma 2,

$$\tilde{A} x_t \geq A x_t - \rho_t(x_t; \delta) \mathbf{1}_U,$$

and so for every i,

$$\langle a, x_t \rangle - \alpha^i \leq \langle \tilde{a}^i, x_t \rangle + \rho_t(x_t; \delta) - \alpha^i \leq \rho_t(x_t; \delta),$$

and the claim follows by maximising over i. ■

The above is enough to enable the argument, which goes along the lines of the proof of logarithmic bounds on $\mathscr{E}_T$ in Theorem 18.

Proof of Theorem 20. As always, we begin by assuming consistency of the confidence sets, which occurs with probability at least $1 - \delta$. Observe that the proof of efficacy can be repeated verbatim from the previous section under consistency. To control the net violations, first recall that by Lemma 28, $\exists i : \langle a^i, x_t \rangle - \alpha^i > \varepsilon \implies \rho_t(x_t; \delta)$. It thus follows that

$$\begin{aligned}
\mathscr{S}_T^{\varepsilon} &= \sum_{t \leq T} (\langle a^i, x_t \rangle - \alpha^i) \mathbb{1}\{\exists i : \langle a^i, x_t \rangle - \alpha^i > \varepsilon\} \\
&\leq \sum_{t \leq T} \rho_t(x_t; \delta) \mathbb{1}\{\rho_t(x_t; \delta) > \varepsilon\} \\
&\leq \sum \rho_t(x_t; \delta)^2 / \varepsilon,
\end{aligned}$$

and the claim follows from Lemma 24. ■

E.3.2. FINITE PRECISION IN CONSTRAINTS

We argue that due to the finite precision in the constraint levels, there exists a minimal error scale for the problem.

Lemma 29 *There exists a constant $\pi > 0$ such that if the confidence sets are consistent, and that the finite-constraint-precision version of* DOSS*(δ) picks an x_t that only activates optimal BISs, but x_t is either infeasible or ineffective, then $\rho_t(x_t; \delta) \geq \pi$.*

Proof. Let I be an optimal BIS that x_t noisily activates. This is again full rank by Assumption 16, and there exists some $\tilde{A} \in \boldsymbol{\mathcal{C}}_t^{\mathsf{P}}$ such that $\tilde{A}(I)x_t = \alpha(I), \tilde{A}(I)x_t \leq \alpha$. As in the proof of Theorem 18, we can restrict attention to $\tilde{A}$ such that $\tilde{A}(I)$ is full-rank, since one such $I, \tilde{A}$ must exist.

Since both $\tilde{A}(I)$ is full rank, we immediately know that $x_t = \tilde{A}(I)^{-1}\alpha(I)$. But then, since there are only a finite number of possible choices for $\tilde{A}(I)$ in P, there are only a finite number of candidate x_t. Let us define $x(\tilde{A}(I)) = \tilde{A}(I)^{-1}\alpha(I)$, and $\mathcal{X}(I) = \{x(\tilde{A}(I))\}$. Since x^* is assumed to be the unique optimum, we know that for each $x \in \mathcal{X}(I)$, it must hold that $\pi(x) := \max\left\{(\langle\theta, x^* - x\rangle, \max_i(\langle a^i, x\rangle - \alpha^i)_+\right\}$ is strictly positive, which in turn yields that

$$\pi_I := \min_{x \in \mathcal{X}(I)} \pi(x) > 0.$$

Of course, we also conclude then that if x_t noisily activates I but is infeasible or suboptimal, then it must be at least π^I-infeasible or π^I-suboptimal, which via Lemma 2 and an argument similar to that in the proof of Lemma 28 yields that $\rho_t(x_t; \delta) \geq \pi_I$.

Of course, since some optimal full rank BIS must be activated, we conclude that if x_t is not the optimum, then

$$\rho_t(x_t; \delta) \geq \pi := \min_{\text{optimal BISs } I} \pi_I,$$

and we are done. ∎

Let us note that the argument above is quite crude, in that we simply take a minimum over all candidates once we establish the finitude of the set of these candidates. A more refined analysis may recover stronger local behaviour by analysing what types of $\tilde{A}(I)$ remain in $\boldsymbol{\mathcal{C}}_t(\delta)$ once enough information has been accumulated, and use this to develop notions of gaps for finite-constraint-precision scenarios that dominate the quantity we have constructed above. We leave this interesting line of study for future work.

In any case, exploiting the above yields the result.

Proof of Theorem 21. Working along the lines of the proof of Theorem 18 yields control in both the efficacy and safety costs accumulated over times t for which a suboptimal BIS was activated of the form $O(\Gamma^1 d^2 \log^2 T)$. Restricting attention then to optimal BISs, by the above, if a suboptimal or infeasible action x_t were picked, then by the above Lemma, $\rho_t(x_t; \delta) \geq \pi$. This lets us repeat the same argument, but now over t for which an optimal BIS was activated, which yields bounds of $O(\pi^{-1} d^2 \log^2(T))$, and the overall costs is bounded by the sum of these two quantities. ∎

E.3.3. FINITE ACTION SETTING.

Let us specify the setting in a little more detail: we are supplied with a finite set $\mathcal{A} \subset \mathbb{R}^d$, and in each round the learner chooses one action $x_t \in \mathcal{A}$. The linear reward and constraint structures are kept identical, and x^* is updated to be the best action in $\mathcal{A}$, i.e.,

$$x^* := \arg\max \langle \theta, x \rangle : Ax \leq \alpha, x \in \mathcal{A}.$$

Note that the known constraints are no longer necessary: if they are given, then we may filter $\mathcal{A}$ before play starts. The gap $\Delta := \min_{x \in \mathcal{A}, x \neq x^*} \max(\langle \theta, x^* - x \rangle, \max_i (\langle a^i, x \rangle - \alpha^i)_+)$ is non-zero simply because each suboptimal arm in $\mathcal{A}$ must be either infeasible, or ineffective, and the minimisation is over a finite set.

The result relies on the following observation, which follows straightforwardly from Lemma 2.

Lemma 30 *If the confidence sets are consistent, and the modified finite-action version of* DOSS *chooses $x_t \neq x_*$ from $\mathcal{A}$, then $\rho_t \geq \Delta$*

Proof of Lemma 30. Notice that the basic result Lemma 2 remains valid in this setting. As a result, if the confidence sets are consistent, then since x_t is permissible, there exists $\tilde{A} \in \boldsymbol{\mathcal{C}}_t$ such that $\tilde{A} x_t \leq \alpha$, and some $\tilde{\theta} \in \mathcal{C}_t^\theta : \langle \tilde{\theta}, x \rangle \geq \langle \theta, x^* \rangle$. Further, either there exists $i : \langle a^i, x_t \rangle \geq \alpha^i + \Delta$ or $\langle \theta, x \rangle \leq \langle \theta, x^* \rangle - \Delta$. But by consistency, $\langle \tilde{a}^i, x_t \rangle \geq \langle a^i, x_t \rangle - \rho_t(x_t; \delta)$ and $\langle \tilde{\theta}, x \rangle \leq \langle \theta, x \rangle + \rho_t(x_t; \delta)$, so either case implies $\rho_t(x_t; \delta) \geq \Delta$, which thus must hold. ■

Proof of Proposition 22. The claim can be shown using Lemma 30 along the lines of the proof of Theorem 20. ■

Appendix F. Proofs of Lower Bounds

We conclude by showing the lower bounds claimed in the main text.

F.1. Proof of Polynomial Lower Bound

We argue Theorem 4 by fleshing out the example developed in § 5. The proof uses techniques that are largely standard in the bandit literature (Lattimore and Szepesvári, 2020, Ch. 24).

Proof of Theorem 4. The instance we consider is

$$\mathcal{X} = [0, 1], \theta^* = 1, a^1 = (1 \pm \kappa)/2, \alpha^1 = 1/4, w_t^i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1), i \in \{0, 1\}$$

for some $\kappa \in (0, 1/4)$. Note that implicitly, the above has the know constraints $-x \leq 0$ and $x \leq 1$. Of course, this one-dimensional construction can be embedded into an arbitrary dimension (for instance, by taking a very skinny box domain, and only enforcing this single unknown constraint).

In the above case, the optimal feasible solutions are $x^+ = \frac{1}{2(1+\kappa)}, x^- = \frac{1}{2(1-\kappa)}$ for these two instance respectively. In addition, both of these two instances are at least $1/8-$well separated. The key observation is the indistinguishability of these two instances with $\ll 1/\kappa^2$ actions.

Indeed, let $\mathbb{P}^+, \mathbb{P}^-$ be the distributions induced by the two problem instances and the learning algorithm. Since in either case, the noise distribution is standard Gaussian, and the reward distributions are identical, it follows that

$$D(\mathbb{P}^+(r_t, s_t^1) \| \mathbb{P}^-(r_t, s_t^1) | x_t = x) = \frac{(\kappa x)^2}{2} \leq \frac{\kappa^2}{2},$$

where we have used standard results about the KL-divergence between two Gaussians. Further, since actions must be causal, and since the noise is independent, we conclude that over the whole trajectory,

$$D(\mathbb{P}^+(\mathcal{H}_T) \| \mathbb{P}^-(\mathcal{H}_T)) \leq \frac{T\kappa^2}{2}.$$

Let $x^{\mathsf{av}} := (x^+ + x^-)/2 = \frac{1}{2(1-\kappa^2)}$. Observe that

- if the ground truth is $a^1 = (1+\kappa)/2$ and $x_t \geq x^{\mathsf{av}}$, then the algorithm incurs an instantaneous safety violation of at least $(1+\kappa)/2 \cdot x^{\mathsf{av}} - 1/4 = \frac{1+\kappa}{2} \cdot \frac{1}{2(1-\kappa^2)} - \frac{1}{4} = \frac{\kappa}{4(1-\kappa)} \geq \frac{\kappa}{4}$;
- if the ground truth is $a^1 = (1-\kappa)/2$ $x_t < x^{\mathsf{av}}$, then the algorithm incurs an instantaneous efficacy regret of at least $\frac{1}{2(1-\kappa)} - \frac{1}{2(1-\kappa^2)} \geq \frac{\kappa}{2}$

Let A be the event $\{\#\{t : x_t \geq x^{\mathsf{av}}\} \geq T/2\}$. Using the Bretagnolle-Huber inequality (Lattimore and Szepesvári, 2020, Thm. 14.2),

$$\mathbb{P}^+(\mathsf{A}) + \mathbb{P}^-(\mathsf{A}^c) \geq \frac{1}{2} \exp\left(D(\mathbb{P}^+(\mathcal{H}_T) \| \mathbb{P}^-(\mathcal{H}_T))\right) \geq \frac{1}{2} \exp(-T\kappa^2/2).$$

Let $\mathscr{E}_T^-$ denote the efficacy regret incurred by the learner under $\mathbb{P}^-$ and $\mathscr{S}_T^+$ denote the safety violation incurred by the learner under $\mathbb{P}^+$. Under the event A, if the true a was $(1+\kappa)/2$, at least $T/2$ rounds incurred a safety regret of at least $\kappa/4$, and so $\mathscr{S}_T^+ \geq \kappa T/8$. Similarly, under A^c, at least $T/2$ rounds had $z_t = -1$, implying that $\mathscr{E}_T^- \geq T\kappa/8$.

But this implies that

$$\max(\mathbb{E}^-(\mathscr{E}_T^-), \mathbb{E}^+(\mathscr{S}_T^+)) \geq \frac{T\kappa}{8} \max(\mathbb{P}^+(\mathsf{A}), \mathbb{P}^-(\mathsf{A}^c)) \geq \frac{T\kappa}{32} \exp(-T\kappa^2/2).$$

For $T \geq 16$, we may choose $\kappa = 1/\sqrt{T} < 1/4$ to conclude that in at least one instance, the safety or efficacy regret incurred must be at least $\sqrt{T}/(32e^{1/2}) \geq \sqrt{T}/64$.

■

F.2. Necessity of Dependence on Gaps.

We conclude the theoretical part of this paper by showing Theorem 19, via a reduction to prior lower bounds on the safe multi-armed bandit problem (Chen et al., 2022).

The safe MAB problem is parametrised by d arms with mean rewards μ_k and mean safety risks ν_k each. The optimal arm, k^* has reward μ_* and the safety risk $ny_* < \alpha$. The associated efficacy and safety gaps are $\Delta_k := (\mu_* - \mu_k)_+$ and $\Gamma_k := (\nu_k - \alpha)_+$. In each round, the leaner is required to select one arm, and observes bounded signals with the above mean for both the rewards and safety. Implicitly, this can be thought of as a linear bandit setting, with the known constraints being that x

lies in a simplex, the reward vector θ, and the constraint vector a. This reduction, however, is not completely correct: in the safe MAB problem, the actions are required to lie entirely on the corner points of the simplex, and we are not allowed to play in the interior. While it is standard to view x as a probability of selecting each arm in a MAB instance, this reduction fails due to the nonlinearity in our metrics. Indeed, the safe MAB problem considers the metrics

$$\mathscr{E}_T^{\mathsf{MAB}} := \sum (\mu^* - \mu_{A_t})_+, \mathscr{S}_T^{\mathsf{MAB}} := \sum (\nu_{A_t} - \alpha)_+.$$

As a result, if the optimum of the SLB problem lies away from the corner points of the simplex, then the SLB problem can incur low regret, while the corresponding MAB actions would incur linear regret. Nevertheless, we shall argue below that for carefully designed instances, a low regret in the linear bandit problem does ensure nontrivial regret in the safe MAB problem.

The main result we shall use is the following, which is a mild variation of Proposition 6 of Chen et al. (2022), and can be shown using their proof.

Lemma 31 *Let $f : \mathbb{N} \to [0, \infty)$ be any function fixed function such that $f(T) \leq T$ for all T. If an algorithm ensures that for every safe MAB instance, suboptimal arms are not played more than $f(T)$ times in expectation, then for every θ, a, there exists a choice of arm distributions for the safe MAB instance for which the means are as described, and the number of times each suboptimal arm k is played is lower bounded in expectation as*

$$\mathbb{E}[N_T^k] \geq \frac{1}{(d(\mu_k \| \mu_*)\mathbb{1}\{\mu_k < \mu_*\} + d(\nu_k \| \alpha)\mathbb{1}\{\nu_k > \alpha\})} \cdot \left((1 - f(T)/T) \log \frac{T}{f(T)} - \log(2) \right),$$

where $d(u\|v)$ is the KL divergence between Bernoulli laws with means u and v. In particular, these distributions are simply Bernoulli laws with the above means.

Our argument for the linear bandit proceeds thus. We shall carefully design a safe linear bandit instance for which we essentially provide multi-armed bandit feedback by using the standard reduction that each coordinate of x_t represents the probability of pulling the corresponding arm. We shall show that in the selected instance, achieving low linear regret ensures that the MAB regret is controlled (although to a weaker level). Then exploiting the above lower bound, we shall argue that the regret of the safe linear bandit cannot be too good, since it would violate the above lower bound.

Proof of Theorem 19. We first carefully describe our main constructions for the SLB and MAB, form a crude bound that allows us to use Lemma 31, and then refine the analysis to show effective lower bounds on the SLB regret.

SLB Instance. We work with $d = 2$ with a single unknown constraint. Let $\theta = (\theta_1, \theta_2)$ and $a^1 = (\alpha, a_2^1)$ be vectors in $[0, 1]^2$ such that $\theta_2 > \theta_1 > 0, a_2^1 > \alpha > 0$ and $\theta_2 \alpha < \theta_1 a_2^1$. The safe bandit instance we design is

$$\max \langle \theta, x \rangle : x_1 \geq 0, x_2 \geq 0, x_1 + x_2 \leq 1, \langle a, x \rangle \leq \alpha,$$

where the last constraint is unknown and the rest are known. Let us call the three known constraints a^2, a^3, a^4. There are 6 BISs, with the associated points and gaps shown in Table 1 below. The only points meeting the constraints a^3 and a^4 is $(0, 1)$, which is infeasible. Note that the situation is highly degenerate at the optimal point is $x^* = (1, 0)$ and three distinct BISs activate it. Nevertheless, each

Table 1: Description of BISs in our construction.

BIS	Activating Point	$\zeta_*(I)$	$\eta_*(I)$
$\{1,2\}$	$(0, \alpha/a_2^1)$	0	$(\theta_1 a_2^1 - \alpha\theta_2)/(a_2^1 + \theta_2)$
$\{1,3\}$	$(1,0)$	0	0
$\{1,4\}$	$(1,0)$	0	0
$\{2,3\}$	$(0,0)$	0	θ_1
$\{2,4\}$	$(1,0)$	0	0
$\{3,4\}$	$\emptyset$	$a_2^1 - \alpha$	0

of these BISs is full rank. Further, since the algorithm ensures that $\mathscr{S}_T$ and $\mathscr{E}_T$ are both $O(\sqrt{T})$ in general, our discussion below is effective.

The gap of this instance is

$$\Gamma := \min\left(\theta_1, a_2^1 - \alpha, \frac{\theta_1 a_2^1 - \alpha\theta_2}{a_2^1 + \theta_2}\right).$$

Our construction requires that this is at least $\Omega(\min(\theta_1, a_2^1 - \alpha))$. This can always be ensured, example, by using the parameterisation $\theta_2 = 2\theta_1, a_2^1 = 4\theta_1, \alpha = \theta_1/2$, whence the expressions work out to

$$a_2^1 - \alpha = 7\theta_1/2, \frac{\theta_1 a_2^1 - \alpha\theta_2}{a_2^1 + \theta_2} = 3\theta_1/5,$$

giving us $\Gamma \geq \theta_1/2$. We further impose the condition $4\theta_1 < 1/4$. Thus, this instance lets us express every value of $\Gamma < 1/32$.

Safe MAB Instance. Let us now describe the associated MAB instance. We work with three arms of means $\mu = (1/2 + \theta_1, 1/2 + \theta_2, 1/2)$ and risks $(1/2 + \alpha, 1/2 + a_2^1, 1/2)$. In each case, the underlying laws are Benoullis with the associated mean, all taken to be independent, which forms the family of instances that underly Lemma 31. The connection to the linear bandit instance is as follows: each time we pick (x_1, x_2), we sample a random variable in $\{1,2,3\}$ according to the pmf $(x_1, x_2, 1 - x_1 - x_2)$, pull the corresponding arm, and then supply the resulting rewards and risks with $1/2$ subtracted to the linear bandit instance. Note that this is an unbiased measurement of the mean for the linear bandit, since

$$\mathbb{E}[R] = x_1 \cdot (1/2 + \theta_1) + x_2 \cdot 1/2 + \theta_2) + x_2 \cdot (1/2) - 1/2 = x_1\theta_1 + x_2\theta_2$$

and similarly for the safety risk. These $1/2$ are added to ensure because then the KL divergences appearing in the bound of Lemma 31 take the form $d(1/2\|1/2 + \theta_1)$ and $d(1/2 + a_2^1\|1/2 + \alpha)$, and the arguments are bounded away from 0 and 1, ensuring that the behaviour for small θ_1 is quadratic rather than the potentially worse dependence near 0 and 1. To ensure this good behaviour, we use that $a_2^1 < 1/4$, due to which $a_2^1 + 1/2 < 13/14$ is bounded away from 1, which is the origin of our condition $\theta_1 \leq 1/16$ in the previous paragraph. The key observation is that in the safe MAB instance, $\mathbb{E}[N_T^2] = \sum x_{t,2}$ and $\mathbb{E}[N_T^3] = \sum(1 - x_{t,1} - x_{t,2})$, where $x_{t,k}$ is the kth component of x_t.

Crude Bound. We first show that as long as the algorithm ensures that $\max(\mathscr{E}_T, \mathscr{S}_T) = O(T^{1-c})$, the play of suboptimal arms in the MAB instance is at least $\Omega(\theta_1^{-2}\log T)$. Fix θ_1 and the above models. Suppose that the safe linear bandit ensures that $\mathscr{E}_T \leq g(T)$ and $\mathscr{S}_T \leq g(T)$ for every

instance, where $g(T) \leq T$ is an arbitrary monotonic function. Let $\zeta > 0$ be a parameter that we will fix later. Then observe that if the linear bandit instance ever plays a point (x_1, x_2) such that

$$\langle a^1, x\rangle \geq \alpha + \zeta \quad \text{or} \quad \langle \theta, x\rangle \leq \theta_1 - \zeta,$$

then it would incur a point wise cost of at least γ in the round, for either $\mathscr{E}_T$ or $\mathscr{S}_T$. This means that the number of rounds in which it plays such points is bounded as $g(T)/\zeta$. So, in at least $\max(T - g(T)/\zeta, 0)$ rounds, the safe linear bandit instance plays in the region

$$P_\zeta := \{\langle a^1, x\rangle \leq \alpha + \zeta, \langle \theta, x\rangle \geq \theta_1 - \zeta, x_1 \geq 0, x_2 \geq 0, x_1 + x_2 \leq 1\}.$$

Now notice that both x_2 and $x_1 + x_2$ are upper bounded in this region. Indeed, the corner points of this region are

$$\left(1 - \frac{\zeta}{\theta_1}, 0\right), \left(1 - \frac{\zeta}{a_2^1 - \alpha}, \frac{\zeta}{a_2^1 - \alpha}\right), \left(1 - \frac{\zeta}{\theta_1}\left\{1 + \frac{\theta_2(\theta_1 + \alpha)}{\theta_1(\theta_1 a_2^1 - \alpha\theta_2)}\right\}, \frac{\theta_1 + \alpha}{\theta_1 a_2^1 - \alpha\theta_2}\frac{\zeta}{\theta_1}\right), (1, 0),$$

and so ensuring that $a_2^1 - \alpha, \theta_1 a_2^1 - \alpha\theta_2 = \Omega(\theta_1)$, we have

$$x \in P_\zeta \implies x_2 \leq \zeta/\theta_1, (1 - x_1 - x_2) = O(\zeta/\theta_1).$$

Of course, outside of P_ζ, $x_2 \leq 1, 1 - x_1 - x_2 \leq 1$. The calculation holds no matter the ζ we chose so long as $\zeta \ll \theta_1$. This means that for every $\zeta = O(\theta_1)$,

$$\mathbb{E}[N_T^2] = \sum \mathbb{E}[x_{t,2}] \leq O(\zeta/\theta_1)T + \frac{g(T)}{\zeta}$$
$$\mathbb{E}[N_T^3] = \sum \mathbb{E}[(1 - x_{t,1} - x_{t,2})] \leq O(\zeta/\theta_1)T + \frac{g(T)}{\zeta},$$

i.e., we have shown that the safe MAB incurs regret bounds of at most $f(T) = O(\zeta T) + g(T)\theta_1/\zeta$ for both $\mathscr{E}_T$ and $\mathscr{S}_T$.

Since $g(T) \leq CT^{1-c}$ for some constants C, c, by taking $\zeta = T^{-c/2}$, for large enough t, we thus have the low-regret bound $\max(\mathbb{E}[\mathscr{E}_T^{\mathsf{MAB}}], \mathbb{E}[\mathscr{S}_T^{\mathsf{MAB}}]) \leq CT^{1-c/2}$. But then, by Lemma 31, it must follow that as $T \to \infty$,

$$\mathbb{E}[N_T^2] \geq \frac{1}{d(1/2 + 4\theta_1 \| 1/2 + \theta_1/2)}\left((1 - o(1))\frac{c}{2}\log T - O(1)\right) = \Omega(\theta_1^{-2}\log T),$$
$$\text{or } \mathbb{E}[N_T^3] \geq \frac{1}{d(1/2 \| 1/2 + \theta_1)}\left((1 - o(1))\frac{c}{2}\log T - O(1)\right) = \Omega(\theta_1^{-2}\log T)$$

To use these bounds effectively, we employ a computer algebra system to argue that[8]

$$\forall \theta_1 \leq 1/16, d(1/2 + 4\theta_1 \| 1/2 + \theta_1/2) \leq 27\theta_1^2, d(1/2 \| 1/2 + \theta_1) \leq 27\theta_1^2.$$

8. Observe that since the divergences considered are minimised to 0 at $\theta_1 = 0$, the local behaviour for small θ_1 is quadratic. Further, the function is smooth in θ_1. Thus, for large enough K, there exists an interval $[0, \theta_1(K)]$ such that for any x in this region, $d(1/2 + ux \| 1/2 + vx) \leq Kx^2$. We simply plugged in various constants for K until we found that $\theta_1(27) \geq 1/16$.

Concretely, then the bounds above yield

$$\mathbb{E}[N_T^2 + N_T^3] \geq \frac{c}{27\theta_1^2}\left((1-o(1))\log T - \frac{1}{c}\log(4)\right),$$

where the $o(1)$ term is $C/T^{c/2}$.

Note again that this bound is effective for our case since the method DOSS does achieve $\max(\mathscr{E}_T, \mathscr{S}_T) = \tilde{O}(\sqrt{T})$ with high probability, in which case we can set $c = 1/2 + \gamma$ for any $\gamma > 0$ in the above.

Lower Bounds on SLB. Let us now come to showing the claims. We select the instance $\theta_2 = 2\theta_1, a_2^1 = 4\theta_1, \alpha = \theta_1/2$. Notice that in this case, the gaps are $(\theta_1, 7\theta_1/2, 3\theta_1/5)$, and so $\Gamma \geq \theta_1/2$. Further, $\theta_1 a_2^1 - \alpha\theta_2 = 3\theta_1$, and so the claim on P_ζ above remains valid for all $\zeta \leq \theta_1$, and so against this instance, the above lower bounds on $\mathbb{E}[N_T^2] + \mathbb{E}[N_T^3]$.

But, observe that for any choice of x_1, x_2, it holds that the instantaneous efficacy regret and safety violations are

$$(\theta_1 - \theta_1 x_1 - \theta_2 x_2)_+ = (\theta_1(1 - x_1 - x_2) - (\theta_2 - \theta_1)x_2)_+ = \theta_1((1 - x_1 - x_2) - x_2)_+$$
$$(\alpha x_1 + a_2^1 x_2 - \alpha)_+ = ((a_2^1 - \alpha)x_2 - \alpha(1 - x_1 - x_2))_+ = \frac{\theta_1}{2}(7x_2 - (1 - x_1 - x_2))_+$$

But notice that the only way both of these quantities are 0 is if $x_2 \geq (1 - x_1 - x_2) \geq 7x_2 \implies x_2 = 1 - x_1 - x_2 = 0 \iff x_1 = 1$. So, in any round such that $x_1 \neq 1$, at least one of these quantities is nonzero. More quantitatively, we have

$$\begin{aligned}
\mathbb{E}[\mathscr{E}_T] + \mathbb{E}[\mathscr{S}_T] &\geq \sum \mathbb{E}[(\theta_1 - \theta_1 x_{t,1} - \theta_2 x_{t,2}) + (\alpha x_{t,1} - a_2^1 x_{t,2} - \alpha)] \\
&\geq \theta_1 \sum \frac{5}{2}\mathbb{E}[x_{t,2}] + \frac{1}{2}\mathbb{E}[(1 - x_{t,1} - x_{t,2})] \\
&\geq \frac{\theta_1}{2}(\mathbb{E}[N_T^2] + \mathbb{E}[N_T^3]) \geq \frac{c(1-o(1)}{54\theta_1}\log(T) - O(1),
\end{aligned}$$

which yields the result upon recalling that $\theta_1 \geq \Gamma \geq \theta_1/2$. ■

Appendix G. Alternative Safety Metrics

We briefly investigate the behaviour of alternative safety metrics of the form

$$\mathscr{S}_T^f := \sum_{t \leq T} f(\max_i(\langle a^i, x_t\rangle - \alpha^i)_+),$$

where f is some increasing h-Hölder continuous map such that $\lim_{x \searrow 0} f(x) = 0$. Note that this section should be read *after* the reader is familiar with our typical proof techniques.

Note that due to our assumption that $\|a^i\| \leq 1, \|x\| \leq 1$, it follows that $(\langle a^i, x_t\rangle - \alpha^i)_+ \leq 2$: since the problem is feasible, and $|\langle a^i, x^*\rangle| \leq 1$, it follows that $-\mathbf{1} \leq Ax^* \leq \alpha$, and so $\langle a^i, x\rangle - \alpha^i \leq 1 - (-1)$. Thus, only the behaviour of f over $[0, 2]$ matters.

Now, since f is Hölder continuous, and $f(0^+) = 0$, its behaviour near 0 is as $f(x) \leq Cx^h$. In this case, we may as well study the behaviour of $f_h := x \mapsto x^h$, which we argue determines the

lower and upper bounds. Before proceeding, note that we may uniformly bound the noise scale by 2, since we know that roundwise inefficacy or constraint violation can be at most 2.

Now, for the penalty $\mathscr{S}_T^h = \sum_{t \le T} (\max_i \langle a^i, x_t \rangle - \alpha^i)_+^h$, observe that if $h \ge 2$, we can direclty bound the behaviour of $\mathscr{S}_T^h$ using

$$\mathscr{S}_T^h \le \sum \rho_t^h(x_t; \delta) \le 2^{h-2} \sum \rho_t^2(x_t; \delta) \le 2^{h-2} \cdot O(d^2 \log^2 T).$$

Thus, the only interesting behaviour is when $h < 2$.

For $h \in (0, 2)$, applying Hölder's inequality with $p = 2/h > 1$, we have

$$\mathscr{S}_T^h \le \sum \rho_t^h(x_t; \delta) \le \left(\sum_t (\rho_t^h)^{2/h} \right)^{h/2} \cdot (\sum 1^{2/(2-h)})^{1-h/2} = T^{1-h/2} \cdot O(d^h \log^h T).$$

We now note that modifying the analysis of §5, this rate of safety decay is tight. Indeed, our construction in that section shows that for $t \le 1/\kappa^2$, one either incurs a roundwise inefficacy of κ or a roundwise violation of κ. Accounting for the power-cost, we get a lower bound of the form

$$\textit{either } \mathscr{E}_T \gtrsim \kappa \cdot \min(\kappa^{-2}, T) \textit{ or } \mathscr{S}_T^h \gtrsim \kappa^h \cdot \min(\kappa^{-2}, T).$$

But again, taking $\kappa = T^{-1/2}$, we find that

$$\textit{either } \mathscr{E}_T \ge \sqrt{T} \textit{ or } \mathscr{S}_T^h \ge T^{1-h/2}.$$

Thus, up to polylog terms, the behaviour of DOSS remains tight in terms of the $\mathscr{S}_T^h$ behaviour, simultaneously for every $h > 0$.

Coming back to general smooth losses, we immediately note that the same analysis extends to any loss that is h-Hölder: using the bound $f(x) \le Cx^h$,

$$\mathscr{S}_T^f \le C \sum_t \rho_t^h(x_t; \delta),$$

and the bound follows. This extends to losses f that are smooth in some interval near 0^+ of the form $(0, k)$. For the upper bound, we may decompose the net violation as

$$\mathscr{S}_T^f \le \sum_t f(\rho_t) \mathbf{1}\{\rho_t > k\} + \sum_t f(\rho_t) \mathbf{1}\{\rho_t \le k\}.$$

The latter term can be dealt with as above, since $f(x) \le Cx^h$ over $(0, k)$, while the former term can be bounded as

$$\sum_t f(\rho_t) \mathbb{1}\{\rho_t \ge k\} \le (\max_{x \in [0,2]} f(x)) \sum_t \rho_t^2 / k^2 = O(k^{-2} d^2 \log^2 T),$$

leading to an additive polylogarithmic overhead beyond the main term. The lower bound also generalises: if on $(0, k)$, f is h-Hölder but not h'-Hölder for any $h' > h$, then there exists some interval $(0, k')$ over which f/x^h remains both lower and upper bounded, and we can employ our lower bound for $\mathscr{S}_T^h$ for $T \gg 1/(k')^2$.

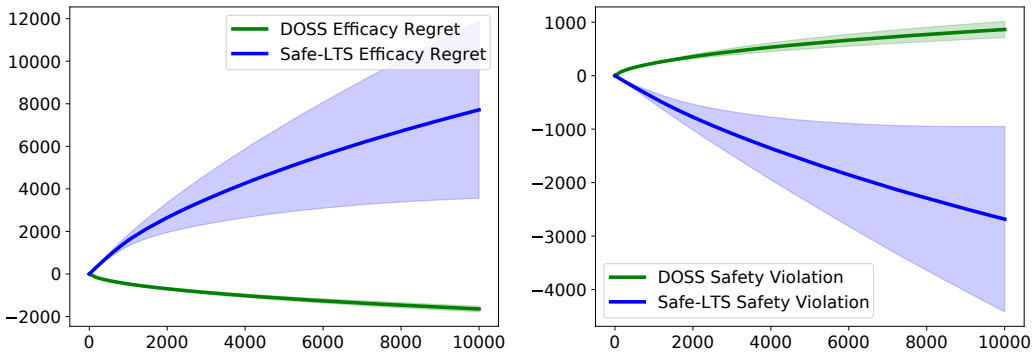

Figure 7: Comparing the behaviour of DOSS and safe-LTS on the instance of Example 7. The left plot shows the *raw* efficacy regret, while the right plot is the *raw* safety violations of the two methods, and each reports means and one-standard deviation confidence regions over 30 runs.. Observe that the efficacy performance of safe-LTS is extremely poor, indicating that the algorithm is far from the boundary of the safe set $\mathcal{S}$ for most of its runs. In contrast, the violation properties of DOSS are well-controlled, and almost four times smaller than the efficacy regret of safe-LTS.

Appendix H. Appendix to the Simulations

DOSS Compares Favourably with Pessimistic-Optimistic Methods. To contextualise our method, we also implement the PO-method *safe-LTS* due to Moradipari et al. (2021) in the instance of Example 7. Instead of the optimistic permissible set $\widetilde{\mathcal{S}}$, safe-LTS constructs a pessimistic set $\Pi_t = \{x : \forall \tilde{A} \in \boldsymbol{C}_t, \tilde{A}x \leq \alpha\}$. Note that with high probability, all points in Π_t must be safe. The method then selects actions optimistically, in this case by exploiting Thompson sampling. Naturally, this method requires the knowledge of a safe point with margin to being with, and we supply the point $x^{\mathsf{s}} = (0, 0)$ to the method, which has the (large) margin $M^{\mathsf{s}} = 1/2$.

Figure 7 compares the behaviour of the *raw efficacy regret* $\sum \langle \theta, x^* - x_t \rangle$ (left) and the *raw safety violation* $\sum \max_i (\langle a^i, x_t \rangle - \alpha^i)$ (right) of Safe-LTS and DOSS (since the efficacy regret of DOSS , and the safety-violations of safe-LTS are both essentially 0, the raw behaviour elucidates more insight). As expected, safe-LTS suffers from 0 safety regret, since it plays in a pessimistic set. However, this is accompanied by a large efficacy regret, with the mean of over 7000 at the horizon $T = 10^4$. This arises due to the extreme conservatism of this method, which is evident from its safety violation property: the method has a strong negative (and decreasing still) violation, indicating that it continues to play deep in the interior of the domain for large T. Indeed, since over the domain, $\langle a, x \rangle - \alpha \in [-0.5, 0.5]$, and since the violation at $T = 10^4$ is roughly -3000, this indicates that with a nontrivial probability, the method remains at least 0.25-separated from the boundary of the safe set.

In comparison, observe that the raw efficacy regret of DOSS is negative, but not nearly as far as the violations of safe-LTS. This indicates that the method is shrinking towards the boundary of the safe set at a much better rate. Of course, this property is similarly illustrated by the violation behaviour: this nearly four times smaller than the efficacy regret of safe-LTS, and concentrates strongly to ≈ 800 at $T = 10^4$.