Constrained Linear Thompson Sampling

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study safe linear bandits (SLBs), where an agent selects actions from a convex set to maximize an unknown linear objective subject to unknown linear constraints in each round. Existing methods for SLBs provide strong regret guarantees, but require solving expensive optimization problems (e.g., second-order cones, NP hard programs). To address this, we propose Constrained Linear Thompson Sampling (COLTS), a sampling-based framework that selects actions by solving perturbed linear programs, which significantly reduces computational costs while matching the regret and risk of prior methods. We develop two main variants: S-COLTS, which ensures zero risk and $\widetilde{O}(\sqrt{d^3T})$ regret given a safe action, and R-COLTS, which achieves $\widetilde{O}(\sqrt{d^3T})$ regret and risk with no instance information. In simulations, these methods match or outperform state of the art SLB approaches while substantially improving scalability. On the technical front, we introduce a novel coupled noise design that ensures frequent 'local optimism' about the true optimum, and a scaling-based analysis to handle the per-round variability of constraints.

1 Introduction

Stochastic bandit problems are a fundamental model for optimising unknown objectives through repeated trials. While single-objective bandit theory is well-developed, real-world learners must also deal with *unknown constraints* at every round of interaction. For instance, in *dose-finding* [AKR21], *micro-grid control* [FLZY22], and *fair recommendation* [Cho+24], a learner must choose actions that maximise reward while never crossing unknown toxicity, voltage, or exposure limits (see §B).

The *safe linear bandit* (SLB) problem models these scenarios in a linear programming (LP) setting: a learner selects actions $\{a_t\}$ from a convex domain $\mathcal{A}$ to optimize an unknown objective vector $\theta_* \in \mathbb{R}^d$ subject to unknown constraints of the form $\Phi_* a \le \alpha$, where $\Phi_* \in \mathbb{R}^{m\times d}$. After each action, the learner observes noisy feedback of the objective $\theta_*^\top a + \mathrm{noise}$ and the constraints $\Phi_* a + \mathrm{noise}$, thus acquiring information to guide future actions. Performance in SLBs is measured via the

$$\textit{regret, } \mathbf{R}_T := \sum_{t\le T} \big(\theta_*^\top(a_* - a_t)\big)_+, \quad \text{and} \quad \textit{risk, } \mathbf{S}_T := \sum_{t\le T} \Big(\max_i \big(\Phi_* a_t - \alpha\big)^i\Big)_+, \tag{1}$$

where a_* is the optimal action under the true (but unknown) constraints, and $(\cdot)_+ := \max(\cdot, 0)$. There are two main notions of safety in SLBs:

- *Hard constraint enforcement*, which requires that with high probability, $\mathbf{S}_T = 0$ for all T. This is only achievable if the learner has prior access to a *known safe action* a_{safe}.
- *Soft constraint enforcement*, which requires $\mathbf{S}_T = o(T)$ with high probability (whp). This is a weaker requirement, but does not need prior information.

A series of recent work [e.g. GCS24; PGB24; AAT19; MAAT21] offers OFUL-style algorithms for SLBs with strong theoretical guarantees. However, these often require the solution of nontrivial optimisation problems (second-order conic programs, and sometimes NP-hard problems) in each round. Our motivation lies in improving this computational cost.

Submitted to the Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025). Do not distribute.

Table 1: COMPARISON OF SLB METHODS. 'Known a_{safe}' means that the method requires an action known a priori to be safe. $\Delta(a) := \theta_*^\top(a_* - a)$ is the reward gap of an action a, and $\Gamma(a) := \min_i(\alpha - \Phi_* a)_+^i$ is its safety margin. $\mathcal{R}(a) := 1 + (\Delta(a)/\Gamma(a))$ if $\Gamma(a) > 0$, and ∞ otherwise. LP is the computation needed to optimize a linear objective with m linear constraints over $\mathcal{A}$ to constant approximation. SOCP is the same with m second-order conic constraints. We write 'NP-hard' if implementing the method needs a solver for an NP-hard problem. OPT-PESS refers to most frequentist hard enforcement methods discussed in §1.1, which have similar costs and bounds; SAFE-LTS is due to [MAAT21]; DOSS and the lower bound are due to [GCS24].

Algorithm	Assumptions	Regret	Risk	Compute at t
OPT-PESS	Known a_{safe}	$\mathcal{R}(a_{\mathsf{safe}}) \cdot \widetilde{O}(\sqrt{d^2T})$	0	NP-hard
Relaxed OPT-PESS	Known a_{safe}	$\mathcal{R}(a_{\mathsf{safe}}) \cdot \widetilde{O}(\sqrt{d^3T})$	0	$d \cdot \mathsf{SOCP} \cdot \log(t)$
SAFE-LTS	Known a_{safe}	$\mathcal{R}(a_{\mathsf{safe}}) \cdot \widetilde{O}(\sqrt{d^3T})$	0	$\mathsf{SOCP} \cdot \log(t)$
S-COLTS	Known a_{safe}	$\mathcal{R}(a_{\mathsf{safe}}) \cdot \widetilde{O}(\sqrt{d^3T})$	0	$\mathsf{LP} \cdot \log(t)$
DOSS	Feasibility	$\widetilde{O}(\sqrt{d^2T})$	$\widetilde{O}(\sqrt{d^2T})$	NP-hard
R-COLTS	Feasibility	$\widetilde{O}(\sqrt{d^3T})$	$\widetilde{O}(\sqrt{d^3T})$	$\mathsf{LP} \cdot \log^2(t)$
LOWER-BOUND	Feasibility	$\max(\mathbf{R}_T, \mathbf{S}_T) = \Omega(\sqrt{T})$, *no matter the instance*;		

Contributions. We introduce a *sampling-based* approach, *COnstrained Linear Thompson Sampling* (COLTS), which adds carefully chosen noise to estimates of both the objective and constraint parameters, and selects actions according to this perturbed program. This allows us to maintain the same order of regret and risk bounds as prior methods, while substantially reducing the complexity of each round. However, just perturbing the program as above does not directly yield good actions, since the perturbed program may be infeasible, or its optimum may be unsafe. We therefore develop two augmentations of COLTS, which address the SLB problem under distinct regimes:

- **S-COLTS** assumes a *given safe action* a_{safe}. Actions are picked by first solving a perturbed LP (while ensuring that a_{safe} is feasible), and then scaling its optimum towards a_{safe} to ensure safety. This yields *zero* risk, and regret $\mathcal{R}(a_{\mathsf{safe}}) \cdot \widetilde{O}(\sqrt{d^3T})$ (see §2, or Table 1 for definition of $\mathcal{R}(a)$).
- **R-COLTS** requires only *feasibility* of the true problem, and operates by sampling $O(\log T)$ perturbed programs, and setting a_t to be the optimiser of the one with largest value. This resampling directly yields optimism, leading to instance-independent $\widetilde{O}(\sqrt{d^3T})$ regret and risk bounds. We additionally argue that under Slater's condition, and with extra exploration, a similar regret and risk guarantee follows without resampling, and so solving only one optimisation per round.

Table 1 summarizes our results in comparison to prior work. Each variant attains regret and risk bounds matching those of prior methods, whilst selecting actions by only optimising over linear constraints (in addition to those due to $\mathcal{A}$). This yields the first efficient method for soft enforcement, and significantly speeds up hard constraint enforcement. Contextual extensions are discussed in §E.

Technical Innovations. The random perturbations in our sampling-based approach cause two challenges that break existing analyses of linear TS: (i) the feasible region fluctuates at each round; and (ii) the true optimum a_* can become infeasible under perturbed constraints, complicating direct analysis. We address these via two key innovations:

A) *Coupled Noise Design.* Independent perturbations of objectives and constraints are difficult to analyze and yield undesirable exponential factors ($e^{\Omega(m)}$). We instead *couple* the perturbations by adding a single random vector ψ to the objective estimate and $-\psi$ to each row of the constraint estimate. This coupling ensures a high *local optimism rate*: with constant probability, the perturbed program is feasible at the true optimum a_*, achieving regret bounds scaling only with $\log(m)$. Empirical studies (§6,J) confirm the advantages of coupled noise.

B) *Scaling and Resampling.* The fluctuating constraints disable both existing analysis frameworks for linear TS: the 'unsaturation' approach of [AG13] and the 'optimism' approach of [AL17]. To analyze S-COLTS, we adapt the unsaturation framework with a new scaling-based trick allowing comparisons across distinct feasible regions. For R-COLTS, we instead use resampling to directly generate optimistic and feasible actions, bypassing these analytic barriers entirely.

1.1 Related Work

Safe Bandits. Safe bandits have been studied under two main notions of constraint enforcement: *soft* [CGS22; GCS24] and *hard* [AAT19; MAAT21; PGBJ21; PGB24; HTA23; HTA24]. Soft enforcement achieves regret and risk bounds of $\widetilde{O}(\sqrt{d^2T})$, with improved instance-specific guarantees for polytopal domains. Hard enforcement achieves zero risk, and regret bounds of $\widetilde{O}(\mathcal{R}(a_{\mathsf{safe}})\sqrt{d^2T})$

but given a safe action a_{safe}. Efficient variants of these methods instead achieve weaker regret bounds of $\tilde{O}(\mathcal{R}(a_{\mathsf{safe}})\sqrt{d^3T})$. In contrast to safe bandits, *bandits with knapsacks* [BKS13; AD16] control aggregate constraints, which is unsuitable for roundwise safety enforcement (see §C).

Computational Complexity. Existing efficient hard-enforcement methods rely on frequentists confidence sets for constraints, which induce m expensive second-order conic (SOC) constraints during action selection [PGBJ21; PGB24; AAT19; MAAT21]. Most variants require solving $2d$ such problems per round, and further suffer from poor numerical conditioning. Our approach, S-COLTS, instead uses perturbations combined with scaling and resampling techniques, requiring only linear constraints per round while maintaining near-optimal guarantees. This scaling approach is related to ROFUL [HTA24] although this prior method uses the inefficient method DOSS as a subroutine.

Notably, no computationally efficient methods have previously been proposed for soft enforcement. The main point of comparison, DOSS need to solve $(2d)^{m+1}$ LPs each round [GCS24]. R-COLTS resolves this gap by sampling $O(\log(t))$ perturbed programs each round. Under mild conditions (Slater's condition), one can further reduce to a single LP per round. See §C for more details.

Thompson Sampling (TS). Frequentist bounds for linear TS were first established by Agrawal & Goyal [AG13] through an 'unsaturation' approach, while Abeille & Lazaric [AL17] developed a related 'global optimism' approach. Neither approach extends to SLBs due the per-round fluctuation of the perturbed constraints, and the ensuing variability of the 'feasible regions' for each round (see §C for more details). We overcome these challenges through our coupled noise design, ensuring frequent optimism, and a novel scaling trick to compare solutions across distinct feasible regions.

The only existing sampling-based treatment of unknown constraints is due to Chen et al. [CGS22] for multi-armed settings, who use posterior quantiles to enforce constraints. Although their method does not scale to continuous action sets, our resampling approach can be interpreted as an efficient, scalable analogue for simultaneously enforcing constraints and optimizing reward indices.

2 Problem Definition and Background

Notation. For a vector v, $\|v\|$ denotes its ℓ_2-norm. For a PSD matrix M, $\|v\|_M := \|M^{1/2}v\|$. $\mathbb{S}^d$ is the unit sphere in $\mathbb{R}^d$. For a matrix M, M^i is the ith row of M. $\mathbf{1}_m$ is the all ones vector in $\mathbb{R}^m$. Also see §A for an extensive glossary of notation used in the paper.

Setup. An instance of a SLB problem is defined by an objective $\theta_* \in \mathbb{R}^d$, a constraint matrix $\Phi_* \in \mathbb{R}^{m\times d}$, constraint levels $\alpha \in \mathbb{R}^m$, a compact *convex* domain $\mathcal{A} \subset \mathbb{R}^d$, and $\delta \in (0,1)$. $\mathcal{A}, \alpha, \delta$ are known to the learner, but θ_* and Φ_* are not. The program of interest is $\max \theta_*^\top a$ s.t. $\Phi_* a \le \alpha, a \in \mathcal{A}$, assumed to be feasible. a_* denotes a(ny) maximiser of this program. The *reward gap* of $a \in \mathcal{A}$ is $\Delta(a) := \theta_*^\top(a_* - a)$, and its *safety margin* is $\Gamma(a) = \min_u (\alpha - \Phi_* a)_+^i$. For infeasible a, $\Gamma(a) = 0$, and Δ may be negative. We set $\mathcal{R}(a) = 1 + {}^{\Delta(a)}/_{\Gamma(a)}$ if $\Gamma(a) > 0$, and ∞ otherwise.

Play. We index rounds by t. At each t, the learner picks $a_t \in \mathcal{A}$, and receives the feedback $R_t = \theta_*^\top a_t + w_t^R$, and $S_t = \Phi_* a_t + w_t^S$, where $w_t^R \in \mathbb{R}$ and $w_t^S \in \mathbb{R}^m$ are noise processes. C_t denotes algorithmic randomness at round t. The historical filtration is $\mathfrak{H}_{t-1} := \sigma(\{(a_s, R_s, S_s, C_s)\}_{s<t})$, and $\mathfrak{G}_t := \sigma(\mathfrak{H}_{t-1} \cup \{(a_t, C_t)\})$. The action a_t must be adapted to $\sigma(\mathfrak{H}_{t-1} \cup \sigma(\{C_t\}))$.

The Soft Enforcement SLB problem demands algorithms that ensure, with high probability, that both the metrics $\mathbf{R}_T$ and $\mathbf{S}_T$ (see (1) grow sublinearly with T.

The Hard Enforcement SLB problem demands algorithms that ensure, with high probability, that $\mathbf{S}_T = 0$ and $\mathbf{R}_T = o(T)$. This is enabled by a safe starting point a_{safe} such that $\Gamma(a_{\mathsf{safe}}) > 0$.

Standard Assumptions. We assume the following standard conditions [e.g. APS11] on the instance $(\theta_*, \Phi_*, \mathcal{A})$, and noise. All subsequent results only hold under these assumptions.

- *Boundedness*: $\|\theta_*\| \le 1$, for each row i, $\|\Phi_*^i\| \le 1$, and $\mathcal{A} \subset \{a : \|a\| \le 1\}$.
- *SubGaussian noise*: $w_t := (w_t^R, (w_t^S)^\top)^\top$ is centred and 1-subGaussian given $\mathfrak{G}_t$, i.e., $\mathbb{E}[w_t|\mathfrak{G}_t] = 0$, and $\forall \lambda \in \mathbb{R}^{m+1}, \mathbb{E}[\exp(\lambda^\top w_t)|\mathfrak{G}_t] \le \exp(\|\lambda^2\|/2)$.

To simplify the form of our bounds, we also assume that $m/\delta = O(\mathrm{poly}(d))$ when stating theorems.

Background. The (1-)RLS estimates for θ_*, Φ_* given the history $\mathfrak{H}_{t-1}$ are

$$\hat{\theta}_t = \arg\min_{\hat{\theta}} \sum_{s<t} (\hat{\theta}^\top a_s - R_s)^2 + \|\hat{\theta}\|^2, \textit{ and } \hat{\Phi}_t = \arg\min_{\hat{\Phi}} \sum_{s<t} \|\hat{\Phi} a_s - S_s\|^2 + \sum_i \|\hat{\Phi}^i\|^2.$$

The standard *confidence sets* [APS11] for (θ_*, Φ_*) are

$$\mathcal{C}_t^\theta(\delta) = \{\widetilde{\theta} : \|\widetilde{\theta} - \hat{\theta}_t\|_{V_t} \le \omega_t(\delta)\}, \text{ and } \mathcal{C}_t^\Phi(\delta) = \{\widetilde{\Phi} : \forall \text{ rows } i, \|\widetilde{\Phi}^i - \hat{\Phi}_t^i\|_{V_t} \le \omega_t(\delta)\},$$

where $V_t := I + \sum_{s<t} a_s a_s^\top$, *and* $\omega_t(\delta) := 1 + \sqrt{1/2 \log((m+1)/\delta) + 1/4 \log(\det V_t)}$. A key standard result states that these confidence sets are *consistent* [APS11].

Lemma 1. *Let the* consistency event at time t *be* $\mathsf{Con}_t(\delta) := \{\theta_* \in \mathcal{C}_t^\theta(\delta), \Phi_* \in \mathcal{C}_t^\Phi\}$, *and let* $\mathsf{Con}(\delta) := \bigcap_{t \geq 1} \mathsf{Con}_t(\delta)$. *Under the standard assumptions, for all* $\delta \in (0,1)$, $\mathbb{P}(\mathsf{Con}(\delta)) \geq 1 - \delta$.

3 The Constrained Linear Thompson Sampling Approach

We begin by describing the COLTS framework. In the frequentist viewpoint, TS is a randomised method for bandits that, at each t, perturbs an estimate of the unknown objective, in a manner sensitive to the historical information $\mathfrak{H}_{t-1}$, and then picks actions by optimising this perturbed objective.

Naturally, then, we will perturb the estimates $\hat{\theta}_t, \hat{\Phi}_t$, for which we use a law μ on $\mathbb{R}^{1 \times d} \times \mathbb{R}^{m \times d}$. For $(\eta, H) \sim \mu$, independent of $\mathfrak{H}_{t-1}$, we define the perturbed parameters

$$\widetilde{\theta}(\eta, t)^\top := \hat{\theta}_t^\top + \omega_t(\delta) \eta V_t^{-1/2} \text{ and } \widetilde{\Phi}(H, t) := \hat{\Phi}_t + \omega_t H V_t^{-1/2}. \tag{2}$$

Notice that these perturbations are aligned with $\mathfrak{H}_{t-1}$ only via the scaling by $\omega_t(\delta) V_t^{-1/2}$. The underlying thesis of the COLTS approach is that for well-chosen μ, the action

$$a(\eta, H, t) = \arg\max\{\widetilde{\theta}(\eta, t)^\top a : \widetilde{\Phi}(H, t) a \leq \alpha, a \in \mathcal{A}\}, \tag{3}$$

if it exists, is a good choice to play, in that it is either underexplored, or nearly safe and optimal. Here we abuse notation, and treat $\arg\max$ as a point function that (measurably) picks any one optimal solution. Two major issues arise with this view. Firstly, the set $\mathcal{A} \cap \{\widetilde{\Phi}(H, t) a \leq \alpha\}$ may be empty for certain H, meaning $a(\eta, H, t)$ need not exist. Secondly, in hard enforcement, $a(\eta, H, t)$ need not actually be safe, and so cannot directly be used. Thus, the main questions are 1) what μ we should use, 2) how we should augment the COLTS principle to design effective algorithms, and 3) how we can analyse these algorithms to prove effectiveness. These questions occupy the rest of this paper.

Before proceeding, however, we observe that if η or H are very large, then they will 'wash out' the 'signal' in $\hat{\theta}_t$ and $\hat{\Phi}_t$, meaning that their size must be contained. We state this as a generic condition.

Definition 2. *Let* $B : (0,1] \to \mathbb{R}_{\geq 0}$ *be a nondecreasing map. A law* μ *on* $\mathbb{R}^{1 \times d} \times \mathbb{R}^{m \times d}$ *is said to satisfy* B*-concentration if* $\forall \xi \in (0,1], \mu\left(\left\{\max(\|\eta\|, \max_{i \in [1:m]} \|H^i\|) \geq B(\xi)\right\}\right) \leq \xi$.

As an example, if marginally each η, H^i were normal, then $B(\xi) = \sqrt{d \log((m+1)/\xi)}$. Henceforth, we will assume that μ satisfies B-concentration for some map B, and define quantities in terms of this B. This condition has the following useful consequence (§F).

Lemma 3. *For* $B : (0,1] \to \mathbb{R}_{\geq 0}$, *and* $t \in \mathbb{N}$, *define* $B_t = 1 + \max(1, B(\delta_t))$, *and* $M_t(a) = B_t \omega_t(\delta) \|a_t\|_{V_t^{-1}}$, *where* $\delta_t = \delta/t(t+1)$. *For all t, let* (η_t, H_t) *be drawn from* μ *independently of* $\mathfrak{H}_{t-1}$ *at time t. If* μ *satisfies* B*-concentration, then with probability at least* $1 - 2\delta$,

$$\forall t, a, \max\left(|(\theta_* - \widetilde{\theta}(\eta_t, t))^\top a|, \max_i |(\widetilde{\Phi}(H_t, t)^i - \Phi_*^i) a|\right) \leq M_t(a).$$

Further, $\sum_{t \leq T} M_t(a_t) \leq B_T \omega_T \cdot O(\sqrt{dT}) \leq B_T \widetilde{O}(\sqrt{d^2 T})$.

4 Hard Constraint Enforcement via Scaling-COLTS

We turn to the problem of hard constraint enforcement of minimising $\mathbf{R}_T$ while ensuring that w.h.p., $\mathbf{S}_T = 0$, using a safe action a_{safe} such that $\Gamma(a_{\mathsf{safe}}) > 0$. We will extend COLTS with a 'scaling heuristic,' that was first proposed in the context of SLBs by Hutchinson et al. [HTA24], who used it to design a (inefficent) method ROFUL.

Algorithm 1 Scaling-COLTS (S-COLTS(μ, δ))

```
Input: a_safe, Γ_0 ∈ [Γ(a_safe)/2, Γ(a_safe)].
for t = 1, 2, ... do
    Draw (η_t, H_t) ~ μ
    if M_t(a_safe) > Γ_0/3 OR a(η_t, H_t, t) does not exist then
        a_t ← a_safe.
    else
        b_t ← a(η_t, H_t, t)
        Compute a_t as in (4).
    Play a_t, observe R_t, S_t, update 𝔥_t.
```

To begin, our method, S-COLTS, draws noise $(\eta_t, H_t) \sim \mu$, and computes the preliminary action $b_t := a(\eta_t, H_t, t)$, assuming for now that this exists. As argued in §3, this action b_t either has low-regret, or is informative. Of course, this b_t need not be safe—we only know via Lemma 3 that $\Phi_* b_t \leq \alpha + M_t(b_t) \mathbf{1}_m$—and so cannot be used for hard enforcement. However, the action a_{safe} is safe, with a large slack of at least $\Gamma(a_{\mathsf{safe}})$ in each constraint. Via linearity, and the convexity of $\mathcal{A}$, this means we can *scale back* b_t towards a_{safe} to find a safe action, i.e., play a_t of the form $(1-\rho_t)a_{\mathsf{safe}} + \rho_t b_t$ for some $\rho_t \in [0,1]$. If ρ_t is not too small, this maintains fidelity with respect to the informative direction b_t, while retaining safety.

Note that if we knew that margin $\Gamma(a_{\mathsf{safe}})$ of a_{safe}, then since

$$\Phi_*(\rho b_t + (1-\rho)a_{\mathsf{safe}}) \le \alpha + (\rho M_t(b_t) - (1-\rho)\Gamma(a_{\mathsf{safe}}))\mathbf{1}_m,$$

we could ensure that $1-\rho_t \le M_t(b_t)/\Gamma(a_{\mathsf{safe}})$, which vanishes for small $M_t(b_t)$. Of course, we do not per se know the value of $\Gamma(a_{\mathsf{safe}})$. However, this can be estimated by repeatedly playing a_{safe}, and maintaining anytime bounds on its risk via a law of iterated logarithms. This results in an estimated value Γ_0 such that $\Gamma(a_{\mathsf{safe}})/2 \le \Gamma_0 \le \Gamma(a_{\mathsf{safe}})$ using $\widetilde{O}(\Gamma(a_{\mathsf{safe}})^{-2})$ rounds. We leave a detailed account of this estimation to §H.1, and henceforth just assume that we know such a value Γ_0.

Define $\widetilde{\theta}_t = \widetilde{\theta}(\eta_t, t)$ and $\widetilde{\Phi}_t = \widetilde{\Phi}(H_t, t)$. A critical observation for S-COLTS is that if $M_t(a_{\mathsf{safe}}) \le \Gamma_0/3$, then the action b_t exists with high probability. Indeed, in this case,

$$\widetilde{\Phi}_t a_{\mathsf{safe}} \le \Phi_* a_{\mathsf{safe}} + M_t(a_{\mathsf{safe}})\mathbf{1}_m \le \alpha - (\Gamma(a_{\mathsf{safe}}) - {\Gamma_0}/{3})\mathbf{1}_m \le \alpha - {2\Gamma(a_{\mathsf{safe}})}/{3}\mathbf{1}_m.$$

Thus, the constraints induced by $\widetilde{\Phi}_t$ are feasible (and so b_t exists). To play a safe action, we set

$$a_t = \mathfrak{a}_t(\rho_t), \textit{ where } \mathfrak{a}_t(\rho) := (1-\rho)a_{\mathsf{safe}} + \rho b_t, \textit{ and} \tag{4}$$
$$\rho_t := \max\{\rho \in [0,1] : \hat{\Phi}_t \mathfrak{a}_t(\rho) + B_t^{-1} M_t(\mathfrak{a}_t(\rho))\mathbf{1}_m \le \alpha\},$$

where $B_t^{-1}M_t(a) = \omega_t \|a\|_{V_t^{-1}}$ is used since $\max_i \|\hat{\Phi}_t^i - \Phi_*^i\| \le \omega_t$ whp. Now, $M_t(a_{\mathsf{safe}}) \le \Gamma_0/3$ ensures that $1-\rho_t \le 6M_t(b_t)/\Gamma(a_{\mathsf{safe}})$, giving roughly the same fidelity as if we knew $\Gamma(a_{\mathsf{safe}})$.

This leaves the law μ undetermined. In §4.1, we first describe an *unsaturation* condition on μ that induces low regret, and then in §4.2, we provide a general construction of unsaturated laws using a local analysis at a_*. This operationalises the S-COLTS design, with regret bounds described in §4.3.

4.1 Analysis of S-COLTS: Unsaturation and Looking-Back

The scaling above direclty ensures the safety of a_t. We now present the main ideas behind a $\widetilde{O}(\sqrt{T})$ regret bound for S-COLTS, leaving detailed proofs to §H. Specifically, we describe how a *look back* method operationalised with an *unsaturation event* yields low-regret, using a *scaling* strategy to handle shifting constraints. Finally, we finally contrast our analysis with prior studies.

Unsaturation. Following [AG13], we say that an action a is *unsaturated* at time t if $\Delta(a) \le M_t(a)$. Now, if b_t is unsaturated, then (as we asserted in §3) it is either informative (large $M_t(b_t)$) or low regret (small $M_t(b_t)$, and so small $\Delta(b_t)$). In fact, if for all t, b_t was unsaturated and $b_t = a_t$, then we would already get a regret bound using Lemma 3: $\sum_t \Delta(a_t) = \sum \Delta(b_t) \le \sum_t M_t(b_t) = \sum M_t(a_t) = \widetilde{O}(\sqrt{T})$. But, b_t need not always be unsaturated (and usually is $\neq a_t$), and we must ensure that enough unsaturated b_t occur, motivating the following definition.

Definition 4. *Let μ be a B-concentrated law. Define the* unsaturation event *at time t as*

$$\mathsf{U}_t(\delta) := \{(\eta, H) : a(\eta, H, t) \textit{ exists, and } \Delta(a(\eta,H,t)) \le M_t(a(\eta,H,t)).$$

For $\chi \in (0,1]$, we say that μ-satisfies χ-unsaturation if for all t such that $\delta/(t(t+1)) \le \chi/2$,

$$\mathbb{P}[\mathsf{U}_t(\delta)|\mathfrak{H}_{t-1}]\mathbb{1}_{\mathsf{Con}_t(\delta)} = \mathbb{E}[\mu(\mathsf{U}_t(\delta))|\mathfrak{H}_{t-1}]\mathbb{1}_{\mathsf{Con}_t(\delta)} \ge (\chi/2)\mathbb{1}_{\mathsf{Con}_t(\delta)}.$$

In words, at all t, given the past, b_t is unsaturated with chance at least $\chi/2$.

Handing the scaling. Our action $a_t = \rho_t b_t + (1-\rho_t)a_{\mathsf{safe}} \neq b_t$. But, by linearity,

$$\Delta(a_t) = \rho_t \Delta(b_t) + (1-\rho_t)\Delta(a_{\mathsf{safe}}).$$

To bound these, we use Lemma 19 (§H.2), which shows that if $M_t(a_{\mathsf{safe}}) \le \Gamma_0/3$ then (i) $(1-\rho_t) \le 6M_t(a_t)/\Gamma(a_{\mathsf{safe}})$, and (ii) $\rho_t M_t(b_t) \le 2M_t(a_t)$. Thus, by (i) and Lemma 3, $\sum(1-\rho_t)\Delta(a_{\mathsf{safe}}) = O({\Delta(a_{\mathsf{safe}})}/{\Gamma(a_{\mathsf{safe}})} \sum M_t(a_t)) = \widetilde{O}(\mathcal{R}(a_{\mathsf{safe}})\sqrt{d^3T})$. Notice that if b_t was always saturated, then via (ii), we would also get regret control using $\sum \rho_t \Delta(b_t) \le \sum \rho_t M_t(b_t) = O(\sum M_t(a_t))$.

Look Back Method. In reality, b_t is not always unsaturated. To handle this, we 'look back' at the *last time* $s < t$ that a b_s was unsaturated. Concretely, define

$$\tau(t) := \inf\{s < t : M_s(a_{\mathsf{safe}}) \le \Gamma_0/3, \Delta(b_s) \le M_s(b_s)\}, \quad \inf\emptyset := 0.$$

We will bound $\Delta(b_t)$ in terms of $\Delta(b_{\tau(t)})$: if unsaturation is frequent, then $\Delta(b_{\tau(t)}) \le M_{\tau(t)}(b_{\tau(t)})$ cannot be much larger than $M_t(b_t)$, yielding regret control. A major issue, however, is that $b_{\tau(t)}$ is not feasible for the perturbed constraints $\widetilde{\Phi}_t$. Thus inequalities like $\widetilde{\theta}_t^\top(b_{\tau(t)} - b_t) \le 0$ may be false, preventing us from using the basic fact that b_t optimises $\widetilde{\theta}_t$ over $\mathcal{A} \cap \{\widetilde{\Phi}_t a \le \alpha\}$. To address this, we introduce a *second, analysis-only, scaling step*. Concretely, for $s < t$, let

$$\bar{b}_{s\to t} := \sigma_{s\to t} b_s + (1 - \sigma_{s\to t}) a_{\mathsf{safe}}, \text{ where } \sigma_{s\to t} = (\Gamma_0/3) \cdot (\Gamma_0/3 + M_t(b_s) + M_s(b_s))^{-1}.$$

Then using Lemma 3 twice, $\widetilde{\Phi}_s b_s \le \alpha \implies \widetilde{\Phi}_t \bar{b}_{s\to t} \le \alpha$. Setting $s = \tau(t)$, we write

$$\rho_t \Delta(b_t) = \rho_t \Delta(\bar{b}_{\tau(t)\to t}) + \rho_t \widetilde{\theta}_t^\top(\bar{b}_{\tau(t)\to t} - b_t) + \rho_t(\theta_* - \widetilde{\theta}_t)^\top(\bar{b}_{\tau(t)\to t} - b_t).$$

Since $\bar{b}_{\tau(t)\to t}$ is feasible, and b_t optimal, for $(\widetilde{\theta}_t, \widetilde{\Phi}_t)$, the second term is ≤ 0. The final term is $\le \rho_t M_t(\bar{b}_{\tau(t)\to t}) + \rho_t M_t(b_t)$, and further, $\rho_t M_t(b_t) \le 2M_t(a_t)$. For the first term,

$$\begin{aligned}\Delta(\bar{b}_{\tau(t)\to t}) &= \sigma_{\tau(t)\to t}\Delta(b_{\tau(t)}) + (1 - \sigma_{\tau(t)\to t})\Delta(a_{\mathsf{safe}}) \\ &\le \sigma_{\tau(t)\to t} M_{\tau(t)}(b_{\tau(t)}) + (1 - \sigma_{\tau(t)\to t})\Delta(a_{\mathsf{safe}}),\end{aligned}$$

using the unsaturation of $b_{\tau(t)}$. This leaves terms involving $\sigma_{\tau(t)\to t}$ and $M_t(b_{\tau(t)})$, which are handled by analysing $\sigma_{\tau(t)\to t}$ similarly to ρ_t above. As shown in §H.2, this leads to

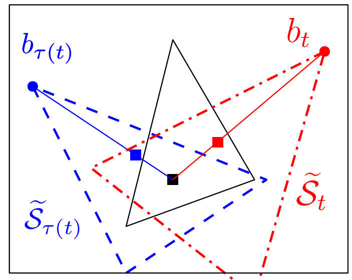

Figure 1: A schematic of the look-back analysis. $\tilde{\mathcal{S}}$ are the perturbed feasible regions. a_t (red box) is a mixture of b_t and a_{safe} (black box). $b_{\tau(t)} \notin \tilde{\mathcal{S}}_t$, and we instead mix it with a_{safe} to produce $\bar{b}_{\tau(t)\to t} \in \tilde{\mathcal{S}}_t$ (blue box).

Lemma 5. *Define $M_0(a) = \Delta(a_0) = B_0 = \omega_0(\delta) = 1$, and for $s < t$, set $J_{s\to t} = 1 + {}^{B_t\omega_t}/_{B_s\omega_s}$. If μ is B-concentrated, then with probability at least $1 - 3\delta$, for all t with $M_t(a_{\mathsf{safe}}) \le \Gamma_0/3$,*

$$\Delta(a_t) \le 6\mathcal{R}(a_{\mathsf{safe}})\left(M_t(a_t) + J_{\tau(t)\to t} M_{\tau(t)}(a_{\tau(t)})\right).$$

Controlling Accumulation Error. By Lemma 3, $\sum M_t(a_t) = \widetilde{O}(\sqrt{T})$. Thus, to bound regret, we are left with $\sum J_{\tau(t)\to t} M_{\tau(t)}(a_{\tau(t)})$. But, due to χ-unsaturation, the time between two unsaturation events is typically only $O(\chi^{-1})$, which enables a martingale analysis in §H.3 to give

Lemma 6. *If μ satisfies χ-unsaturation, then with probability at least $1 - \delta$, for all $T \ge \sqrt{2/\chi}$,*

$$\sum_{t \le T, M_t(a_{\mathsf{safe}}) \le \Gamma_0/3} J_{\tau(t)\to t} M_{\tau(t)}(a_{\tau(t)}) \le 10\omega_t B_T \cdot \chi^{-1} \cdot (\log(1/\delta) + \sum_{t \le T} \|a_t\|_{V_t^{-1}}).$$

Finally, by Lemma 3, $\sum \|a_t\|_{V_t^{-1}} = \widetilde{O}(\sqrt{T})$ as well, concluding the regret analysis.

Novelty Relative to Prior Work. In unconstrained linear TS, Agrawal and Goyal [AG13] compare a_t to a *minimum-norm unsaturated action* $\bar{a}_t$ (see Lemma 4 in their arxiv version [AG12]), using $\widetilde{\theta}_t^\top(\bar{a}_t - a_t) \le 0$ and $\Delta(\bar{a}_t) \le M_t(\bar{a}_t)$ to bound their $\Delta(a_t)$. Therein, χ-unsaturation is essentially used to show $M_t(\bar{a}_t) \le \chi^{-1} M_t(a_t)$. Our analysis sees three differences.

- *Scaled Actions*: $b_t \ne a_t$, which is managed by expanding the linear blend and analysing ρ_t.
- *Look-back Analysis*: instead of a reference action at time t, we look back to the last unsaturated $b_{\tau(t)}$. This, in our opinion, yields a more intuitive analysis, but is not an essential step.
- *Fluctuating Constraints*: The reference action $b_{\tau(t)}$ may be infeasible for $\widetilde{\Phi}_t$, breaking the comparison of values between it and b_t. This would also affect the equivalent of $\bar{a}_t$, and causes the prior analysis to break down. We handle this via a second analysis-only scaling step.

The technical gap from [HTA24] is similar: a_* may be infeasible for $\widetilde{\Phi}_t$, which breaks their analysis.

4.2 The Coupled Noise Design

§4.1 shows that χ-unsaturation yields control on the regret. To operationalise this, we need to design well-concentrated laws with good unsaturation. In single-objective TS, unsaturation is enabled via anticoncentration of the ηs, and a good balance is attained by, e.g., $\mathrm{Unif}(\sqrt{3d}\mathbb{S}^d)$ or $\mathscr{N}(0, I_d)$.

A natural guess with unknown constraints is to sample both η and each row of H from such a law. However, the unsaturation rate under such a design is difficult to control well. The main issue arises from maintaining feasibility with respect to all m constraints under perturbation, since each such perturbation gets an independent shot at shaving away some unsaturated actions, suggesting that χ decays as $e^{-\Omega(m)}$ and indeed, experimentally, increase in m may lead to at least a polynomial decay in the unsaturated rate with such independent noise (see §J.3). We sidestep this issue by *coupling* the perturbations of the reward and constraints, as encapsulated below.

Lemma 7. *Let $\bar{B} \in \{(0,1] \to \mathbb{R}_{\geq 0}\}$ be a map, and $p \in (0,1]$. Let ν be a law on $\mathbb{R}^{d\times 1}$ such that*

$$\forall u \in \mathbb{R}^d, \nu(\{\zeta : \zeta^\top u \geq \|u\|) \geq p, \text{ and } \forall \xi \in (0,1], \nu(\{\zeta : \|\zeta\| > \bar{B}(\xi)\}) \leq \xi.$$

Let μ be the law of $\zeta \mapsto (\zeta^\top, -\mathbf{1}_m \zeta^\top)$ for $\zeta \sim \nu$. Then μ is p-unsaturated and $\bar{B}$-concentrated.

Our proof of this lemma, executed in §G, is based upon analysing the *local optimism event* at a_*:

$$\mathsf{L}_t(\delta) := \{(\eta, H) : \widetilde{\theta}(\eta, t)^\top a_* \geq \theta_*^\top a_*, \widetilde{\Phi}(H, t) a_* \leq \alpha\}. \tag{5}$$

Notice that L_t demands that the perturbation is such that a_* remains feasible with respect to $\widetilde{\Phi}_t$, and its value at $\widetilde{\theta}_t$ increases beyond $\theta_*^\top a_*$, in other words, the perturbed program is optimistic *at* a_*. Our proof first directly analyses a_* under the perturbations to show that $\mathbb{P}[\mathsf{L}_t(\delta)|\mathfrak{H}_{t-1}]\mathbb{1}_{\mathsf{Con}_t(\delta)} \geq p\mathbb{1}_{\mathsf{Con}_t(\delta)}$, i.e., frequent local optimism. This enables an argument due to [AG13]: since a_* is unsaturated ($\Delta(a_*) = 0$), and, w.h.p. the perturbed reward of any saturated action is dominated by that of a_*, it follows that $\mathsf{L}_t(\delta) \subset \mathsf{U}_t(\delta)$, yielding lower bounds on $\mu(\mathsf{U}_t(\delta))$.

We note that the conditions of Lemma 7 are the same as those used for unconstrained linear TS in prior work [AG13; AL17], and so this generic result extends this unconstrained guarantee to the constrained setting. In our bounds, we will set μ to be the law induced by the coupled design with $\nu = \mathrm{Unif}(\sqrt{3d}\mathbb{S}^d)$, which is 0.14-unsaturated, and B-concentrated for $B(\xi) = \sqrt{3d}$ (§G.1).

4.3 Regret Bounds for S-COLTS

With the pieces in place, we state and discuss our main result, which is formally proved in §H.

Theorem 8. *Let μ be the law induced by $\mathrm{Unif}(\sqrt{3d}\mathbb{S}^d)$ under the coupled noise design. Then* S-COLTS$(\mu, \delta/3)$ *ensures that with probability at least $1-\delta$, for all T, it holds that*

$$\mathbf{S}_T = 0 \quad \text{and} \quad \mathbf{R}_T = \mathcal{R}(a_{\mathsf{safe}}) \cdot \widetilde{O}(\sqrt{d^3 T + d^2 T \log(m/\delta)}) + \widetilde{O}(d^2 \Delta(a_{\mathsf{safe}}) \Gamma(a_{\mathsf{safe}})^{-2}).$$

Comparison of Regret Bounds to Prior Results. As noted in §1.1, prior inefficient hard enforcement SLB methods attain regret $\widetilde{O}(\mathcal{R}(a_{\mathsf{safe}})\sqrt{d^2T})$, while efficient methods attain regret $\widetilde{O}(\mathcal{R}(a_{\mathsf{safe}})\sqrt{d^3T})$. Our results above recover the latter bounds. The loss of $\sqrt{d}$ relative to inefficient methods is expected since it appears in all known efficient linear bandit methods (without or without unknown constraints). The $\Omega(\sqrt{T})$ dependence is necessary (even with instance-specific information) [GCS24] as is the additive $\Delta(a_{\mathsf{safe}})/\Gamma(a_{\mathsf{safe}})^2$ term [PGBJ21]. Thus, S-COLTS recovers previously known guarantees using sampling rather than frequentist bounds.

Computational Aspects. An advantage of S-COLTS is that it only optimises over linear constraints (beyond those of $\mathcal{A}$), instead of SOC constraints of the form $\{\forall i \in [1:m], \hat{\Phi}_t^i a + \omega_t(\delta)\|a\|_{V_t^{-1}} \leq \alpha^i\}$ imposed by prior methods. While convex, these m SOC constraints can have a palpable practical slowdown on the time needed for optimisation, especially as m grows (over $\mathcal{A} = [0, 1/\sqrt{d}]^d$, with the modest $d = m = 9$ we see a $> 5\times$ speedup, and with $d = 2, m = 100$, a $18\times$ speedup, in §6). In particular, when $\mathcal{A}$ is a polyhedron, S-COLTS can be implemented with just linear programming.

We explicitly note that S-COLTS is efficient for convex $\mathcal{A}$. The dominating step is the computation of b_t, which can be carried out to an approximation of $1/t$ with no loss in Theorem 8. With, say, interior point methods, this needs $O(\mathsf{LP} \cdot \log(t))$ computation at round t, where LP is the computation needed to optimise $\max\{\theta^\top a : \Phi a \leq \alpha, a \in \mathcal{A}\}$ to constant error [BV04].

Practical Choice of Noise. It has long been understood that while existing theoretical techniques for analysing linear TS need large noise (with $B(\xi) = \Theta(\sqrt{d})$), in practice much smaller noise (e.g., $\mathrm{Unif}(\mathbb{S}^d)$ with $B(\xi) = \Theta(1)$) typically retains a large enough rate of unsaturation, and significantly improve regret (although not in the worst-case [HB20]). Our practical recommendation is to indeed use such a small noise, which we find to be effective in simulations (§6). We underscore that no matter the noise used, the risk guarantee for S-COLTS is maintained.

5 Soft Constraint Enforcement with Resampling-COLTS

Given an action a_{safe} with positive safety margin, S-COLTS ensures strong safety and good regret. This section studies scenarios where we do not know such an a_{safe}. In this case, it is impossible to ensure that $\mathbf{S}_T = 0$, and we instead show $\widetilde{O}(\sqrt{T})$ bounds on $\mathbf{S}_T$, following prior work [GCS24].

S-COLTS uses forced exploration of a_{safe} to ensure the feasibility of perturbed programs. However, the local optimism underlying our proof of Lemma 7 gives a different way to achieve this. Indeed, the event $\mathsf{L}_t(\delta)$ of (5) implies that a_* is feasible, and so $a(\eta, H, t)$ exists. Thus, if $\mathbb{P}[\mathsf{L}_t(\delta)|\mathfrak{H}_{t-1}] \geq \pi$, then we can just resample the noise $O(\log(t))$ times and end up with feasibility. In fact, even more is true: since $\widetilde{\theta}^\top a_* \geq \theta_*^\top a_*$ under L_t, resampling $\pi^{-1}\Theta(\log(t))$ times ensures not only feasibility, but also *optimism* of the 'best' perturbed optimum. The R-COLTS method is based on this observation.

Concretely, we parametrise R-COLTS with μ, δ and a resampling order r. At each time t, we sample $I_t = 1 + \lceil r \log{}^{t(t+1)}/\delta \rceil$ independent (η, H) from μ, optimise each perturbed program, and pick the optimiser of the one with largest value as a_t. If all are infeasible, we just set $a_t = a_{t-1}$ (picking a_0 arbitrarily). We let $\widetilde{\theta}_t$ denote the objective vector of this 'winning' perturbed program: in the notation of Alg. 2, $\widetilde{\theta}_t = \widetilde{\theta}(\eta_{i_{*,t}}, t)$. The main idea is captured in the following simple lemma.

Lemma 9. *Let $\pi \in (0, 1]$, and suppose μ satisfies $\mathbb{1}_{\mathsf{Con}_t(\delta)}\mathbb{E}[\mu(\mathsf{L}_t(\delta))|\mathfrak{H}_{t-1}] \geq \pi\mathbb{1}_{\mathsf{Con}_t(\delta)}$ for every t. If $r \geq \pi^{-1}$, then with probability at least $1 - 2\delta$, at all t, the actions a_t and perturbed objective $\widetilde{\theta}_t$ selected by R-COLTS(μ, r, δ) are optimistic, i.e., they satisfy that $\theta_*^\top a_* \leq \widetilde{\theta}_t^\top a_t$.*

The 'local optimism condition' on μ above is reminiscent of the global optimism condition of Abeille & Lazaric [AL17], and indeed the same result holds under a global optimism assumption with unknown constraints. However, the analysis in this prior work does not extend to unknown constraints due to its reliance of convexity (§1.1), and resampling bypasses this issue. See §D for more details.

Lemma 9 enables the use of standard optimism based regret analyses [e.g. APS11]. By operationalising the condition on μ via the coupled design in §4.2, we show

Theorem 10. *If μ is the law induced by $\mathrm{Unif}(\sqrt{3d}\mathbb{S}^d)$ under the coupled design of Lemma 7, then with probability at least $1 - \delta$, R-COLTS$(\mu, 4, \delta/2)$ ensures that for all T,*

$$\max(\mathbf{S}_T, \mathbf{R}_T) = \widetilde{O}(\sqrt{d^3T + d^2T\log(m/\delta)}).$$

Algorithm 2 Resampling-COLTS (R-COLTS(μ, r, δ))

Input: μ, δ, 'resampling order' $r \in \mathbb{N}$
Initialise: $I_t \leftarrow 1 + \lceil r \log{}^{t(t+1)}/\delta \rceil$
for $t = 1, 2, \dots$ **do**
  for $i = 1, 2, \dots, I_t$ **do**
    Draw $(\eta_{i,t}, H_{i,t}) \sim \mu$.
    if $a(\eta_{i,t}, H_{i,t}, t)$ exists **then**
      $K(i,t) \leftarrow \widetilde{\theta}(\eta_{i,t}, t)^\top a(\eta_{i,t}, H_{i,t}, t)$
    else
      $K(i,t) \leftarrow -\infty$
  if $\max K(i,t) = -\infty$ **then**
    $a_t \leftarrow a_{t-1}$.
  else
    $i_{*,t} \leftarrow \arg\max_i K(i,t)$,
    $a_t \leftarrow a(\eta_{i_{*,t},t}, H_{i_{*,t},t}, t)$.
    $\widetilde{\theta}_t \leftarrow \widetilde{\theta}(\eta_{i_{*,t},t}, t)$.
  Play a_t, observe R_t, S_t, update $\mathfrak{H}_t$.

Instance-Independent Regret Bound. The above result limits both regret and risk to $\widetilde{O}(\sqrt{d^3T})$, with no instance-specific terms, unlike $\mathcal{R}(a_{\mathsf{safe}})$ in S-COLTS. In particular, this bound holds even if $\max_a \Gamma(a) = 0$, i.e., the problem is marginally feasible. This result is directly comparable to the $\widetilde{O}(\sqrt{d^2T})$ bound on both regret and risk under the DOSS method [GCS24], and loses a $\sqrt{d}$-factor relative to this, a loss that appears in all known efficient linear bandit methods.

Computational Costs. R-COLTS with μ as above solves $\sim 4\log(t^2/\delta)$ optimisations of $\widetilde{\theta}_t^\top a$ over $\{\widetilde{\Phi}_t a \leq \alpha\} \cap \mathcal{A}$. Again, Theorem 10 is resilient to approximation of, say, $1/t$, and so this takes $O(\mathsf{LP} \cdot \log^2 t)$ computation per round, a factor of $\log(t)$ slower than S-COLTS, but still efficient in the practical regime of $\log(T/\delta) = O(\mathrm{poly}(d, m))$. The main point of comparison, however, is DOSS, which instead needs to solve $(2d)^{m+1}$ such programs, and so uses $(2d)^{m+1}\mathsf{LP} \cdot \log(t)$ computation per round. R-COLTS is practically *much faster* even for small domains with long horizons—for instance, with $T = 1/\delta = 10^{10}, 4\log(t^2/\delta) \leq (2d)^{m+1}$ for all $d \geq 4, m \geq 2$.

Relationship to Posterior Quantile Indices and Safe MABs. The resampling approach executed in R-COLTS is closely related to the posterior-quantile approach of the BAYESUCB method [KCG12], wherein it is proposed to use a quantile of the arm posteriors as a reward index instead of a frequentist upper confidence bound. Indeed, we can compute such a quantile in a randomised way by taking many samples from the posterior of each arm, and then picking the largest of the samples as the reward index. Most pertinently, this approach was proposed for safe multi-armed bandits [CGS22], wherein this posterior quantile index is used to decide on the 'plausible safety' of putative actions. The same work further argued that the usual *single-sample* TS cannot obtain sublinear regret in safe MABs. The R-COLTS approach can be viewed as an efficient extension of this principle to linear bandits with continuum actions, and differs by directly optimising the indices under each draw, and then picking the largest, instead of performing an untenable per-arm posterior quantile computation.

R-COLTS Without Resampling. Given the lack of a safe action to play, one cannot direct establish the feasibility of the perturbed programs by contracting the confidence radius of a single action as in S-COLTS. However, if we introduce a small amount of 'flat' exploration whenever V_t is 'small', then this ensures that any a with $\Gamma(a) > 0$ will eventually be strictly feasible under perturbations. If such a exists, we only need a single noise draw to attain feasibility, and can bootstrap the scaling analysis of S-COLTS to show bounds. We term this method 'exploratory-COLTS', or E-COLTS, and specify and analyse it in §I.2. This results in the following soft-enforcement guarantee.

Theorem 11. *If μ is the law induced by $\mathrm{Unif}(\sqrt{3d}\mathbb{S}^d)$ under the coupled noise design, then the* E-COLTS$(\mu, \delta/3)$ *method of Algorithm 3 ensures that with probability at least $1-\delta$, for all T,*

$$\mathbf{S}_T = \widetilde{O}(\sqrt{d^3T}) + \min_a \widetilde{O}\Big(\frac{d^3\|a\|^4}{\kappa^2\Gamma(a)^4}\Big), \text{ and } \mathbf{R}_T = \min_{a:\Gamma(a)>0}\Big\{\mathcal{R}(a)\widetilde{O}(\sqrt{d^3T}) + \widetilde{O}\Big(\frac{d^3\|a\|^4}{\kappa^2\Gamma(a)^4}\Big)\Big\},$$

where κ is a constant depending on the geometry of $\mathcal{A}$.

Relative to R-COLTS, the above guarantees are instance-dependent, and are only nontrivial if $\max_a \Gamma(a) > 0$, i.e., the Slater parameter of the optimisation problem induced by $\theta_*, \Phi_*, \mathcal{A}$ is nonzero. The advantage of E-COLTS lies in its reduced computation. Comparing to S-COLTS, the above loses the strong $\mathbf{S}_T = 0$ safety, but improves regret by adapting to the best possible $\mathcal{R}(a)$.

6 Simulations

We give a brief summary of our simulations, leaving most details, and well as deeper investigation of our methods to §J. In all cases, we utilise the coupled noise design, driven with the (uninflated) noise $\nu = \mathrm{Unif}(0.5 \cdot \mathbb{S}^d)$, in accordance with the discussion in §4. The same noise is used for SAFE-LTS.

Resampling tradeoff in R-COLTS. For $d = 9$, we optimise $\theta_* = \mathbf{1}_d/\sqrt{d}$ over $\mathcal{A} = [0, 1/\sqrt{d}]^d$, with a 9×9 constraint matrix (i.e., $m = 9$). In this case, the action 0 is feasible, and so R-COLTS without any resampling is effective. Since $a = 0$ has a nontrivial safety margin, R-COLTS, even without resampling, is effective for this problem. This is borne out in Table 2, which shows regret and risk at the terminal time T. We see that resampling slightly worsens risk, but significantly improves regret (although with diminishing returns). Further, both regret and risk are far below the $\sqrt{d^2T}$ scale expected from our bounds. We note that while a single iteration of R-COLTS takes ~ 1ms, since $(2d)^{m+1} > 10^{12}$, this would take *years* for DOSS, and so we do not implement it. In any case, note that the computational advantage of R-COLTS is extremely strong.

Table 2: $\mathbf{R}_T$ and $\mathbf{S}_T$ at $T = 5\cdot 10^4$ for R-COLTS with $1, 2, 3$ samples per round (100 trials).

Samples	$\mathbf{R}_T$	$\mathbf{S}_T$
1	658 ± 170	2891 ± 171
2	397 ± 116	3126 ± 137
3	301 ± 102	3266 ± 172

Significant Computational Advantage and Regret Parity/Improvement of S-COLTS. We compare S-COLTS with the hard enforcement method SAFE-LTS [MAAT21], which has been shown to match the performance of alternate such methods, while being faster. Both methods are run on the $d = m = 9$ instance above, with $a_{\mathsf{safe}} = 0$. As expected, both never play unsafe actions. Further (Fig. 2, left), S-COLTS achieves an *improvement* in regret relative to SAFE-LTS, while reducing wall-clock time by a $5.1\times$. To gain a deeper understanding of S-COLTS's computational advantage, we investigate the same with growing $m \in \{1, 10, 20, \dots, 100\}$ constraints for a simple $d = 2$ setting (see §J.2.1 for the setup). In this problem, the benefit is even starker (Fig. 2, right). For $m \geq 10$, the regret of SAFE-LTS is $2-4\times$ larger than that of S-COLTS, i.e., the latter has much better regret ($m = 1$ has wide confidence bands for the ratio, but mean ~ 1.5) Further, the computational costs of SAFE-LTS relative to S-COLTS grow roughly linearly, starting from $\approx 1.3\times$ for $m = 1$ to $> 18\times$ at $m = 100$.

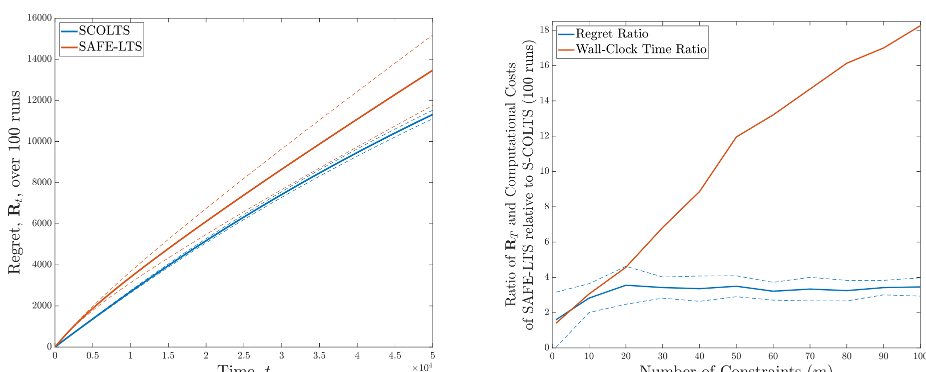

Figure 2: COMPUTATIONAL AND REGRET COMPARISONS OF S-COLTS AND SAFE-LTS. *Left.* Regret traces in the $d = 9$ instance (one-sigma error bars); S-COLTS mildly improves regret, and is $5\times$ faster. *Right.* Relative performance as m is varied in the $d = 2$ instance. The speedup of S-COLTS grows linearly with m from $1.3\times$ to $> 18\times$. Further, for $m \geq 10$, the regret of S-COLTS is 2-3$\times$ smaller than that of SAFE-LTS

References

[AAT19] Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. "Linear stochastic bandits under safety constraints". In: *arXiv preprint arXiv:1908.05814* (2019) (cit. on pp. 1–3, 13).

[AD14] Shipra Agrawal and Nikhil R Devanur. "Bandits with concave rewards and convex knapsacks". In: *Proceedings of the fifteenth ACM conference on Economics and computation*. 2014, pp. 989–1006 (cit. on p. 13).

[AD16] Shipra Agrawal and Nikhil Devanur. "Linear contextual bandits with knapsacks". In: *Advances in Neural Information Processing Systems* 29 (2016), pp. 3450–3458 (cit. on p. 3).

[AG12] Shipra Agrawal and Navin Goyal. "Thompson Sampling for Contextual Bandits with Linear Payoffs". In: *arXiv preprint arXiv:1209.3352 version 4* (2012). *Note: This version of the preprint, dated to 2014, improves upon the version of the paper published at ICML 2013, showing* $\sqrt{d^3 T}$ *regret instead of* $\sqrt{d^4 T}$. *Most of the text cites the ICML version, but attributes this stronger bound to it. This version is cited only to refer to specific details in this paper.* (cit. on p. 6).

[AG13] Shipra Agrawal and Navin Goyal. "Thompson sampling for contextual bandits with linear payoffs". In: *International conference on machine learning*. PMLR. 2013, pp. 127–135 (cit. on pp. 2, 3, 5–7, 14–16).

[AK08] Baruch Awerbuch and Robert Kleinberg. "Online linear optimization and adaptive routing". In: *Journal of Computer and System Sciences* 74.1 (2008), pp. 97–114 (cit. on p. 30).

[AKR21] Maryam Aziz, Emilie Kaufmann, and Marie-Karelle Riviere. "On multi-armed bandit designs for dose-finding trials". In: *Journal of Machine Learning Research* 22.14 (2021), pp. 1–38 (cit. on pp. 1, 13).

[AL17] Marc Abeille and Alessandro Lazaric. "Linear Thompson sampling revisited". In: *Electronic Journal of Statistics* 11.2 (2017), pp. 5165–5197. DOI: 10.1214/17-EJS1341SI. URL: https://doi.org/10.1214/17-EJS1341SI (cit. on pp. 2, 3, 7, 8, 14, 15, 38).

[AML24] Amirhossein Afsharrad, Ahmadreza Moradipari, and Sanjay Lall. "Convex methods for constrained linear bandits". In: *2024 European Control Conference (ECC)*. IEEE. 2024, pp. 2111–2118 (cit. on p. 13).

[APS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. "Improved algorithms for linear stochastic bandits". In: *Advances in neural information processing systems* 24 (2011), pp. 2312–2320 (cit. on pp. 3, 4, 8, 17).

[BGCG23] Debangshu Banerjee, Avishek Ghosh, Sayak Ray Chowdhury, and Aditya Gopalan. "Exploration in linear bandits with rich action sets and its implications for inference". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 8233–8262 (cit. on p. 31).

[BKS13] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. "Bandits with knapsacks". In: *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE. 2013, pp. 207–216 (cit. on pp. 3, 13).

[Bot+22] Alessandro Bottero, Carlos Luis, Julia Vinogradska, Felix Berkenkamp, and Jan R Peters. "Information-theoretic safe exploration with Gaussian processes". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 30707–30719 (cit. on p. 13).

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004 (cit. on p. 7).

[Cam+22] Romain Camilleri, Andrew Wagenmaker, Jamie Morgenstern, Lalit Jain, and Kevin Jamieson. "Active Learning with Safety Constraints". In: *arXiv preprint arXiv:2206.11183* (2022) (cit. on p. 13).

[CGS22] Tianrui Chen, Aditya Gangrade, and Venkatesh Saligrama. "Strategies for safe multi-armed bandits with logarithmic regret and risk". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 3123–3148 (cit. on pp. 2, 3, 8).

[Cho+24] Alex Chohlas-Wood, Madison Coots, Henry Zhu, Emma Brunskill, and Sharad Goel. "Learning to Be Fair: A Consequentialist Approach to Equitable Decision Making". In: *Management Science* (2024) (cit. on pp. 1, 13).

[CVA20] Alexandra Carpentier, Claire Vernade, and Yasin Abbasi-Yadkori. "The elliptical potential lemma revisited". In: *arXiv preprint arXiv:2010.10182* (2020) (cit. on p. 17).

[DHK08] Varsha Dani, Thomas P Hayes, and Sham M Kakade. "Stochastic linear optimization under bandit feedback". In: *Conference on Learning Theory*. 2008 (cit. on pp. 13, 30).

[FLZY22] Liang Feng, Wei Liu, Zhenyu Zhang, and Haoyang Yu. "Stability-Constrained Reinforcement Learning for Real-Time Voltage Control". In: *arXiv preprint arXiv:2203.12345* (2022). URL: https://arxiv.org/abs/2203.12345 (cit. on pp. 1, 13).

[GCS24] Aditya Gangrade, Tianrui Chen, and Venkatesh Saligrama. "Safe Linear Bandits over Unknown Polytopes". In: *The Thirty Seventh Annual Conference on Learning Theory*. PMLR. 2024, pp. 1755–1795 (cit. on pp. 1–3, 7, 8, 14, 31, 35).

[GGSS24] Aditya Gangrade, Aditya Gopalan, Venkatesh Saligrama, and Clayton Scott. "Testing the Feasibility of Linear Programs with Bandit Feedback". In: *Proceedings of the 41st International Conference on Machine Learning*. 2024 (cit. on p. 14).

[HB20] Nima Hamidi and Mohsen Bayati. "On worst-case regret of linear thompson sampling". In: *arXiv preprint arXiv:2006.06790* 472 (2020) (cit. on p. 7).

[HRMS21] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. "Time-uniform, nonparametric, nonasymptotic confidence sequences". In: *The Annals of Statistics* 49.2 (2021) (cit. on p. 19).

[HTA23] 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*. PMLR. 2023, pp. 497–508 (cit. on p. 2).

[HTA24] Spencer Hutchinson, Berkay Turan, and Mahnoosh Alizadeh. "Directional optimism for safe linear bandits". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2024, pp. 658–666 (cit. on pp. 2–4, 6, 13).

[KCG12] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. "On Bayesian upper confidence bounds for bandit problems". In: *Artificial intelligence and statistics*. PMLR. 2012, pp. 592–600 (cit. on p. 8).

[KS19] Julian Katz-Samuels and Clayton Scott. "Top feasible arm identification". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 1593–1601 (cit. on p. 13).

[LM00] Beatrice Laurent and Pascal Massart. "Adaptive estimation of a quadratic functional by model selection". In: *Annals of statistics* (2000), pp. 1302–1338 (cit. on p. 19).

[MAAT21] Ahmadreza Moradipari, Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. "Safe linear thompson sampling with side information". In: *IEEE Transactions on Signal Processing* (2021) (cit. on pp. 1–3, 9, 13, 35).

[PGB24] Aldo Pacchiano, Mohammad Ghavamzadeh, and Peter Bartlett. "Contextual Bandits with Stage-wise Constraints". In: *arXiv preprint arXiv:2401.08016* (2024) (cit. on pp. 1–3, 13, 16).

[PGBJ21] Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. "Stochastic bandits with linear constraints". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 2827–2835 (cit. on pp. 2, 3, 7, 13, 31).

[SGBK15] Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. "Safe exploration for optimization with Gaussian processes". In: *International conference on machine learning*. PMLR. 2015, pp. 997–1005 (cit. on p. 13).

A Glossary

Symbol	Explanation	Expression/Comments
(θ_*, Φ_*)	True objective/constraints	$\in \mathbb{R}^{d\times 1} \times \mathbb{R}^{m\times d}$
α	Constraint level	$\in \mathbb{R}^{m\times 1}$
$\mathcal{A}$	Action domain	
a_*	Optimal action for (θ_*, Φ_*)	$\arg\max\{\theta_*^\top a : a \in \mathcal{A}, \Phi_* a \le \alpha\}$
$K(\theta, \Phi)$	Value function	$\sup\{\theta^\top a : a \in \mathcal{A}, \Phi a \le \alpha\}$, $-\infty$ if $\{\Phi a \le \alpha\} \cap \mathcal{A} = \varnothing$.
$\Delta(a)$	Reward gap	$\theta_*^\top (a_* - a)$
$\Gamma(a)$	Safety margin	$\min_i((\alpha - \Phi_* a)^i)_+$
$\mathcal{R}(a)$	Gap-margin ratio	$1 + (\Delta(a)/\Gamma(a))$
	Estimation and Signal	
$\mathfrak{H}_{t-1}$	Historical filtration	See §2
$\hat\theta_t, \hat\Phi_t$	RLS-estimates of parameters	See §2
V_t	Action second moment	$I + \sum_{s<t} a_s a_s^\top$
$\omega_t(\delta)$	Confidence radius	See §2
$\mathcal{C}_t^\theta, \mathcal{C}_t^\Phi$	Confidence sets for θ_*, Φ_*	
$\mathsf{Con}_t(\delta)$	Consistency event at time t	$\{\theta_* \in \mathcal{C}_t^\theta(\delta), \Phi_* \in \mathcal{C}_t^\Phi(\delta)\}$
$\mathsf{Con}(\delta)$	Overall consistency	$\bigcap_{t\ge 1} \mathsf{Con}_t(\delta)$
	COLTS in general	
μ	Perturbation law	Distribution on $\mathbb{R}^{1\times d} \times \mathbb{R}^{m\times d}$
(η, H)	Perturbation noise	$\sim \mu$, independently of $\mathfrak{H}_{t-1}$
$\widetilde\theta(\eta, t)$	Pertrubed objective	$\hat\theta_t + \omega_t(\delta)\eta V_t^{-1/2}$
$\widetilde\Phi(H, t)$	Perturbed constraint	$\hat\Phi_t + \omega_t(\delta) H V_t^{-1/2}$.
$B(\xi)$	Tail bound on $\\|\eta\\|, \max_i \\|H^i\\|$	
B_t	Noise radius bound	$\max(1, B(\delta_t))$, where $\delta_t = \delta/(t^2+t)$.
$M_t(a)$	Perturbation scale at a	$B_t \omega_t \\|a\\|_{V_t^{-1}}$
$a(\eta, H, t)$	Perturbed optimum	See (3)
$\mathsf{U}_t(\delta)$	Unsaturation event	$\{(\eta, H) : \Delta(a(\eta, H, t)) \le M_t(a(\eta, H, t))\}$
χ	Unsaturation rate	
$\mathsf{L}_t(\delta)$	Local optimism event	$\{(\eta, H) : \widetilde\theta(\eta, t)^\top a_* \ge \theta_*^\top a_*, \widetilde\Phi(H, t) a_* \le \alpha\}$
π	Local optimism rate	
	Coupled Noise Design	
ν	Baseline perturbation law	Supported on $\mathbb{R}^{d\times 1}$
ζ	Generic draw from ν	$\zeta \sim \nu$, independent of $\mathfrak{H}_{t-1}$
$\bar B$	Tail bound for ν	$\nu(\\|\zeta\\| > \bar B(\xi)) \le \xi$
p	Anticoncentration parameter for ν	$\inf_u \nu(\zeta^\top u > \\|u\\|) \ge p$
$(\zeta^\top, -\mathbf{1}_m \zeta^\top)$	Coupled noise induced by ν	i.e., draw ζ, set $\eta = \zeta^\top$ and $H = -\mathbf{1}_m \zeta^\top$.
	S-COLTS	
a_{safe}	A priori given safe action	$\Gamma(a_{\mathsf{safe}}) > 0$.
Γ_0	Reference margin (see §H.1) for estimation)	$\Gamma_0 \ge \Gamma(a_{\mathsf{safe}})/2$ and $\Gamma_0 \le \Gamma(a_{\mathsf{safe}})$
(η_t, H_t)	Perturbation noise at t	
$\widetilde\theta_t, \widetilde\Phi_t$	Perturbed parameters at t	$\widetilde\theta_t = \widetilde\theta(\eta_t, t), \widetilde\Phi_t = \widetilde\Phi(H_t, t)$
b_t	Preliminary action at time t (if exists)	$b_t = a(\eta_t, H_t, t)$
$\mathfrak{a}(\rho)$	ρ-mixture of b_t and a_{safe}	$\mathfrak{a}(\rho) = \rho b_t + (1-\rho) a_{\mathsf{safe}}$
ρ_t	Largest ρ with safe $\mathfrak{a}(\rho)$	See (4); $a_t = \mathfrak{a}(\rho_t)$.
$\tau(t)$	Look-back time	Lemma 5
	E-COLTS	
(η_t, H_t)	Perturbation noise draws at time t	$(\eta_t, H_t) \sim t$
κ	Goodness factor of exploratory policy	See §I.2
u_t	Number of exploration steps up to time t	$u_t \approx B_t \omega_t \sqrt{dt}$
	R-COLTS	
r	Resampling parameter	
I_t	Number of resamplings at time t	$I_t = \lceil r \log(1/\delta_t) \rceil + 1$.
$(\eta_{i,t}, H_{i,t})$	ith draw of noise perturbation at time t	$\sim \mu$ independently
$K(i, t)$	Value under perturbation	$K(\widetilde\theta(\eta_{i,t}, t), \widetilde\Phi(H_{i,t}, t))$
$i_{*,t}$	Best index at time t	$\arg\max_i K(i, t)$
a_t	Action picked	$a_t = a(\eta_{i_{*,t}}, H_{i_{*,t}}, t)$
$\widetilde\theta_t$	Objective for $i_{*,t}$	$\widetilde\theta_t = \widetilde\theta(\eta_{i_{*,t}}, t)$.

B Examples of Real-World Domains where the Safe Linear Bandit Problem Applies

Table 3: **Mapping real domains to the bandit linear programming.** In all three cases the reward is linear in an unknown parameter vector θ_*, and the safety/fairness predicate is an *unknown linear inequality* $\Phi_* a \le \alpha$. Feedback noise in both rewards and constraints arises through environmental or individual fluctuations.

Domain (ref.)	Action $a \in \mathcal{A} \subset \mathbb{R}^d$	Reward $\theta_*^\top a$ + noise	Constraints
Dose-finding [AKR21]	One-hot vector for d discrete dose levels	θ_*^i = patient-level efficacy probability at dose i	Φ_*^i = toxicity of dose i; constraint so that $P(\text{toxic}\|\text{dose}) \le \alpha$
Voltage-constrained micro-grid [FLZY22]	Active/reactive power set-point $[P, Q]^\top$ for each bus	θ_*^i = locational marginal price vector	Φ_* = linearised network power-flow imposing nodal-voltage constraints under variable demand
Fair Reccommendation in A/B testing [Cho+24]	Distribution over d items or policies	θ_*^i = revenue of item i	Φ_*^i = encoding group attributes and costs; constraints demand fair exposure for each group

C Further Related Work

Distinction of Safe Bandits From BwK. BwK settings are concerned with aggregate cost metrics of the form $\mathbf{A}_T := \max_i (\sum \alpha - \Phi_* a_t)^i$, without the $(\cdot)_+$ nonlinearity in $\mathbf{S}_T$ [e.g. AD14; BKS13]. This simple change has a drastic effect, in that BwK algorithms can 'bank' violation by playing very safe actions for some rounds, and then 'spend' it to gain high reward, without any net penalty in $\mathbf{A}_T$. This is appropriate for modeling aggregate cost constraints (monetary/energy/et c.), but is evidently inappropriate to model safety constraints where feasibility violation in any round cannot be offset by acting safely in another round. Notice that such behaviour is precluded by the ramp nonlinearity in $\mathbf{R}_T, \mathbf{S}_T$: playing too-conservatively does not decrease $\mathbf{S}_T$, while any violation of constraints is accumulated, and similarly, playing suboptimally causes $\mathbf{R}_T$ to rise, but playing an over-aggressive action with negative $\Delta(a)$ does not reduce $\mathbf{R}_T$.

Pure Exploration in Safe Bandits. While our paper focuses on controlling regret and risk, naturally the safe bandit problem can be studied in the pure-exploration sense. These are studied in both the 'soft enforcement' sense, in which case methods can explore both within and outside the feasible region and return actions that are ε-safe and ε-optimal [e.g., Cam+22; KS19], and the 'hard enforcement', wherein exploratory actions must be restricted to the feasible region [e.g., SGBK15; Bot+22].

More Details on Computational Costs of Prior Methods. Most frequentist confidence-set based hard enforcement methods pick actions by solving the program

$$\max_{\theta \in \mathcal{C}_t^\theta, a \in \mathcal{A}} \theta^\top a \text{ s.t. } \forall \Phi \in \mathcal{C}_t^\Phi, \Phi a \le \alpha.$$

Assuming, for simplicity, that $a_{\mathsf{safe}} = 0$, due to the structure of the confidence sets the above constraint translates to

$$\forall i \in [1:m], \hat{\Phi}_t^i a + \omega_t(\delta) \|a\|_{V_t^{-1}} \mathbf{1}_m \le \alpha.$$

Notice that this constitutes m different second-order conic constraints. In fact, as discussed in §I.2, we expect V_t^{-1} to have condition number scaling as $\Omega(t^{1/4})$, which adds further computational burdens to optimising under such constraints.

Of course, as written, the above program is nonconvex due to the objective $\theta^\top a$. This can be addressed via a standard 'ℓ_1-relaxation [DHK08], which reduces the problem to solving $2d$ optimisation problems with linear objectives and the above SOC constraints, while weakening regret to $\widetilde{O}(\mathcal{R}(a_{\mathsf{safe}})\sqrt{d^3 T})$. This characterises the costs of most of these 'optimistic-pessimistic' methods [e.g. PGBJ21; PGB24; AAT19]. Afsharrad et al. give a systematic and detailed account of these considerations [AML24]. There are two exceptions. The SAFE-LTS method of Moradipari et al. [MAAT21] uses sampling to select the objective, but still imposes the same SOC constraints, thus needing only one optimisation each round. The ROFUL method of Hutchinson et al.[HTA24] instead first picks an action according to (the NP-hard to implement method) DOSS, and then scales it towards a_{safe} as in S-COLTS. Of course, note that S-COLTS samples only one set of *linear* constraints each round, and is efficient. There are also analytical differences between ROFUL and S-COLTS, as discussed in §4.

Turning to soft enforcement, as we mentioned in the main text, no efficient method is known. The main method herein for linear bandits is DOSS [GCS24], which instead picks actions by solving

$$\max_{\theta \in \mathcal{C}_t^\theta, a \in \mathcal{A}} \theta^\top a \text{ s.t. } \exists \Phi \in \mathcal{C}_t^\Phi : \Phi a \le \alpha.$$

This $\exists$ operator renders this problem much more challenging, since now the constraint works out to the union of polytopes

$$\bigcup_{A \in \mathcal{C}_t^\Phi} \mathcal{A} \cap \{\Phi a \le \alpha\},$$

which is highly nonconvex, and hard to condense or relax. Indeed, Gangrade et al. [GCS24] propose using a similar ℓ_1-relaxation as discussed above for both the objective and the constraints, but this now leads to $(2d)^{m+1}$-extreme points of the confidence sets (accounting for both θ and the m-rows of Φ), leading to $(2d)^{m+1} \cdot \mathsf{LP} \cdot \log(t)$ compute needed per round. In contrast, R-COLTS uses $\mathsf{LP} \cdot \log^2(t)$ compute, and E-COLTS uses only $\mathsf{LP} \cdot \log(t)$ compute.

More Details on the Failure of Prior Thompson Sampling Analyses. §4 discusses the point where the prior unsaturation-based analysis of linear TS due to [AG13] breaks down in the presence of unknown constraints in some detail. For the optimism-based analysis of [AL17], we only briefly touch upon this in §5, and give a more detailed look in §D. This section serves as a brief summary of the latter.

The analysis of Abeille and Lazaric relies on the convexity of the value function $J(\theta) := \max_{a \in \mathcal{A}} \theta^\top a$ to both analyse the roundwise regret ($\Delta(a_t)$) and to establish the frequency of a certain 'global optimism' event (see §D. With unknown constraints, the corresponding object of interest is the value function $K(\theta, \Phi) := \sup\{\theta^\top a : a \in \mathcal{A}, \Phi a \le \alpha\}$. This map is *not* convex in Φ, which causes both of these steps to break down. R-COLTS avoids this issue by resampling. It is also possible to give an analysis of S-COLTS (and E-COLTS) within the optimism framework, although this again utilises a scaling trick to bypass the same issue. Of course, we also establish optimism in a convexity-free way by analysing the local behaviour at a_*.

Finding a Feasible Point. Notice that since there are plenty of polynomial time methods for hard enforcement in SLBs (even though the prior methods impose SOC constraints), in principle one can develop efficient soft-enforcement methods with regret scaling inversely in $\max_a \Gamma(a)$ by first discovering an action that has $\Gamma(a) \ge \text{const.} \cdot \max_a \Gamma(a)$, and then plugging this into a hard enforcement method. In this case, the exploration time would be random, but a constant, so the net risk would ostensibly be $O(1)$ as T explodes, far below our $\sqrt{T}$ bounds, making the performance close to that of hard enforcement.[1] Unfortunately, there is no *efficient* method to discover such an action. Indeed, the closest method one can find in the literature is a feasibility *test* due to Gangrade et al. [GGSS24], which can be extended to such an estimator, but this test uses DOSS-like optimisation to select actions, and needs to solve $(2d)^m$ optimisation problems each round. Our coupled noise design should have implications for this problem, but we do not pursue this direction further.

D Local Optimism, Global Optimism, and Unsaturation

In §5, we (implicitly) defined a local-optimism condition on the perturbation law μ in the statement of Lemma 9, which is compared to a 'global optimism' condition suggested by the prior work of Abeille & Lazaric [AL17]. To further contextualise these, let us explicitly define them.

Definition 12. *Let $K(\theta, \Phi) := \sup\{\theta^\top a : a \in \mathcal{A}, \Phi a \le \alpha\}$ denote the value function of optimising the objective θ under constraint matrix Φ over $\mathcal{A}$, with the convention that $\sup \emptyset = -\infty$. Recall that the* local optimism event *at a_* is*

$$\mathsf{L}_t(\delta) := \{(\eta, H) : \widetilde{\theta}(\eta, t)^\top a_* \ge \theta_*^\top a_*, \widetilde{\Phi}(H, t) a_* \le \alpha\},$$

where a_* is the constrained optimum for the true parameters (θ_*, Φ_*). Further, define the global optimism event

$$\mathsf{G}_t(\delta) := \{(\eta, H) : K(\widetilde{\theta}(\eta, t), \widetilde{\Phi}(H, t)) \ge \theta_*^\top a_* = K(\theta_*, \Phi_*)\}.$$

[1] note that there is a cost, though: as stated before, the regret would scale inversely in the Slater gap, and until the safe point is discovered, would grow linearly.

For $\pi \in (0, 1]$, we say that a law μ on (η, H) satisfies π-local optimism if

$$\forall t, \mathbb{E}[\mu(\mathsf{L}_t(\delta))|\mathfrak{H}_{t-1}]\mathbb{1}_{\mathsf{Con}_t(\delta)} \geq \pi \mathbb{1}_{\mathsf{Con}_t(\delta)},$$

and similarly, that μ satisfies π-global optimism if

$$\forall t, \mathbb{E}[\mu(\mathsf{G}_t(\delta))|\mathfrak{H}_{t-1}]\mathbb{1}_{\mathsf{Con}_t(\delta)} \geq \pi \mathbb{1}_{\mathsf{Con}_t(\delta)}.$$

Notice that G demands perturbations such that after optimising the perturbed parameters, the value of the resulting program is larger than $\theta_*^\top a_*$, while L demands the stronger condition that a_* is feasible, and its value increases. Evidently, $\mathsf{L} \subset \mathsf{G}$, and so π-local optimism of μ implies π-global optimism. Naturally, the entirety of §5 follows if we have a globally optimistic μ instead of locally optimistic μ. We presented this section with L_t instead due to limited space in the main text.

As discussed in §4.2, we will also show, in §G, $\mathsf{L}_t(\delta) \cap \mathsf{Con}_t(\delta) \subset \mathsf{U}_t(\delta) \cap \mathsf{Con}_t(\delta)$, i.e., when consistency holds, local optimism implies unsaturation. Thus, L_t links the global-optimism based framework of [AL17], and the unsaturation based framework of [AG13]. Nevertheless, technically, these are distinct events.

Let us briefly note that the prior work [AG13] essentially passes through the same strategy as us when establishing a good unsaturation rate, in that they argue that local-optimism holds frequently (although they do not consider unknown constraints, so their argument does not extend to our setting). On the other hand, [AL17] presents a convexity-based proof of frequent global optimism for linear TS without unknown constraints, while immediately breaks in our setting because $K(\theta, \Phi)$ is nonconvex in Φ. We also reiterate that our coupled noise design of §4.2 essentially takes the same conditions on perturbations used in these prior works, and extends them to produce the *same* bounds on unsaturation or global-optimism rates by arguing that local-optimism holds. This means that these prior results do not capture the prevalence of these events beyond local optimism. Our simulations in §J suggest that this leaves a significant amount of performance on the table, capturing which theoretically would require deeper understanding of $\mathsf{U}_t \setminus \mathsf{L}_t$ and $\mathsf{G}_t \setminus \mathsf{L}_t$.

Role of These Conditions in Our Work. To analyse S-COLTS and E-COLTS, we used a look-back approach enabled by the unsaturation condition, while to analyse R-COLTS, we relied on a direct use of the optimism condition. It turns out that the unsaturation condition is not effective at capturing at least our strategy for analysing the resampling-based strategy R-COLTS. The reason is that while the resampling will ensure that at least one of the optima of attaining the various $K(i,t)$ values will be unsaturated, we have no guarantee that the procedure we take of picking the $i_{*,t}$ that maximises $K(i,t)$ will choose an unsaturated action. On the other hand, the optimism condition *can* be used to analyse S-COLTS and E-COLTS directly (see §H.5), but a direct execution of the previous optimism based approach [AL17] fails due to the lack of convexity of the map $K(\theta, \Phi)$. Instead, we have to directly analyse expressions of the form $\mathbb{E}[|K(\widetilde{\theta}, \widetilde{\Phi}) - K(\widetilde{\theta}', \widetilde{\Phi}')| \mid \mathfrak{H}_{t-1}]$, where $(\widetilde{\theta}, \widetilde{\Phi})$ and $(\widetilde{\theta}', \widetilde{\Phi}')$ are iid draws of the perturbation at tie t. Under the assumption that there is an action with positive safety margin with small M_t, this can be executed via a similar scaling-based analysis, albeit at a loss of some factors in the regret bound (§H.5). In our opinion the unsaturation based look-back analysis of $\Delta(a_t)$ is conceptually clearer, and we chose to present it in the main instead.

Nevertheless, in terms of their explanatory power, neither condition dominates the other. Indeed, in simulations, we find both cases where unsaturation is frequent but global optimism is not, and cases where global optimism is frequent but unsaturation is not.[2] Of course, in our analysis, both of these are connected by local optimism as detailed above, which is rendered frequent through our coupled design. Nevertheless, the local optimism rate can be significantly smaller than the unsaturation and global optimism rates, particularly when the noise is shrunk far below the theoretically analysed setting of $\Theta(\sqrt{d})$-scale noise (see §J). These observations again hint that developing a tight theory of linear TS (both with and without unknown constraints) requires a deeper understanding of the portion of these events that do not intersect with local optimism.

[2] This is most pertinent for the setting where we drive the perturbations with independent noise, where in §J.3 we observed that the unsaturation rate decayed with m, but the global optimism rate did not. Indeed, this is what prompted us to write the optimism-based analysis of §H.5.

E An Informal Discussion of Contextual Safe Linear Bandits

Rather than static bandit problems, most practical scenarios are contextual, wherein the learner observes some side information x_t before choosing an action, and this side information affects the reward and constraint structure at time t. A common setting to model this [PGB24; AG13] is to assume that there is a known feature map $\varphi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ such that the reward and constraints at time t are of the form

$$\theta_*^\top \varphi(x_t, a) \text{ and } \Phi_* \varphi(x_t, a) \le \alpha.$$

Throughout, we assume the same feedback structure, i.e., noisy measurements of $\theta_*^\top \varphi(x_t, a_t)$ and $\Phi_* \varphi(x_t, a_t)$. Naturally, regret is compared to the optimal policy $\mathscr{A}_* : \mathcal{X} \to \mathcal{A}$, where

$$\mathscr{A}_*(x) = \arg\max \theta_*^\top \varphi(x, a) : \Phi_* \varphi(x, a) \le \alpha, a \in \mathcal{A}.$$

It should be noted that the Lemma 1 on consistency, and the elliptical potential lemma (Lemma 13) continue to hold, with V_t replaced by $I + \sum_{s \le t} \varphi(x_s, a_s)\varphi(x_s, a_s)^\top$, and a_t by $\varphi(x_t, a_t)$. Notationally, we extend $\Delta(a), \Gamma(a)$ to $\Delta(x, a) = \theta_*^\top(\varphi(x, \mathscr{A}_*(x)) - \varphi(x, a))$ and $\Gamma(x, a) = \max_i((\alpha - \Phi_* \varphi(x, a))^i)_+$.

A key observation is that our result on the frequency of the local optimism persists in this contextual setting. Under the hood, this essentially shows that at any t, and for any vector φ,

$$\mathbb{P}\left\{(\eta, H) : \widetilde{\theta}^\top \varphi \ge \theta_*^\top \varphi, \widetilde{\Phi} \varphi \le \Phi_* \varphi \middle| \mathfrak{H}_{t-1}\right\} \mathbb{1}_{\mathsf{Con}_t(\delta)} \ge \pi \mathbb{1}_{\mathsf{Con}_t(\delta)},$$

where $\pi \ge 0.28$ for the coupled noise driven by $\mathrm{Unif}(\sqrt{3d}\mathbb{S})^d$. Consequently, frequent local optimism follows in the contextual setting by using this result for $\varphi(x_t, \mathscr{A}_*(x_t))$ at time t.

The above observation means that using the same coupled noise lets us extend the results of Theorem 10 on the regret of R-COLTS to the contextual case with only cosmetic changes in the analysis. This holds no matter how the sequence x_t is selected, as long as the noise remains conditionally centred and subGaussian given a_t, x_t, the algorithmic randomness, and the history. Note, however, that the optimisation over a may become harder due to the feature map φ, and efficiency requires further structural assumptions on φ.

Focusing now on S-COLTS, let us first note that if we were given a safe action a_{safe} that was safe no matter the context, i.e., such that $\inf_x \Gamma(x, a_{\mathsf{safe}}) \ge \Gamma_{\text{safe}} > 0$, and φ were 'nice' in terms of $a \in \mathcal{A}$,[3] then as long as we *know* Γ_{safe} a priori, no real change is required, and the guarantees of Theorem 8 for S-COLTS extend to the contextual setting,[4] since we can again guarantee the frequent choice of unsaturated actions through our persistent local optimism property. We note that previous works on safe contextual bandits [PGB24] assume exactly this existence of an 'always very safe' action. Nevertheless, this structure is unrealistic: practically, safety should depend strongly on the context, and it is unlikely that a single action would always be safe, let alone have a large safety margin. A more natural assumption is that instead of a single safe action, we are given a safe policy $\mathscr{A}_{\text{safe}} : \mathcal{X} \to \mathcal{A}$. Here, again, if we know that $\inf_x \Gamma(x, \mathscr{A}_{\text{safe}}(x)) \ge \Gamma_{\text{safe}} > 0$, and we know the value of Γ_{safe}, then we are good to go, although this is a strong assumption. Without knowing this value, we need to be able to determine a good estimate of $\Gamma(x_t, \mathscr{A}_{\text{safe}}(x_t))$ in order to appropriately ensure feasiblity of perturbed programs, and to scale back the actions b_t. This can be a challenging task, especially if x_t varies in an adversarial way, and structures enabling such estimation must be assumed.[5] Finally, note that even if we were given $\Gamma(x, \mathscr{A}_{\text{safe}}(x))$ as a function explicitly, the easily forthcoming regret bounds rather pessimistically scale with $(\inf_x \Gamma(x, \mathscr{A}_{\text{safe}}(x)))^{-1}$, and do not capture how variation in this margin with x can be used to limit regret. A (at least somewhat) different analysis is needed to express this in a clear way. Resolving such limitations is an important open problem in the theory of SLBs.

This lacuna also affects the E-COLTS method of §I.2, but to a lesser extent. Sticking with 'nice' feature maps, again, if there *exists* an action that is always safe, i.e., if $\max_a \min_x \Gamma(x_t, a_t) > 0$, then the guarantees of Theorem 11 extend with arbitrary context sequence. Without this guarantee,

[3]We essentially need a way to efficiently select an action a such that $\varphi(x_t, a) = \rho\varphi(x_t, b_t) + (1 - \rho)\varphi(x_t, a_{\mathsf{safe}})$, so that safety can still be attained by mixing with a_{safe}.

[4]upto replacing $\Delta(a_{\mathsf{safe}})$ by 1

[5]For instance, if x_t were drawn in some static randomised way, and Γ were sufficiently simple, then we could learn $\Gamma(x\mathscr{A}_{\text{safe}}(x))$ using regression techniques.

the main gap is the exploration policy being utilised, which must be adapted to attain a good coverage over $\{\varphi(x,a)\}$ even as x_t varies. Given such a policy, however, the results of Theorem 11 again extend to the contextual case with arbitrary x_t.

F Some Basic Tools For the Analysis

We begin with some standard tools that are repeatedly utilised in the analysis. The first of these, termed the *elliptical potential lemma* offers generic control on the accumulation of $\|a_t\|_{V_t^{-1}}$.

Lemma 13. *[APS11; CVA20] For any sequence of actions $\{a_t\} \subset \{\|a\| \le 1\}$, and any t,*

$$\sum_{s\le t} \|a_s\|^2_{V_s^{-1}} \le 2d\log(1+t/d), \text{ and } \sum_{s\le t} \|a_s\|_{V_s^{-1}} \le \sqrt{2dt\log(1+t/d)}.$$

Further, for all $t, \delta, \omega_t(\delta) \le 1 + \sqrt{1/2\log((m+1)/\delta) + d/2\log(1+t/d)}$.

We further explicitly write the following instantiation of the Cauchy-Schwarz inequality pertinent to our setting.

Lemma 14. *For any positive definite matrix V. For pair of tuples (θ,Φ) and $(\widetilde{\theta},\widetilde{\Phi})$ lying in $\mathbb{R}^d \times \mathbb{R}^{m\times d}$ and any $a \in \mathbb{R}^d$, it holds that*

$$\max\left(|(\theta-\widetilde{\theta})^\top a|, \max_i |(\Phi^i - \widetilde{\Phi}^i)a|\right) \le \max(\|\widetilde{\theta}-\theta\|_V, \max_i \|\widetilde{\Phi}^i - \Phi^i\|_V)\cdot \|a\|_{V^{-1}}.$$

Proof. Notice that $(\widetilde{\theta}-\theta)^\top a = (\widetilde{\theta}-\theta)^\top V^{1/2}V^{-1/2}a \le \|(V^{1/2}(\widetilde{\theta}-\theta)\| \cdot \|V^{-1/2}a\|$. The claim follows by first repeating the same observation for each $(\Phi^i - \widetilde{\Phi}^i)$ (adjusting for the fact that these are row-vectors), and then recalling that (for column vectors) $\|a\|_M = \|M^{1/2}a\|$ by definition. □

This immediately yields a proof of the concentration statement of Lemma 3, which motivated the definition of $M_t(a)$.

Proof of Lemma 3. Notice that by a union bound

$$\mathbb{P}(\exists t : \max(\|\eta_t\|, \max_i \|H_t^i\|) > B(\delta_t)) \le \sum_t \delta_t = \delta.$$

Now assume that $\max(\|\eta_t\|, \max_i \|H_t^i\|) \le B(\delta_t)$, and that the consistency event $\mathsf{Con}_t(\delta)$ holds. Then, via the triangle inequality,

$$\|\widetilde{\theta}(\eta_t,t) - \theta_*\|_{V_t} \le \|\widetilde{\theta}(\eta_t,t) - \hat{\theta}_t\|_{V_t} + \|\hat{\theta}_t - \theta_*\|_{V_t}.$$

Of course, given $\mathsf{Con}_t(\delta)$, the second term is smaller than $\omega_t(\delta)$. For the first, expanding the definition of $\widetilde{\theta}(\cdot,\cdot)$, we find that

$$\|\widetilde{\theta}(\eta_t,t) - \hat{\theta}_t\|_{V_t} = \omega_t(\delta)\eta_t V_t^{-1/2}\|_{V_t} = \|\omega_t(\delta)\eta_t V_t^{-1/2}\cdot V_t^{1/2}\| \le \omega_t(\delta)\|\eta_t\|,$$

and of course, $\|\eta_t\| \le B(\delta_t)$ by our assumption above. Thus, given the concentration assumption on $\|\eta_t\|$s and $\mathsf{Con}_t(\delta)$, for any t, it holds that

$$\|\widetilde{\theta}(\eta_t,t) - \theta_*\|_{V_t} \le (1+B(\delta_t))\omega_t(\delta) \le B_t\omega_t(\delta).$$

Of course, entirely the same applies to $\|\widetilde{\Phi}(H_t,t)^i - \Phi_*^i\|_{V_t}$, with η replaced by H_t^i. The claim now follows by Lemma 14 and the fact that $\mathsf{Con}(\delta) := \bigcap \mathsf{Con}_t(\delta)$ has chance at least $1-\delta$. □

G Analysis of the Coupled Noise Design

We will first execute the strategy described in §4.2 to show that under the conditions of Lemma 7, local optimism is frequent. We will then use this to show the frequency of unsaturation.

Lemma 15. *Let $p \in (0,1]$, and let ν be a law on $\mathbb{R}^{d\times 1}$ such that*

$$\forall u \in \mathbb{R}^d, \nu(\{\zeta : \zeta^\top u \geq \|u\|) \geq p.$$

Let μ be the the pushforward of ν under the map $\zeta \mapsto (\zeta^\top, -\mathbf{1}_m\zeta^\top)$. Then, for all t, $\mathbb{1}_{\mathsf{Con}_t(\delta)}\mathbb{E}[\mu(\mathsf{L}_t(\delta))|\mathfrak{H}_{t-1}] \geq p\mathbb{1}_{\mathsf{Con}_t(\delta)}$, where $\mathsf{L}_t(\delta)$ is the local optimism event (5).

Proof. Observe that under a draw from μ, for all t, we have

$$\widetilde{\theta}^\top := (\widetilde{\theta}(\eta,t))^\top = \hat{\theta}_t^\top + \omega_t(\delta)\zeta^\top V_t^{-1/2}$$
$$\widetilde{\Phi} := \widetilde{\Phi}(H,t) = \hat{\Phi}_t - \mathbf{1}_m(\omega_t(\delta)\zeta^\top V_t^{-1/2}).$$

Further, recall that if the event $\mathsf{Con}_t(\delta)$ occurs, then, for all a,

$$\hat{\theta}_t^\top a \geq \theta_*^\top a + \omega_t(\delta)\|V_t^{-1/2}a\|, \text{ and } \hat{\Phi}_t a \leq \Phi_* a + \mathbf{1}_m(\omega_t(\delta)\|V_t^{-1/2}a\|,$$

where we have the Cauchy-Schwarz inequality, and the fact that $\|a\|_{V_t^{-1}} = \|V_t^{-1/2}a\|$. Thus, assuming $\mathsf{Con}_t(\delta)$, for any action a, we find that

$$\widetilde{\theta}^\top a \geq \theta_*^\top a + \omega_t(\delta)\left(\zeta^\top V_t^{-1/2}a - \|V_t^{-1/2}a\|\right),$$
$$\widetilde{\Phi} a \geq \Phi_* a + \mathbf{1}_m\omega_t(\delta)\left(\zeta^\top V_t^{-1/2}a - \|V_t^{-1/2}a\|\right).$$

Now, set $a = a_*$, and suppose that $\zeta^\top V_t^{-1/2}a_* \geq \|V_t^{-1/2}a_*\|$. Then we can conclude that

$$\widetilde{\theta}^\top a_* \geq \theta_*^\top a_* \text{ and } \widetilde{\Phi} a_* \leq \Phi_* a_* \leq \alpha,$$

the final inequality holding since a_* is of course feasible for the program it optimises. Of course, by definition, this means that the ensuing noise η, H lie in the event $\mathsf{L}_t(\delta)$

Now, it only remains to argue that $\zeta^\top V_t^{-1/2}a_* \geq \|\zeta^\top V_t^{-1/2}a_*\|$ happens with large chance given $\mathfrak{H}_{t-1}$. But notice that both $V_t^{-1/2}$ and (the constant) a_* are $\mathfrak{H}_{t-1}$-measurable, and so are constant given it. It follows thus that

$$\mathbb{E}[\nu(\{\zeta : \zeta^\top V_t^{-1/2}a_* > \|V_t^{-1/2}a_*\|\}) \mid \mathfrak{H}_{t-1}] \geq \inf_{u\in\mathbb{R}^d} \nu(\{\zeta^\top u > \|u\|\}) \geq p. \qquad \square$$

To finish the proof of frequent unsaturation, we only need to determine that this local optimism induces unsaturation in the actions.

Proof of Lemma 7. Fix a t, and assume consistency. Suppose that $\max(\|\eta_t\|, \max_i \|H_t^i\|) \leq B(\delta_t)$. Note that given $\mathsf{Con}_t(\delta)$, this with chance at least $1-\delta_t$. As a consequence, for any action $a \in \mathfrak{S}_t := \{a : \Delta(a) > M_t(a)\}$, by following the proof of Lemma 3 we can conclude that

$$\widetilde{\theta}(\eta,t)^\top a \leq \theta_*^\top a + M_t(a) = \theta_*^\top a_* - \Delta(a) + M_t(a) < \theta_*^\top a_*.$$

Now, suppose that the drawn ζ induces local optimism. We claim that then all saturated actions are suboptimal. Indeed, by the above, each unsaturated action satisfies $\widetilde{\theta}(\eta,t)^\top a < \theta_*^\top a_*$. But $\widetilde{\theta}(\eta,t)^\top a_* \geq \theta_*^\top a_*$, and further $\widetilde{\Phi}(H,t)a_* \leq \alpha$, means that there is an action that is feasible for the perturbed program with value strictly larger than that attained by any saturated action, i.e., any member of $\mathfrak{S}_t$. It thus follows that the optimum $a(\eta,H,t) \in \mathfrak{S}_t^c = \{a : \Delta(a) \leq M_t(a)\}$.

Now, we know from Lemma 15 that given $\mathfrak{H}_{t-1}$, our assumptions of $\mathsf{Con}_t(\delta)$ and the norm-control on $\|\eta_t\|, \max_i \|H_t^i\|$ imply that local optimism occurs with chance at least p. Since these events occur with chance at least $1-\delta_t$, this means that unsaturation occurs with chance at least $p - \delta_t$. Since definition 4 restricts attention to $t : \delta_t \leq p/2$, the statement follows. $\square$

G.1 Bounds for Simple Reference Laws

We argue that both the standard Gaussian, and the uniform law of the sphere of radius $\sqrt{3d}$ yield effective noise distributions for our coupled design.

For the Gaussian, recall that if $Z \sim \mathscr{N}(0, I_d)$, then $\|Z\|^2$ is distributed as a χ^2-random variable. A classical subexponential concentration argument [e.g. LM00, Lemma 1] yields that for any x,

$$\mathbb{P}(\|Z\|^2 \geq d + 2\sqrt{dx} + 2x) \leq e^{-x}.$$

Note that $(d + 2\sqrt{dx} + 2x) \leq (\sqrt{d} + \sqrt{2x})^2$, and hence taking $x = \log(1/\xi)$ in the above yields that $B(\xi) \leq \sqrt{d} + \sqrt{2\log(1/\xi)}$. Further, due to the isotropicity of Z, $Z^\top u/\|u\| \overset{\text{law}}{=} Z_1 \sim \mathscr{N}(0,1)$, and thus $\pi \geq 1 - \Phi(1) \geq 0.158\dots$.

Further, notice that if $Z \sim \mathscr{N}(0, I_d)$, then $Y := \sqrt{3d}Z/\|Z\| \sim \mathrm{Unif}(\sqrt{3d} \cdot \mathbb{S}^d)$, and by isotropicity, for any u, $Y^\top u/\|u\| \overset{\text{law}}{=} Y_1$. As a result,

$$\begin{aligned}\mathbb{P}(Y^\top u/\|u\| \geq 1) = \mathbb{P}(Y_1 \geq 1) = \frac{1}{2}\mathbb{P}(Y_1^2 \geq 1)\\ = \frac{1}{2}\mathbb{P}((3d-1)Z_1^2 \geq \sum_{i=2}^d Z_i^2) \geq \frac{1}{2}\mathbb{P}(Z_1^2 \geq 1) \cdot \mathbb{P}(\sum_{i=2}^d Z_i^2 \geq 3d-1).\end{aligned}$$

But notice that $d - 1 + 2\sqrt{(d-1)\cdot d/3} + 2d/3 \leq 3d - 1$, and thus, $\mathbb{P}(\sum_{i=2}^d Z_i^2 \geq 3d-1) \leq \exp(-d/3)$. Invoking the bound on $\mathbb{P}(Z_1 \geq 1) = \frac{1}{2}\mathbb{P}(|Z_1| \geq 1)$ above, we conclude that $\pi \geq 0.15 \cdot (1 - e^{-d/3})$. Of course, $\|Y\| = \sqrt{3d}$ surely, giving the B expression.

We note that while the above only shows a $0.15(1 - e^{-d/3})$ bound on the anticoncentration of the uniform law on $\sqrt{3d}\mathbb{S}^d$, it is a simple matter of simulation to find that this is actually larger than 0.28 for all d - for small dimensions, the bound turns out to be very loose, while as d diverges, this converges from above towards the chance that a standard Gaussian exceeds $1/\sqrt{3}$, which is $0.2818\dots$.

H The Analysis of S-COLTS

We move on to the analysis of S-COLTS. Before proceeding, we recall that in our presentation of S-COLTS in Algorithm 1, we assumed access to a quantity $\Gamma_0 \in [\Gamma(a_{\mathsf{safe}})/2, \Gamma(a_{\mathsf{safe}})]$. We will first address how to obtain such a quantity by repeatedly playing $a_t = a_{\mathsf{safe}}$, and characterise how long this takes. For completenesss, the cost of this will be incorporated into our regret bound.

Beyond this, we need to characterise the subsequent time spent playing a_{safe} due to $M_t(a_{\mathsf{safe}})$ being large, and to prove the look-back bound of Lemma 5, along with the characterisation of $\sum M_{\tau(t)}(a_{\tau(t)})$ offered in Lemma 6. We will analyse these results in order, and finally show Theorem 8 using these results.

H.1 Identifying Γ_0 and Sampling Rate of a_{safe}

We first discuss the determination of Γ_0. There are two main points to make: how to ensure a correct value of Γ_0, and how many rounds of exploration this costs. To this end, we first recall the following nonasymptotic law of iterated logarithms [e.g. HRMS21].

Lemma 16. *Let $\{\mathfrak{F}_t\}$ be a filtration, and let $\{\xi_t\}$ be a process such that each ξ_t is $\mathfrak{F}_t$-measurable, and is further conditionally centred and 1-subGaussian given $\mathfrak{F}_{t-1}$. Then*

$$\forall \delta \in (0,1], \mathbb{P}(\exists t : |Z_t| > \mathrm{LIL}(t,\delta)) \leq \delta,$$

where $Z_t := \sum_{s \leq t} \xi_t$, and

$$\mathrm{LIL}(t,\delta) := \sqrt{4t\log\frac{\max(1,\log(t))}{\delta}}.$$

With this in hand, the determination of Γ_0 proceeds thus: we repeatedly play a_{safe}, and maintain the running average $\mathrm{Av}_t = \sum_{s \le t} (\alpha - S_s)/t$. Further, we maintain the upper and lower bounds

$$u_t^i := \mathrm{Av}_t + \mathrm{LIL}(t, \delta/m)/t, \ell_t^i := \mathrm{Av}_t - \mathrm{LIL}(t, \delta)/t.$$

We stop at the first time when $\forall i, \ell_t^i \ge u_t^i/2$, and set $\Gamma_0 = \min_i \ell_t^i$. This stopping time is denoted T_0.

Let us first show that this procedure is correct, and bound the size of T_0.

Lemma 17. *Under the procedure specified above, it holds with probability at least* $1 - \delta$ *that*

$$\Gamma_0 \in [\Gamma(a_{\mathsf{safe}})/2, \Gamma(a_{\mathsf{safe}})]$$

and that

$$T_0 \le \frac{8}{\Gamma(a_{\mathsf{safe}})^2} \log(8/(\delta \Gamma(a_{\mathsf{safe}})^2))$$

Proof. Notice that we can write

$$\mathrm{Av}_t = \alpha - \Phi_* a_{\mathsf{safe}} + \sum_{s \le t} w_s^S/t.$$

For succinctness, let us write $\Gamma = \alpha - \Phi_* a_{\mathsf{safe}}$. Now, by our assumption on the noise w_t^S, we observe that each coordinate of w_t^S constitutes an adapted, centred, and 1-subGaussian process. Applying Lemma 16 along with a union bound over the coordinates then tells us that with probability at least $1 - \delta$,

$$\forall t, |\mathrm{Av}_t - \Gamma| \le \mathrm{LIL}(t, \delta/m)/t \cdot \mathbf{1}_m.$$

As a consequence, at all t, we have

$$u_t \ge \Gamma \ge \ell_t,$$

where u_t is the vector with ith coordinate u_t^i, and similarly for ℓ_t. It follows thus that at the stopping time T_0,

$$\forall i, \ell_{T_0}^i \ge u_{T_0}^i/2 \implies \ell_{T_0} \ge \Gamma/2.$$

Of course, a fortiori, it follows that $\Gamma_0 = \min_i \ell_t^i \ge \min_i \Gamma^i/2 = \Gamma(a_{\mathsf{safe}})/2$. Further, of course, $\Gamma_0 \le \Gamma(a_{\mathsf{safe}})$ follows as well, since $\forall t \min_i \ell_t^i \le \min_i(\Gamma^i) = \Gamma(a_{\mathsf{safe}})$.

It only remains to control T_0. To this end, notice that for all t

$$\ell_t = \mathrm{Av}_t - \mathrm{LIL}(t, m/\delta)/t \cdot \mathbf{1}_m \ge \Gamma - 2\mathrm{LIL}(t, m/\delta)/t \cdot \mathbf{1}_m,$$

and similarly,

$$u_t \le \Gamma + 2\mathrm{LIL}(t, m/\delta)/t \cdot \mathbf{1}_m.$$

Of course, then $\ell_t^i > u_t^i/2$ for all t such that

$$\forall i, \Gamma^i - \mathrm{LIL}(t, m/\delta)/t \ge \Gamma^i/2 + \mathrm{LIL}(t, m/\delta)/2t \iff \Gamma^i > 3\mathrm{LIL}(t, m/\delta)/t.$$

It follows thus that

$$T_0 \le \inf\{t : t\Gamma(a_{\mathsf{safe}}) \ge 3\mathrm{LIL}(t, m/\delta)\}.$$

By a simple inversion, this can be bounded as

$$T_0 \le \inf\{t : t > 8/\Gamma(a_{\mathsf{safe}})^2 \log(1/\delta) \text{ and } t > 8/\Gamma(a_{\mathsf{safe}})^2 \log(1 + \log(t))\},$$

which is bounded as

$$T_0 \le \frac{8}{\Gamma(a_{\mathsf{safe}})^2} \log(8/(\delta \Gamma(a_{\mathsf{safe}})^2)).$$

□

Number of Times a_{safe} ***is sampled after*** T_0**.** Given the behaviour of Γ_0 above, we can further bound the number of times a_{safe} is played after determining Γ_0.

Lemma 18. *For any* $\Gamma_0 > 0$*, and* T*, the number of times* S-COLTS *plays* a_{safe} *because* $M_t(a_{\mathsf{safe}}) > \Gamma_0/3$ *is bounded as* $\frac{9\omega_T^2 B_T^2}{\Gamma_0^2} + 1$.

Proof. Let n_t denote the total number of times a_{safe} has been played up to time t. Then, of course, $V_t \succcurlyeq I + n_t a_{\mathsf{safe}} a_{\mathsf{safe}}^\top$. Now, recall that for symmetric positive definite matrices A, B, it holds that $A \succcurlyeq B \iff B^{-1} \succcurlyeq A^{-1}$.[6] Thus, we have

$$M_t(a_{\mathsf{safe}}) \le \omega_t B_t \sqrt{a_{\mathsf{safe}}^\top (I + n_t a_{\mathsf{safe}} a_{\mathsf{safe}}^\top)^{-1} a_{\mathsf{safe}}}.$$

Now, by the Sherman-Morrisson formula,

$$a_{\mathsf{safe}} (I + n_t a_{\mathsf{safe}} a_{\mathsf{safe}}^\top)^{-1} a_{\mathsf{safe}} = \|a_{\mathsf{safe}}\|^2 - \frac{a_{\mathsf{safe}}^\top (n_t a_{\mathsf{safe}} a_{\mathsf{safe}}^\top) a_{\mathsf{safe}}}{1 + n_t \|a_{\mathsf{safe}}\|^2} = \frac{\|a_{\mathsf{safe}}\|^2}{1 + n_t \|a_{\mathsf{safe}}\|^2} \le \frac{1}{n_t}.$$

It follows thus that

$$M_t(a_{\mathsf{safe}}) \le \frac{\omega_t B_t}{\sqrt{n_t}}.$$

Thus $M_t(a_{\mathsf{safe}}) > \Gamma_0/3$ if and only if

$$n_t \le \frac{9\omega_t^2 B_t^2}{\Gamma_0^2}.$$

Of course, each time this occurs, n_t is increased by one. Consequently, the number of times a_{safe} is played by time t is at most

$$\frac{9\omega_T^2 B_T^2}{\Gamma_0^2} + 1. \qquad \square$$

Note that since $(\omega_T B_T)^2 = \Theta(d^2 + d\log(m/\delta))$ with our choice of the coupled noise driven by $\mathrm{Unif}(\sqrt{3d}\mathbb{S}^d)$, the bound above due to playing a_{safe} due to too large an $M_t(a_{\mathsf{safe}})$ outstrips the bound on T_0 above as long as $\log(1/\Gamma(a_{\mathsf{safe}})) = o(d^2)$, as is to be expected.

H.2 Proof of the Look-Back Bound

The main text provides a brief sketch of the approach. We will flesh out these details, as well as fill in the omitted aspects of the bound. To this end, we first state a result lower bounding ρ_t.

Lemma 19. *Assume that $\Gamma_0 \in [\Gamma(a_{\mathsf{safe}})/2, \Gamma(a_{\mathsf{safe}})]$, and that both $\mathsf{Con}(\delta) = \bigcap \mathsf{Con}_t(\delta)$ and the event of Lemma 3 hold true. Then for all t such that $M_t(a_{\mathsf{safe}}) \le \Gamma_0/3$, it holds that*

$$\rho_t \ge \frac{\Gamma(a_{\mathsf{safe}})}{\Gamma(a_{\mathsf{safe}}) + 3M_t(b_t)}$$

and

$$\rho_t \ge \frac{2M_t(a_{\mathsf{safe}})}{2M_t(a_{\mathsf{safe}}) + M_t(b_t)}.$$

A fortiori, each of the following bounds is true:

$$(1-\rho_t)M_t(a_{\mathsf{safe}}) \le M_t(a_t),$$
$$\rho_t M_t(b_t) \le 2M_t(a_t), \text{ and}$$
$$(1-\rho_t)\Gamma(a_{\mathsf{safe}}) \le 6M_t(a_t).$$

Proof. Recall that ρ_t is the largest ρ in $[0,1]$ such that

$$\hat{\Phi}_t(\rho b_t + (1-\rho)a_{\mathsf{safe}}) + \omega_t(\delta)\|\rho b_t + (1-\rho)a_{\mathsf{safe}}\|_{V_t^{-1}} \mathbf{1}_m \le \alpha.$$

So, if we demonstrate a $\rho_0 \le 1$ that satisfies this inequality, then $\rho_t \ge \rho_0$.

[6] In more technical terms, inversion is monotone decreasing in the Loewner sense. A simple way to see this is to define $C = B^{-1/2}AB^{-1/2}$. Then $A \succcurlyeq B \implies C \succcurlyeq I$ (really iff), since for any x, $(B^{-1/2}x)^\top A(B^{-1/2}x) \ge (B^{-1/2}x)^\top B(B^{-1/2}x) \iff x^\top Cx \ge x^\top x$. Using this for $y = C^{-1/2}x$ then gives $x^\top x = (C^{-1/2}x)^\top C(C^{-1/2}x) \ge (C^{-1/2}x)^\top (C^{-1/2}x) = x^\top C^{-1}x$. Since $C^{-1} = B^{1/2}A^{-1}B^{1/2}$ (direct multiplication), the same trick yields $x^\top B^{-1}x = (B^{-1/2}x)^\top (B^{-1/2}x) \ge x^\top B^{-1/2}(B^{1/2}A^{-1}B^{1/2})B^{-1/2}x$, or in other words, $B^{-1} \succcurlyeq A^{-1}$.

First note that under the assumption $M_t(a_{\mathsf{safe}}) \le \Gamma_0/3$, we know that

$$\widetilde{\Phi}_t a_{\mathsf{safe}} \le \alpha - \Gamma(a_{\mathsf{safe}})\mathbf{1}_m + \Gamma_0/3 \cdot \mathbf{1}_m \le \alpha - 2\Gamma(a_{\mathsf{safe}})/3 \cdot \mathbf{1}_m,$$

and thus b_t exists since the program defining it is feasible. Now,

$$\begin{aligned}\hat{\Phi}_t a_{\mathsf{safe}} + \omega_t(\delta)\|a_{\mathsf{safe}}\|_{V_t^{-1}}\mathbf{1}_m &= \hat{\Phi}_t a_{\mathsf{safe}} + \frac{M_t(a_{\mathsf{safe}})}{B_t}\mathbf{1}_m \\ &\le \alpha - \Gamma(a_{\mathsf{safe}})\mathbf{1}_m + \frac{2M_t(a_{\mathsf{safe}})}{B_t}\mathbf{1}_m \le \alpha - \frac{2\Gamma(a_{\mathsf{safe}})}{3}\mathbf{1}_m,\end{aligned}$$

using the consistency of the confidence sets (and the Cauchy-Schwarz inequality), along with the fact that $B_t = 1 + \max(1, B(\delta_t)) \ge 2$. Further,

$$\hat{\Phi}_t b_t + \omega_(\delta)\|b_t\|_{V_t^{-1}} \le \widetilde{\Phi}_t b_t + \frac{B_t - 1}{B_t} M_t(b_t)\mathbf{1}_m + \frac{1}{B_t} M_t(b_t)\mathbf{1}_m \le \alpha + M_t(b_t)\mathbf{1}_m.$$

Therefore,

$$\begin{aligned}&\hat{\Phi}_t(\rho b_t + (1-\rho)a_{\mathsf{safe}}) + \omega_t(\delta)\|\rho b_t + (1-\rho)a_{\mathsf{safe}}\|_{V_t^{-1}} \\ &\le \rho\left(\hat{\Phi}_t b_t + \frac{M_t(b_t)}{B_t}\mathbf{1}_m\right) + (1-\rho)\left(\hat{\Phi}_t a_{\mathsf{safe}} + \frac{M_t(a_{\mathsf{safe}})}{B_t}\mathbf{1}_m\right) \\ &\le \alpha + \left(\rho M_t(b_t) - (1-\rho)\Gamma(a_{\mathsf{safe}})/3\right)\mathbf{1}_m.\end{aligned}$$

It is straightforward to find that the additive term above is nonpositive for $\rho_0 = \frac{\Gamma(a_{\mathsf{safe}})}{\Gamma(a_{\mathsf{safe}}) + 3M_t(b_t)}$, and thus $\rho_t \ge \frac{\Gamma(a_{\mathsf{safe}})}{\Gamma(a_{\mathsf{safe}}) + 3M_t(b_t)}$.

Further, since $M_t(a_{\mathsf{safe}}) \le \Gamma(a_{\mathsf{safe}})/3$, we also have

$$\alpha - 2\Gamma(a_{\mathsf{safe}})/3 \le \alpha - 2M_t(a_{\mathsf{safe}}).$$

Thus, we can also write

$$\hat{\Phi}_t a_{\mathsf{safe}} + M_t(a_{\mathsf{safe}})/B_t\mathbf{1}_m \le \alpha - 2M_t(a_{\mathsf{safe}})\mathbf{1}_m,$$

and carrying out the same procedure then shows that

$$\rho_t \ge \frac{2M_t(a_{\mathsf{safe}})}{2M_t(a_{\mathsf{safe}}) + M_t(b_t)}.$$

To draw the final conclusions, first observe that

$$1 - \rho_t \le \frac{M_t(b_t)}{2M_t(a_{\mathsf{safe}}) + M_t(b_t)} \implies 2(1-\rho_t)M_t(a_{\mathsf{safe}}) \le \rho_t M_t(b_t) \le M_t(a_t) + (1-\rho_t)M_t(a_{\mathsf{safe}}),$$

where we used the fact that $\rho_t b_t = a_t - (1-\rho_t)a_{\mathsf{safe}}$, and that M_t is a scaling of a norm. It follows that $(1-\rho_t)M_t(a_{\mathsf{safe}}) \le M_t(a_t)$, and of course, that $\rho_t M_t(b_t) \le 2M_t(a_t)$. Further, by a similar calculation,

$$(1-\rho_t) \le \frac{3M_t(b_t)}{\Gamma(a_{\mathsf{safe}}) + 3M_t(b_t)} \implies (1-\rho_t)\Gamma(a_{\mathsf{safe}}) \le 3\rho_t M_t(b_t) \le 6M_t(a_t). \qquad \square$$

Proving the Look-Back Bound. The above control on $(1-\rho_t)$ is natural in light of terms of the form $(1-\rho_t)\Delta(a_{\mathsf{safe}})$ appearing in the bound as sketched in the main text. Let us now complete this argument.

Proof of Lemma 5. We assume $\Gamma_0 \in [\Gamma(a_{\mathsf{safe}})/2, \Gamma(a_{\mathsf{safe}})]$, and that the event of Lemma 3 holds, as well as $\mathsf{Con}(\delta)$. Together these occur with chance at least $1 - 3\delta$.

Now, we begin as in the main text, by observing that

$$\Delta(a_t) = \Delta(\rho_t b_t + (1-\rho_t)a_{\mathsf{safe}}) = \rho_t\Delta(b_t) + (1-\rho_t)\Delta(a_{\mathsf{safe}}).$$

Let $s < t$ be such that $M_s(a_{\mathsf{safe}}) \le \Gamma_0/3$ as well. Then we further know that

$$\widetilde{\Phi}_s b_s \le \alpha \implies \widetilde{\Phi}_t b_s \le \alpha + (M_t(b_s) + M_s(b_s))\mathbf{1}_m.$$

As a consequence, for

$$\sigma_{s\to t} := \frac{\Gamma(a_{\mathsf{safe}})}{\Gamma(a_{\mathsf{safe}}) + 3(M_t(b_s) + M_s(b_s))},$$

we have

$$\widetilde{\Phi}_t(\sigma_{s\to t} b_s + (1-\sigma_{s\to t})a_{\mathsf{safe}}) \le \alpha + \left(\sigma_{s\to t}(M_t(b_s) + M_s(b_s)) - \frac{2(1-\sigma_{s\to t})\Gamma(a_{\mathsf{safe}})}{3}\right)\mathbf{1}_m \le \alpha.$$

Define $\bar{b}_{s\to t} = \sigma_{s\to t} b_s + (1-\sigma_{s\to t})a_{\mathsf{safe}}$. By the above observation, $\bar{b}_{s\to t}$ is feasible for $\widetilde{\Phi}_t$, and therefore $\widetilde{\theta}_t^\top \bar{b}_{s\to t} \le \widetilde{\theta}_t^\top b_t$. To use this, we note that

$$\begin{aligned}\Delta(b_t) &= \Delta(\bar{b}_{s\to t}) + \theta_*^\top(\bar{b}_{s\to t} - b_t) = \Delta(\bar{b}_{s\to t}) + \widetilde{\theta}_t^\top(\bar{b}_{s\to t} - b_t) + (\widetilde{\theta}_t - \theta_*)^\top(\bar{b}_{s\to t} - b_t)\\ &\le \Delta(\bar{b}_{s\to t}) + \widetilde{\theta}_t^\top(\bar{b}_{s\to t} - b_t) + M_t(\bar{b}_{s\to t}) + M_t(b_t),\end{aligned}$$

where we first use Lemma 3, and then bound $M_t(\bar{b}_{s\to t} - b_t)$ by using the fact that M_t is a norm. The second term above is of course nonpositive, and so can be dropped while retaining the upper bound. Further,

$$\Delta(\bar{b}_{s\to t}) = \sigma_{s\to t}\Delta(b_s) + (1-\sigma_{s\to t})\Delta(a_{\mathsf{safe}}).$$

This leaves us with the bound

$$\begin{aligned}\Delta(a_t) \le\ &(1-\rho_t + \rho_t(1-\sigma_{s\to t}))\Delta(a_{\mathsf{safe}})\\ &+ \rho_t\left(\sigma_{s\to t}\Delta(b_s) + M_t(b_t) + \sigma_{s\to t}M_t(b_s) + (1-\sigma_{s\to t})M_t(a_{\mathsf{safe}})\right),\end{aligned}$$

where we used the triangle inequality and the fact that M_t is a scaling of a norm to write the final two terms. We will, of course, evaluate this at $s = \tau(t)$. In the subsequent, we will just write τ instead of $\tau(t)$ for the sake of reducing the density of notation. Using the fact that $\Delta(b_\tau) \le M_\tau(b_\tau)$, we set up the basic bound

$$\begin{aligned}\Delta(a_t) \le\ &(1-\rho_t + \rho_t(1-\sigma_{\tau\to t}))\Delta(a_{\mathsf{safe}})\\ &+ \rho_t M_t(b_t) + \rho_t(\sigma_{\tau\to t}(M_\tau(b_\tau) + M_t(b_\tau)) + (1-\sigma_{\tau\to t})M_t(a_{\mathsf{safe}})).\end{aligned}$$

Now, first observe that by Lemma 19,

$$(1-\rho_t)\Delta(a_{\mathsf{safe}}) \le 6\frac{\Delta(a_{\mathsf{safe}})}{\Gamma(a_{\mathsf{safe}})}M_t(a_t),$$

and further

$$\rho_t M_t(b_t) \le 2M_t(a_t).$$

We are left with terms scaling with $\sigma_{\tau\to t}$ or $(1-\sigma_{\tau\to t})$. For this, we first observe that

$$M_t(b_\tau) = B_t\omega_t\|b_\tau\|_{V_t^{-1}} \le \frac{B_t\omega_t}{B_\tau\omega_\tau}\cdot B_\tau\omega_\tau\|b_\tau\|_{V_\tau^{-1}} = \frac{B_t\omega_t}{B_\tau\omega_\tau}\cdot M_\tau(b_\tau),$$

where we use the fact that V_t is nondecreasing (in the positive definite ordering). Let us abbreviate $J_{\tau\to t} := 1 + {B_t\omega_t}/{(B_\tau\omega_\tau)}$. Upon observing that $\rho_t \le 1$, to finish the argument, we only need to control

$$(1-\sigma_{\tau\to t})(\Delta(a_{\mathsf{safe}}) + M_t(a_{\mathsf{safe}})) + J_{\tau\to t}\sigma_{\tau\to t}M_\tau(b_\tau).$$

Now, notice that since $M_\tau(a_{\mathsf{safe}}) \le \Gamma_0/3$,

$$\sigma_{\tau\to t} = \frac{\Gamma(a_{\mathsf{safe}})}{\Gamma(a_{\mathsf{safe}}) + 3(M_t(b_\tau) + M_\tau(b_\tau))} \le \frac{\Gamma(a_{\mathsf{safe}})}{\Gamma(a_{\mathsf{safe}}) + 3M_\tau(b_\tau)} \le \rho_\tau \le \frac{2M_\tau(a_\tau)}{M_\tau(b_\tau)},$$

where we invoke Lemma 19 for the final two inequalities. Thus, we find that

$$\sigma_{\tau\to t}J_{\tau\to t}M_\tau(b_\tau) \le J_{\tau\to t}\cdot\rho_\tau M_\tau(b_\tau) \le 2J_{\tau\to t}M_\tau(a_\tau).$$

This leaves us with the term $(1-\sigma_{\tau\to t})(\Delta(a_{\mathsf{safe}})+M_t(a_{\mathsf{safe}}))$. To bound this, observe that

$$(1-\sigma_{\tau\to t}) = \frac{3(M_t(b_\tau)+M_\tau(b_\tau))}{\Gamma(a_{\mathsf{safe}})+3(M_t(b_\tau)+M_\tau(b_\tau))}$$
$$\implies (1-\sigma_{\tau\to t})\Gamma(a_{\mathsf{safe}}) = 3\sigma_{\tau\to t}(M_t(b_\tau)+M_\tau(b_\tau)) \le 3\sigma_{\tau\to t}J_{\tau\to t}M_\tau(b_\tau).$$

Recall from the discussion above that $\sigma_{\tau\to t}M_\tau(b_\tau) \le \rho_\tau M_\tau(b_\tau) \le 2M_\tau(a_\tau)$. Using this, and the fact that $M_t(a_{\mathsf{safe}}) \le \Gamma(a_{\mathsf{safe}})/3$ then yields

$$(1-\sigma_{\tau\to t})(\Delta(a_{\mathsf{safe}})+M_t(a_{\mathsf{safe}})) \le 6J_{\tau\to t}\frac{\Delta(a_{\mathsf{safe}})}{\Gamma(a_{\mathsf{safe}})}M_\tau(a_\tau) + 2J_{\tau\to t}M_\tau(a_\tau).$$

Putting everything together, then, we conclude that

$$\Delta(a_t) \le 6\frac{\Delta(a_{\mathsf{safe}})}{\Gamma(a_{\mathsf{safe}})}\left(M_t(a_t)+J_{\tau\to t}M_\tau(a_\tau)\right) + 2M_t(a_t) + 4J_{\tau\to t}M_\tau(a_\tau),$$

which of course implies the bound we set out to show. □

H.3 Controlling Accumulation in the Look-Back Bound

We proceed to control the accumulation of the look-back terms.

Proof of Lemma 6. Since B_t and ω_t are nondecreasing, for any $s \le t \le T$, we have

$$(1+(B_t\omega_t(\delta)/B_s\omega_s(\delta)))M_s(a_s) = (B_s\omega_s(\delta)+B_t\omega_t(\delta))\|a_s\|_{V_s^{-1}} \le 2B_T\omega_T(\delta)\|a_s\|_{V_s^{-1}}.$$

Let $\mathcal{T}_T = \{t \le T : M_t(a_{\mathsf{safe}}) \le \Gamma_0/3\}$, and $\mathcal{U}_T = \{s \in \mathcal{T}_T : \Delta(b_s) \le M_t(b_s)\}$. Then notice that

$$\sum_{t\in\mathcal{T}_T}\|a_{\tau(t)}\|_{V_{\tau(t)}^{-1}} = \sum_{s\in\mathcal{U}_T}L_s\|a_s\|_{V_s^{-1}},$$

where $L_s = |\{t \in \mathcal{T}_T : \tau(t) = s\}|$ is the number of times s serves as $\tau(t)$ for some t. But this is the same as the time (restricted to $\mathcal{T}_T$) between s and the *next* member of $\mathcal{U}_T$, i.e., the length of the 'run' of the method playing saturated actions (plus one).

At this point, a weaker bound of the form $\frac{2}{\chi}\log(T^2/\delta)\sum_{s\in\mathcal{U}_T}\|a_s\|_{V_s^{-1}}$ is straightforward: each round has at least a chance $\chi/2$ of picking a saturated b_t, and so the chance that the kth such run has length greater than $\frac{2}{\chi}\log(k(k+1)/\delta)$ is at most $\delta/k(k+1)$. Since there are at most T runs up to time T, union bounding over this gives $\max_{\mathcal{U}_T} L_s \le 1 + 2\log(T(T+1)/\delta)/\chi$.

The rest of this proof is devoted to give a more refined martingale analysis that saves upon the multiplicative $\log(T)$ term above. We encapsulate this as an auxiliary Lemma below.

Lemma 20. *In the setting of Lemma 6, it holds that with probability at least* $1-\delta$,

$$\sum_{s\in\mathcal{U}_T}L_s\|a_s\|_{V_s^{-1}} \le \frac{5}{\chi}\Big(\sum_{s\in\mathcal{U}_T}\|a_s\|_{V_s^{-1}} + \log(1/\delta)\Big)$$

This result is shown below. Assuming this result, the original claim follows immediately, since due to the nonnegativity of $\|\cdot\|_\cdot$, $\sum_{s\in\mathcal{U}_T}\|a_s\|_{V_s^{-1}} \le \sum_{t\le T}\|a_t\|_{V_t^{-1}}$. □

To finish the argument, we move on to showing the auxiliary lemma described above.

Proof of Lemma 20. We work with the reduction to $\sum_{s\in\mathcal{U}_T}\|a_s\|_{V_s^{-1}}$ established above. Let us denote $\zeta_i = \inf\{t > \zeta_{i-1} : M_t(a_{\mathsf{safe}}) \le \Gamma_0/3, \Delta(b_t) \le M_t(b_t)\}$ as the times that an unsaturated action is picked, with $\zeta_0 := 0$—for $i : \zeta_i \le T$, these are precisely the elements of $\mathcal{U}_T$. Notice that this $\{\zeta_i\}$ is a sequence of stopping times adapted to the history $\{\mathfrak{H}_t\}$. Let us further denote $L_i = (\zeta_{i+1} - \zeta_i)$, for $i \ge 0$ (this corresponds to L_s, where $s = \zeta_i$). The object we need to control is

$$\sum_{i:\zeta_i\le T}L_iX_i,$$

where $X_i = \|a_{\zeta_i}\|_{V_{\zeta_i}^{-1}} \in [0,1]$, the lower bound being since X_i is a norm, and the upper bound since $V_{\zeta_i} \succcurlyeq I$. For notational convenience, we always set $X_0 = 1$. Now, to control this, let us first pass to the associated sigma algebrae of the ζ_i past, denoted as

$$\mathfrak{G}_i := \zeta(\mathfrak{H}_{\zeta_i}).$$

Notice that since ζ_i is nondecreasing, we know that $\{\mathfrak{G}_i\}$ forms a filtration. Of course, by definition, X_i are adapted to $\mathfrak{G}_i$, while L_i are adapted to $\mathfrak{G}_{i+1}$. We further know that L_i is the time (including ζ_i) between ζ_i and ζ_{i+1}. But then for each $t > \zeta_i$, $P(\zeta_{i+1} = t|\zeta_{i+1} > t-1, \mathfrak{H}_{t-1}) \geq \chi/2$. As a result, these L_is are conditionally stochastically domainted by a geometric random variable, i.e.,

$$\mathbb{P}(L_i > 1 + k|\mathfrak{G}_i) \leq (1-\chi/2)^k.$$

This in turn implies that for any λ small enough,

$$\mathbb{E}[e^{\lambda(L_i-1)X_i}|\mathfrak{G}_i] \leq \frac{\chi/2}{1-(1-\chi/2)e^{\lambda X_i}}.$$

In the subsequent, we will need to select a λ that is independent of all of these L_i, X_i. To ensure that the calculation makes sense, we ensure that $(1-\chi/2)e^{\lambda} \leq 1$ (which suffices since $0 \leq X_i \leq 1$). Let us define $F_i(\lambda) := -\log((1-(1-\chi/2)e^{\lambda X_i})/(\chi/2))$. Then by the above calculation, we find that the process $\{M_i\}$ with $M_0 := 1$ and

$$M_i := \exp\left(\lambda \sum (L_i - 1)X_i - \sum F_i(\lambda)\right)$$

is a nonnegative supermartingale with respect to the filtration $\{\mathfrak{F}_i\}$ with $\mathfrak{F}_i = \mathfrak{G}_{i+1}\}$ and $\mathfrak{F}_0$ defined to be the trivial sigma algebra. Thus, by Ville's inequality, $\mathbb{P}(\exists i : M_i > 1/\delta) \leq \delta$. Taking logarithms, we find that with probability at least $1-\delta$, it holds that

$$\forall n, \sum_{i \leq n} L_i X_i \leq \sum_{i \leq n} X_i + \frac{\log(1/\delta)}{\lambda} + \sum_{i \leq n} \frac{F_i(\lambda)}{\lambda},$$

as long as $0 < \lambda < -\log(1-\chi/2)$. All we need now is a convenient bound on $F_i(\lambda)$ and a judicious choice of λ. To this end, we observe the following simple result.

Lemma 21. *For any constant $u \in (0,1)$, consider the map $f(x) := -\log \frac{1-ue^x}{1-u}$ over the domain $[0, -\log(u))$. Then for all $x \in [0, -\frac{1}{2}\log(u)]$, we have*

$$f(x) \leq \frac{\sqrt{u}}{1-\sqrt{u}} x.$$

Proof. Observe that

$$f'(x) = \frac{ue^x}{1-ue^x} = \frac{e^f}{1-u}(1-e^{-f}(1-u)) = \frac{e^f}{1-u} - 1 \geq 0$$

The inequalities above arise since $e^f = \frac{1-u}{1-ue^x} > 1-u$ using the fact that $e^x \geq 1$. By taking another derivative, we may see that f' itself is an increasing function. Now, suppose $g(x)$ satisfies

$$g(0) = f(0) = 0, \text{ and } \forall x, g'(x) = f'(-1/2\log(u)) = \frac{\sqrt{u}}{1-\sqrt{u}}.$$

Then since $g'(x) \geq f'(x)$ for all $x \in [0, -\frac{1}{2}\log(u)]$, by the fundamental theorem of calculus it follows that for all $x \leq -\frac{1}{2}\log u$, $f(x) = \int_0^x f' \leq \int_0^x g' = g(x)$. □

Now, of course, $F_i(\lambda) = f(\lambda X_i)$, with $u = 1-\chi/2$. Then setting $\lambda = -\frac{1}{2}\log(1-\chi/2)$, we have

$$\forall n, \sum_{i \leq n} L_i X_i \leq \sum X_i + \frac{\log(1/\delta)}{-\log(1-\chi/2)/2} + \sum_{i \leq n} \frac{\sqrt{1-\chi/2}}{1-\sqrt{1-\chi/2}} X_i.$$

To get the form needed, we observe that

$$\frac{\sqrt{1-v}}{1-\sqrt{1-v}} \le \frac{2}{v} \iff (2+v)^2(1-v) \le 4 \iff -v^3 - 3v^2 \le 0,$$

and of course $-\log(1-v)/2 \ge v/2$. Plugging in $v = \chi/2 > 0$, we end up at

$$\forall n, \sum_{i \le n} L_i X_i \le \left(1 + \frac{4}{\chi}\right) \sum_{i \le n} X_i + \frac{4}{\chi} \log(1/\delta).$$

Note that no explicit n-dependent term appears in the above. This makes sense: we essentially have the X_is acting as 'time steps', and so $\sum L_i X_i$ should behave as $(1 + 2/\chi) \sum X_i + O(\sqrt{\sum X_i \log(1/\delta)/\chi} + \log(1/\delta))$, via a Bernstein-type computation. In our case, the square root terms do not meaningfully help the solution,[7] and so we just pick a convenient λ instead.[8] Now, going back to our original object of study, we have $L_i = L_{\zeta_i}, X_i = \|a_{\zeta_i}\|_{V_{\zeta_i}^{-1}}$, and these ζ_is are precisely the members of $\mathcal{U}_T$, so we conclude that

$$\forall T, \sum_{s \in \mathcal{U}_T} L_s X_s \le \frac{5}{\chi} \left(\sum_{s \in \mathcal{U}_t} X_s + \log(1/\delta) \right). \qquad \square$$

H.4 Regret and Risk Bounds for S-COLTS

With the above pieces in place, we move on to showing the final bounds on the behaviour of S-COLTS.

Proof of Theorem 8. We first argue the safety properties. Firstly, in the exploration phase, as well as to explore, we repeatedly play a_{safe}. But this is, by definition, safe, and so accrues no safety cost. When not playing a_{safe}, the selected action a_t at time t satisfies

$$\hat{\Phi}_t a_t + \omega_t(\delta) \|a_t\|_{V_t^{-1}} \mathbf{1}_m \le \alpha.$$

But, given the consistency event $\mathsf{Con}_t(\delta)$,

$$\forall a, \Phi_* a \le \hat{\Phi}_t a + \omega_t(\delta) \|a\|_{V_t^{-1]}},$$

and so $\Phi_* a_t \le \alpha$. Since $\mathsf{Con}(\delta) := \bigcap \mathsf{Con}_t(\delta)$ holds with chance at least $1-\delta$, it follows that a_t is safe at every t, and a fortiori, $\mathbf{S}_T = 0$ for every T.

Let us turn to the regret analysis. Fix any T. We break the regret analysis into four pieces: the regret accrued over the initial exploration, that accrued after this phase, but when $M_t(a_{\mathsf{safe}}) > \Gamma_0/3$, and over the time $\mathcal{T}_T := \{t \ge T_0 : M_t(a_{\mathsf{safe}}) \le \Gamma_0/3\}$, and finally the regret incurred up to the time $\inf\{t : \delta_t > \chi/2\}$.

The last of these is the most trivial to handle: the number of such rounds is bounded as $\sqrt{2\delta/\chi}$, and the regret in any round is at most 2.

For the first case, Lemma 17 ensures that with probability at least $1-\delta$, this phase has length at most

$$\frac{8}{\Gamma(a_{\mathsf{safe}})^2} \log(8/(\delta\Gamma(a_{\mathsf{safe}})^2)),$$

and further, the output Γ_0 is at least $\Gamma(a_{\mathsf{safe}})/2$ at the end. Using this to instantiate Lemma 18, we further find that the number of times a_{safe} is selected beyond this initial exploration is in total bounded as

$$1 + \frac{36\omega_T^2 B_T^2}{\Gamma(a_{\mathsf{safe}})^2}.$$

Together these contribute at most

$$\Delta(a_{\mathsf{safe}}) \cdot \frac{44\omega_T^2 B_T^2}{\Gamma(a_{\mathsf{safe}})^2} \log(8/\delta\Gamma(a_{\mathsf{safe}})^2)$$

[7] since there will always be an additive $\log(1/\delta)$ and $\frac{1}{\chi}\sum X_i$ term

[8] and in the process, avoid the subtleties of the dependence of λ on the X_is if we optimised it

to the regret.

This leaves us with the times at which $M_t(a_{\mathsf{safe}}) \le \Gamma_0/3$, for which we apply Lemma 5, along with the control of Lemma 6 to find that the net regret accrued thus is bounded as

$$O\left(\left(1 + \frac{\Delta(a_{\mathsf{safe}})}{\Gamma(a_{\mathsf{safe}})}\right) B_T \omega_T(\delta) \cdot \frac{5}{\chi}\left(\sum_{t \le T} \|a_t\|_{V_t^{-1}} + \log(1/\delta)\right)\right).$$

To complete the book-keeping, the probabilistic events required for this are the consistency of the confidence sets, that for all $t, \max(\|\eta_t\|, \max_i \|H_t^i\|)$ is bounded by $B(\delta_t)$, and of course the bound on the times between unsaturated b_t being constructed from Lemma 6. Together, these occur with chance at least $1 - 3\delta$, and putting the same together with the stopping time bound, we conclude that with chance at least $1 - 4\delta$, S-COLTS(μ, δ) satisfies the regret bound

$$\mathbf{R}_T \le \left(1 + \frac{\Delta(a_{\mathsf{safe}})}{\Gamma(a_{\mathsf{safe}})}\right) \widetilde{O}\left(\frac{\omega_t(\delta) B_T}{\chi} \sum_{t \le T} \|a_t\|_{V_t^{-1}}\right) + \frac{\Delta(a_{\mathsf{safe}})}{\Gamma(a_{\mathsf{safe}})} \cdot \widetilde{O}\left(\frac{\omega_T^2 B_T^2}{\Gamma(a_{\mathsf{safe}})}\right) + \sqrt{\frac{8\delta}{\chi}}.$$

Now, invoking Lemma 13, we can bound $\omega_T(\delta) = \widetilde{O}(\sqrt{d} + \log(m/\delta))$, and $\sum \|a_t\|_{V_t^{-1}} = \widetilde{O}(\sqrt{dT})$. Finally, for the law μ induced via the coupled noise design by $\mathrm{Unif}(\sqrt{3d}\mathbb{S}^d)$, we further know that $B_T = O(\sqrt{d})$ and $\chi \ge 0.28$. Of course, for this noise, $B_t = \sqrt{3d}$ with certainty, which boosts the probability above to $1 - 3\delta$. The claim thus follows for S-COLTS$(\mu, \delta/3)$. □

H.5 An Optimism-Based Analysis of S-COLTS

We analyse S-COLTS under the assumption that μ satisfies B-concentration and π-global optimism (Definition 12). We shall be somewhat informal in executing this.

Setting Up. We first note that regret accrued over rounds in which $M_t(b_t) > \Gamma(a_{\mathsf{safe}})/3$ and $M_t(a_{\mathsf{safe}}) \le \Gamma_0/3$ is small. Indeed,

$$\begin{aligned}\sum_{t \in \mathcal{T}_T} \mathbb{1}\{M_t(b_t) > \Gamma(a_{\mathsf{safe}})/3\} &\le \frac{9}{\Gamma(a_{\mathsf{safe}})^2} \sum_{t \in \mathcal{T}_T} M_t(b_t)^2 \\ &\le \frac{16}{\Gamma(a_{\mathsf{safe}})^2} \sum_{t \in \mathcal{T}_T} \frac{M_t(a_t)^2}{\rho_t^2} = \widetilde{O}\left(\frac{d^5}{\Gamma(a_{\mathsf{safe}})^4}\right),\end{aligned}$$

where $\mathcal{T}_T = \{t : M_t(a_{\mathsf{safe}}) \le \Gamma_0/3\}$, and we used the bound on $M_t(b_t)$ from Lemma 19, along with the fact that since $b_t \in \mathcal{A}$, $M_t(b_t) \le B_t \omega_t = \widetilde{O}(d)$, which in turn implies that $\rho_t \ge \Gamma(a_{\mathsf{safe}})/\widetilde{\Omega}(d)$. Naturally, this additive term is much weaker than that seen in Theorem 8. Nevertheless, the optimism-based framework does recover a similar main term. In particular we will show a regret bound of $\widetilde{O}(\Gamma(a_{\mathsf{safe}})^{-1}\sqrt{d^3 T})$

The point of the above condition is that (using Lemma 19), if $M_t(b_t) \le \Gamma(a_{\mathsf{safe}})/3$, then $\rho_t \ge \frac{1}{2}$. We will repeatedly use this fact in the subsequent.

Now, we begin similarly to the previous analysis by using

$$\Delta(a_t) = (1 - \rho_t)\Delta(a_{\mathsf{safe}}) + \rho_t \Delta(b_t) \le (1 - \rho_t)\Delta(a_{\mathsf{safe}}) + \Delta(b_t).$$

The first term is well-controlled, as detailed in the proof of Lemma 5. So, we only need to worry about $\sum \Delta(b_t)$. Notice that for this it suffices to control $\sum \mathbb{E}[\Delta(b_t)|\mathfrak{H}_{t-1}]$. Indeed, $\Delta(b_t) \le 1$ (and if it is ≤ 0, we can just drop it from the sum, i.e., we could study $(\Delta(b_t))_+$ instead with no change in the argument), so the difference $\sum_{t \le T} \Delta(b_t) - \mathbb{E}[\Delta(b_t)|\mathfrak{H}_{t-1}]$ is a martingale with increments lying in $[-1, 1]$, and the LIL (Lemma 16) ensures that for all T simultaneously, the difference between these is $O(\sqrt{T \log(\log(T)/\delta)})$ with chance at least $1 - \delta$.

From the above, then, we can restrict attention to t such that $M_t(a_{\mathsf{safe}}) \le \Gamma_0/3, \rho_t \ge \frac{1}{2}$. Finally, recalling the notation $K(\theta, \Phi) = \max\{\theta^\top a : a \in \mathcal{A}, \Phi a \le \alpha\}$ from Definition 12, we observe that

$$\begin{aligned}\Delta(b_t) &= \theta_*^\top a_* - \widetilde{\theta}_t^\top b_t + (\widetilde{\theta}_t - \theta_*)^\top b_t \\ &\le K(\theta_*, \Phi_*) - K(\widetilde{\theta}_t, \widetilde{\Phi}_t) + M_t(b_t) \\ &\le K(\theta_*, \Phi_*) - K(\widetilde{\theta}_t, \widetilde{\Phi}_t) + 4M_t(a_t),\end{aligned}$$

where we used Lemma 3, and Lemma 19 along with the fact that $\rho_t \ge 1/2$. Now note that the final term above is summable to $\widetilde{O}(\sqrt{d^3T})$. Thus, it equivalently suffices to analyse the behaviour of $\mathbb{E}_{t-1}[K(\theta_*, \Phi_*) - K(\widetilde{\theta}_t, \widetilde{\Phi}_t)|\mathfrak{H}_{t-1}]$. In order to do so, we begin with a 'symmetrisation' lemma.

Lemma 22. *Let $(\widetilde{\theta}_t, \widetilde{\Phi}_t)$ and $(\bar{\theta}_t, \bar{\Phi}_t)$ denote two independent copies of parameter perturbations at time t. Let $\mathbb{E}_{t-1}[\cdot] := \mathbb{E}[\cdot \mid \mathfrak{H}_{t-1}]$. If μ satisfies π-global optimism, then*

$$\mathbb{1}_{\mathsf{Con}_t(\delta)}\mathbb{E}_{t-1}[(K(\theta_*, \Phi_*) - K(\widetilde{\theta}_t, \widetilde{\Phi}_t)] \le \mathbb{1}_{\mathsf{Con}_t(\delta)} \cdot \frac{1}{\pi}\mathbb{E}_{t-1}[|K(\widetilde{\theta}_t, \widetilde{\Phi}_t) - K(\bar{\theta}_t, \bar{\Phi}_t)|].$$

Proof. Let $\bar{\mathsf{G}} := \{K(\bar{\theta}_t, \bar{\Phi}_t) \ge K(\theta_*, \Phi_*)\}$. Since $K(\theta_*, \Phi_*)$ is a constant, and since $(\widetilde{\theta}_t, \widetilde{\Phi}_t)$ are independent of $(\bar{\theta}_t, \bar{\Phi}_t)$ given $\mathfrak{H}_{t-1}$, we conclude that

$$\mathbb{E}_{t-1}[K(\theta_*, \Phi_*) - K(\widetilde{\theta}_t, \widetilde{\Phi}_t)] = \mathbb{E}_{t-1}[K(\theta_*, \Phi_*) - K(\widetilde{\theta}_t, \widetilde{\Phi}_t) \mid \bar{\mathsf{G}}].$$

But given $\bar{\mathsf{G}}$, $K(\theta_*, \Phi_*) \le K(\bar{\theta}_t, \bar{\Phi}_t)$, and so

$$\begin{aligned}\mathbb{E}_{t-1}[K(\theta_*, \Phi_*) - K(\widetilde{\theta}_t, \widetilde{\Phi}_t)] &\le \mathbb{E}_{t-1}[K(\bar{\theta}_t, \bar{\Phi}_t) - K(\widetilde{\theta}_t, \widetilde{\Phi}_t) \mid \bar{\mathsf{G}}] \\ &\le \mathbb{E}_{t-1}[|K(\bar{\theta}_t, \bar{\Phi}_t) - K(\widetilde{\theta}_t, \widetilde{\Phi}_t)| \mid \bar{\mathsf{G}}].\end{aligned}$$

Finally, for any nonnegative random variable X, and any event E, it holds that

$$\mathbb{E}_{t-1}[X|\mathsf{E}]\mathbb{E}_{t-1}[\mathbb{1}_{\mathsf{E}}] = \mathbb{E}_{t-1}[X\mathbb{1}_{\mathsf{E}}] \le \mathbb{E}_{t-1}[X].$$

The claim follows upon taking $X = |K(\bar{\theta}_t, \bar{\Phi}_t) - K(\widetilde{\theta}_t, \widetilde{\Phi}_t)|$, $\mathsf{E} = \bar{\mathsf{G}}$, and recognising that due to π-optimism, $\bar{\mathsf{G}}$ satisfies $\mathbb{E}_{t-1}[\mathbb{1}_{\bar{\mathsf{G}}}]\mathbb{1}_{\mathsf{Con}_t} \ge \pi\mathbb{1}_{\mathsf{Con}_t}$. □

The main question now becomes controlling how far the deviations in K can go. We control this using a similar scaling trick as in the proof of Lemma 5.

For the sake of clarity, we will denote the optimiser of $K(\widetilde{\theta}_t, \widetilde{\Phi}_t)$ as $\tilde{b}_t$ (instead of just b_t as in the rest of the text), and similarly that of $K(\bar{\theta}_t, \bar{\Phi}_t)$ as $\bar{b}_t$. Our goal is to control (the conditional mean of)

$$|\bar{\theta}_t^\top \bar{b}_t^\top - \widetilde{\theta}_t^\top \tilde{b}_t|.$$

Naturally, the core issue remains that $\bar{b}_t$ and $\tilde{b}_t$ are optima in distinct feasible sets, and so it is hard to, e.g., compare $\widetilde{\theta}_t^\top \tilde{b}_t$ and $\widetilde{\theta}_t^\top \bar{b}_t$. To this end, we observe that

$$\bar{\Phi}_t\bar{b}_t \le \alpha \implies \Phi_* b_t \le \alpha + M_t(\bar{b}_t)\mathbf{1}_m \implies \widetilde{\Phi}_t\bar{b}_t \le \alpha + 2M_t(b_t)\mathbf{1}_m,$$

as long as consistency and the boundedness of the noise norms holds (which occurs with high probability). Using this and the fact that $\widetilde{\Phi}_t a_{\mathsf{safe}} \le \alpha - 2\Gamma(a_{\mathsf{safe}})/3\mathbf{1}_m$, we find that

$$\widetilde{\Phi}_t(\bar{\sigma}_t\bar{b}_t + (1-\bar{\sigma}_t)a_{\mathsf{safe}}) \le \alpha, \text{ where } \bar{\sigma}_t = \frac{\Gamma(a_{\mathsf{safe}})}{\Gamma(a_{\mathsf{safe}}) + 3M_t(\bar{b}_t)}.$$

Thus, we may write

$$\bar{\theta}_t^\top \bar{b}_t - \widetilde{\theta}_t^\top \tilde{b}_t = (1-\bar{\sigma}_t)\bar{\theta}^\top \bar{b}_t + \bar{\sigma}_t(\bar{\theta}_t - \widetilde{\theta}_t^\top)\bar{b}_t + \widetilde{\theta}_t^\top(\bar{\sigma}_t\bar{b}_t - \tilde{b}_t).$$

Above, the third term is nonpositive, while the second term may be bounded by $2\bar{\sigma}_t M_t(\bar{b}_t)$, which can further be bounded by $8M_t(\bar{a}_t)$ upon recalling that $\rho_t(\bar{b}_t) \ge \frac{1}{2}$ and the bound on $\rho_t M_t(b_t)$ in Lemma 19. This leaves the first term. It is tempting to bound this directly via $\bar{\theta}_t^\top \bar{b}_t \le \|\bar{\theta}_t\|\|\bar{b}_t\|$, but notice that the former can be as large as $B_t \sim \sqrt{d}$. Instead, we can use the related bound

$$(1-\bar{\sigma}_t)(\bar{\theta}^\top \bar{b}_t) \le (1-\bar{\sigma}_t)M_t(\bar{b}_t) + (1-\bar{\sigma}_t)\theta_*^\top \bar{b}_t.$$

Now notice that $(1-\bar{\sigma}_t) \le 1$, and $M_t(\bar{b}_t) \le 4M_t(\bar{a}_t)$ controls the first term. Similarly, $\theta_*^\top \bar{b}_t \le 1$ (both have norm bounded by 1), so the second term is bounded by $1 - \bar{\sigma}_t \le \frac{3M_t(\bar{b}_t)}{\Gamma(a_{\mathsf{safe}})} \le 12\frac{M_t(\bar{a}_t)}{\Gamma(a_{\mathsf{safe}})}$. Putting these together, we conclude that

$$(1-\bar{\sigma}_t)(\bar{\theta}_t^\top \bar{b}_t) \le 4M_t(\bar{a}_t) + \frac{12M_t(\bar{a}_t)}{\Gamma(a_{\mathsf{safe}})},$$

which in turn yields the bound

$$K(\bar{\theta}_t, \bar{\Phi}_t) - K(\widetilde{\theta}_t, \widetilde{\Phi}_t) \le 12 M_t(\bar{a}_t) + \frac{12 M_t(\bar{a}_t)}{\Gamma(a_{\mathsf{safe}})} \le \frac{24 M_t(\bar{a}_t)}{\Gamma(a_{\mathsf{safe}})}.$$

Of course, switching the roles of $(\bar{\theta}_t, \bar{\Phi}_t)$ and $(\widetilde{\theta}_t, \widetilde{\Phi}_t)$, we have an analogous bound on $K(\widetilde{\theta}_t, \widetilde{\Phi}_t) - K(\bar{\theta}_t, \bar{\Phi}_t)$. Putting these together, we conclude that

$$|K(\bar{\theta}_t, \bar{\Phi}_t) - K(\widetilde{\theta}_t, \widetilde{\Phi}_t)| \le \frac{24(M_t(\bar{a}_t) + M_t(\tilde{a}_t))}{\Gamma(a_{\mathsf{safe}})}.$$

Finally, notice that $\bar{a}_t$, $\tilde{a}_t$, and the actually selected action a_t all have the same distribution given $\mathfrak{H}_{t-1}$. We can thus conclude that

$$\mathbb{E}_{t-1}[|K(\bar{\theta}_t, \bar{\Phi}_t) - K(\widetilde{\theta}_t, \widetilde{\Phi}_t)|] \le 48 \mathbb{E}_{t-1}\left[\frac{M_t(a_t)}{\Gamma(a_{\mathsf{safe}})}\right].$$

With this in hand, the issue returns to one of concentration. We know that $\sum M_t(a_t)$ is $\widetilde{O}(\sqrt{d^3 T})$, and each $M_t(a_t)$ is bounded as $O(d)$ and so $\sum M_t(a_t) - \mathbb{E}_{t-1}[M_t(a_t)]$ enjoys concentration at the scale $d\mathrm{LIL}(T, \delta) = \widetilde{O}(\sqrt{d^2 T}) = o(\sqrt{d^3 T})$. Thus, passing back to the the unconstrained sums, we end up with a bound of the form

$$\mathbf{R}_T = \widetilde{O}\left(\Gamma(a_{\mathsf{safe}})^{-1}\sqrt{d^3 T}\right) + \widetilde{O}(d^5 \Gamma(a_{\mathsf{safe}})^{-4}).$$

The main loss in the main term above is that instead of a $\Delta(a_{\mathsf{safe}})/\Gamma(a_{\mathsf{safe}})$, we just have a $\Gamma(a_{\mathsf{safe}})^{-1}$ term in the bound. This can be lossy, e.g., when a_{safe} is very close to a_*, but in the regime $\Delta(a_{\mathsf{safe}}) = \Omega(1)$, it recovers essentially the same guarantees as Theorem 8, albeit with a weaker additive term.

I The Analysis of Soft Constraint Enforcement Methods.

I.1 The Analysis of R-COLTS

Let us first show the optimism result for R-COLTS

Proof of Lemma 9. Fix any t, and assume $\mathsf{Con}_t(\delta)$. For each $i \in [1 : I_t]$, we know that $K(i,t) := K(\widetilde{\theta}(i,t), \widetilde{\Phi}(i,t)) \ge \widetilde{\theta}(i,t)^\top a_* \ge \theta_*^\top a_*$ whenever the event L occurs, and thus this inequality holds with chance at least π in every round. Since the draws are all independent given $\mathfrak{H}_{t-1}$, the chance that $\max K(i,t) < \theta_*^\top a_*$ is at most $(1-\pi)^{I_t} \le \exp(-\log(1/\delta_t) r \cdot \pi) \le \delta_t = {}^{\delta}/{}_{t(t+1)}$. Thus, if we assume that $\mathsf{Con}(\delta) := \bigcap \mathsf{Con}_t(\delta)$ holds true, the chance that at any t, $K(i_t, t) < \theta_*^\top a_*$ is at most $\sum \delta_t = \delta$. By Lemma 1, $\mathsf{Con}(\delta)$ holds with chance at least $1 - \delta$, and we are done. □

Of course, the above proof, and thus the statement of this Lemma, holds verbatim if we replace L_t by G_t (Definition 12).

With the optimism result of Lemma 9, the argument underlying Theorem 10 is extremely standard.

Proof of Theorem 10. Assume consistency, and that at every t, $\widetilde{\theta}_t^\top a_t \ge \theta_*^\top a_*$. Since we sample at most $2 + r\log(1/\delta_t)$ programs in round t, we further know that with probability at least $1 - \delta$,

$$\forall t, \max_i \left(\max \|\eta(i,t)\|, \max_j \|H_t^j(i,t)\| \right) \| \le \beta_t := B(\delta_t/(2 + r\log(1/\delta_t)).$$

Assume that this too occurs, and define $\tilde{M}_t(a) = \omega_t(\delta)(1+\beta_t)\|a\|_{V_t^{-1}}$. Then, using consistency,

$$\theta_*^\top a_t \ge \widetilde{\theta}_t^\top a_t - \tilde{M}_t(a_t), \Phi_* a_t \le \widetilde{\Phi}_t a_t + \tilde{M}_t(a_t)\mathbf{1}_m \le \alpha + \tilde{M}_t(a_t)\mathbf{1}_m.$$

So, the safety risk is bounded as

$$\mathbf{S}_T \le \sum_t \tilde{M}_t(a_t) \le \omega_T(\delta)(1+\beta_T)\sum_{t \le T} \|a_t\|_{V_t^{-1}}.$$

Further,

$$\theta_*^\top a_* - \theta_*^\top a_t \le \theta_*^\top a_* - \widetilde{\theta}_t^\top a_t + \tilde{M}_t(a_t),$$

which implies that

$$\mathbf{R}_T \le \omega_T(\delta)(1 + \beta_T \sum_{t \le T} \|a_t\|_{V_t^{-1}}$$

as well. Now Lemma 13 controls $\omega_T \sum_{t\le T} \|a_t\|_{V_t^{-1}}$ to $\widetilde{O}(\sqrt{d^2T})$, and for our selected noise, the coupled design driven by $\mathrm{Unif}(\sqrt{3d}\mathbb{S}^d)$, we have $B(\cdot) = \sqrt{3d}$ independently of t, r, δ, and thus $\beta_T = \sqrt{3d}$. The events needed to show the above were the consistency, the concentration of the sampled noise to β_t at each time t, and the optimism event of Lemma 9. Again, the second happens with certainty for us, and so the above bounds hold at all T with chance at least $1 - 2\delta$. Consequently, the result was stated for R-COLTS$(\mu, r, \delta/2)$. □

I.2 The Exploratory-COLTS Method

As discussed in §5, the Exploratory COLTS, or E-COLTS method, augments COLTS with a low-rate of flat exploration, and exploits the resulting (eventual) perturbed feasibility of actions with nontrivial safety margin to bootstrap the scaling-based analysis of S-COLTS to a soft-enforcement result without resampling.

The main distinction lies, of course, in the fact that in the soft enforcement setting, we do not have access to a given safe action a_{safe}. To motivate the method, let us consider how S-COLTS uses the knowledge of a_{safe}. This occurs in three ways: to ensure the existence of $a(\eta_t, H_t, t)$, to compute the action a_t from this, and to enable the look-back analysis of Lemma 5. The second use is easy to address: we will simply play $a_t = a(\eta_t, H_t, t)$ if it exists. The key observation is that rather than explicit knowledge of any one particular safe action, as long as *some action* a exists such that $M_t(a) \le \Gamma(a)/3$, the entirely of the first and third uses can be recovered, and so the machinery of §4 can be enabled.

Forced Exploration. We enable the *eventual* existence of such actions by introducing a small rate of *forced exploration* in our method E-COLTS. Concretely, we demand a 'κ-good' exploration policy over $\mathcal{A}$, i.e., one such that after N exploratory actions $e_1, \cdots, e_N$, we are assured that $\sum e_i e_i^\top \succcurlyeq \kappa \lfloor N/d \rfloor I_d$, where $\kappa > 0$ is a constant. This can, e.g., be done by playing the elements of a barycentric spanner of $\mathcal{A}$ in round-robin [AK08; DHK08]. The resulting κ is a geometric property of $\mathcal{A}$, and we note that κ only enters the analysis, not the algorithm.

Algorithm 3 Exploratory-COLTS (E-COLTS(μ, δ))

Input: μ, δ, exploration policy.
Initialise: $u_0 \leftarrow 0, B_t \leftarrow 1 + B(\delta_t)$
for $t = 1, 2, \dots$ **do**
  Draw $(\eta_t, H_t) \sim \mu$.
  if $u_{t-1} \le B_t\omega_t(\delta)\sqrt{dt}$ OR $a(\eta_t, H_t, t)$ does not exist **then**
    Pick a_t via exploration policy.
    $u_t \leftarrow u_{t-1} + 1$.
  else
    $a_t \leftarrow a(\eta_t, H_t, t), u_t \leftarrow u_{t-1}$.
  Play a_t, observe R_t, S_t, update $\mathfrak{H}_t$.

Let us call a time step t where the exploratory policy is executed an 'E-step'. In E-COLTS, we ensure tha tat any t, at least $B_t\omega_t\sqrt{dt}$ such E-steps have been performed, and if not, we force an E-step. Note that we expect that the majority of the learning process occurs at steps other than E-steps, since this is where the informative action $a(\eta_t, H_t, t)$ is played. Consequently, we will call such steps 'L-steps'.

By our requirement of enough E-steps, at any L-step t, the sample second moment matrix V_t satsifies $V_t \succcurlyeq \kappa B_t \omega_t \sqrt{t/d} I_d$, and so,

$$\forall a, M_t(a) \le \psi(t) := \left(\frac{dB_t^2\omega_t^2}{\kappa^2 t}\right)^{1/4} \cdot \|a\|.$$

This means that at such t, any a with $\Gamma(a) > 2\psi(t)/3$ satisfies $M_t(a) \le \Gamma(a)/3$, and so $a(\eta_t, H_t, t)$ exists, and we may use the analysis of §4 for such a.

Regret Bound. The above insight is the main driver of the result of Theorem 11, which we show in §I.2.1 to follow. Recall that this states that under the E-COLTS strategy, executed with a μ constructed through the coupled noise design with base measure $\mathrm{Unif}(\sqrt{3d}\mathbb{S}^d)$, the risk and regret satisfy, with high probability, the bounds

$$\mathbf{S}_T = \widetilde{O}(\sqrt{d^3T}) + \min_a \widetilde{O}\Big(\frac{d^3\|a\|^4}{\kappa^2\Gamma(a)^4}\Big), \textit{ and}$$
$$\mathbf{R}_T = \min_{a:\Gamma(a)>0}\left\{\mathcal{R}(a)\widetilde{O}(\sqrt{d^3T}) + \widetilde{O}\Big(\frac{d^3\|a\|^4}{\kappa^2\Gamma(a)^4}\Big)\right\},$$

where κ is precisely the 'goodness-factor' of the exploratory policy. Let us briefly discuss this result.

Risk bound. Unlike S-COLTS, E-COLTS suffers nontrivial risk, which is unavoidable due to the lack of knowledge of a_{safe} [PGBJ21]. The $\widetilde{O}(\sqrt{d^3T})$ risk above above is comparable to the $\widetilde{O}\sqrt{d^2T})$ risk of the prior soft enforcement method DOSS [GCS24], with a $\sqrt{d}$ loss again attributable to efficiency. Note that compared to R-COLTS, the risk bound is essentially the same, but now incurs an extra additive term scaling, essentially, with $(\max_a \Gamma(a))^{-4}$. Thus, a nontrivial risk bound is only shown if this maximum is strictly positive, i.e., under Slater's condition. Nevertheless, the term is additive, and scales with T only logarithmically (through a dependence on $\omega_t(\delta)$, and so in typical scenarios is not expected to dominate as T diverges, although the fourth-power dependence on this quantity would increase the 'burn-in' time of this result.

Regret bound. As discussed in §5, the main term of the regret bound above improves over that of S-COLTS, since it *minimises* over $\mathcal{R}(a)$, rather than working with the arbitrary $\mathcal{R}(a_{\mathsf{safe}})$. Note that finding the minimiser of $\mathcal{R}$ may be challenging, but E-COLTS nevertheless adapts to this. However, the additive lower-order term is larger than in S-COLTS due to the 'flat' exploration of E-COLTS, and its practical effect is unclear. In simple simulations, we do observe a significant regret improvement (§J). We note that the κ-good exploration condition only affects the lower order term in $\mathbf{R}_T$, although again the fourth order dependence on $\Gamma(a)$ is nontrivial. Of course, relative to E-COLTS, the result suffers from an instance-dependence, and again, unless Slater's condition is satisfied, it is ineffective.

Practical Role of Forced Exploration. E-COLTS uses forced exploration to ensure that V_t is large, which leads to both feasibility of the perturbed program, and the scaling-based analysis. In practice, however, one expects that low-regret algorithms satisfy $\max_a \|a\|_{V_t^{-1}} \lesssim t^{-1/4}\|a\|$ directly, the idea being that actions with larger V_t^{-1}-norm represent underexplored directions that would naturally be selected (recent work has made strides towards actually proving such a result, although it does not quite get there [BGCG23]). Thus we believe that this forced exploration can practically be omitted except when the perturbed program is infeasible. Indeed, in simulations, we find that this strategy already has good regret (§J).

I.2.1 The Analysis of E-COLTS

We will essentially reuse our analysis of S-COLTS, with slight variations.

Proof of Theorem 11. We will first discuss the bound on the regret. Throughout, we assume consistency, and the noise concentration event of Lemma 3. We will further just write ω_t instead of $\omega_t(\delta)$. Recall the terminology that every t in which we pick an action according to the exploratory policy is called an 'E-step', and every other step an 'L-step'. Here E and L stand for exploration and learning respectively, the idea being that the former constitute the basic exploration required to enable feasibility under perturbations, and so the main learning process occurs in L-steps.

Note that the number of E-steps up to time t is explicitly delineated to be at most $\lceil B_t\omega_t\sqrt{dt}\rceil$. Using the κ-good assumption, then, we find that at every L-step,

$$V_t \succcurlyeq \kappa B_t\omega_t\sqrt{t/d}I \iff (\kappa B_t\omega_t\sqrt{t/d})^{-1}I \succcurlyeq V_t^{-1}.$$

Now, fix any action a_0 with $\Gamma(a_0) > 0$. Then notice that at any L-step,

$$\|a_0\|^2_{V_t^{-1}} \le \frac{\sqrt{d}\|a_0\|^2}{\kappa B_t\omega_t\sqrt{t}} \implies M_t(a_0)^2 \le \frac{B_t\omega_t\|a_0\|^2}{\kappa}\cdot\sqrt{d/t}.$$

Thus, for all

$$t \ge t_0(a_0) := \inf\left\{t : \frac{3^4 d\|a_0\|^4 B_t^2\omega_t^2}{\kappa^2\Gamma(a_0)^4} \le t\right\}$$

that are L-steps, we know that as long as the noises η_t, H_t satisfy the bound of Lemma 3, $\widetilde{\Phi}_t a_0 \le \alpha - 2\Gamma(a_0)/3\mathbf{1}_m$. Note that since $\omega_t^2 \le d\log(t) + \log(m/\delta)$, and since under our choice of coupled noise, $B_t = \sqrt{3d}$ for all t, we can conclude that

$$\begin{aligned} t_0(a_0) &\le \frac{Cd^3\|a_0\|^4}{\kappa^2\Gamma(a_0)^4}\log\frac{Cd^3\|a_0\|^4}{\kappa^2\Gamma(a_0)^4} + \frac{Cd^2\|a_0\|^4\log(m/\delta)}{\kappa^2\Gamma(a_0)^4}\log\frac{Cd^2\|a_0\|^4\log(m/\delta)}{\kappa^2\Gamma(a_0)^4} \\ &= \widetilde{O}\left(\frac{d^3\|a_0\|^4}{\kappa^2\Gamma(a_0)^4}\right), \end{aligned}$$

where C is some large enough constant ($C = 4\cdot 81$ suffices). This implies that at all $t > t_0(a_0)$ at which the number of E-steps, u_t, is large enough, the perturbed program is feasible, and a_t exists. Thus, after this time, no extraneous E-steps are accrued due to infeasibility of the perturbed program.

At this point we apply the proof of Lemma 5, with $\rho_t = 1$. Let

$$\tau = \tau(t) = \sup\{s \le t : \Delta(a_s) \le M_t(a_s), M_t(a_0) \le \Gamma(a_0)/3\}.$$

Now, a_τ need not be feasible for $\widetilde{\Phi}_t$, but we know that $\widetilde{\Phi}_\tau a_\tau \le \alpha \implies \widetilde{\Phi}_t a_\tau \le \alpha + M_t(a_\tau) + M_\tau(a_\tau)$. So for

$$\sigma_{\tau\to t} := \frac{\Gamma(a_0)}{\Gamma(a_0) + 3(M_t(a_\tau) + M_\tau(a_\tau))},$$

we know that

$$\widetilde{\Phi}_t(\sigma_{\tau\to t}a_\tau + (1-\sigma_{\tau\to t})a_0) \le \alpha.$$

Let $\bar{a}_{\tau\to t} := \sigma_{\tau\to t}a_\tau + (1-\sigma_{\tau\to t})a_0$. Then we can write

$$\begin{aligned} \Delta(a_t) &= \Delta(\bar{a}_{\tau\to t}) + \theta_*^\top(\bar{a}_{\tau\to t} - a_t) \\ &\le \Delta(\bar{a}_{\tau\to t}) + \widetilde{\theta}_t^\top(\bar{a}_{\tau\to t} - a_t) + M_t(a_t) + M_t(\bar{a}_{\tau\to t}) \\ &\le \sigma_{\tau\to t}\Delta(a_\tau) + (1-\sigma_{\tau\to t})\Delta(a_0) + M_t(a_t) + \sigma_{\tau\to t}M_t(a_\tau) + (1-\sigma_{\tau\to t})M_t(a_0) \\ &\le (1-\sigma_{\tau\to t})\Delta(a_0) + M_t(a_t) + \sigma_{\tau\to t}(M_t(a_\tau) + M_\tau(a_\tau)) + (1-\sigma_{\tau\to t})M_t(a_0), \end{aligned}$$

where in the end we used the fact that $\Delta(a_\tau) \le M_\tau(a_\tau)$. Now,

$$1 - \sigma_{\tau\to t} \le \frac{3(M_t(a_\tau) + M_\tau(a_\tau))}{\Gamma_0},$$

and of course $M_t(a_0) \le \Gamma_0$. We end up with a bound of the form

$$\Delta(a_t) \le C\left(1 + \frac{\Delta(a_0)}{\Gamma(a_0)}\right)(M_t(a_t) + M_\tau(a_\tau) + M_t(a_\tau)),$$

which is essentially the same as that of Lemma 5. Given this, we can immediately invoke Lemma 6 (appropritaely modifying by $a_{\mathsf{safe}} \to a_0$ and $\Gamma_0 \to \Gamma(a_0)$). We end up with the control that

$$\sum_{t\le T, M_t(a_0)\le\Gamma_0/3} \Delta(a_t) = \widetilde{O}\left(\left(1 + \frac{\Delta(a_0)}{\Gamma(a_0)}\right)\cdot\frac{B_T\omega_T}{\chi}\cdot\sum_{t\le T}\|a_t\|_{V_t^{-1}}\right).$$

For our choice of noise (being the coupled design executed with $\nu = \mathrm{Unif}(\sqrt{3d}\mathbb{S}^d)$, we have $\chi = \Omega(1), B = O(\sqrt{d})$, and so this can be bounded as

$$\sum_{t\le T, M_t(a_0)\le\Gamma_0/3} \Delta(a_t) = \widetilde{O}\left(\mathcal{R}(a_0)d^3T\right).$$

Of course, the above holds true for all $t > t_0(a_0)$ that were not E-steps. Before $t_0(a_0)$, we may bound the per-round regret by 2. Finally, we are left with the E-steps after the time $t_0(a_0)$. Since, as argued above, no extraneous E-steps due to the infeasiblity of perturbed programs occur, we can then, for $T \ge t_0(a_0)$, simply bound the total number of E-steps by $1 + B_T\omega_T\sqrt{dT}$, and accrue roudwise regret of at most 2 in these steps. With our chosen noise, $B_t = O(\sqrt{d})$, this cost is $\widetilde{O}(\sqrt{d^3T})$. Summing these three contributions, and invoking the bound on $t_0(a_0)$ finishes the argument upon recognizing that a_0 is arbitrary, and so we may minimise over it.

Turning now to the risk, first observe that for any $t > T_0 := \min_a t_0(a)$, there exists at least one action such that $M_t(a) \le \Gamma(a)/3$, and so the perturbed program is always feasible, i.e., $a(\eta_t, H_t, t)$ exists. Now, consider subsequent times. Observe that in L-steps, since $\widetilde{\Phi}_t a_t \le \alpha$, we know by Lemma 3 that $\Phi_* a_t \le \alpha + M_t(a_t)\mathbf{1}_m$, assuming consistency and the concentration of $\max(\|\eta_t\|, \max_i \|H_t^i\|)$. Thus, in L-steps, the risk accrued at any time is at most $M_t(a_t)$. On the other hand, in E-steps, the risk accumulated can be bounded by just 1 (using the boundedness of Φ_* and $\mathcal{A}$, and so we only need to work out the total number of these. But after time T_0 such an E-step only occurs to make sure that the net number of E-steps is at least $B_t \omega_t \sqrt{dt}$, and so the total number of such steps is at most $B_T \omega_T \sqrt{dT}$.

Putting these together, we conclude that the net risk accrued is bounded as

$$\mathbf{S}_T \le T_0 + \sum_{\substack{T_0 \le t \le T \\ t \text{ is an E-step}}} 1 + \sum_{\substack{t \le T, \\ t \text{ is an L-step}}} M_t(a_t) = B_T \cdot \widetilde{O}(\sqrt{d^2 T} + \min_a \widetilde{O}\left(\frac{dB_t^2 \omega_t^2 \|a\|^4}{\kappa^2 \Gamma(a)^4}\right).$$

Invoking Lemma 13, as well as the fact that $B_T = B_t = \sqrt{3d}$ for our noise design, the claim follows.

Finally, let us account for the probabilistic conditions needed: we need the concentration event of Lemma 6 to hold for the regret bound, and the consistency and noise-boundedness events for both. Of course, the second is not actually needed, since our noise is bounded always. Together, then, these occur with chance at least $1 - 2\delta$ under our noise design. Of course, then, passing to E-COLTS$(\mu, \delta/2)$ yields the claimed result. □

J Simulation Study

We conduct simulation studies to investigate the behaviour of E-COLTS/R-COLTS, and of S-COLTS. We first study the soft and hard constraint enforcement problems with our coupled noise design. After this, we investigate the behaviour of COLTS methods using independent (or decoupled) noise in §J.3. All experiments were executed on a consumer-grade laptop computer running a Ryzen-5 chip, in the MATLAB environment, and the total time of all experiments ran to about 8 hours.

J.1 Soft Constraint Enforcement

We begin with studying the behaviour of the soft constraint enforcement strategies E-COLTS and R-COLTS. Throughout, we treat E-COLTS as R-COLTS$(\mu, 0, \delta)$, with no exploration.

Setting. We set Φ_* to be a certain 9×9 directed adjacency matrix, A, obtained from https://sparse.tamu.edu/vanHeukelum/cage4, which is a $\approx 60\%$ populated matrix with $d = m = 9$. The rows of Φ_* were normalised to have norm 1. We study the problem of optimising $\theta_* = \mathbf{1}_d/\sqrt{d}$ over $\mathcal{A} = [0, 1/\sqrt{d}]^d$, and enforce the unkonwn constraints $\Phi_* a \le 0.8 \cdot {}^{1}/\sqrt{d}$. We note that the action 0 is always safe, no matter the $\widetilde{\Phi}_t$. This choice is intentional, in that it lets us avoid the inconvenient fixed exploration present in E-COLTS and S-COLTS. Throughout, we set $\delta = 0.1$.

As stated above, for the bulk of this section, we will implement E-COLTS without forced exploration. Indeed, this is not required since 0 is always feasible, as discussed above. This can equivalently be interpreted as R-COLTS with the resampling parameter $r = 0$.

Effect of Noise Rate. As previously noted, in linear TS, small perturbation noise—of the scale 1 rather than $\Theta(\sqrt{d})$—retains sufficient rates of global optimism and unsaturation to enable good regret behaviour. Note that such a small noise directly reduces B_T, and thus we would expect it to improve our regret behaviour by a factor of about $\sqrt{d}$. In order to exploit this, we begin by conducting pilot experiments with our coupled noise design to determine a reasonable noise scale for us to use.

Concretely, we drive our coupled noise design with the laws $\nu_\gamma = \mathrm{Unif}(\gamma \cdot \mathbb{S}^d)$, and run E-COLTS without exploration for 10^3 steps 100 times. In each run, we simply record whether (i) global optimism; (ii) local optimism; and (iii) unsaturation held, and estimate their rates simply as the fraction of time over the run that this property was true. We construct these rate estimates for $\gamma \in [\sqrt{3d}^{-3}, \sqrt{3d}]$, specifically evaluating the same for 41 values of γ chosen over an exponential grid (i.e., so that $\log(\gamma)$ has a constant step). Figure 3 shows the resulting estimates.

The core observation is that global optimism and unsaturation rates are already at ~ 1 for $\log(\gamma) \approx -1$, indicating good performance with this noise. Note that while such performance with small noise has been previously observed for linear TS without unknown constraints, we are unaware if an explicit observation of these rates as above has been performed. Of course, proving these properties at such small γ is an open question, and we also note that our estimates above are not quite correct, since they integrate the events across time, while their rates could vary with t. In any case, the main upshot for this is that in our subsequent experiments, *we work with* $\gamma = 0.5$ *instead of* $\sqrt{3d} \approx 5.2$.

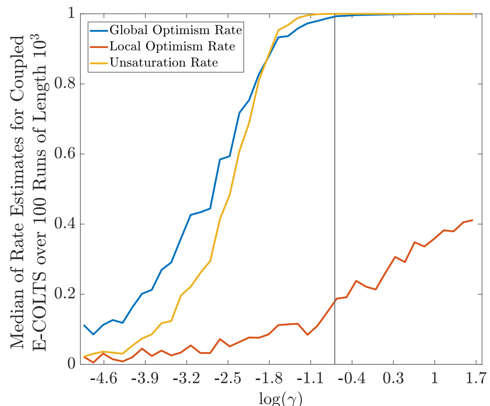

Figure 3: Behaviour of the Global and Local Optimism Rates, as well as the Unsaturation Rate. The black vertical line lies at $\gamma = 0.5$, the value selected for subsequent experimentation. The largest studied value is at $\sqrt{3d}$, which has logarithm about 1.65 Observe that the global optimism and unsaturation rates are significant, and in particular ≈ 1 for $\gamma = 0.5$, far below $\sqrt{3d} \approx 5.2$.

The Behaviour of E-COLTS and R-COLTS. We now study R-COLTS and E-COLTS over the long horizon $T = 5 \cdot 10^4$. We execute R-COLTS with zero resamplings (i.e., E-COLTS with no exploration), and then one and finally two resamplings in each round, all driven by the coupled perturbation noise with $\nu_{0.5}$.

On DOSS. We note that DOSS is not implemented. E-COLTS runs in $\sim 10^{-3}$s per round on our machine. (Relaxed)-DOSS is totally impractical: $(2d)^{m+1} > 10^{12}$, and so it needs $> 10^9$s, i.e., years, per round!

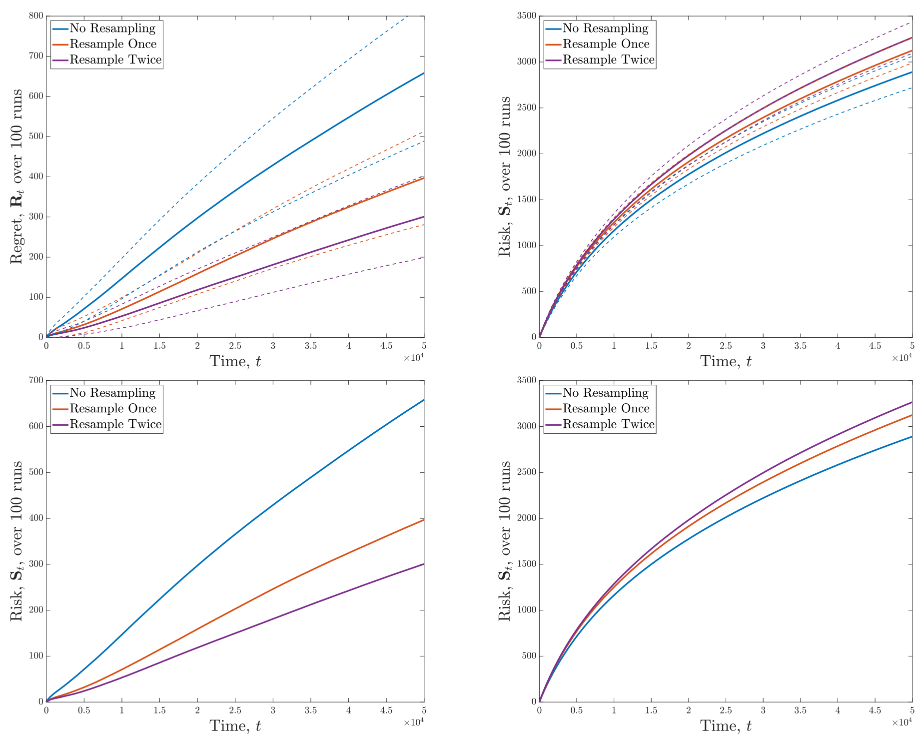

Figure 4: Regret (left) and Risk (right) of R-COLTS with zero, one, and two resamplings per round. Top includes one-sigma error bars, and for clarity, the bottom figures omit them. Note that the regret behaviour is an order of magnitude smaller than the scale $\sqrt{d^2T} \approx 6600$, while the risk behaviour is about a factor of half of this. We further observe that resampling improves regret signficantly, while only hurting the risks slightly, although this effect appears to decelerate as resampling is increased.

Observations. Figure 4 shows the observed regret and risk traces over 100 runs. The observed regret behaviour is very strong: even without resampling, the terminal median regret of ~ 600 is closer to $\sqrt{T \log T} \approx 750$ than to $\sqrt{d^2 T \log(T)} \approx 6600$. The risk behaviour is more significant, but still half this scale. The observation of $\mathbf{R}_T$ suggests that a stronger regret bound may hold for E-COLTS and R-COLTS, which is in line with the stronger instance-specific regret behaviour of the optimism-based method DOSS [GCS24]. Proving this is an interesting open problem.

These simulations thus bear out the strong performance of E-COLTS/R-COLTS with $r = 0$. Further, as we add resampling, risk degrades mildly, but the regret improves significantly, although the returns diminish with more resampling. This suggests that practically, a few resamplings in R-COLTS are enough to extract most of the advantage. Interestingly, resampling has a palpable effect even though the optimism rate is nearly one!

J.2 Hard Constraint Enforcement

Next, we investigate the behaviour of S-COLTS over the same instance, supplied with the data $a_{\mathsf{safe}} = 0$. The natural point of comparison to S-COLTS is the SAFE-LTS algorithm [MAAT21], which operates in $O(\mathsf{SOCP} \log t)$ computation per round.[9]

Concretely, we again drive this method with $\nu_{0.5}$ as before. For SAFE-LTS, we sample a perturbed objective vector with the same noise scale, and otherwise optimise over the second order conic constraints as detailed in §4.3. In both cases, we used the library methods `linprog` and `coneprog` provided by MATLAB to implement these methods.[10] Note that these methods are specifically tailored to linear and conic programming respectively. As before, we repeat runs of length $T = 5 \cdot 10^4$ for a total of 100 runs.

Strong Safety Behaviour. We note that in all of our runs, we did not observe any constraint violation from either S-COLTS or SAFE-LTS, despite the fact that we executed these methods with $\delta = 0.1$. This suggests both that in practice, the parameter δ can be relaxed (which would yield mild improvements in regret), and in any case verifies the strong safety properties of these methods.

Comparison of Regret. We show the regret traces over the 100 runs in Figure 5. We observe that S-COLTS has a slightly improved regret performance relative to SAFE-LTS, which may be attributed to the selection of stronger exploratory directions through solving the perturbed program.

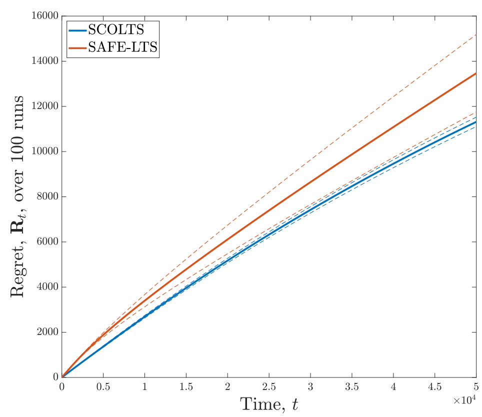

Figure 5: Regret Behaviour of S-COLTS and SAFE-LTS on the same instance as previous figures (one-sigma error curves). We note that S-COLTS offers a mild improvement in regret over SAFE-LTS. However, this comes with a $5\times$ reduction in net computational time per round, which is the main advantage of S-COLTS.

Computational Speedup. In wall-clock terms, each iteration of SAFE-LTS is about $5.2\times$ slower than that of S-COLTS on this 9 dimensional instance with 9 unknown constraints (over $5 \cdot 10^6$

[9] We do not implement other prior methods for SLBs, mainly because SAFE-LTS has previously been seen to have similar behaviour, and be about $2d = 18$ times faster than these methods. Of course, we also did not implement DOSS as a comparison for the soft constriant enforcement methods since it is impractical to execute for $d = m = 9$.

[10] Of course, R-COLTS/E-COLTS were also implemented using `linprog`.

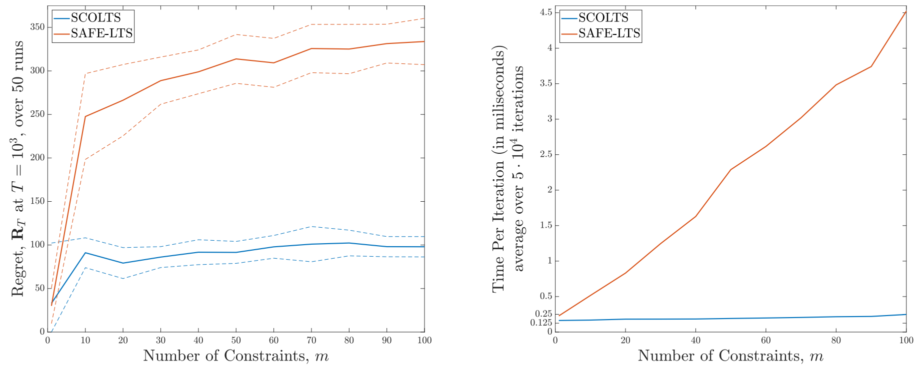

Figure 6: Comparisons of the regret (left, one-sigma error curves) and computational costs (right) of S-COLTS and SAFE-LTS in the $d = 2$ instance as m varies. This is the same setting as Figure 2, right, but presented separately rather than as a ratio. The left plots the regrets at time $T = 10^3$ over 50, and the right plots the wall-clock time per iteration on our resources in milliseconds. S-COLTS needs $0.14 - 0.25$ milliseconds per iteration, while SAFE-LTS needs > 4.5 at $m = 100$. At the same time, for $m \geq 10$, the regret of S-COLTS is about $3\times$ smaller.

total iterations, S-COLTS took about 0.22ms per iteration, while SAFE-LTS took about 1.16ms), a significant computational advantage even in this modest parameter setup.

High Level Conclusions. The main takeaway from this set of experiments is that S-COLTS offers tangible benefits in computational time relative to SAFE-LTS (and a fortiori, to other pessimism-optimism based frequentist hard constraint enforcement methods), while even obtaining a slight improvement in the regret behaviour. This demonstrates the utility of S-COLTS over these prior methodologies, and suggests that it is the natural approach that should be used in practice.

J.2.1 Investingating Behaviour with Increasing m

Of course, the computational problem of optimising m SOC constraints becomes harder as m grows, and so we expect that the computational advantage of S-COLTS over SAFE-LTS would grow with m.[11] To investigate this hypothesis more closely, we turn to a slightly different setup.

Setup. We set $d = 2, \theta_* = (1, 0), \mathcal{A} = [-1/\sqrt{d}, 1/\sqrt{d}]^d$. For $m \geq 3$, we impose m unknown constraints such that the feasible region forms a regular m-gon with one vertex at $(0.2/\sqrt{2}, 0)$. This allows us to systematically increase m (to very high values) without incurring significant computational costs. We investigate the behaviour of S-COLTS and E-COLTS on this setup with the coupled noise design as in the previous section ($\gamma = 0.5$) for $m \in \{10, 20, \cdots, 100\}$. We also execute this for $m = 1$, where a single constraint passing through the same vertex is enforced. In all cases, we set $a_{\mathsf{safe}} = 0$, which is always feasible.

Strong Computatational Speedup. As seen in Figure 6, S-COLTS has a strong computational advatange, which further grows with m. In particular, at $m = 1$, S-COLTS is about $1.3\times$ faster to execute than SAFE-LTS, while for $m = 100$, this advantage grows to $18\times$.

Improved Regret Performance.[12] Further, instead of the small gain seen in the previous section, in this problem S-COLTS has a strong statistical advantage relative to SAFE-LTS for even moderate m. Indeed, while at $m = 1$, its regret is about 10% larger than that of SAFE-LTS, for larger m, its regret is many times *smaller*. In particular, for $m \geq 10$, we found that the regret of S-COLTS is roughly $3\times$ smaller (ranging between $2.7\times$ and $3.4\times$.).

Takeaways. This investigation further bolsters the strong advantage of S-COLTS over SAFE-LTS. Note

[11]Note that it may be possible to mitigate this somewhat by instead imposing the convex constraint $\max_i (\hat{\Phi}_t a - \alpha)^i + \|a\|_{V_t}^{-1} \leq 0$ to exploit that the same matrix V_t^{-1} appears in all constraints. However, the gradient computation of this map still grows with m, so the overall picture is unclear. Of course, imposing only m linear constraints is bound to be faster.

[12]Note: for the regret ratio in Figure 2, we perform 100 separate runs with both methods, and compute the ratio of regret for the two methods in each. That figure reports the mean over this data - in this case, the expected mean is ~ 1.5 at $m = 1$, but with wide confidence intervals (CIs). For $m \geq 10$, the lower confidence bounds all exceed 2. At $m = 1$, the mean regret of SAFE-LTS is about $0.91\times$ that of S-COLTS, with strongly overlapping CIs.

that alternative confidence-set based hard enforcement methods are at least $2d$ times slowed than SAFE-LTS, meaning that the computational advantage of S-COLTS is even stronger relative to these methods. For large m, this appears to be accompanied by a large statistical advantage, making this the natural method in applications of SLBs.

J.3 Simulation Study on the Behaviour of the Decoupled Noise

Finally, we investigate the behaviour of the COLTS framework under the decoupled noise design, wherein, instead of setting $H = -\mathbf{1}_m \eta$, we draw η, and each row of H, independently from ν_γ. The main impetus behind this, of course, is that this decoupled design is a natural choice to execute COLTS, although it is contraindicated by the analysis tools available to us.

Behaviour of Event Rates with γ. To begin with, Figure 7 shows the global optimism, local optimism, and unsaturation rates with this decoupled noise for the same instance as previously studied. Observe first that the decoupled noise design does experience a slight decrease in each of these rates compared to those seen in Figure 3. However, this effect is relatively mild, and in particular, we can see that the unsaturation rate is already up to nearly one at our previously selected value of $\gamma = 0.5$. This suggests that the decoupled noise would do nearly as well as the coupled noise in this case.

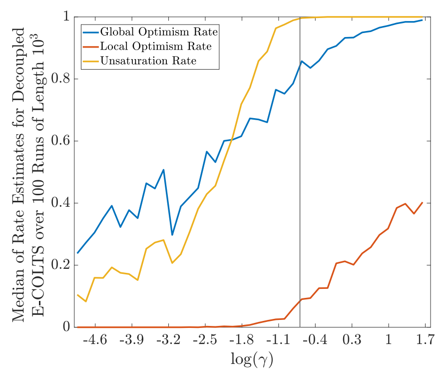

Figure 7: Behaviour of the Global Optimism, Local Optimism, and Unsaturation Rates with γ for the Decoupled Noise in the setting of Figure 3. Observe that while these rates decay somewhat with respect to the coupled noise, they are still strong, and especially for large γ are nearly as good as with the coupled noise.

Behaviour of Regret and Risk. To further investigate the above claim, we execute E-COLTS without exploration (or equivalently, R-COLTS with $r = 0$) driven with this decoupled noise over the longer horizon $T = 5 \cdot 10^4$. The resulting regret and risks are plotted in Figure 8, along with the same for E-COLTS with coupled noise. Observe that the decoupled noise sees a significant loss of about $3\times$ in regret, but sees a gain of about $1.5\times$ in risk. Heuristically, we may think of the decoupled noise as behaving as if the noise is coupled but "shrunk", so that the behaviour of the risk is improved, but the behaviour of the regret worsens.

Practically speaking, our recommendation remains to use the coupled noise design, in that it attains higher rates of explanatory events, and carries theoretical guarantees. Nevertheless, establishing that $\mathbf{R}_T$ and $\mathbf{S}_T$ do scale sublinearly with the decoupled noise design, as is evident from Figure 8, is an interesting open problem.

J.3.1 Investigation of Rates with Increasing m

Of course, the main obstruction with the use of the decoupled noise in §4.2 was to do with many constraints. Indeed, it should be clear that under this decoupled noise, the local optimism rate must decay exponentially with m, since if any row of $\widetilde{\Phi}_t$ is perturbed so that a_* violates its constraints, local optimism would fail (and this would occcur with a constant chance, no matter the estimates).

To probe whether this indeed occurs, we simulate the behaviour of E-COLTS with the coupled and decoupled noise designs on a simplified setup.

Setup. We again take the $d = 2$ polygonal constraints investigated in §J.2.1. We investigate the behaviour of E-COLTS with both the coupled and decoupled noise designs on this instance as

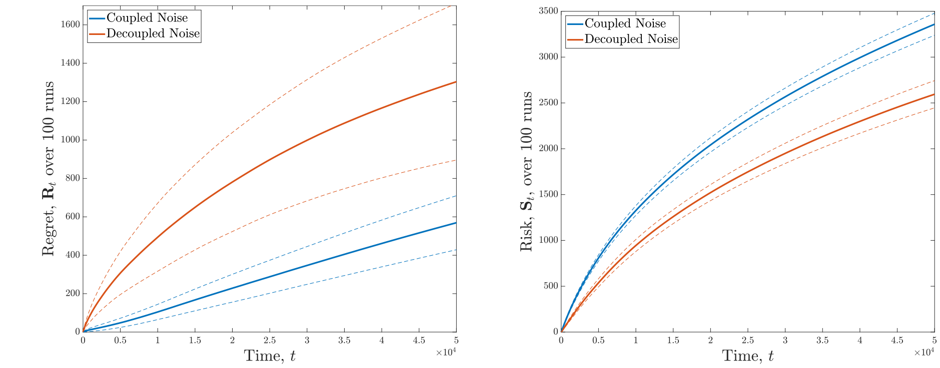

Figure 8: Behaviour of regret (left) and risk (right) for E-COLTS executed with the decoupled noise compared with E-COLTS executed with the coupled noise (one-sigma error bars). Observe that the regret behaviour sharply deteriorates, while the risk behaviour slightly improves for the decoupled noise design. Heuristically, this suggests that the decoupled noise behaves 'like' the coupled noise, but with a smaller value of γ.

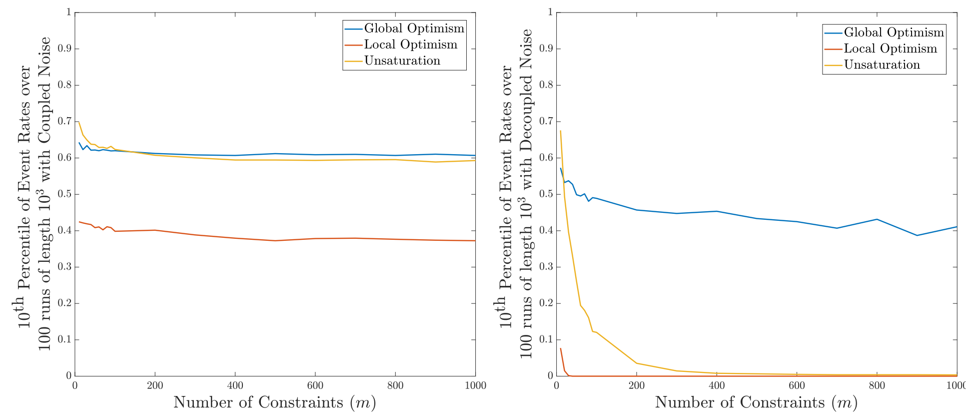

Figure 9: Behaviour of the rates of global and local optimism, and of unsaturation, in the polygonal instances as the number of constraints is increased for the coupled (left) and decoupled (right) noise designs driven by $\mathrm{Unif}(\mathbb{S}^2)$. Observed that the behaviour of these is stable with m for the coupled design, but for the decoupled design, the local optimism and unsaturation rate decay with m. Surprisingly, the global optimism rate remains stable even for the decoupled noise design.

$m \in \{10, 20, \ldots, 100\} \cup \{200, 300, \ldots, 1000\}$, thus letting us probe an extremely high number of unknown constraints.

Observations. There are two main observations of Figure 9. Firstly, note that as shown in the main text, the rates of optimism and unsaturation under the coupled noise design are stable, and do not meaningfully vary with m after it has grown at least slightly.

On the other hand, under the decoupled noise design, the local optimism rate clearly crashes exponentially. The unsaturation rate has a slower but evident decay: roughly, this is as $m^{-1.3}$ for $m \le 100$, and appears to be exponential for large m. However, surprisingly, the *global optimism* rate remains stable (although lower than the same with the coupled design). This shows that there are situations with low-regret where frequent global optimism would be the 'correct' explanation for good performance of methods like S-COLTS or E-COLTS (indeed, this is what prompted us to write the optimism based analysis of these methods in §H.5). Note however that *proving* that global optimism is frequent under the decoupled design is an open problem. In fact, with unknown constraints, we do not know of any method to deal with global optimism lower bounds that does not pass through local optimism, since the approach of Abeille & Lazaric [AL17] relies on convexity properties of the value function in terms of the unknown parameters, which fails in this case.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: as the abstract and introduction specify, we propose and analyse algorithms for SLBs based on the COLTS framework.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Theoretical limitations of results, and their context, is extensively discussed, along with open problems specified (latter mostly in the appendix).

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Complete proofs are available in the appendix. We also specify the main technical ideas in the main text.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: the algorithms and tools used, the computational resources needed, and the settings investigated, are all explicitly described in §J

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: the methods implemented use standard estimates in bandits, along with library linear programming routines. We believe that this is easy to implement.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: See §J

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: one-sigma error bars are reported throughout, except for iteration time measurement in Fig. 6 (which were recorded in bulk over 10^3 iterations for each methods), and for the event rate estimates in Fig. 3 and Fig. 7 (where we report 10th percentiles of integrated averages of values in $(0, 1)$ over 100 runs, which are smaller than the mean minus one-sigma lower confidence bounds, and so better capture our focus for these values).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: see §J

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: No data drawn from human subjects was used in this research. The paper is theoretical, and focuses on developing algorithms for safe linear bandits. As such, societal impacts are related to downstream applications of this generic underlying method, and so are associated with work applying ours as a subroutine, rather than directly stemming from our contributions.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: as discussed in the answer to the previous point, potential social impacts (positive or negative) would stem from downstream applications that use bandit linear programming routines that we have designed. We decline to prognosticate such applications.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: no such data or models are released via this paper.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: the code was entirely written bythe authors. The sole data used is the matrix Φ_* as specified in §J, which is accompanied by appropriate attribution in the same section.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: no new assets are released.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: we did not do this.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: no research with human subjects or crowdsourcing were employed.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and ference, workshop, or preprint repository) and already published papers are allowed.

      we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were not used in any part of the preparation of this paper.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.