Safe Linear Bandits with Unknown Safety Gaps

Maoli Liu
The Chinese University of Hong Kong
mlliu@cse.cuhk.edu.hk

Zhuohua Li*
Xidian University
lizhuohua@xidian.edu.cn

Zeyu Zhang
The Chinese University of Hong Kong
zyzhang@cse.cuhk.edu.hk

Xiangxiang Dai
The Chinese University of Hong Kong
xxdai23@cse.cuhk.edu.hk

John C.S. Lui
The Chinese University of Hong Kong
cslui@cse.cuhk.edu.hk

Abstract

We study stochastic linear bandits whose rewards and linear safety constraint depend on the same unknown parameter, requiring every chosen action to be safe at every round with high probability, even though the safe decision set is initially unknown. Prior work achieves $\widetilde{O}(\sqrt{T})$ regret when the safety gap, i.e., the constraint slack at the optimal action, is strictly positive and known to the learner. However, the best previously established gap-agnostic guarantee is only $\widetilde{O}(T^{2/3})$, both in the zero-gap regime, and when the gap is positive but unknown. This raises two natural questions: is the $\widetilde{O}(T^{2/3})$ bound tight in the zero-gap case, and is knowledge of the gap necessary for $\widetilde{O}(\sqrt{T})$ regret when the gap is positive? We answer both questions. First, we establish an $\Omega(T^{2/3})$ lower bound for zero-gap instances, showing that the existing upper bound is tight in its dependence on the horizon. Second, we propose EPOCH-SLUCB, an algorithm that attains $\widetilde{O}(\sqrt{T})$ regret for every fixed positive safety gap, without requiring its value to be known. Numerical simulations illustrate the empirical performance of EPOCH-SLUCB.

1 Introduction

The *stochastic linear bandit* [11, 1] is a fundamental model for sequential decision-making under uncertainty. At each round $t \in [T]$, a learner chooses an action $\boldsymbol{x}_t \in \mathcal{D} \subset \mathbb{R}^d$ and observes a noisy reward $y_t = \boldsymbol{\theta}^\top \boldsymbol{x}_t + \eta_t$, where $\boldsymbol{\theta}$ is an unknown but fixed parameter and η_t is a noise term. The learner's objective is to minimize the cumulative regret over T rounds, i.e., the cumulative reward gap between the chosen actions and the optimal action. This setting captures applications in personalized recommendation [20], crowdsourcing task assignment [18], and dynamic pricing [10].

In many sequential decision problems, however, reward maximization alone is not sufficient. The learner must ensure that all chosen actions satisfy a hard *safety constraint* throughout the horizon with high probability. For example, demand response in power systems requires chosen load profiles to respect operational limits [4], and parameter tuning in robotics must avoid settings that could damage hardware [9]. In such settings, even a single unsafe action can be unacceptable, ruling out exploration strategies that gather information by trying actions whose safety cannot yet be certified.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Regret bounds in T and Δ, with d and other problem constants fixed.

Result	Setting	Requires known Δ?	Regret
Safe-LUCB [4]	$\Delta > 0$	Yes	$\widetilde{O}(\sqrt{T} + \Delta^{-2})$
Safe-LUCB [4]	$\Delta \geq 0$	No	$\widetilde{O}(T^{2/3})$
EPOCH-SLUCB, Theorem 5.4	$\Delta > 0$	No	$\widetilde{O}(\sqrt{T} + \Delta^{-3})$
Lower bound, Theorem 4.1	$\Delta = 0$	N/A	$\Omega(T^{2/3})$

To formalize this requirement, Amani et al. [4] introduced the *safe linear stochastic bandit*, which augments the standard linear bandit with a linear safety constraint. The learner must ensure that $\boldsymbol{\theta}^\top \boldsymbol{B}\boldsymbol{x}_t \leq c_{\text{safe}}$ holds simultaneously for all rounds t, with high probability. The matrix $\boldsymbol{B}$ and threshold c_{safe} are known, but both the reward and the safety constraint depend on the same unknown parameter $\boldsymbol{\theta}$. As a result, the safe action set is itself initially unknown. The learner must therefore infer which actions are safe from noisy reward observations while seeking actions with high reward.

For this model, Amani et al. [4] proposed Safe-LUCB, which combines an initial pure-exploration phase with a subsequent UCB rule. The regret guarantees of Safe-LUCB depend on the *safety gap* $\Delta := c_{\text{safe}} - \boldsymbol{\theta}^\top \boldsymbol{B}\boldsymbol{x}^*$, the constraint slack at the optimal safe action $\boldsymbol{x}^*$. For a fixed $\Delta > 0$ whose value is known to the learner, Safe-LUCB attains $\widetilde{O}(\sqrt{T})$ regret. However, the best established gap-agnostic guarantee is only $\widetilde{O}(T^{2/3})$, both when $\Delta = 0$ and when $\Delta > 0$ but its value is unknown. This leaves two questions open: whether the $\widetilde{O}(T^{2/3})$ rate is unavoidable when $\Delta = 0$, and whether prior knowledge of Δ is necessary for $\widetilde{O}(\sqrt{T})$ regret when $\Delta > 0$.

We answer both questions in this paper. First, we establish an $\Omega(T^{2/3})$ lower bound for instances with $\Delta = 0$ (Theorem 4.1), showing that the existing $\widetilde{O}(T^{2/3})$ upper bound is tight in its dependence on the horizon. Second, we propose EPOCH-SLUCB, an epoch-based algorithm that attains $\widetilde{O}(\sqrt{T})$ regret for every fixed $\Delta > 0$, without requiring its value to be known (Theorem 5.4). EPOCH-SLUCB thus matches the dependence on T of Safe-LUCB with knowledge of Δ. Finally, we provide numerical simulations illustrating the performance of EPOCH-SLUCB relative to Safe-LUCB baselines.

The technical intuition behind our results is that the safety gap Δ determines where the optimal safe action lies relative to the safety boundary, and our lower and upper bounds exploit this geometry in complementary ways. For our lower bound, when $\Delta = 0$, the optimal safe action $\boldsymbol{x}^*$ lies on the boundary of the safe action set. A learner that moves too aggressively toward $\boldsymbol{x}^*$ may violate safety under a nearby parameter, whereas a learner that stays too far inside the safe action set incurs regret. Our construction exploits this tension using two parameters that are statistically hard to distinguish from observations generated by safe actions but have different optimal safe actions, yielding the $\Omega(T^{2/3})$ lower bound. For the upper bound, EPOCH-SLUCB constructs certified safe sets from confidence bounds without taking Δ as input. Across epochs, the algorithm uses progressively finer safety margins. The confidence bounds eventually become tight enough to certify $\boldsymbol{x}^*$ as safe. From that epoch onward, the regret of optimistic action selections can be controlled by standard linear-bandit confidence bounds. The geometric design of the epoch schedule keeps the cumulative pure exploration across all epochs at $\widetilde{O}(\sqrt{T})$ as well, so the total bound is $\widetilde{O}(\sqrt{T})$ for every fixed $\Delta > 0$, attained without sacrificing per-round safety.

Together, these results show that the zero-gap case is intrinsically hard, whereas a fixed positive gap whose value is unknown still permits $\widetilde{O}(\sqrt{T})$ regret. Table 1 summarizes the regret bounds. The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 introduces the safe linear bandit model and the safety gap. Section 4 states the lower bound, and Section 5 presents EPOCH-SLUCB and its regret guarantee for unknown positive gaps. Section 6 sketches the lower-bound proof, Section 7 gives numerical simulations, and Section 8 concludes.

2 Related work

The work most directly related to ours is Amani et al. [4], who introduced the safe linear bandit model with mean reward $\boldsymbol{\theta}^\top \boldsymbol{x}_t$ and safety constraint $\boldsymbol{\theta}^\top \boldsymbol{B}\boldsymbol{x}_t \leq c_{\text{safe}}$, where the same unknown parameter $\boldsymbol{\theta}$

governs both the objective and the constraint and $\boldsymbol{B}$ is known. The safety constraint must be satisfied at all rounds with high probability. They propose Safe-LUCB and their analysis gives a $\widetilde{O}(\sqrt{T})$ regret bound for a fixed positive safety gap Δ known to the learner, but only a $\widetilde{O}(T^{2/3})$ bound in the zero-gap or gap-agnostic regime. We revisit the same model, establish a matching $\Omega(T^{2/3})$ lower bound at $\Delta = 0$, and propose EPOCH-SLUCB, which achieves $\widetilde{O}(\sqrt{T})$ regret for every fixed $\Delta > 0$ without taking Δ as input. Our lower bound builds on the information-theoretic lower-bound tradition for stochastic linear bandits [11, 19], with the additional difficulty that the learner must satisfy the per-round safety requirement.

Subsequent work studies nearby safe bandit formulations in several directions. Amani et al. [5] extend both the reward model and the safety constraint to generalized linear forms with known link functions. Moradipari et al. [22] introduce Safe-LTS, a safe linear Thompson sampling algorithm that uses side information about the safety constraint, and prove high-probability safety together with $\widetilde{O}(d^{3/2}\sqrt{T})$ regret. Hutchinson et al. [15] study safe linear bandits with noisy constraint observations and develop directional optimism under geometric conditions such as separated instances and finite star-convex action sets. Gangrade et al. [13] study safe linear bandits over unknown polytopes with stochastic feedback on both rewards and safety risks, and bound both efficacy regret and cumulative safety violations. Pacchiano et al. [23, 24] study bandit models with stagewise linear cost constraints, reward and cost observations, and safety guarantees formulated either in expectation or with high probability. Khezeli and Bitar [17] study a formulation in which each played arm must have expected reward above a prescribed safety threshold with high probability, starting from a known safe baseline arm. These works are complementary to ours because they either provide direct observations or side information about safety values, assume different initial knowledge, or use a different safety criterion. We keep the bandit feedback model of Amani et al. [4], where safety must be inferred from reward observations and the known matrix $\boldsymbol{B}$, and every chosen action must be safe with high probability.

Another related line of work studies conservative and resource-constrained bandits. Conservative bandit algorithms require the learner to maintain performance relative to a known baseline, either cumulatively or at each round [27, 16, 14, 21]. Bandits with knapsacks and related resource models instead constrain total resource consumption over the horizon [7, 3]. These cumulative formulations permit compensation across rounds, so conservative choices later can offset earlier reward loss or resource use. Stagewise conservative models remove this time averaging, but their safety criterion is still reward relative to a baseline rather than an unknown linear constraint on the chosen action. In our setting, the unknown linear constraint must hold for each selected action with high probability. An exploratory action that may violate the constraint cannot be justified by conservative behavior later.

Safe optimization and safe reinforcement learning give broader frameworks for safety under uncertainty. Gaussian process methods expand a safe set using smoothness assumptions [25, 8, 9], while safe exploration has also been studied in Markov decision processes and reinforcement learning [26, 2, 12, 6]. These settings differ from ours in feedback, dynamics, and safety models.

3 Problem Formulation

Notation. We adopt the following notation throughout the paper. For $n \in \mathbb{N}_+$, let $[n] = \{1, \dots, n\}$. For a vector $\boldsymbol{x} \in \mathbb{R}^d$, $\|\boldsymbol{x}\|_2$ denotes the Euclidean norm and $\boldsymbol{x}^\top$ the transpose. For a matrix $\boldsymbol{M}$, $\|\boldsymbol{M}\|$ denotes the spectral norm. For a positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$, the weighted norm is $\|\boldsymbol{x}\|_{\boldsymbol{A}} = \sqrt{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}$, and we write $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ for its smallest and largest eigenvalues. We use $\widetilde{O}(\cdot)$ to hide polylogarithmic factors. For scalars a, b, we write $a \wedge b = \min\{a, b\}$. For probability measures P and Q, we write $\mathrm{KL}(P\|Q)$ for their Kullback–Leibler divergence. For $p, q \in [0, 1]$, we write $\mathrm{kl}(p, q) = p \log(p/q) + (1 - p) \log((1 - p)/(1 - q))$ for the Kullback–Leibler divergence between Bernoulli distributions with means p and q, with the usual boundary conventions.

We consider the safe linear stochastic bandit model of Amani et al. [4]. A learner interacts with the environment over T rounds. Let $\mathcal{D} \subset \mathbb{R}^d$ be a known convex compact decision set. At each round $t \in [T]$, the learner selects an action $\boldsymbol{x}_t \in \mathcal{D}$ and observes a reward $y_t = \boldsymbol{\theta}^\top \boldsymbol{x}_t + \eta_t$, where $\boldsymbol{\theta} \in \mathbb{R}^d$ is an unknown parameter and η_t is a noise term. The action $\boldsymbol{x}_t$ must additionally satisfy a linear safety constraint $\boldsymbol{\theta}^\top \boldsymbol{B}\boldsymbol{x}_t \le c_{\mathrm{safe}}$, where $\boldsymbol{B} \in \mathbb{R}^{d\times d}$ and $c_{\mathrm{safe}} > 0$ are known to the learner. The learner aims to maximize cumulative reward while ensuring safety at every round with high probability.

An action $\boldsymbol{x} \in \mathcal{D}$ is safe under parameter $\boldsymbol{\theta}$ if $\boldsymbol{\theta}^\top \boldsymbol{B}\boldsymbol{x} \le c_{\text{safe}}$. Correspondingly, the safe set is

$$\mathcal{D}^s(\boldsymbol{\theta}) = \{\boldsymbol{x} \in \mathcal{D} : \boldsymbol{\theta}^\top \boldsymbol{B}\boldsymbol{x} \le c_{\text{safe}}\}.$$

Since $\boldsymbol{\theta}$ is unknown and the learner observes only rewards, the learner does not know $\mathcal{D}^s(\boldsymbol{\theta})$.

The optimal safe action is defined as $\boldsymbol{x}^*(\boldsymbol{\theta}) \in \arg\max_{\boldsymbol{x} \in \mathcal{D}^s(\boldsymbol{\theta})} \boldsymbol{\theta}^\top \boldsymbol{x}$. When clear from context, we drop the argument $\boldsymbol{\theta}$ and write $\mathcal{D}^s$ and $\boldsymbol{x}^*$. Maximizing cumulative reward is equivalent to minimizing the cumulative pseudo-regret $R_T := \sum_{t=1}^{T} \left(\boldsymbol{\theta}^\top \boldsymbol{x}^* - \boldsymbol{\theta}^\top \boldsymbol{x}_t\right)$. It measures the cumulative reward gap between $\boldsymbol{x}^*$ and the actions chosen by the learner.

Given a confidence level $\delta \in (0, 1)$, we say that a policy is δ-safe with respect to $\boldsymbol{\theta}$ if

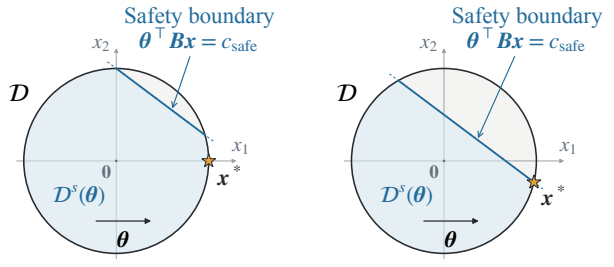
$$\mathbb{P}_{\boldsymbol{\theta}} \left(\boldsymbol{\theta}^\top \boldsymbol{B}\boldsymbol{x}_t \le c_{\text{safe}},\ \forall t \in [T]\right) \ge 1 - \delta.$$

The goal is to design δ-safe policies with sublinear cumulative pseudo-regret.

An important quantity in this problem is the *safety gap*, defined as

$$\Delta := c_{\text{safe}} - \boldsymbol{\theta}^\top \boldsymbol{B}\boldsymbol{x}^*,$$

which measures the slack between the optimal safe action $\boldsymbol{x}^*$ and the safety boundary. Since $\boldsymbol{x}^* \in \mathcal{D}^s(\boldsymbol{\theta})$, we have $\Delta \ge 0$. When $\Delta > 0$, the safety constraint is satisfied strictly at $\boldsymbol{x}^*$, which may lie on the boundary of $\mathcal{D}$. When $\Delta = 0$, $\boldsymbol{x}^*$ lies on the safety boundary. These two cases are illustrated in Figure 1. Unless stated otherwise, the learner does not know Δ.



(a) Positive gap, $\Delta > 0$. (b) Zero gap, $\Delta = 0$.

Figure 1: Geometry of the safety gap. In (a), $\boldsymbol{x}^*$ lies on the boundary of $\mathcal{D}$ with $\Delta > 0$. In (b), it lies on the safety boundary.

We adopt the following assumptions from Amani et al. [4].

Assumption 3.1 (Sub-Gaussian noise)**.** For all t, the noise η_t is conditionally zero-mean and R-sub-Gaussian, i.e., $\mathbb{E}[\eta_t | \boldsymbol{x}_{1:t}, \eta_{1:t-1}] = 0$ and $\mathbb{E}[e^{\lambda \eta_t} | \boldsymbol{x}_{1:t}, \eta_{1:t-1}] \le \exp(\lambda^2 R^2/2)$ for all $\lambda \in \mathbb{R}$.

Assumption 3.2 (Boundedness)**.** There exist constants $S, L > 0$ such that $\|\boldsymbol{\theta}\|_2 \le S$, $\|\boldsymbol{x}\|_2 \le L$ for all $\boldsymbol{x} \in \mathcal{D}$. Moreover, $|\boldsymbol{\theta}^\top \boldsymbol{x}| \le 1$ for all $\boldsymbol{x} \in \mathcal{D}$.

Assumption 3.3 (Convex decision set)**.** The decision set $\mathcal{D}$ is a convex body in $\mathbb{R}^d$ containing the origin in its interior.

Assumption 3.3 together with $c_{\text{safe}} > 0$ ensures that the safe set $\mathcal{D}^s(\boldsymbol{\theta})$ is non-empty for every $\boldsymbol{\theta}$, since the origin is always safe.

4 Lower Bounds

In this section, we establish an $\Omega(T^{2/3})$ lower bound for $\Delta = 0$ in dimension $d = 2$. This matches the $\widetilde{O}(T^{2/3})$ upper bound of Safe-LUCB [4], showing that the rate is tight in T in this regime.

Theorem 4.1 (Lower bound)**.** *There exist a universal constant $c_0 > 0$, a matrix $\boldsymbol{B} \in \mathbb{R}^{2\times 2}$, and a safety threshold $c_{\text{safe}} > 0$ such that, for every integer $T \ge 1$, there exist a decision set $\mathcal{D}_T \subset \mathbb{R}^2$ and a parameter set $\Theta_T = \{\boldsymbol{\theta}_+, \boldsymbol{\theta}_-\} \subset \mathbb{R}^2$ with the following properties. Each instance $(\boldsymbol{B}, \mathcal{D}_T, c_{\text{safe}}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta_T$, with independent standard Gaussian noise, satisfies Assumptions 3.1–3.3 and has zero safety gap. For these instances, for every $\delta \in (0, 1/4]$, any policy π satisfying*

$$\mathbb{P}^{\pi}_{\boldsymbol{\theta}} \left(\boldsymbol{\theta}^\top \boldsymbol{B}\boldsymbol{x}_t \le c_{\text{safe}},\ \forall t \in [T]\right) \ge 1 - \delta \qquad \text{for each } \boldsymbol{\theta} \in \Theta_T$$

also satisfies $\max_{\boldsymbol{\theta} \in \Theta_T} \mathbb{E}^{\pi}_{\boldsymbol{\theta}} [R_T(\boldsymbol{\theta})] \ge c_0 T^{2/3}$.

The $\Omega(T^{2/3})$ lower bound reflects an information-theoretic limitation. In our construction, an action that is safe and nearly optimal under one parameter can be unsafe under the other. Intuitively, the learner faces a fundamental trade-off. Gathering information in the safety-relevant direction requires playing actions that are informative about the boundary, but per-round safety restricts how aggressively the learner can use such actions before the boundary is known. The proof formalizes this trade-off by constructing two zero-gap instances that are difficult to distinguish from reward observations collected under the safety constraint. Balancing the regret from conservative play against the precision required for safety certification then yields the $T^{2/3}$ rate. Theorem 4.1 answers the first question from the introduction. The proof is sketched in Section 6 and detailed in Appendix A.

Remark 4.2 (Dependence on dimension d). Theorem 4.1 shows that the $T^{2/3}$ dependence is unavoidable even in dimension $d = 2$. By embedding the two-dimensional hard instance into the first two coordinates of a d-dimensional problem, the $\Omega(T^{2/3})$ lower bound extends to every $d \geq 2$. Thus, the result establishes tightness in the dependence on the horizon T, but does not characterize the optimal dependence on the dimension d.

Remark 4.3 (Comparison with Pacchiano et al. [24]). Pacchiano et al. [24] study contextual bandits with stagewise linear cost constraints, where the learner observes both reward and cost signals and the cost constraint depends on a parameter separate from the reward parameter. Their $\Omega(\sqrt{T})$ lower bound applies to a different feedback model. In the safe linear bandit model considered here, the learner does not observe a separate cost signal. Safety must be inferred from reward observations through the shared parameter $\boldsymbol{\theta}$ and the known matrix $\boldsymbol{B}$. The $\Omega(T^{2/3})$ lower bound therefore isolates the cost of certifying a boundary action when safety is observed only indirectly.

5 Algorithm for Unknown Positive Gap

In this section, we turn to the regime $\Delta > 0$. We propose EPOCH-SLUCB in Algorithm 1 and show that it achieves $\widetilde{O}(\sqrt{T})$ regret for every fixed positive Δ without requiring Δ as input.

In our model, the learner cannot observe safety values directly and must instead assess safety through an estimate of $\boldsymbol{\theta}$. Reward observations yield a confidence set for $\boldsymbol{\theta}$, which provides an upper bound on $\boldsymbol{\theta}^\top \boldsymbol{B}\boldsymbol{x}$. An action can be certified safe only when this upper bound is at most c_{safe}. If the safety gap Δ were known, the learner could choose an exploration length that drives the safety uncertainty below Δ, ensuring that $\boldsymbol{x}^*$ is eventually included in the certified safe set. However, without knowing Δ, the learner cannot tell how much exploration is enough to certify $\boldsymbol{x}^*$. Stopping too early may exclude $\boldsymbol{x}^*$, while continuing too long creates unnecessary regret. EPOCH-SLUCB addresses this difficulty by running the algorithm at a sequence of progressively finer certification scales.

Forced exploration to estimate $\boldsymbol{\theta}$. Since the true safe set $\mathcal{D}^s(\boldsymbol{\theta})$ is unknown, the algorithm first uses a subset of $\mathcal{D}$ determined by the safety constraint and the boundedness assumption. Define

$$\mathcal{D}^w = \{\boldsymbol{x} \in \mathcal{D} : \max_{\|\boldsymbol{\nu}\|_2 \leq S} \boldsymbol{\nu}^\top \boldsymbol{B}\boldsymbol{x} \leq c_{\mathrm{safe}}\} = \{\boldsymbol{x} \in \mathcal{D} : \|\boldsymbol{B}\boldsymbol{x}\|_2 \leq c_{\mathrm{safe}}/S\}.$$

Every action in $\mathcal{D}^w$ is safe for all parameters satisfying Assumption 3.2, rather than only for the true parameter $\boldsymbol{\theta}$. Thus, when uncertainty about $\boldsymbol{\theta}$ prevents the learner from certifying actions outside $\mathcal{D}^w$, this set provides a safe source of reward observations for estimating $\boldsymbol{\theta}$. EPOCH-SLUCB uses this set for forced exploration, drawing independent samples from a fixed distribution ρ on $\mathcal{D}^w$. For $\boldsymbol{U} \sim \rho$, define $\lambda_- := \lambda_{\min}(\mathbb{E}[\boldsymbol{U}\boldsymbol{U}^\top]) > 0$.

Remark 5.1 (Role of λ_-). By definition, every unit vector $\boldsymbol{u}$ satisfies $\mathbb{E}[(\boldsymbol{u}^\top \boldsymbol{U})^2] \geq \lambda_-$. Thus, $\lambda_- > 0$ ensures that exploration samples provide variation in every direction. A known positive lower bound on λ_- can replace its exact value throughout the algorithm and regret bound. For invertible $\boldsymbol{B}$, the construction of Amani et al. [4, Remark 1] chooses $0 < \varepsilon \leq c_{\mathrm{safe}}/S$ such that $\{\boldsymbol{x} : \|\boldsymbol{B}\boldsymbol{x}\|_2 = \varepsilon\} \subseteq \mathcal{D}^w$, draws $\boldsymbol{Z}$ uniformly from the unit sphere, and sets $\boldsymbol{U} = \varepsilon \boldsymbol{B}^{-1}\boldsymbol{Z}$. This gives $\lambda_- = \varepsilon^2/(d\|\boldsymbol{B}\|^2)$. Alternatively, uniform sampling from a known Euclidean ball of radius r centered at the origin and contained in $\mathcal{D}^w$ gives $\lambda_- = r^2/(d+2)$, without requiring $\boldsymbol{B}$ to be invertible. Such a ball exists by Assumption 3.3 and $c_{\mathrm{safe}} > 0$.

The next lemma states that, after sufficiently many samples from $\mathcal{D}^w$, the resulting Gram matrix is well conditioned with high probability.

Lemma 5.2 (Gram matrix of samples from $\mathcal{D}^w$, Amani et al. [4]). *Let $T' \in \mathbb{N}_+$ and let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{T'}$ be independent samples from ρ. For any $\lambda_{\mathrm{ridge}} > 0$ and $\delta' \in (0,1)$, if $T' \geq \frac{8L^2}{\lambda_-} \log\left(\frac{d}{\delta'}\right)$, then, with*

probability at least $1-\delta'$*,*

$$\lambda_{\min}\left(\lambda_{\text{ridge}}\boldsymbol{I}+\sum_{i=1}^{T'}\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)\geq\lambda_{\text{ridge}}+\frac{\lambda_- T'}{2}.$$

Confidence set for $\boldsymbol{\theta}$. Following standard linear bandit algorithms [1], EPOCH-SLUCB estimates $\boldsymbol{\theta}$ by ridge least squares and constructs an ellipsoidal confidence set. Fix a regularization parameter $\lambda_{\text{ridge}}>0$. After t observations, let $\boldsymbol{M}_t=\lambda_{\text{ridge}}\boldsymbol{I}+\sum_{s=1}^{t}\boldsymbol{x}_s\boldsymbol{x}_s^\top$, $\boldsymbol{b}_t=\sum_{s=1}^{t}\boldsymbol{x}_s y_s$, and $\hat{\boldsymbol{\theta}}_t=\boldsymbol{M}_t^{-1}\boldsymbol{b}_t$, with $\boldsymbol{M}_0=\lambda_{\text{ridge}}\boldsymbol{I}$ and $\boldsymbol{b}_0=\boldsymbol{0}$.

Lemma 5.3 (Confidence set from Abbasi-Yadkori et al. [1])**.** *Suppose Assumptions 3.1 and 3.2 hold. For any $\delta\in(0,1)$, let*

$$\beta_t(\delta)=R\sqrt{d\log\left(\frac{1+tL^2/\lambda_{\text{ridge}}}{\delta}\right)}+\sqrt{\lambda_{\text{ridge}}}S,\qquad t=0,\dots,T.$$

Then, with probability at least $1-\delta$, it holds that $\|\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}\|_{\boldsymbol{M}_t}\leq\beta_t(\delta)$ for all $t\leq T$.

Using the confidence radius $\beta_t(\delta/2)$, let $\mathcal{C}_t=\{\boldsymbol{\nu}:\|\boldsymbol{\nu}-\hat{\boldsymbol{\theta}}_t\|_{\boldsymbol{M}_t}\leq\beta_t(\delta/2)\}$. Then, with probability at least $1-\delta/2$, we have $\boldsymbol{\theta}\in\mathcal{C}_t$ for all $t\leq T$.

Safety certification. The confidence set gives a natural way to certify safety. Define the certified safe set as

$$\mathcal{D}_t^{\text{safe}}=\{\boldsymbol{x}\in\mathcal{D}:\max_{\boldsymbol{\nu}\in\mathcal{C}_t}\boldsymbol{\nu}^\top\boldsymbol{B}\boldsymbol{x}\leq c_{\text{safe}}\}.$$

When $\boldsymbol{\theta}\in\mathcal{C}_t$, every action in $\mathcal{D}_t^{\text{safe}}$ is safe for the true parameter. The algorithm optimizes only over actions that are safe for every plausible value of $\boldsymbol{\theta}$.

Algorithm 1: EPOCH-SLUCB for unknown positive gap

Input: $T,\lambda_{\text{ridge}},\{\beta_t(\delta/2)\}_{t=0}^{T},\delta,\rho,\lambda_-$

Initialization: $\mathcal{D}^w=\{\boldsymbol{x}\in\mathcal{D}:\max_{\|\boldsymbol{\nu}\|_2\leq S}\boldsymbol{\nu}^\top\boldsymbol{B}\boldsymbol{x}\leq c_{\text{safe}}\}$, $t=0$, $\tau=0$

1 $t_\delta=\frac{8L^2}{\lambda_-}\log\left(\frac{2d(3+\log_2 T)}{\delta}\right)$

2 **for** $k=0,1,\dots$ **do**

3 $\quad\epsilon_k=2^{-k/2}$, $T_k=\frac{2\|\boldsymbol{B}\|^2L^2\beta_T^2(\delta/2)}{\lambda_-\epsilon_k^2}$, $H_k=2^kT_k$

4 $\quad$**while** $t<H_k$ **do**

5 $\quad\quad t=t+1$

6 $\quad\quad$**if** $\tau<T_k$ *or* $t\leq t_\delta$ **then**

7 $\quad\quad\quad\tau=\tau+1$

8 $\quad\quad\quad$Draw $\boldsymbol{x}_t\sim\rho$ and observe reward y_t

9 $\quad\quad$**else**

10 $\quad\quad\quad\boldsymbol{M}_{t-1}=\lambda_{\text{ridge}}\boldsymbol{I}+\sum_{i=1}^{t-1}\boldsymbol{x}_i\boldsymbol{x}_i^\top$, $\boldsymbol{b}_{t-1}=\sum_{i=1}^{t-1}\boldsymbol{x}_i y_i$, $\hat{\boldsymbol{\theta}}_{t-1}=\boldsymbol{M}_{t-1}^{-1}\boldsymbol{b}_{t-1}$

11 $\quad\quad\quad\mathcal{C}_{t-1}=\{\boldsymbol{\nu}:\|\boldsymbol{\nu}-\hat{\boldsymbol{\theta}}_{t-1}\|_{\boldsymbol{M}_{t-1}}\leq\beta_{t-1}(\delta/2)\}$

12 $\quad\quad\quad\mathcal{D}_{t-1}^{\text{safe}}=\{\boldsymbol{x}\in\mathcal{D}:\max_{\boldsymbol{\nu}\in\mathcal{C}_{t-1}}\boldsymbol{\nu}^\top\boldsymbol{B}\boldsymbol{x}\leq c_{\text{safe}}\}$

13 $\quad\quad\quad\boldsymbol{x}_t=\arg\max_{\boldsymbol{x}\in\mathcal{D}_{t-1}^{\text{safe}}}\max_{\boldsymbol{\nu}\in\mathcal{C}_{t-1}}\boldsymbol{\nu}^\top\boldsymbol{x}$

14 $\quad\quad\quad$Observe reward y_t

15 $\quad\quad$**if** $t\geq T$ **then**

16 $\quad\quad\quad$Terminate the algorithm

EPOCH-SLUCB. Algorithm 1 combines the above components using an epoch schedule, without requiring knowledge of Δ.

The initial exploration length t_δ is chosen using Lemma 5.2 (Line 1). At the start of epoch k, the algorithm sets the certification scale ϵ_k, the cumulative exploration target T_k, and the cumulative

round budget H_k (Line 3). The counter τ records the number of forced exploration samples. While $\tau < T_k$ or $t \le t_\delta$, the algorithm samples from ρ and increments τ. Otherwise, it chooses the action with the largest optimistic reward over $\mathcal{D}_{t-1}^{\mathrm{safe}}$ (Line 13).

The joint choice of the exploration targets T_k and round budgets H_k removes the need to know Δ. The scale ϵ_k acts as a working safety margin. Early epochs use larger margins and shorter cutoffs, while later epochs use smaller margins and allow the certified safe set to approach the true safe boundary. Once $2\epsilon_k \le \Delta$, the optimal safe action $\boldsymbol{x}^*$ can be included in the certified safe set. Before this happens, the cutoffs H_k keep the time spent at overly conservative scales under control.

Theorem 5.4 (Upper bound of EPOCH-SLUCB)**.** *Suppose $\Delta > 0$. For a confidence parameter $\delta \in (0,1)$, with probability at least $1-\delta$, all selected actions are safe and the regret of Algorithm 1 satisfies*

$$R_T \le 2t_\delta + 8\frac{\|\boldsymbol{B}\| L \beta_T(\frac{\delta}{2})}{\sqrt{\lambda_-}}\sqrt{T} + 2\beta_T(\frac{\delta}{2})\sqrt{2dT\log\Big(1+\frac{2TL^2}{d(\lambda_- t_\delta + 2\lambda_{\mathrm{ridge}})}\Big)} + 9\frac{\|\boldsymbol{B}\|^2 L^2 \beta_T^2(\frac{\delta}{2})}{c_{\mathrm{safe}}\,\lambda_-\,\Delta^3},$$

where $t_\delta = \frac{8L^2}{\lambda_-}\log\Big(\frac{2d(3+\log_2 T)}{\delta}\Big)$ as in Line 1, and $\lambda_- = \lambda_{\min}(\mathbb{E}[\boldsymbol{U}\boldsymbol{U}^\top])$ for $\boldsymbol{U} \sim \rho$.

The bound separates the regret incurred by forced exploration from the regret incurred by optimistic action selection. The first two terms, $2t_\delta + 8\|\boldsymbol{B}\| L \beta_T(\delta/2)\sqrt{T/\lambda_-}$, come from samples drawn from $\mathcal{D}^w$. The last two terms come from optimistic action selection. The third term is the standard linear bandit regret contribution from the uncertainty terms on optimistic rounds, after samples from $\mathcal{D}^w$ make the Gram matrix well conditioned. It follows from the elliptical potential lemma [1, Lemma 11] and the Cauchy–Schwarz inequality. The final term bounds the additional comparison loss before $\boldsymbol{x}^*$ is certified. Considering only the dependence on ϵ_k, each epoch has at most $O(\epsilon_k^{-4})$ rounds and the comparison term per optimistic round is $O(\epsilon_k)$, giving at most $O(\epsilon_k^{-3})$ per epoch. Summing over epochs with $2\epsilon_k > \Delta$ yields the Δ^{-3} dependence. The full proof is given in Appendix B.

For any fixed $\Delta > 0$, EPOCH-SLUCB achieves the same $\widetilde{O}(\sqrt{T})$ dependence on T as Safe-LUCB [4] without requiring prior knowledge of Δ. The two sampling constructions in Remark 5.1 give $\lambda_- = \Theta(1/d)$ when ε and $\|\boldsymbol{B}\|$ are fixed in the ellipsoidal construction, or r is fixed in the ball construction. With other problem parameters fixed, our bound becomes $\widetilde{O}(d\sqrt{T} + d^2/\Delta^3)$, whereas Theorem 2 of Amani et al. [4] gives $\widetilde{O}(d\sqrt{T} + d^2/\Delta^2)$ for known $\Delta > 0$. When Δ is small, however, the additive Δ^{-3} term can be large at finite horizons. Improving this dependence on Δ remains an open problem for future work.

Computational implementation. Since $\mathcal{C}_{t-1}$ is an ellipsoid, the optimistic action selection in Algorithm 1 is equivalent to

$$\begin{aligned} \max_{\boldsymbol{x}\in\mathcal{D}} \quad & \hat{\boldsymbol{\theta}}_{t-1}^\top \boldsymbol{x} + \beta_{t-1}(\delta/2)\|\boldsymbol{x}\|_{\boldsymbol{M}_{t-1}^{-1}} \\ \text{subject to} \quad & \hat{\boldsymbol{\theta}}_{t-1}^\top \boldsymbol{B}\boldsymbol{x} + \beta_{t-1}(\delta/2)\|\boldsymbol{B}\boldsymbol{x}\|_{\boldsymbol{M}_{t-1}^{-1}} \le c_{\mathrm{safe}}. \end{aligned}$$

This problem is generally nonconvex, and our regret analysis assumes an exact maximizer, as in Safe-LUCB [4]. Solving this problem with objective error at most $e_t \ge 0$, while maintaining feasibility, adds at most $\sum_{t\in\mathcal{O}} e_t$ to the regret bound, where $\mathcal{O}$ denotes the optimistic rounds.

6 Proof Sketch of Theorem 4.1

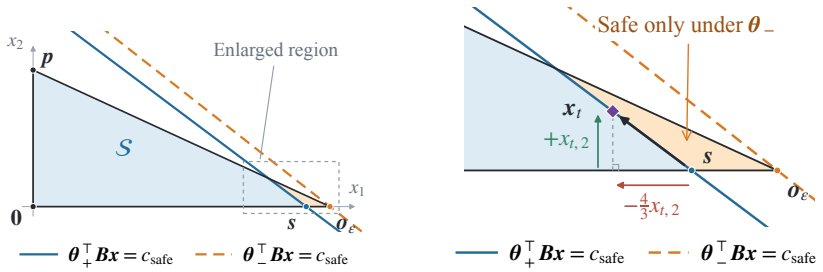
The lower bound is driven by two zero-gap instances that are hard to distinguish from safe reward observations but have different safety boundaries. Fix an integer $T \ge 1$, set $\varepsilon = \frac{1}{4}T^{-1/3}$, and consider

$$\boldsymbol{\theta}_+ = (1,1), \qquad \boldsymbol{\theta}_- = (1, 1-\varepsilon), \qquad \boldsymbol{B} = \begin{pmatrix} 1 & 1 \\ 1/2 & 1 \end{pmatrix}, \qquad c_{\mathrm{safe}} = 3/2,$$

so that the two safety constraints are

$$\boldsymbol{\theta}_+^\top \boldsymbol{B}\boldsymbol{x} = \frac{3}{2}x_1 + 2x_2 \le \frac{3}{2}, \qquad \boldsymbol{\theta}_-^\top \boldsymbol{B}\boldsymbol{x} = \Big(\frac{3}{2} - \frac{\varepsilon}{2}\Big)x_1 + (2-\varepsilon)x_2 \le \frac{3}{2}.$$

Let $\mathcal{S}$ be the convex hull of $\mathbf{0}$, $\boldsymbol{o}_\varepsilon = (3/(3-\varepsilon), 0)$, and $\boldsymbol{p} = (0, 1/2)$. The action set is a ρ-thickening of $\mathcal{S}$, with $\rho = \varepsilon^2/48$. This thickening is used to ensure that the origin lies in the interior of the decision set, as required by Assumption 3.3. The point $\boldsymbol{s} = (1, 0)$ lies on the $\boldsymbol{\theta}_+$-safety boundary, whereas $\boldsymbol{o}_\varepsilon$ lies on the $\boldsymbol{\theta}_-$-safety boundary and is unsafe under $\boldsymbol{\theta}_+$. In both instances, the optimal safe action lies on an active safety constraint, so the safety gap is zero. Figure 2 shows the triangle $\mathcal{S}$ before thickening.



(a) Two safety boundaries.

(b) Exploration under $\boldsymbol{\theta}_+$.

Figure 2: Geometry of the lower bound instance construction.

The geometry captures the following difficulty. If the true parameter is $\boldsymbol{\theta}_-$, then the action $\boldsymbol{o}_\varepsilon$ is feasible and improves the reward by at least order ε compared with actions that remain inside the $\boldsymbol{\theta}_+$-safe set. A safe learner therefore needs evidence favoring $\boldsymbol{\theta}_-$ before exploiting this reward advantage. This evidence is hard to obtain from safe observations. The two parameters differ only in the second coordinate, so the mean rewards under $\boldsymbol{\theta}_+$ and $\boldsymbol{\theta}_-$ differ by εx_2 at action $\boldsymbol{x}$. The two competing boundary points $\boldsymbol{s}$ and $\boldsymbol{o}_\varepsilon$ both have $x_2 = 0$, so reward observations at these points do not distinguish the two parameters. In the second case below, the learner leaves the $\boldsymbol{\theta}_+$-safe set with probability at least $1/2$ under $\boldsymbol{\theta}_-$, but at most $\delta \le 1/4$ under $\boldsymbol{\theta}_+$. This difference in probabilities implies that the expected sum of $x_{t,2}^2$ under $\boldsymbol{\theta}_+$ before exit is $\Omega(1/\varepsilon^2)$, as shown in Equation (3) below and Appendix A, Equation (10). Under $\boldsymbol{\theta}_+$, collecting such information is costly: the safety constraint forces x_1 to decrease when x_2 is used, moving the learner away from the reference point $\boldsymbol{s}$. Thus the construction creates a trade-off between staying conservative, which costs $T\varepsilon$ regret under $\boldsymbol{\theta}_-$, and gathering information, which costs $1/\varepsilon^2$ regret under $\boldsymbol{\theta}_+$.

We now turn this geometry into a testing argument. Let $\mathbb{P}_\sigma$ and $\mathbb{E}_\sigma$ denote probability and expectation under $\boldsymbol{\theta}_\sigma$, for $\sigma \in \{+, -\}$. Define the first time the learner exits the $\boldsymbol{\theta}_+$-safe set by

$$\tau = \inf\{t \le T : \boldsymbol{x}_t \notin \mathcal{D}^s(\boldsymbol{\theta}_+)\} \wedge (T+1), \qquad E = \{\tau \le T\}.$$

Safety under $\boldsymbol{\theta}_+$ implies $\mathbb{P}_+(E) \le \delta$. The rest of the proof separates two possibilities according to how often the same exit event E occurs under $\boldsymbol{\theta}_-$.

First suppose $\mathbb{P}_-(E) < 1/2$. Then, with constant probability under $\boldsymbol{\theta}_-$, the learner never leaves the more conservative $\boldsymbol{\theta}_+$-safe set. On this set, every action is separated from the $\boldsymbol{\theta}_-$-optimal safe reward by at least order ε: $\boldsymbol{\theta}_-^\top \boldsymbol{x}^*(\boldsymbol{\theta}_-) - \boldsymbol{\theta}_-^\top \boldsymbol{x} \ge \varepsilon/6$ for every $\boldsymbol{x} \in \mathcal{D}^s(\boldsymbol{\theta}_+)$. For any action in $\mathcal{D}$, the instantaneous regret under $\boldsymbol{\theta}_-$ is at least $-\sqrt{2}\rho$, as shown in Case 1 of Appendix A. Since $\mathbb{P}_-(E^c) > 1/2$, the choices $\varepsilon = T^{-1/3}/4$ and $\rho = \varepsilon^2/48$ give $\mathbb{E}_-[R_T] \ge T\varepsilon/12 - \sqrt{2}\rho T \ge T^{2/3}/96$.

It remains to handle the opposite case, $\mathbb{P}_-(E) \ge 1/2$. Here the exit event E distinguishes the two environments, since it is unlikely under $\boldsymbol{\theta}_+$ but occurs with constant probability under $\boldsymbol{\theta}_-$. This is the informative branch. The learner can only make this event likely under $\boldsymbol{\theta}_-$ if the reward observations before exit contain enough information about the second coordinate of $\boldsymbol{\theta}$. Let $\mathcal{G}_\tau$ denote the history containing all actions and rewards before τ, together with the exit action if $\tau \le T$, and let $P_\sigma^{\mathcal{G}_\tau}$ be its law under $\boldsymbol{\theta}_\sigma$. The indicator $\mathbf{1}_E$ is determined by this history and has distribution $\text{Bernoulli}(\mathbb{P}_\sigma(E))$ under $\boldsymbol{\theta}_\sigma$. Applying the data processing inequality [19, Exercise 14.10] to this indicator gives

$$\mathrm{KL}\left(P_+^{\mathcal{G}_\tau} \middle\| P_-^{\mathcal{G}_\tau}\right) \ge \mathrm{kl}(\mathbb{P}_+(E), \mathbb{P}_-(E)) \ge \mathrm{kl}(1/4, 1/2) > 1/10, \tag{1}$$

using $\mathbb{P}_+(E) \le \delta \le 1/4$, $\mathbb{P}_-(E) \ge 1/2$, and the fact that $\mathrm{kl}(p, q)$ decreases in p and increases in q for $p < q$. The action selection rule is the same under both environments. With independent standard

Gaussian noise, the two conditional reward distributions have variance 1 and means differing by $\varepsilon x_{t,2}$, hence KL divergence $\varepsilon^2 x_{t,2}^2/2$. Only rewards from rounds $t < \tau$ enter the stopped history. The KL chain rule [19, Exercises 15.7 and 15.8] therefore gives

$$\mathrm{KL}\Big(P_+^{\mathcal{G}_\tau} \Big\| P_-^{\mathcal{G}_\tau}\Big) = \frac{\varepsilon^2}{2}\,\mathbb{E}_+\left[\sum_{t<\tau} x_{t,2}^2\right]. \tag{2}$$

Combining Equations (1) and (2) gives

$$\mathbb{E}_+\left[\sum_{t<\tau} x_{t,2}^2\right] \geq \frac{1}{5\varepsilon^2}. \tag{3}$$

This information requirement becomes regret under $\boldsymbol{\theta}_+$. Indeed, by the definition of τ, every action $\boldsymbol{x}_t$ with $t < \tau$ belongs to $\mathcal{D}^s(\boldsymbol{\theta}_+)$. The safety constraint $\frac{3}{2}x_{t,1} + 2x_{t,2} \leq \frac{3}{2}$ gives $x_{t,1} \leq 1 - \frac{4}{3}x_{t,2}$, as illustrated in Figure 2(b). Since $\boldsymbol{s} = (1,0)$ is safe under $\boldsymbol{\theta}_+ = (1,1)$, the optimal safe reward is at least 1, so the instantaneous regret is at least $x_{t,2}/3$. Moreover, thickening gives $-\rho \leq x_{t,2} \leq 1/2 + \rho < 3/4$, which implies $x_{t,2} \geq x_{t,2}^2 - 2\rho$. Consequently, $\boldsymbol{\theta}_+^\top \boldsymbol{x}^*(\boldsymbol{\theta}_+) - \boldsymbol{\theta}_+^\top \boldsymbol{x}_t \geq x_{t,2}^2/3 - 2\rho/3$. The full derivation is given in Appendix A before Equation (11). For any action in $\mathcal{D}$, the instantaneous regret under $\boldsymbol{\theta}_+$ is at least $-(\varepsilon/(3-\varepsilon) + \sqrt{2}\rho)$. Post-exit rounds occur only on E, with $\mathbb{P}_+(E) \leq \delta$. Combining Equation (3) with the regret bounds before and after τ gives

$$\mathbb{E}_+[R_T] \geq \frac{1}{15\varepsilon^2} - \frac{2\rho T}{3} - \left(\frac{\varepsilon}{3-\varepsilon} + \sqrt{2}\rho\right)\delta T.$$

Using $\varepsilon = T^{-1/3}/4$, $\rho = \varepsilon^2/48$, and $\delta \leq 1/4$ gives $\mathbb{E}_+[R_T] \geq T^{2/3}$ for every $T \geq 1$.

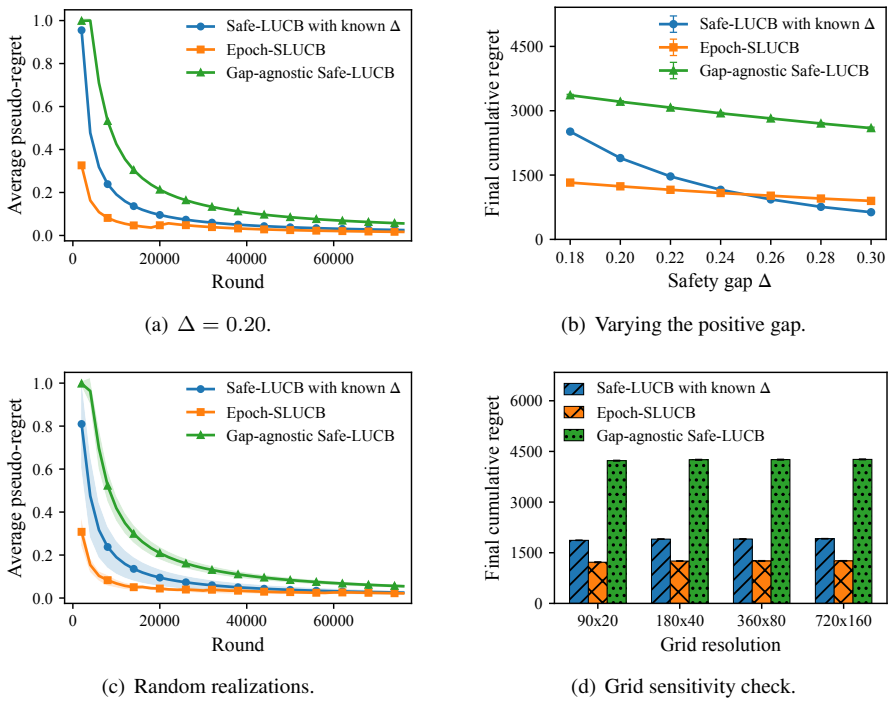
The two cases give the trade-off $T\varepsilon$ versus $1/\varepsilon^2$. Our choice of ε balances these terms, yielding the $\Omega(T^{2/3})$ lower bound. The construction is stated before normalization. The Gaussian noise satisfies Assumption 3.1. We scale both $\boldsymbol{\theta}_\pm$ and c_{safe} by $1/2$ to satisfy Assumption 3.2. This leaves the safe sets unchanged and preserves the $\Omega(T^{2/3})$ lower bound. The full proof is given in Appendix A.

7 Experiments

In this section, we validate the performance of EPOCH-SLUCB when the safety gap is positive but unknown. We compare it with two safe benchmarks from Amani et al. [4]: Safe-LUCB run with the true value of Δ, and the gap-agnostic version of Safe-LUCB corresponding to the $\widetilde{O}(T^{2/3})$ guarantee.

The main experiment uses a two-dimensional unit-ball instance, discretized by a polar grid. We set $\boldsymbol{\theta} = (1,0)$, $\boldsymbol{x}^* = (1,0)$, $\boldsymbol{B} = \begin{pmatrix} 0.4 & 0.6 \\ 0.8 & -0.2 \end{pmatrix}$, and $c_{\mathrm{safe}} = \boldsymbol{\theta}^\top \boldsymbol{B}\boldsymbol{x}^* + \Delta$. The instance is chosen so that $\boldsymbol{x}^* \notin \mathcal{D}^w$, so the optimal action cannot be used before the learner collects enough safe information to certify it. Rewards have Gaussian noise with standard deviation 0.05. All methods use the same grid, regularization, confidence radius, and noise level. We set the ridge regularization to $\lambda_{\mathrm{ridge}} = 0.5$ and the confidence parameter to $\delta = 0.1$, using $\beta_t(\delta/2)$ with $\delta/2 = 0.05$. Uniform sampling from the grid actions in $\mathcal{D}^w$ gives $\lambda_- = 0.0814$ in Figure 3(a). In Figure 3(b), λ_- ranges from 0.0759 to 0.1134, and in the random realizations of Figure 3(c), it ranges from 0.0655 to 0.1251. Across the grids in Figure 3(d), it ranges from 0.0808 to 0.0829. We report pseudo-regret and safety violations, where a violation means $\boldsymbol{\theta}^\top \boldsymbol{B}\boldsymbol{x}_t > c_{\mathrm{safe}}$.

Figure 3(a) compares the three algorithms for $\Delta = 0.20$ and $T = 76000$, averaged over 10 random seeds. EPOCH-SLUCB has the lowest mean final pseudo-regret on this instance, followed by Safe-LUCB with known Δ and gap-agnostic Safe-LUCB. The kinks in the EPOCH-SLUCB curves reflect switches between forced exploration and optimistic action selection, which incur different per-round pseudo-regret. Figure 3(b) varies Δ from 0.18 to 0.30, with $T = 50000$ and 10 random seeds per gap. All methods improve as the gap grows, because x^* can be certified with less precision. EPOCH-SLUCB has lower mean final pseudo-regret than Safe-LUCB with known Δ for gaps from 0.18 to 0.24, while the ordering reverses from 0.26 to 0.30. Both have lower regret than gap-agnostic Safe-LUCB throughout. Figure 3(c) evaluates the algorithms on 10 random problem instances with $\Delta = 0.20$. Each instance is generated by drawing random $\boldsymbol{\theta}$ and $\boldsymbol{B}$, and then setting c_{safe} to obtain the target gap. EPOCH-SLUCB has the lowest mean final pseudo-regret across these realizations.



(a) $\Delta = 0.20$.

(b) Varying the positive gap.

(c) Random realizations.

(d) Grid sensitivity check.

Figure 3: Simulation results. (a) Average per-step pseudo-regret over $T = 76000$ rounds, $\Delta = 0.20$. (b) Cumulative pseudo-regret over $T = 50000$ rounds as Δ varies. (c) Average per-step pseudo-regret over 10 random problem realizations with $\Delta = 0.20$, with bands showing one standard deviation across realizations. (d) Cumulative pseudo-regret as the polar grid varies.

Figure 3(d) uses the same instance as Figure 3(a) on polar grids from 90×20 to 720×160, using the same three random seeds at each resolution. Using denser grids than the 180×40 grid in Figure 3(a) changes final regret by less than one percent. Across all reported experiments, the number of safety violations is zero. Appendix C provides cumulative pseudo-regret plots corresponding to Figures 3(a) and (c) and additional experiments in dimensions $d = 5, 10, 20$.

8 Discussion and Conclusion

We study safe linear bandits with an unknown safety gap, and our results answer the two questions left open by Amani et al. [4]. We establish an $\Omega(T^{2/3})$ lower bound for zero-gap instances, matching the existing upper bound in T up to logarithmic factors. We also propose EPOCH-SLUCB, which achieves $\widetilde{O}(\sqrt{T})$ regret for every fixed $\Delta > 0$ without knowing the value of Δ, while maintaining per-round safety.

We discuss the limitations of our work. The analysis assumes that $\boldsymbol{B}$, c_{safe}, and a bound on $\|\boldsymbol{\theta}\|_2$ are known, and that reward and safety are governed by the same unknown parameter. EPOCH-SLUCB requires sampling from $\mathcal{D}^w$ and optimization over certified safe sets. Our experiments are synthetic and illustrate the positive-gap behavior suggested by Theorem 5.4. Our work also leaves several directions for future work. Natural extensions include models with generalized link functions, contextual actions, or mild model misspecification. While our lower bound establishes tightness in the dependence on T, identifying the optimal dependence on the dimension d remains open.

Acknowledgments and Disclosure of Funding

The authors thank the anonymous reviewers for the helpful comments and suggestions. The work of Zhuohua Li was supported in part by the Xiaomi Young Talent Program and the Fundamental Research Funds for the Central Universities under Grant QTZX26133. The work of Xiangxiang Dai was supported in part by the National Natural Science Foundation of China (625B2163). The work of John C.S. Lui was supported in part by the RGC SRFS2122-4S02 and GRF-14202925.

References

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

[2] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pages 22–31. Pmlr, 2017.

[3] Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. *Advances in neural information processing systems*, 29, 2016.

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

[5] Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Generalized linear bandits with safety constraints. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3562–3566. IEEE, 2020.

[6] Sanae Amani, Christos Thrampoulidis, and Lin Yang. Safe reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 243–253. PMLR, 2021.

[7] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.

[8] Felix Berkenkamp, Matteo Turchetta, Angela Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. *Advances in neural information processing systems*, 30, 2017.

[9] Felix Berkenkamp, Andreas Krause, and Angela P Schoellig. Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics. *Machine learning*, 112(10): 3713–3747, 2023.

[10] Maxime C Cohen, Ilan Lobel, and Renato Paes Leme. Feature-based dynamic pricing. *Management Science*, 66(11):4921–4943, 2020.

[11] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, number 101, pages 355–366, 2008.

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

[13] Aditya Gangrade, Tianrui Chen, and Venkatesh Saligrama. Safe linear bandits over unknown polytopes. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1755–1795. PMLR, 2024.

[14] Evrard Garcelon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Matteo Pirotta. Improved algorithms for conservative exploration in bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3962–3969, 2020.

[15] Spencer Hutchinson, Berkay Turan, and Mahnoosh Alizadeh. Directional optimism for safe linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 658–666. PMLR, 2024.

[16] Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. *Advances in neural information processing systems*, 30, 2017.

[17] Kia Khezeli and Eilyan Bitar. Safe linear stochastic bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10202–10209, 2020.

[18] Akshay Krishnamurthy, Alekh Agarwal, and Miro Dudik. Contextual semibandits via supervised learning oracles. *Advances In Neural Information Processing Systems*, 29, 2016.

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

[20] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

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

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

[23] Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. In *International conference on artificial intelligence and statistics*, pages 2827–2835. PMLR, 2021.

[24] Aldo Pacchiano, Mohammad Ghavamzadeh, and Peter Bartlett. Contextual bandits with stage-wise constraints. *Journal of machine learning research*, 26(170):1–57, 2025.

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

[26] Matteo Turchetta, Felix Berkenkamp, and Andreas Krause. Safe exploration in finite markov decision processes with gaussian processes. *Advances in neural information processing systems*, 29, 2016.

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

A Proof of Theorem 4.1

Theorem 4.1 (Lower bound)**.** *There exist a universal constant $c_0 > 0$, a matrix $\boldsymbol{B} \in \mathbb{R}^{2\times 2}$, and a safety threshold $c_{\mathrm{safe}} > 0$ such that, for every integer $T \geq 1$, there exist a decision set $\mathcal{D}_T \subset \mathbb{R}^2$ and a parameter set $\Theta_T = \{\boldsymbol{\theta}_+, \boldsymbol{\theta}_-\} \subset \mathbb{R}^2$ with the following properties. Each instance $(\boldsymbol{B}, \mathcal{D}_T, c_{\mathrm{safe}}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta_T$, with independent standard Gaussian noise, satisfies Assumptions 3.1–3.3 and has zero safety gap. For these instances, for every $\delta \in (0, 1/4]$, any policy π satisfying*

$$\mathbb{P}_{\boldsymbol{\theta}}^{\pi}\left(\boldsymbol{\theta}^\top \boldsymbol{B}\boldsymbol{x}_t \leq c_{\mathrm{safe}},\ \forall t \in [T]\right) \geq 1-\delta \qquad \textit{for each } \boldsymbol{\theta} \in \Theta_T$$

also satisfies $\max_{\boldsymbol{\theta}\in\Theta_T} \mathbb{E}_{\boldsymbol{\theta}}^{\pi}\left[R_T(\boldsymbol{\theta})\right] \geq c_0 T^{2/3}$.

Proof. **Instance Construction.** Fix an integer $T \geq 1$. We first give an unscaled construction and enforce boundedness at the end of the proof. Let $\varepsilon = 1/4T^{-1/3}$, $\rho = \varepsilon^2/48$, and $\sigma \in \{+,-\}$. Define the two parameters, the constraint matrix, and safety threshold

$$\boldsymbol{\theta}_+ = (1,1),\ \boldsymbol{\theta}_- = (1, 1-\varepsilon),\ \boldsymbol{B} = \begin{pmatrix} 1 & 1 \\ \frac{1}{2} & 1 \end{pmatrix},\ c_{\mathrm{safe}} = \frac{3}{2}.$$

The unknown parameter $\boldsymbol{\theta}$ belongs to the set $\Theta = \{\boldsymbol{\theta}_+, \boldsymbol{\theta}_-\}$. Let $\mathcal{S}$ be the convex hull of $\mathbf{0}$, $\boldsymbol{o}_\varepsilon = (3/(3-\varepsilon), 0)$, and $\boldsymbol{p} = (0, 1/2)$. We define the convex compact decision set as $\mathcal{D}_\rho = \mathcal{S} + \rho\,\mathbb{B}_2^2$, where $\mathbb{B}_2^2 := \{\boldsymbol{x} \in \mathbb{R}^2 : \|\boldsymbol{x}\|_2 \leq 1\}$ is the closed unit ball. The thickening by $\rho\,\mathbb{B}_2^2$ ensures that $\mathbf{0} \in \mathrm{int}(\mathcal{D}_\rho)$, so that Assumption 3.3 is satisfied. For convenience, we denote the point $\boldsymbol{s} := (1,0) \in \mathcal{D}_\rho$.

At each round the learner selects $\boldsymbol{x}_t \in \mathcal{D}_\rho$ and observes $y_t = \boldsymbol{\theta}^\top \boldsymbol{x}_t + \eta_t$ where the noise variables η_t are independent and standard Gaussian, satisfying Assumption 3.1.

For each $\boldsymbol{\theta}_\sigma$, the safe decision set is

$$\mathcal{D}_\rho^s(\boldsymbol{\theta}_\sigma) := \{\boldsymbol{x} \in \mathcal{D}_\rho : \boldsymbol{\theta}_\sigma^\top \boldsymbol{B}\boldsymbol{x} \leq c_{\mathrm{safe}}\}.$$

Define the optimal safe action as $\boldsymbol{x}_\sigma^* := \arg\max_{\boldsymbol{x}\in\mathcal{D}_\rho^s(\boldsymbol{\theta}_\sigma)} \boldsymbol{\theta}_\sigma^\top \boldsymbol{x}$ and the optimal reward as $V_\sigma := \boldsymbol{\theta}_\sigma^\top \boldsymbol{x}_\sigma^*$. The regret is

$$R_T(\boldsymbol{\theta}_\sigma) := \textstyle\sum_{t=1}^T \big(V_\sigma - \boldsymbol{\theta}_\sigma^\top \boldsymbol{x}_t\big).$$

Under $\boldsymbol{\theta}_+$ and $\boldsymbol{\theta}_-$, the safety constraint $\boldsymbol{\theta}_\sigma^\top \boldsymbol{B}\boldsymbol{x} \leq c_{\mathrm{safe}}$ takes the form

$$\boldsymbol{\theta}_+^\top \boldsymbol{B}\boldsymbol{x} = \tfrac{3}{2}x_1 + 2x_2 \leq \tfrac{3}{2}, \qquad \boldsymbol{\theta}_-^\top \boldsymbol{B}\boldsymbol{x} = (\tfrac{3}{2} - \tfrac{\varepsilon}{2})x_1 + (2-\varepsilon)x_2 \leq \tfrac{3}{2}.$$

Since $\boldsymbol{\theta}_+^\top \boldsymbol{B}\boldsymbol{s} = 3/2$ and $\boldsymbol{\theta}_+^\top \boldsymbol{B}\boldsymbol{o}_\varepsilon = \frac{3}{2}\cdot\frac{3}{3-\varepsilon} > 3/2$, $\boldsymbol{s}$ is a safe action while $\boldsymbol{o}_\varepsilon$ is unsafe under $\boldsymbol{\theta}_+$. On the other hand, $\boldsymbol{\theta}_-^\top \boldsymbol{B}\boldsymbol{s} = \frac{3-\varepsilon}{2} < 3/2$ and $\boldsymbol{\theta}_-^\top \boldsymbol{B}\boldsymbol{o}_\varepsilon = (\frac{3}{2} - \frac{\varepsilon}{2})\cdot\frac{3}{3-\varepsilon} = c_{\mathrm{safe}}$, so $\boldsymbol{o}_\varepsilon$ lies on the safety boundary under $\boldsymbol{\theta}_-$. In particular, we have $V_+ \geq \boldsymbol{\theta}_+^\top \boldsymbol{s} = 1$ and $V_- \geq \boldsymbol{\theta}_-^\top \boldsymbol{o}_\varepsilon = 3/(3-\varepsilon)$.

We now show the safety gap $\Delta_\sigma = c_{\mathrm{safe}} - \boldsymbol{\theta}_\sigma^\top \boldsymbol{B}\boldsymbol{x}_\sigma^* = 0$ for both environments $\sigma \in \{+,-\}$. Since $\boldsymbol{\theta}_\sigma^\top \boldsymbol{x}$ is linear and $\mathcal{D}_\rho = \mathcal{S} + \rho\,\mathbb{B}_2^2$, the unconstrained maximizer over $\mathcal{D}_\rho$ is

$$\boldsymbol{x}_\sigma^{\mathrm{uc}} := \boldsymbol{o}_\varepsilon + \rho\,\frac{\boldsymbol{\theta}_\sigma}{\|\boldsymbol{\theta}_\sigma\|_2},$$

since $\boldsymbol{o}_\varepsilon$ maximizes $\boldsymbol{\theta}_\sigma^\top \boldsymbol{z}$ over $\mathcal{S}$. From $\boldsymbol{\theta}_\sigma^\top \boldsymbol{B}\boldsymbol{o}_\varepsilon \geq c_{\mathrm{safe}}$ and $\boldsymbol{B}^\top\boldsymbol{\theta}_\sigma$ having strictly positive entries,

$$\boldsymbol{\theta}_\sigma^\top \boldsymbol{B}\boldsymbol{x}_\sigma^{\mathrm{uc}} \geq c_{\mathrm{safe}} + \rho\,\frac{(\boldsymbol{B}^\top\boldsymbol{\theta}_\sigma)^\top\boldsymbol{\theta}_\sigma}{\|\boldsymbol{\theta}_\sigma\|_2} > c_{\mathrm{safe}},$$

so $\boldsymbol{x}_\sigma^{\mathrm{uc}} \notin \mathcal{D}_\rho^s(\boldsymbol{\theta}_\sigma)$. By convexity of $\mathcal{D}_\rho$, any safe point at which the constraint is inactive can be perturbed toward $\boldsymbol{x}_\sigma^{\mathrm{uc}}$ to increase the linear objective $\boldsymbol{\theta}_\sigma^\top \boldsymbol{x}$ while remaining safe. Hence the safety constraint must be active at $\boldsymbol{x}_\sigma^*$, giving $\Delta_\sigma = 0$.

Intuition. Before the formal argument, we outline the source of the $T^{2/3}$ lower bound, which arises from a fundamental tension between safe exploration and regret minimization under zero safety gap. The two instances $\boldsymbol{\theta}_+$ and $\boldsymbol{\theta}_-$ differ only in the second coordinate by ε, and the optimal safe action under each lies on its respective safety boundary: near $\boldsymbol{s} = (1,0)$ under $\boldsymbol{\theta}_+$, and near $\boldsymbol{o}_\varepsilon = (\frac{3}{3-\varepsilon}, 0)$ under $\boldsymbol{\theta}_-$. Crucially, $\boldsymbol{o}_\varepsilon$ is strictly unsafe under $\boldsymbol{\theta}_+$. This forces the learner into a dilemma:

- To achieve low regret under $\boldsymbol{\theta}_-$, the learner must play actions near $\boldsymbol{o}_\varepsilon$, which lies outside $\mathcal{D}_\rho^s(\boldsymbol{\theta}_+)$. If the learner remains within $\mathcal{D}_\rho^s(\boldsymbol{\theta}_+)$, it incurs a per-round reward gap of at least $\varepsilon/6$ under $\boldsymbol{\theta}_-$.
- To distinguish $\boldsymbol{\theta}_+$ from $\boldsymbol{\theta}_-$, the learner must explore in the second coordinate, where the two parameters differ. Under $\boldsymbol{\theta}_+$, the unthickened optimal boundary point $\boldsymbol{s} = (1, 0)$ has zero second coordinate, while thickening perturbs the optimal safe action by at most $O(\rho)$. Thus such exploration incurs a quadratic regret cost, yielding an $\Omega(1/\varepsilon^2)$ contribution up to the lower-order $O(\rho T)$ correction caused by thickening.

The choice $\varepsilon = \Theta(T^{-1/3})$ balances the two costs $T\varepsilon$ and $1/\varepsilon^2$. We now formalize this dichotomy via a case split based on whether the learner's policy exits $\mathcal{D}_\rho^s(\boldsymbol{\theta}_+)$ with non-trivial probability under $\boldsymbol{\theta}_-$.

Stopping Time and Case Split. Define the stopping time

$$\tau := \inf\{t \in [T] : \boldsymbol{x}_t \notin \mathcal{D}_\rho^s(\boldsymbol{\theta}_+)\} \wedge (T+1)$$

and the event $E := \{\tau \le T\}$. Thus $t < \tau$ means that the action at round t lies in $\mathcal{D}_\rho^s(\boldsymbol{\theta}_+)$, whereas $\tau \le T$ means that the policy π exits $\mathcal{D}_\rho^s(\boldsymbol{\theta}_+)$ by time T. The safety condition under $\boldsymbol{\theta}_+$ gives

$$\mathbb{P}^\pi_{\boldsymbol{\theta}_+}(E) \le \delta.$$

For brevity, write $\mathbb{P}_\pm := \mathbb{P}^\pi_{\boldsymbol{\theta}_\pm}$ and $\mathbb{E}_\pm := \mathbb{E}^\pi_{\boldsymbol{\theta}_\pm}$.

We consider two cases based on $\mathbb{P}_-(E)$.

Case 1: $\mathbb{P}_-(E) < 1/2$**.** We show $\mathbb{E}_-[R_T(\boldsymbol{\theta}_-)] \ge T^{2/3}/96$.

We first quantify the regret incurred by staying inside $\mathcal{D}_\rho^s(\boldsymbol{\theta}_+)$ under $\boldsymbol{\theta}_-$. Observe that

$$\boldsymbol{\theta}_- = (1, 1-\varepsilon) = \frac{1+3\varepsilon}{4}(1,0) + \frac{1-\varepsilon}{2}(\frac{3}{2}, 2).$$

For $\boldsymbol{x} = (x_1, x_2)$, we have

$$\boldsymbol{\theta}_-^\top \boldsymbol{x} = \frac{1+3\varepsilon}{4} x_1 + \frac{1-\varepsilon}{2} \boldsymbol{\theta}_+^\top \boldsymbol{B}\boldsymbol{x}.$$

For $\boldsymbol{x} \in \mathcal{D}_\rho = \mathcal{S} + \rho\mathbb{B}_2^2$, we can write $\boldsymbol{x} = \boldsymbol{z} + \boldsymbol{w}$ with $\boldsymbol{z} = (z_1, z_2) \in \mathcal{S}$ and $\|\boldsymbol{w}\|_2 \le \rho$. By the definition of $\mathcal{S}$, we have $z_2 \ge 0$, thus $x_2 \ge -\rho$. By the safety constraint, for $\boldsymbol{x} \in \mathcal{D}_\rho^s(\boldsymbol{\theta}_+)$, we have

$$\frac{3}{2} x_1 + 2x_2 \le \frac{3}{2},\ x_1 \le 1 - \frac{4}{3}x_2 \le 1 + \frac{4}{3}\rho.$$

Moreover, $\boldsymbol{\theta}_+^\top \boldsymbol{B}\boldsymbol{x} \le 3/2$ if $\boldsymbol{x} \in \mathcal{D}_\rho^s(\boldsymbol{\theta}_+)$. Therefore, we have

$$\boldsymbol{\theta}_-^\top \boldsymbol{x} \le \frac{1+3\varepsilon}{4}(1 + \frac{4}{3}\rho) + \frac{3(1-\varepsilon)}{4} \le 1 + \frac{4}{3}\rho.$$

Since $V_- \ge 3/(3-\varepsilon) \ge 1 + \varepsilon/3$, we have

$$V_- - \boldsymbol{\theta}_-^\top \boldsymbol{x} \ge \frac{\varepsilon}{3} - \frac{4}{3}\rho = \frac{\varepsilon}{3} - \frac{\varepsilon^2}{36} \ge \frac{\varepsilon}{6}, \tag{4}$$

for $\varepsilon \le 1$.

On the event $E^c = \{\tau = T+1\}$, $\boldsymbol{x}_t \in \mathcal{D}_\rho^s(\boldsymbol{\theta}_+)$ for $t \le T$. By Equation (4),

$$R_T(\boldsymbol{\theta}_-) = \sum_{t=1}^{T} (V_- - \boldsymbol{\theta}_-^\top \boldsymbol{x}_t) \ge \frac{T\varepsilon}{6}. \tag{5}$$

Next, we bound the instantaneous regret uniformly over $\boldsymbol{x} \in \mathcal{D}_\rho$ on the event E. Recall that the unconstrained maximizer of $\boldsymbol{\theta}_-^\top \boldsymbol{x}$ over $\mathcal{D}_\rho$ is

$$\boldsymbol{x}_-^{\mathrm{uc}} = \boldsymbol{o}_\varepsilon + \rho \frac{\boldsymbol{\theta}_-}{\|\boldsymbol{\theta}_-\|_2}.$$

and hence

$$\boldsymbol{\theta}_-^\top \boldsymbol{x}_t \le \max_{\boldsymbol{x}\in\mathcal{D}_\rho} \boldsymbol{\theta}_-^\top \boldsymbol{x} = \boldsymbol{\theta}_-^\top \boldsymbol{o}_\varepsilon + \rho\|\boldsymbol{\theta}_-\|_2 = \frac{3}{3-\varepsilon} + \rho\|\boldsymbol{\theta}_-\|_2.$$

Since $V_- \ge \boldsymbol{\theta}_-^\top \boldsymbol{o}_\varepsilon$ and $\|\boldsymbol{\theta}_-\|_2 < \sqrt{2}$, it follows that

$$V_- - \boldsymbol{\theta}_-^\top \boldsymbol{x}_t \ge -\rho\|\boldsymbol{\theta}_-\|_2 \ge -\rho\sqrt{2}.$$

Therefore, on the event E, we have

$$R_T(\boldsymbol{\theta}_-) = \sum_{t=1}^{T} \big(V_- - \boldsymbol{\theta}_-^\top \boldsymbol{x}_t\big) \ge -T\rho\sqrt{2}. \tag{6}$$

Combining Equations (5) and (6), we obtain

$$\begin{aligned}
\mathbb{E}_-[R_T(\boldsymbol{\theta}_-)] &= \mathbb{E}_-\big[R_T(\boldsymbol{\theta}_-)\mathbf{1}_{E^c}\big] + \mathbb{E}_-\big[R_T(\boldsymbol{\theta}_-)\mathbf{1}_E\big] \\
&\ge \frac{T\varepsilon}{6}\,\mathbb{P}_-(E^c) - T\rho\sqrt{2}\,\mathbb{P}_-(E) \\
&\ge \frac{T\varepsilon}{12} - T\rho\sqrt{2}.
\end{aligned}$$

Substituting $\rho = \varepsilon^2/48$, we get

$$T\rho\sqrt{2} = \frac{\sqrt{2}}{48}\,\varepsilon^2 T \le \frac{\sqrt{2}}{48}\,\varepsilon T < \frac{T\varepsilon}{24}.$$

Since $\varepsilon = 1/4T^{-1/3}$, it follows that

$$\mathbb{E}_-[R_T(\boldsymbol{\theta}_-)] \ge \frac{T\varepsilon}{24} \ge \frac{1}{96}\,T^{2/3}. \tag{7}$$

Case 2: $\mathbb{P}_-(E) \ge 1/2$**.** We show $\mathbb{E}_+[R_T(\boldsymbol{\theta}_+)] \ge T^{2/3}$.

Let $\mathcal{G}_t = \sigma(\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_{t-1}, y_{t-1}, \boldsymbol{x}_t)$ for $t \le T$, and let $\mathcal{G}_{T+1} = \sigma(\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_T, y_T)$. Thus $\mathcal{G}_t$ is the history just after the action at round t is selected and before its reward is observed. Let $P_\sigma^{\mathcal{G}_\tau}$ denote the law of the stopped history under $\boldsymbol{\theta}_\sigma$. The event $E = \{\tau \le T\}$ is $\mathcal{G}_\tau$-measurable, since it is determined once the first exit action is selected. Under $\boldsymbol{\theta}_\sigma$, the indicator $\mathbf{1}_E$ has distribution $\mathrm{Bernoulli}(\mathbb{P}_\sigma(E))$. Applying the data processing inequality [19, Exercise 14.10] to the mapping from the stopped history to $\mathbf{1}_E$ gives

$$\mathrm{KL}(P_+^{\mathcal{G}_\tau} \| P_-^{\mathcal{G}_\tau}) \ge \mathrm{kl}(\mathbb{P}_+(E), \mathbb{P}_-(E)) \ge \mathrm{kl}(1/4, 1/2) > \frac{1}{10} \tag{8}$$

using $\mathbb{P}_+(E) \le \delta \le 1/4$, $\mathbb{P}_-(E) \ge 1/2$, and the fact that the Bernoulli KL divergence $\mathrm{kl}(p, q)$ is decreasing in p and increasing in q for $p < q$.

On the other hand, the two environments $\boldsymbol{\theta}_+$ and $\boldsymbol{\theta}_-$ differ only in the second coordinate and $\boldsymbol{\theta}_+ - \boldsymbol{\theta}_- = (0, \varepsilon)$. The action-selection kernel is the same under the two environments, so it contributes no conditional KL. The stopped history contains rewards only from rounds $t < \tau$. At each of these rounds, given the same history and action $\boldsymbol{x}_t$, the reward distributions under $\boldsymbol{\theta}_+$ and $\boldsymbol{\theta}_-$ are Gaussian with variance 1 and mean difference $\varepsilon x_{t,2}$, so their KL divergence is $\varepsilon^2 x_{t,2}^2/2$. Applying the KL chain rule to the stopped history (see Exercises 15.7 and 15.8 in Lattimore and Szepesvári [19]) gives

$$\mathrm{KL}(P_+^{\mathcal{G}_\tau} \| P_-^{\mathcal{G}_\tau}) = \frac{\varepsilon^2}{2}\mathbb{E}_+\Big[\sum_{t=1}^{\tau-1} x_{t,2}^2\Big]. \tag{9}$$

Combining Equations (8) and (9):

$$\mathbb{E}_+\Big[\sum_{t=1}^{\tau-1} x_{t,2}^2\Big] \ge \frac{1}{5\varepsilon^2}. \tag{10}$$

Equation (10) shows that distinguishing $\boldsymbol{\theta}_+$ from $\boldsymbol{\theta}_-$ requires a nontrivial amount of exploration in the second coordinate before the stopping time τ. We next convert this exploration cost into regret under $\boldsymbol{\theta}_+$.

For pre-exit rounds $t < \tau$, we have $\boldsymbol{x}_t \in \mathcal{D}_\rho^s(\boldsymbol{\theta}_+)$, thus

$$\frac{3}{2}\, x_{t,1} + 2x_{t,2} \le \frac{3}{2},\ x_{t,1} \le 1 - \frac{4}{3}x_{t,2}.$$

Therefore, we have

$$\boldsymbol{\theta}_+^\top \boldsymbol{x}_t = x_{t,1} + x_{t,2} \le 1 - \frac{x_{t,2}}{3}.$$

Since $V_+ \ge 1$,

$$V_+ - \boldsymbol{\theta}_+^\top \boldsymbol{x}_t \ge \frac{x_{t,2}}{3}.$$

Note that $\boldsymbol{x}_t \in \mathcal{D}_\rho = \mathcal{S} + \rho\mathbb{B}_2^2$. For any point $\boldsymbol{z} \in \mathcal{S}$, we have $0 \le z_2 \le 1/2$. Thus, $-\rho \le x_{t,2} \le \rho + 1/2 < 3/4$. For every $u \in [-\rho, 3/4]$, it holds that $u \ge u^2 - 2\rho$. Therefore,

$$V_+ - \boldsymbol{\theta}_+^\top \boldsymbol{x}_t \ge \frac{x_{t,2}^2}{3} - \frac{2\rho}{3},$$

and

$$\sum_{t=1}^{\tau-1}(V_+ - \boldsymbol{\theta}_+^\top \boldsymbol{x}_t) \ge \frac{1}{3}\sum_{t=1}^{\tau-1} x_{t,2}^2 - \frac{2\rho T}{3}. \tag{11}$$

Now we consider post-exit rounds $\tau \le t \le T$, which occur only on the event E. We again bound the instantaneous regret uniformly over $\boldsymbol{x} \in \mathcal{D}_\rho$. The unconstrained maximum of $\boldsymbol{\theta}_+^\top \boldsymbol{x}$ over $\mathcal{D}_\rho$ is $3/(3-\varepsilon) + \rho\sqrt{2}$. Therefore,

$$V_+ - \boldsymbol{\theta}_+^\top \boldsymbol{x}_t \ge V_+ - \max_{\boldsymbol{x}\in\mathcal{D}_\rho} \boldsymbol{\theta}_+^\top \boldsymbol{x} \ge 1 - \frac{3}{3-\varepsilon} - \rho\sqrt{2} = -(\frac{\varepsilon}{3-\varepsilon} + \rho\sqrt{2}).$$

Since $\mathbb{P}_+(E) \le \delta$, it follows that

$$\begin{aligned}\mathbb{E}_+\left[\mathbf{1}_E \sum_{t=\tau}^{T}\left(V_+ - \boldsymbol{\theta}_+^\top \boldsymbol{x}_t\right)\right] &\ge -\left(\frac{\varepsilon}{3-\varepsilon} + \rho\sqrt{2}\right) T\mathbb{P}_+(E) \\ &\ge -\left(\frac{\varepsilon}{3-\varepsilon} + \rho\sqrt{2}\right)\delta T.\end{aligned} \tag{12}$$

Using the decomposition

$$R_T(\boldsymbol{\theta}_+) = \sum_{t=1}^{\tau-1}\left(V_+ - \boldsymbol{\theta}_+^\top \boldsymbol{x}_t\right) + \mathbf{1}_E \sum_{t=\tau}^{T}\left(V_+ - \boldsymbol{\theta}_+^\top \boldsymbol{x}_t\right),$$

and combining Equations (10), (11), and (12), we obtain

$$\begin{aligned}\mathbb{E}_+[R_T(\boldsymbol{\theta}_+)] &\ge \frac{1}{3}\mathbb{E}_+\left[\sum_{t=1}^{\tau-1} x_{t,2}^2\right] - \frac{2\rho T}{3} - \left(\frac{\varepsilon}{3-\varepsilon} + \rho\sqrt{2}\right)\delta T \\ &\ge \frac{1}{15\varepsilon^2} - \frac{2\rho T}{3} - \left(\frac{\varepsilon}{3-\varepsilon} + \rho\sqrt{2}\right)\delta T.\end{aligned} \tag{13}$$

Since $\varepsilon = \frac{1}{4}T^{-1/3}$, $\rho = \varepsilon^2/48$, and $\delta \le 1/4$, we have $\frac{1}{15\varepsilon^2} = \frac{16}{15}T^{2/3}$, $\rho T = T^{1/3}/768$, and $\frac{\varepsilon}{3-\varepsilon}\delta T \le \frac{1}{44}T^{2/3}$. Using $T^{1/3} \le T^{2/3}$ for $T \ge 1$ gives

$$\mathbb{E}_+[R_T(\boldsymbol{\theta}_+)] \ge \left(\frac{16}{15} - \frac{1}{44} - \frac{1}{1152} - \frac{\sqrt{2}}{3072}\right) T^{2/3} \ge T^{2/3}. \tag{14}$$

Therefore, from Equations (7) and (14), there exists $\boldsymbol{\theta} \in \Theta$ such that

$$\mathbb{E}_{\boldsymbol{\theta}}^{\pi}[R_T(\boldsymbol{\theta})] \ge c_1 T^{2/3},$$

for some constant $c_1 > 0$.

It remains only to enforce the boundedness requirements in Assumption 3.2. Take $\kappa = 1/2$ and define the scaled parameters and threshold by $\bar{\boldsymbol{\theta}}_\sigma = \kappa\boldsymbol{\theta}_\sigma$ and $\bar{c}_{\mathrm{safe}} = \kappa c_{\mathrm{safe}} = 3/4$, leaving $\boldsymbol{B}$, $\mathcal{D}_\rho$, and the noise distribution unchanged. For every $T \geq 1$, we have $\|\boldsymbol{\theta}_\sigma\|_2 \leq \sqrt{2}$ and $\|\boldsymbol{x}\|_2 \leq 12/11 + 1/768 < \sqrt{2}$ for all $\boldsymbol{x} \in \mathcal{D}_\rho$ and $\sigma \in \{+, -\}$. Hence $\|\kappa\boldsymbol{\theta}_\sigma\|_2 \leq 1$ and $|\kappa\boldsymbol{\theta}_\sigma^\top \boldsymbol{x}| \leq 1$ by the Cauchy–Schwarz inequality.

Since $\kappa > 0$, this scaling leaves the safe sets, the optimal safe actions, the stopping event E, and the zero-gap property unchanged. Moreover, the optimal rewards and all instantaneous reward gaps are multiplied by κ.

Let $\bar{\mathbb{P}}_\sigma$ and $\bar{\mathbb{E}}_\sigma$ denote probability and expectation under the scaled environment $\bar{\boldsymbol{\theta}}_\sigma$. Let $\bar{P}_\sigma^{\mathcal{G}_\tau}$ denote the corresponding law of the stopped history. We repeat the preceding case split under these scaled environments.

In Case 1, if $\bar{\mathbb{P}}_-(E) < 1/2$, the preceding geometric argument gives

$$\bar{\mathbb{E}}_- \left[R_T(\bar{\boldsymbol{\theta}}_-)\right] \geq \frac{\kappa}{96} T^{2/3}.$$

In Case 2, if $\bar{\mathbb{P}}_-(E) \geq 1/2$, the stopped KL identity becomes

$$\mathrm{KL}\left(\bar{P}_+^{\mathcal{G}_\tau} \middle\| \bar{P}_-^{\mathcal{G}_\tau}\right) = \frac{\kappa^2\varepsilon^2}{2} \bar{\mathbb{E}}_+ \left[\sum_{t=1}^{\tau-1} x_{t,2}^2\right].$$

The same data-processing argument therefore gives $\bar{\mathbb{E}}_+ \left[\sum_{t=1}^{\tau-1} x_{t,2}^2\right] \geq \frac{1}{5\kappa^2\varepsilon^2}$. Since the regret inequalities are multiplied by κ, we obtain

$$\bar{\mathbb{E}}_+ \left[R_T(\bar{\boldsymbol{\theta}}_+)\right] \geq \frac{1}{15\kappa\varepsilon^2} - \frac{2\kappa\rho T}{3} - \kappa\left(\frac{\varepsilon}{3-\varepsilon} + \rho\sqrt{2}\right)\delta T.$$

Since $1/\kappa \geq \kappa$, the same estimates used in Equation (14) show that the right-hand side is at least $\kappa T^{2/3}$ for every $T \geq 1$.

The Gaussian noise distribution is unchanged by the scaling and remains 1-sub-Gaussian. By the choice of κ, the scaled construction satisfies Assumptions 3.1–3.3 and retains the zero-gap property.

Combining the two scaled cases, for every integer $T \geq 1$,

$$\max_{\sigma\in\{+,-\}} \bar{\mathbb{E}}_\sigma \left[R_T(\bar{\boldsymbol{\theta}}_\sigma)\right] \geq \frac{\kappa}{96} T^{2/3} = \frac{1}{192} T^{2/3}.$$

Set $\mathcal{D}_T = \mathcal{D}_\rho$ and $\Theta_T = \{\bar{\boldsymbol{\theta}}_+, \bar{\boldsymbol{\theta}}_-\}$. The matrix $\boldsymbol{B}$ defined above and the safety threshold $\bar{c}_{\mathrm{safe}} = 3/4$ are independent of T. These choices prove the theorem with $c_0 = 1/192$. □

B Proof of Theorem 5.4

Theorem 5.4 (Upper bound of EPOCH-SLUCB)**.** *Suppose $\Delta > 0$. For a confidence parameter $\delta \in (0, 1)$, with probability at least $1 - \delta$, all selected actions are safe and the regret of Algorithm 1 satisfies*

$$R_T \leq 2t_\delta + 8\frac{\|\boldsymbol{B}\| L \beta_T(\frac{\delta}{2})}{\sqrt{\lambda_-}}\sqrt{T} + 2\beta_T(\frac{\delta}{2})\sqrt{2dT\log\Big(1 + \frac{2TL^2}{d(\lambda_- t_\delta + 2\lambda_{\mathrm{ridge}})}\Big)} + 9\frac{\|\boldsymbol{B}\|^2 L^2 \beta_T^2(\frac{\delta}{2})}{c_{\mathrm{safe}}\,\lambda_-\,\Delta^3},$$

where $t_\delta = \frac{8L^2}{\lambda_-}\log\Big(\frac{2d(3+\log_2 T)}{\delta}\Big)$ as in Line 1, and $\lambda_- = \lambda_{\min}(\mathbb{E}[\boldsymbol{U}\boldsymbol{U}^\top])$ for $\boldsymbol{U} \sim \rho$.

We prove the regret bound for Algorithm 1.

For convenience, let

$$A := \frac{2\|\boldsymbol{B}\|^2 L^2 \beta_T^2(\delta/2)}{\lambda_-}, \qquad \epsilon_k = 2^{-k/2}.$$

Thus $T_k = A\epsilon_k^{-2} = A2^k$ and $H_k = \epsilon_k^{-2} T_k = A2^{2k}$.

Let $\mathcal{E}_1$ be the confidence event $\{\boldsymbol{\theta} \in \mathcal{C}_t,\ \forall t \le T\}$. Applying Lemma 5.3 at confidence level $\delta/2$ gives $\mathbb{P}(\mathcal{E}_1) \ge 1 - \delta/2$.

We next bound the smallest eigenvalue of $\boldsymbol{M}_{t-1}$ simultaneously for all optimistic rounds $t \le T$. Let K be the number of epochs entered before termination, so the entered epochs are indexed by $0, \dots, K-1$.

For the union bound, we only count epochs containing optimistic rounds. Every such epoch has $H_k > 1$. Since $H_{k+1} = 4H_k$ and $H_k < T$ except possibly in the last such epoch, there are fewer than $2 + \log_4 T$ such epochs. Let $\mathcal{E}_2$ be the event that the eigenvalue lower bounds from Lemma 5.2 hold simultaneously for the first t_δ forced exploration samples and for the first T_k forced exploration samples in each of these epochs with $T_k > t_\delta$. Recall that

$$t_\delta = \frac{8L^2}{\lambda_-} \log\left(\frac{2d(3 + \log_2 T)}{\delta}\right).$$

By the choice of t_δ, each application of Lemma 5.2 has failure probability at most $\delta/(2(3 + \log_2 T))$. A union bound over fewer than $3 + \log_2 T$ applications gives $\mathbb{P}(\mathcal{E}_2) \ge 1 - \delta/2$.

On any optimistic round $t \le T$ in epoch k, the algorithm has already drawn at least $\max\{t_\delta, T_k\}$ forced exploration samples. The matrix $\boldsymbol{M}_{t-1}$ includes the contributions of these samples, and the remaining observations add positive semidefinite matrices. Thus, on $\mathcal{E}_2$,

$$\boldsymbol{M}_{t-1} \succeq \left(\lambda_{\mathrm{ridge}} + \frac{\lambda_- T_k}{2}\right) \boldsymbol{I}. \tag{15}$$

Let $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$. Then $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$, and we work on $\mathcal{E}$ throughout the rest of the proof.

We first bound the regret from forced exploration. Let N_T be the number of rounds up to time T in which the algorithm samples from $\mathcal{D}^w$. By construction, $N_T \le \max\{T_{K-1}, t_\delta\}$ when $K \ge 1$. If $K = 1$, then $T \le H_0 = A$, so $N_T \le T \le 2\sqrt{AT}$. If $K \ge 2$, then epoch $K-2$ has ended before the horizon, so $H_{K-2} < T$. Since $T_{K-1} = A2^{K-1} = 2\sqrt{AH_{K-2}}$, we have $T_{K-1} < 2\sqrt{AT}$, and hence $N_T \le t_\delta + 2\sqrt{AT}$. Each forced exploration round has regret at most 2, so these rounds contribute at most

$$2N_T \le 2t_\delta + 8\frac{\|\boldsymbol{B}\| L \beta_T(\delta/2)}{\sqrt{\lambda_-}}\sqrt{T}.$$

It remains to control optimistic rounds. Consider such a round t in epoch k. Since the forced exploration condition is inactive, the algorithm has drawn at least T_k samples from $\mathcal{D}^w$ and $t > t_\delta$. Therefore, by Equation (15),

$$\lambda_{\min}(\boldsymbol{M}_{t-1}) \ge \lambda_{\mathrm{ridge}} + \frac{\lambda_- T_k}{2} \ge \frac{\|\boldsymbol{B}\|^2 L^2 \beta_T^2(\delta/2)}{\epsilon_k^2}.$$

Consequently,

$$\beta_{t-1}(\delta/2)\|\boldsymbol{B}\boldsymbol{x}^*\|_{\boldsymbol{M}_{t-1}^{-1}} \le \epsilon_k.$$

Moreover, for every $\boldsymbol{\nu} \in \mathcal{C}_{t-1}$, since $\boldsymbol{\theta} \in \mathcal{C}_{t-1}$,

$$\|\boldsymbol{\nu} - \boldsymbol{\theta}\|_{\boldsymbol{M}_{t-1}} \le \|\boldsymbol{\nu} - \hat{\boldsymbol{\theta}}_{t-1}\|_{\boldsymbol{M}_{t-1}} + \|\hat{\boldsymbol{\theta}}_{t-1} - \boldsymbol{\theta}\|_{\boldsymbol{M}_{t-1}} \le 2\beta_{t-1}(\delta/2).$$

Hence,

$$\boldsymbol{\nu}^\top \boldsymbol{B}\boldsymbol{x}^* \le \boldsymbol{\theta}^\top \boldsymbol{B}\boldsymbol{x}^* + 2\beta_{t-1}(\delta/2)\|\boldsymbol{B}\boldsymbol{x}^*\|_{\boldsymbol{M}_{t-1}^{-1}} \le \boldsymbol{\theta}^\top \boldsymbol{B}\boldsymbol{x}^* + 2\epsilon_k.$$

If $2\epsilon_k \le \Delta$, then $\boldsymbol{x}^* \in \mathcal{D}_{t-1}^{\mathrm{safe}}$. Indeed, for every $\boldsymbol{\nu} \in \mathcal{C}_{t-1}$,

$$\boldsymbol{\nu}^\top \boldsymbol{B}\boldsymbol{x}^* \le c_{\mathrm{safe}} - \Delta + 2\epsilon_k \le c_{\mathrm{safe}}.$$

If $2\epsilon_k > \Delta$, set $\boldsymbol{z}_k = c_{\mathrm{safe}}\boldsymbol{x}^*/(c_{\mathrm{safe}} - \Delta + 2\epsilon_k)$. Since $2\epsilon_k > \Delta$, the coefficient $c_{\mathrm{safe}}/(c_{\mathrm{safe}} - \Delta + 2\epsilon_k)$ belongs to $(0,1)$. Since $\mathbf{0} \in \mathcal{D}$ and $\mathcal{D}$ is convex, $\boldsymbol{z}_k \in \mathcal{D}$. Moreover,

$$\max_{\boldsymbol{\nu} \in \mathcal{C}_{t-1}} \boldsymbol{\nu}^\top \boldsymbol{B}\boldsymbol{z}_k \le \frac{c_{\mathrm{safe}}}{c_{\mathrm{safe}} - \Delta + 2\epsilon_k}\left(\boldsymbol{\theta}^\top \boldsymbol{B}\boldsymbol{x}^* + 2\epsilon_k\right) \le c_{\mathrm{safe}},$$

so $\boldsymbol{z}_k \in \mathcal{D}_{t-1}^{\text{safe}}$. By the definition of $\boldsymbol{z}_k$ and Assumption 3.2,

$$\boldsymbol{\theta}^\top(\boldsymbol{x}^* - \boldsymbol{z}_k) = \frac{2\epsilon_k - \Delta}{c_{\text{safe}} - \Delta + 2\epsilon_k}\boldsymbol{\theta}^\top \boldsymbol{x}^* \le \frac{2\epsilon_k - \Delta}{c_{\text{safe}}}.$$

Thus, for every optimistic round in epoch k, there is an action $\boldsymbol{z} \in \mathcal{D}_{t-1}^{\text{safe}}$ such that

$$\boldsymbol{\theta}^\top(\boldsymbol{x}^* - \boldsymbol{z}) \le \frac{2\epsilon_k - \Delta}{c_{\text{safe}}}\mathbf{1}\{2\epsilon_k > \Delta\}.$$

Let $\widetilde{\boldsymbol{\theta}}_t \in \mathcal{C}_{t-1}$ attain the optimistic value used by the algorithm. Since $\boldsymbol{\theta} \in \mathcal{C}_{t-1}$, optimism gives $\widetilde{\boldsymbol{\theta}}_t^\top \boldsymbol{x}_t \ge \boldsymbol{\theta}^\top \boldsymbol{z}$. Therefore,

$$\boldsymbol{\theta}^\top(\boldsymbol{x}^* - \boldsymbol{x}_t) \le \frac{2\epsilon_k - \Delta}{c_{\text{safe}}}\mathbf{1}\{2\epsilon_k > \Delta\} + 2\beta_{t-1}(\delta/2)\|\boldsymbol{x}_t\|_{\boldsymbol{M}_{t-1}^{-1}}. \tag{16}$$

We now sum the two terms in Equation (16). Let $\mathcal{O}$ be the set of optimistic rounds. All optimistic rounds follow the first t_δ forced exploration samples. On $\mathcal{E}_2$, these samples and the choice of t_δ ensure $\|\boldsymbol{x}_t\|^2_{\boldsymbol{M}_{t-1}^{-1}} \le L^2/(\lambda_{\text{ridge}} + \lambda_- t_\delta/2) \le 1$ for every $t \in \mathcal{O}$. Applying the elliptical potential lemma [1, Lemma 11] to observations after these initial samples bounds the sum of squared norms. Together with the Cauchy–Schwarz inequality and $|\mathcal{O}| \le T$, this gives

$$\begin{aligned}\sum_{t\in\mathcal{O}} \|\boldsymbol{x}_t\|_{\boldsymbol{M}_{t-1}^{-1}} &\le \sqrt{|\mathcal{O}|\sum_{t\in\mathcal{O}} \|\boldsymbol{x}_t\|^2_{\boldsymbol{M}_{t-1}^{-1}}} \\ &\le \sqrt{2dT\log\left(1 + \frac{2TL^2}{d(\lambda_- t_\delta + 2\lambda_{\text{ridge}})}\right)}.\end{aligned}$$

Using $\beta_{t-1}(\delta/2) \le \beta_T(\delta/2)$, the sum of the second term in Equation (16) over $t \in \mathcal{O}$ is bounded by

$$2\beta_T(\delta/2)\sqrt{2dT\log\left(1 + \frac{2TL^2}{d(\lambda_- t_\delta + 2\lambda_{\text{ridge}})}\right)}.$$

We now bound the sum of the first term in Equation (16). Let $k_\Delta := \min\{k \ge 0 : 2\epsilon_k \le \Delta\}$. Only epochs $k < k_\Delta$ contribute, each with at most $H_k = A\epsilon_k^{-4}$ rounds. For $k_\Delta \ge 1$, we have $2\epsilon_{k_\Delta} \le \Delta < 2\epsilon_{k_\Delta - 1}$, so $\Delta < 2$ and $2 \le \Delta 2^{k_\Delta/2} < 2\sqrt{2}$. Hence,

$$\begin{aligned}\sum_{k<k_\Delta} H_k\frac{2\epsilon_k - \Delta}{c_{\text{safe}}} &= \frac{A}{c_{\text{safe}}}\sum_{k<k_\Delta}\frac{2\epsilon_k - \Delta}{\epsilon_k^4} \\ &= \frac{A}{c_{\text{safe}}}\left[\frac{2((2\sqrt{2})^{k_\Delta} - 1)}{2\sqrt{2} - 1} - \frac{\Delta(4^{k_\Delta} - 1)}{3}\right] \\ &= \frac{A}{c_{\text{safe}}\Delta^3}\left[\frac{2(\Delta 2^{k_\Delta/2})^3}{2\sqrt{2} - 1} - \frac{(\Delta 2^{k_\Delta/2})^4}{3}\right] + \frac{A}{c_{\text{safe}}}\left(\frac{\Delta}{3} - \frac{2}{2\sqrt{2} - 1}\right) \\ &\le \frac{A}{c_{\text{safe}}\Delta^3}\max_{2\le u\le 2\sqrt{2}}\left(\frac{2u^3}{2\sqrt{2} - 1} - \frac{u^4}{3}\right) \\ &\le \frac{A}{c_{\text{safe}}\Delta^3}\cdot\frac{729}{16(2\sqrt{2} - 1)^4} \\ &< \frac{9A}{2c_{\text{safe}}\Delta^3} = 9\frac{\|\boldsymbol{B}\|^2 L^2\beta_T^2(\delta/2)}{c_{\text{safe}}\lambda_-\Delta^3}.\end{aligned}$$

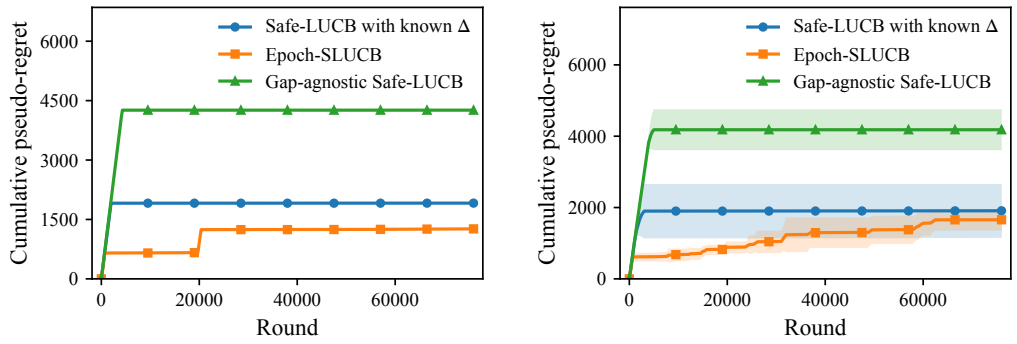
Combining these bounds gives the stated regret bound.

It remains only to verify safety. Actions sampled from $\mathcal{D}^w$ are safe by definition. Optimistic actions are selected from $\mathcal{D}_{t-1}^{\text{safe}}$, which is safe for every parameter in $\mathcal{C}_{t-1}$. Since $\boldsymbol{\theta} \in \mathcal{C}_{t-1}$ on $\mathcal{E}$, every selected action is safe on $\mathcal{E}$. Thus the algorithm is safe with probability at least $1 - \delta$.

C Additional experiments

C.1 Cumulative pseudo-regret

Figure 4 shows cumulative pseudo-regret for the same runs as Figures 3(a) and (c).



(a) $\Delta = 0.20$. (b) Random realizations.

Figure 4: Cumulative pseudo-regret. Shaded bands show one sample standard deviation.

C.2 Experiments in higher dimensions

We evaluate the three algorithms in dimensions $d = 5, 10, 20$, with 10 random problem realizations per dimension, $T = 150000$, and $\Delta = 0.20$. We use $L = S = 1$, Gaussian noise with standard deviation 0.05, and the same ridge regularization and confidence parameters as in Section 7.

For each realization, draw a random orthogonal matrix $\boldsymbol{Q}$ and write $\boldsymbol{q}_1, \boldsymbol{q}_2$ for its first two columns. Set $\boldsymbol{\theta} = \boldsymbol{x}^* = \boldsymbol{q}_1$, $c_{\text{safe}} = 0.6$, and

$$\boldsymbol{B} = \boldsymbol{Q}\,\mathrm{diag}\left(\begin{pmatrix} 0.4 & \sqrt{0.33} \\ -\sqrt{0.33} & 0.4 \end{pmatrix}, 0.7\boldsymbol{I}_{d-2}\right)\boldsymbol{Q}^\top.$$

Then $\|\boldsymbol{B}\| = 0.7$ and $\boldsymbol{\theta}^\top \boldsymbol{B}\boldsymbol{x}^* = 0.4$, giving the prescribed safety gap.

The finite candidate set contains $15d$ actions. Its $5d$ exploration actions are $c_{\text{safe}}\boldsymbol{B}^{-1}\boldsymbol{u}$, where $\boldsymbol{u}$ ranges over five independent random orthonormal bases. The other $10d$ actions are $\cos(2\pi j/(10d))\boldsymbol{q}_1 + \sin(2\pi j/(10d))\boldsymbol{q}_2$, for $j = 0, \ldots, 10d-1$. The optimal action $\boldsymbol{x}^*$ is included and lies outside $\mathcal{D}^w$. Forced exploration samples uniformly from the $5d$ exploration actions, giving $\lambda_- = c_{\text{safe}}^2/(d\|\boldsymbol{B}\|^2) = 36/(49d)$. Optimistic action selection enumerates the certified candidates. For each realization, the algorithms share the same reward noise sequence and exploration random seed.

Table 2: Cumulative pseudo-regret and runtime at $T = 150000$ with $\Delta = 0.20$. Entries are means, with standard errors in parentheses, over 10 realizations per dimension.

d	Algorithm	Cumulative pseudo-regret	Runtime (s)
5	Safe-LUCB with known Δ	866.71 (15.52)	5.43 (0.13)
	EPOCH-SLUCB	2172.62 (37.47)	5.26 (0.09)
	Gap-agnostic Safe-LUCB	4890.34 (87.97)	5.25 (0.10)
10	Safe-LUCB with known Δ	2428.85 (20.03)	6.60 (0.10)
	EPOCH-SLUCB	3092.31 (25.14)	6.48 (0.08)
	Gap-agnostic Safe-LUCB	7079.14 (60.07)	6.53 (0.19)
20	Safe-LUCB with known Δ	6888.97 (39.00)	9.15 (0.05)
	EPOCH-SLUCB	4521.51 (25.49)	9.21 (0.03)
	Gap-agnostic Safe-LUCB	10302.96 (59.88)	9.14 (0.10)

Table 2 reports the results. EPOCH-SLUCB has lower mean final pseudo-regret than gap-agnostic Safe-LUCB in all three dimensions. Compared with Safe-LUCB with known Δ, its mean final

pseudo-regret is higher for $d = 5, 10$ and lower for $d = 20$. No safety violations were observed in any of the 90 runs.

We report the running time of each algorithm over $T = 150000$ rounds. The order of the three algorithms was rotated across realizations.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's two main contributions: an $\Omega(T^{2/3})$ lower bound when the safety gap is zero and an epoch-based algorithm that achieves $\tilde{O}(\sqrt{T})$ regret for any fixed positive safety gap without knowing its value. These claims match Theorems 4.1 and 5.4, and the empirical results are presented as illustrative simulations rather than as the main source of the theoretical claims.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: The paper discusses limitations in Section 8, including the reliance on the safe linear bandit model of Amani et al. [4], the assumptions that B, c_{safe}, and a norm bound on θ are known, and the requirement of sampling from $\mathcal{D}^w$ and optimizing over certified safe sets. It also notes that the experiments are synthetic and that the optimal dependence on the dimension remains open.

   Guidelines:

   - The answer [N/A] 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: The paper states the model assumptions in Section 3 and states the main theoretical results in Theorems 4.1 and 5.4. Proof sketches are provided in the main text where appropriate, and the complete proofs are given in Appendices A and B.

   Guidelines:

   - The answer [N/A] 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: Section 7 describes the simulation setup, including the problem instances, horizons, noise level, safety gaps, number of random seeds, and grid sensitivity checks. We also provide code in the supplemental material with the implementation details and scripts needed to reproduce the reported simulations.

   Guidelines:

   - The answer [N/A] 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: [Yes]

   Justification: We provide code in the supplemental material to reproduce the reported simulations. The experiments use synthetic problem instances generated by the provided scripts, and no external dataset is required.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 7 specifies the synthetic experimental setup, including the problem instances, safety gaps, horizons, noise level, number of random seeds, evaluation metrics, and grid sensitivity checks. Additional implementation details and scripts are provided in the supplemental code.

   Guidelines:

   - The answer [N/A] 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: Section 7 reports averages over multiple random seeds or random problem realizations, and Figure 3 (c) includes bands showing one standard deviation across 10 random problem realizations. The experiments are synthetic and intended to illustrate the behavior predicted by the theory rather than to establish a benchmark comparison on real data.

Guidelines:

- The answer [N/A] 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: The experiments were run on a 2021 14-inch MacBook Pro with an Apple M1 Pro chip (8-core CPU, 14-core GPU), 16GB unified memory, and 512GB SSD. No GPU acceleration or large-scale compute was required.

Guidelines:

- The answer [N/A] 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: The research is theoretical and uses only synthetic simulations, without human subjects, private data, scraped datasets, or deployed systems. To the best of our knowledge, the work conforms with the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] 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: [Yes]

    Justification: The work is foundational research on safe sequential decision-making. Its potential positive impact is to improve learning algorithms for settings where unsafe actions are costly, while potential negative impact could arise from deploying such algorithms in real safety-critical systems without validating the model assumptions, confidence calibration, and implementation details.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release data or models with a high risk for misuse. The experiments use synthetic simulations, and the supplemental material contains only code for reproducing these simulations.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not use external datasets, pre-trained models, or existing code assets.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not introduce new datasets, models, or benchmarks.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects, so IRB approval or an equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] 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 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 rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this paper does not involve LLMs.

    Guidelines:

    - The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.