Safe Linear Stochastic Bandits

Kia Khezeli, Eilyan Bitar
School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, USA
{kk839, eyb5}@cornell.edu

Abstract

We introduce the safe linear stochastic bandit framework—a generalization of linear stochastic bandits—where, in each stage, the learner is required to select an arm with an expected reward that is no less than a predetermined (safe) threshold with high probability. We assume that the learner initially has knowledge of an arm that is known to be safe, but not necessarily optimal. Leveraging on this assumption, we introduce a learning algorithm that systematically combines known safe arms with exploratory arms to safely expand the set of safe arms over time, while facilitating safe greedy exploitation in subsequent stages. In addition to ensuring the satisfaction of the safety constraint at every stage of play, the proposed algorithm is shown to exhibit an expected regret that is no more than $O(\sqrt{T}\log(T))$ after T stages of play.

1 Introduction

We investigate the role of safety in constraining the design of learning algorithms within the classical framework of linear stochastic bandits (Dani, Hayes, and Kakade 2008; Rusmevichientong and Tsitsiklis 2010; Abbasi-Yadkori, Pál, and Szepesvári 2011). Specifically, we introduce a family of *safe linear stochastic bandit problems* where—in addition to the typical goal of designing learning algorithms that minimize regret—we impose a constraint requiring that an algorithm's stagewise expected reward remains above a predetermined safety threshold with high probability at every stage of play. In the proposed framework, we assume that a "safe" baseline arm is initially known, and consider a class of safety thresholds that are defined as fixed cutbacks on the expected reward of the known baseline arm. Accordingly, an algorithm that is deemed to be safe cannot induce stagewise rewards that dip below the baseline reward by more than a fixed amount. Critically, the assumption of a known baseline arm—and the limited capacity for exploration implied by the class of safety thresholds considered—can be leveraged on to initially guide the exploration of allowable arms by playing combinations of the baseline arm and exploratory arms in a manner that expands the set of safe arms over time, while simultaneously preserving safety at every stage of play.

There are a variety of real-world applications that might benefit from the design of stagewise-safe online learning algorithms (Khezeli and Bitar 2017; Li et al. 2019; Sui et al. 2015). Most prominently, clinical trials have long been used as a motivating application for the multi-armed bandit (Berry and Pearson 1985) and linear bandit (Dani, Hayes, and Kakade 2008) frameworks. However, as pointed out by (Villar, Bowden, and Wason 2015): "Despite this apparent near-perfect fit between a real-world problem and a mathematical theory, the MABP has yet to be applied to an actual clinical trial." One could argue that the ability to provide a learning algorithm that is guaranteed to be stagewise safe has the potential to facilitate the utilization of bandit models and algorithms in clinical trials. More concretely, consider the possibility of using the linear bandit framework to model the problem of optimizing a combination of d candidate treatments for a specific health issue. In this context, an "arm" represents a mixture of treatments, the "unknown reward vector" encodes the effectiveness of each treatment, and the "reward" represents a patient's response to a chosen mixture of treatments. In terms of the safety threshold, it is natural to select the "baseline arm" to be the (possibly suboptimal) combination of treatments possessing the largest reward known to date. As it is clearly unethical to prescribe a treatment that may degrade a patient's health, the stagewise safety constraint studied in this paper can be interpreted as a requirement that a patient's response to a chosen treatment must be arbitrarily close to that of the baseline treatment, if not better.

1.1 Contributions

In this paper, we propose a new learning algorithm that is tailored to the safe linear bandit framework. The proposed algorithm, which we call the *Safe Exploration and Greedy Exploitation* (SEGE) algorithm, is shown to exhibit near-optimal expected regret, while guaranteeing the satisfaction of the proposed safety constraint at every stage of play. Initially, the SEGE algorithm performs safe exploration by combining the baseline arm with a random exploratory arm that is constrained by an "exploration budget" implied by the stagewise safety constraint. Over time, the proposed algorithm systematically expands the family of safe arms in this manner to include new safe arms with expected rewards that exceed the baseline reward level. Exploitation under

the SEGE algorithm is based on the certainty equivalence principle. That is, the algorithm constructs an "estimate" of the unknown reward parameter, and selects an arm that is optimal for the given parameter estimate. The SEGE algorithm only plays the certainty equivalent (i.e., greedy) arm when it is safe—a condition that is determined according to a lower confidence bound on its expected reward. Moreover, the proposed algorithm balances the trade-off between exploration and exploitation by controlling the rate at which information is accumulated over time, as measured by the growth rate of the minimum eigenvalue of the so-called information matrix.[1] More specifically, the SEGE algorithm guarantees that the minimum eigenvalue of the information matrix grows at a rate ensuring that the expected regret of the algorithm is no greater than $O(\sqrt{T}\log(T))$ after T stages of play. This regret rate is near optimal in light of $\Omega(\sqrt{T})$ lower bounds previously established in the linear stochastic bandit literature (Dani, Hayes, and Kakade 2008; Rusmevichientong and Tsitsiklis 2010).

1.2 Related Literature

There is an extensive literature on linear stochastic bandits. For this setting, several algorithms based on the principle of Optimism in the Face of Uncertainty (OFU) (Dani, Hayes, and Kakade 2008; Rusmevichientong and Tsitsiklis 2010; Abbasi-Yadkori, Pál, and Szepesvári 2011) or Thompson Sampling (Agrawal and Goyal 2013) have been proposed. Although such algorithms are known to be near-optimal under various measures of regret, they may fail in the safe linear bandit framework, as their (unconstrained) approach to exploration may result in a violation of the stagewise safety constraints considered in this paper.

In the context of multi-armed bandits, there is a related stream of literature that focuses on the design of "risk-sensitive" learning algorithms by encoding risk in the performance objectives according to which regret is measured (Cassel, Mannor, and Zeevi 2018; David et al. 2018). Typical risk measures that have been studied in the multi-armed bandit literature include Mean-Variance (Sani, Lazaric, and Munos 2012; Vakili and Zhao 2016), Value-at-Risk (Vakili and Zhao 2015), and Conditional Value-at-Risk (Galichet, Sebag, and Teytaud 2013). Although such risk-sensitive algorithms are inclined to exhibit reduced volatility in the cumulative reward that is received over time, they are not constrained in a manner that explicitly limits the stagewise risk of the reward processes that they induce.

Closer to the setting studied in this paper is the conservative bandit framework (Wu et al. 2016; Kazerouni et al. 2017), which incorporates explicit safety constraints on the reward process induced by the learning algorithm. However, in contrast to the stagewise safety constraints considered in this paper, conservative bandits encode their safety requirements in the form of constraints on the cumulative rewards received by the algorithm. Along a similar line of research, (Sun, Dey, and Kapoor 2017) investigate the design of learning algorithms for risk-constrained contextual bandits that balance a tradeoff between cumulative constraint violation and regret. Given the cumulative nature of the safety constraints considered by the aforementioned algorithms, they cannot be directly applied to the stagewise safe linear bandit problem considered in this paper. In Section 6.3, we provide a simulation-based comparison between the SEGE algorithm and the Conservative Linear Upper Confidence Bound (CLUCB) algorithm (Kazerouni et al. 2017) to more clearly illustrate the potential weaknesses and strengths of each approach.

We close this section by mentioning another closely related body of work in the online learning literature that investigates the design of stagewise-safe algorithms for a more general class of smooth reward functions (Sui et al. 2015; 2018; Usmanova, Krause, and Kamgarpour 2019). Although the proposed algorithms are shown to respect stagewise safety constraints that are similar in spirit to the class of safety constraints considered in this paper, they lack formal upper bounds on their cumulative regret.

1.3 Organization

The remainder of the paper is organized as follows. We introduce pertinent notation in Section 2. In Section 3, we define the safe linear stochastic bandit problem. In Section 4, we introduce the Safe Exploration and Greedy Exploitation (SEGE) algorithm. We present our main theoretical findings in Section 5, and close the paper with a simulation study of the SEGE algorithm in Section 6. Due to space limitations, all mathematical proofs for the theoretical results contained in this paper are included in the appendix of the arXiv version of this paper (Khezeli and Bitar 2019).

2 Notation

We denote the standard Euclidean norm of a vector $x \in \mathbb{R}^d$ by $\|x\|$ and define its weighted Euclidean norm as $\|x\|_S = \sqrt{x^\top S x}$ where $S \in \mathbb{R}^{d\times d}$ is a given symmetric positive semidefinite matrix. We denote the inner product of two vectors $x, y \in \mathbb{R}^d$ by $\langle x, y\rangle = x^\top y$. For a square matrix $A \in \mathbb{R}^{d\times d}$, we denote its minimum and maximum eigenvalues by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively.

3 Problem Formulation

In this section, we introduce the safe linear stochastic bandit model considered in this paper. Before doing so, we review the standard model for linear stochastic bandits on which our formulation is based.

3.1 Linear Bandit Model

Linear stochastic bandits belong to a class of sequential decision-making problems in which a learner (i.e., decision-maker) seeks to maximize an unknown linear function using noisy observations of its function values that it collects over multiple stages. More precisely, at each stage $t = 1, 2, \ldots$, the learner is required to select an arm (i.e., action) X_t from

[1] We note that a closely related class of learning algorithms, which explicitly control the rate of information gain in this manner, have been previously studied in the context of dynamic pricing algorithms for revenue maximization (den Boer and Zwart 2013; Keskin and Zeevi 2014).

a compact set $\mathcal{X} \subset \mathbb{R}^d$ of allowable arms, which is assumed to be an ellipsoid of the form

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid (x - \bar{x})^\top H^{-1} (x - \bar{x}) \leq 1 \right\}, \quad (1)$$

where $\bar{x} \in \mathbb{R}^d$ and $H \in \mathbb{R}^{d \times d}$ is a symmetric and positive definite matrix. In response to the particular arm played at each stage t, the learner observes a reward Y_t that is induced by the stochastic linear relationship:

$$Y_t = \langle X_t, \theta^* \rangle + \eta_t. \quad (2)$$

Here, the noise process $\{\eta_t\}_{t=1}^{\infty}$ is assumed be a sequence of independent and zero-mean random variables, and, critically, the reward parameter $\theta^* \in \mathbb{R}^d$ is assumed to be fixed and unknown. This a priori uncertainty in the reward parameter gives rise to the need to balance the exploration-exploitation trade-off in adaptively guiding the sequence of arms played in order to maximize the expected reward accumulated over time.

Admissible Policies and Regret. We restrict the learner's decisions to those which are causal in nature. That is to say, at each stage t, the learner is required to select an arm based only on the history of past observations $H_t = (X_1, Y_1, \ldots, X_{t-1}, Y_{t-1})$, and on an external source of randomness encoded by a random variable U_t. The random process $\{U_t\}_{t=1}^{\infty}$ is assumed to be independent across time, and independent of the random noise process $\{\eta_t\}_{t=1}^{\infty}$. Formally, an *admissible policy* is a sequence of functions $\pi = \{\pi_t\}_{t=1}^{\infty}$, where each function π_t maps the information available to the learner at each stage t to a feasible arm $X_t \in \mathcal{X}$ according to $X_t = \pi_t(H_t, U_t)$.

The performance of an admissible policy after T stages of play is measured according to its *expected regret*,[2] which equals the difference between the expected reward accumulated by the optimal arm and the expected reward accumulated by the given policy after T stages of play. Formally, the expected regret of an admissible policy is defined as

$$R_T = \sum_{t=1}^{T} \langle X^*, \theta^* \rangle - \mathbb{E}\left[\sum_{t=1}^{T} \langle X_t, \theta^* \rangle \right], \quad (3)$$

where expectation is taken with respect to the distribution induced by the underling policy, and $X^* \in \mathcal{X}$ denotes the *optimal arm* that maximizes the expected reward at each stage of play given knowledge of the reward parameter θ^*, i.e.,

$$X^* = \text{argmax}_{x \in \mathcal{X}} \langle x, \theta^* \rangle. \quad (4)$$

At a minimum, we seek policies exhibiting an expected regret that is sublinear in the number of stages played T. Such policies are said to have *no-regret* in the sense that $\lim_{T \to \infty} R_T / T = 0$. To facilitate the design and theoretical analysis of such policies, we adopt a number of technical assumptions, which are standard in the literature on linear stochastic bandits, and are assumed to hold throughout the paper.

[2]It is worth noting, that in the context of linear stochastic bandits, *expected regret* is equivalent to *expected pseudo-regret* due to the additive nature of the noise process (Abbasi-Yadkori, Pál, and Szepesvári 2011).

Assumption 1 *The unknown reward parameter is bounded according to $\|\theta^*\| \leq S$, where $S > 0$ is a known constant.*

Assumption 1 will prove essential to the design of policies that *safely explore* the parameter space in a manner ensuring that the expected reward stays above a predetermined (safe) threshold with high probability at each stage of play. We refer the reader to Definition 1 for a formal definition of the particular safety notion considered in this paper.

Assumption 2 *Each element of $\{\eta_t\}_{t=1}^{\infty}$ is assumed to be σ_η-sub-Gaussian, where $\sigma_\eta \geq 0$ is a fixed constant. That is,*

$$\mathbb{E}\left[\exp(\gamma \eta_t)\right] \leq \exp\left(\gamma^2 \sigma_\eta^2 / 2\right)$$

for all $\gamma \in \mathbb{R}$ and $t \geq 1$.

Assumptions 1 and 2, together with the class of admissible policies considered in this paper, enable the utilization of existing results that provide an explicit characterization of confidence ellipsoids for the unknown reward parameter based on a ℓ_2-regularized least-squares estimator (Abbasi-Yadkori, Pál, and Szepesvári 2011). Such confidence regions play a central role in the design of no-regret algorithms for the linear stochastic bandits (Dani, Hayes, and Kakade 2008; Rusmevichientong and Tsitsiklis 2010; Abbasi-Yadkori, Pál, and Szepesvári 2011).

3.2 Safe Linear Bandit Model

In what follows, we introduce the framework of *safe linear stochastic bandits* studied in this paper. Loosely speaking, an admissible policy is said to be *safe* if the expected reward $\mathbb{E}[Y_t \mid X_t] = \langle X_t, \theta^* \rangle$ that it induces at each stage t is guaranteed to stay above a given reward threshold with high probability.[3] More formally, we have the following definition.

Definition 1 (Stagewise Safety Constraint) *Let $b \in \mathbb{R}$ and $\delta \in [0, 1]$. An admissible policy π—or equivalently the arm X_t that it induces—is defined to be (δ, b)-*safe at stage t *if*

$$\mathbb{P}\left(\langle X_t, \theta^* \rangle \geq b\right) \geq 1 - \delta, \quad (5)$$

where the probability is calculated according to the distribution induced by the policy π.

The stagewise safety constraint requires that the expected reward at stage t exceed the *safety threshold* $b \in \mathbb{R}$ with probability no less than $1 - \delta$, where $\delta \in [0, 1]$ encodes the *maximum allowable risk* that the learner is willing to tolerate.

Clearly, without making additional assumptions, it is not possible to design policies that are guaranteed to be safe according to (5) given arbitrary safety specifications. We circumvent this obvious limitation by giving the learner access to a *baseline arm* with a known lower bound on its expected reward. We formalize this assumption as follows.

Assumption 3 (Baseline Arm) *We assume that the learner knows a deterministic baseline arm $X_0 \in \mathcal{X}$ satisfying*

$$\langle X_0, \theta^* \rangle \geq b_0,$$

[3]To simplify the exposition, we will frequently refer to $\mathbb{E}[Y_t \mid X_t]$—the expected reward conditioned on the arm X_t—as the *expected reward*, unless it is otherwise unclear from the context.

where $b_0 \in \mathbb{R}$ is a known lower bound on its expected reward.

We note that it is straightforward to construct a baseline arm satisfying Assumption 3 by leveraging on the assumed boundedness of the unknown reward parameter as specified by Assumption 1. In particular, any arm $X_0 \in \mathcal{X}$ and its corresponding "worst-case" reward given by $b_0 = \min_{\|\theta\| \leq S} \langle X_0, \theta \rangle = -S\|X_0\|$ are guaranteed to satisfy Assumption 3.

With Assumption 3 in hand, the learner can leverage on the baseline arm to initially guide its exploration of allowable arms by playing combinations of the baseline arm and carefully designed exploratory arms in a manner that safely expands the set of safe arms over time. Plainly, the ability to safely explore in the vicinity of the baseline arm is only possible under stagewise safety constraints defined in terms of safety thresholds satisfying $b < b_0$. Under such stagewise safety constraints, the difference in rewards levels $b_0 - b$ can be interpreted as a stagewise "exploration budget" of sorts, as it reflects the maximum relative loss in expected reward that the learner is willing to tolerate when playing arms that deviate from the baseline arm. Naturally, the larger the exploration budget, the more aggressively can the learner explore. With the aim of designing safe learning algorithms that leverage on this simple idea, we will restrict our attention to stagewise safety constraints that are specified in terms of safety thresholds satisfying $b < b_0$.

Before proceeding, we briefly summarize the framework of *safe linear stochastic bandits* considered in this paper. Given a baseline arm satisfying Assumption 3, the learner is initially required to fix a safety threshold that satisfies $b < b_0$. At each subsequent stage $t = 1, 2, \ldots$, the learner must select a risk level $\delta_t \in [0, 1]$ and a corresponding arm $X_t \in \mathcal{X}$ that is (δ_t, b)-safe. The learner aims to design an admissible policy that minimizes its expected regret, while simultaneously ensuring that all arms played satisfy the stagewise safety constraints. In the following section, we propose a policy that is guaranteed to both exhibit no-regret and satisfy the safety constraint at every stage of play.

Relationship to Conservative Bandits. We briefly discuss the relationship between the safety constraints considered in this paper and the conservative bandit framework orginally studied by (Wu et al. 2016) in the context of multi-armed bandits, and subsequently extended to the setting of linear bandits by (Kazerouni et al. 2017). In contrast to the stagewise safety constraints considered in this paper, conservative bandits encode their safety requirements in the form of constraints on the cumulative expected rewards received by a policy. Specifically, given a baseline arm satisfying Assumption 3, an admissible policy is said to respect the safety constraint defined in (Kazerouni et al. 2017) if

$$\mathbb{P}\left(\sum_{k=1}^{t} \langle X_k, \theta^* \rangle \geq (1-\alpha) \sum_{k=1}^{t} b_0, \ \forall\, t \geq 1\right) \geq 1 - \delta, \tag{6}$$

where $\delta \in [0, 1]$ and $\alpha \in (0, 1)$. Here, the parameter α encodes the maximum fraction of the cumulative baseline rewards that the learner is willing to forgo over time. In this context, smaller values of α imply greater levels of conservatism (safety). It is straightforward to show that conservative performance constraints of the form (6) are a special case of the class of stagewise safety constraints considered in Definition 1. In particular, if we set the safety threshold according to $b = (1-\alpha)b_0$, and let $\{\delta_t\}_{t=1}^{\infty}$ be any summable sequence of risk levels satisfying $\sum_{t=1}^{\infty} \delta_t \leq \delta$, then any admissible policy that is (δ_t, b)-safe for each stage $t \geq 1$ also satisfies the conservative performance constraint (6).

4 A Safe Linear Bandit Algorithm

In this section, we propose a new algorithm, which we call the *Safe Exploration and Greedy Exploitation* (SEGE) algorithm, that is guaranteed to be safe in every stage of play, while exhibiting a near-optimal expected regret. Before proceeding with a detailed description of the proposed algorithm, we briefly summarize the basic elements underpinning its design. Initially, the SEGE algorithm performs safe exploration by playing convex combinations of the baseline arm and random exploratory arms in a manner that satisfies Definition 1. Through this process of exploration, the SEGE algorithm is able to expand the family of safe arms to incorporate new arms that are guaranteed to outperform the baseline arm with high probability. Among all safe arms available to the algorithm at any given stage of play, the arm with the largest lower confidence bound on its expected reward is used as the basis for safe exploration. The SEGE algorithm performs exploitation by playing the certainty equivalent (greedy) arm based on a ℓ_2-regularized least-squares estimate of the unknown reward parameter. The SEGE algorithm only plays the greedy arm when it is safe, i.e., when a lower confidence bound on its expected reward exceeds the given safety threshold. Critically, the proposed algorithm balances the trade-off between exploration and exploitation by explicitly controlling the growth rate of the so-called information matrix (cf. Eq. (8)) in a manner that ensures that the expected regret of the SEGE algorithm is no greater than $O(\sqrt{T}\log(T))$ after T stages of play. The pseudocode for the SEGE algorithm is presented in Algorithm 1.

In the following section, we introduce a regularized least-squares estimator that will serve as the foundation for the proposed learning algorithm.

4.1 Regularized Least Squares Estimator

The ℓ_2-regularized least-squares estimate of the unknown reward parameter θ^* based on the information available to the algorithm up until and including stage t is defined as

$$\widehat{\theta}_t = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \sum_{k=1}^{t} (Y_k - \langle X_k, \theta \rangle)^2 + \lambda \|\theta\|^2 \right\}.$$

Here, $\lambda > 0$ denotes a user-specified regularization parameter. It is straightforward to show that

$$\widehat{\theta}_t = V_t^{-1} \sum_{k=1}^{t} X_k Y_k, \tag{7}$$

where

$$V_t = \lambda I + \sum_{k=1}^{t} X_k X_k^\top. \tag{8}$$

Throughout the paper, we will frequently refer to the matrix V_t as the *information matrix* at each stage t.

The following result taken from (Abbasi-Yadkori, Pál, and Szepesvári 2011, Theorem 2) provides an ellipsoidal characterization of a confidence region for the unknown reward parameter based on the regularized least-squares estimator (7). It is straightforward to verify that the conditions of (Abbasi-Yadkori, Pál, and Szepesvári 2011, Theorem 2) are satisfied under the standing assumptions of this paper.

Theorem 1 *For any admissible policy and $\delta \in (0,1)$, it holds that*

$$\mathbb{P}\left(\theta^* \in \mathcal{C}_t(\delta),\ \forall t \geq 1\right) \geq 1-\delta,$$

where the confidence set $\mathcal{C}_t(\delta)$ is defined as

$$\mathcal{C}_t(\delta) = \left\{\theta \in \mathbb{R}^d \ :\ \|\widehat{\theta}_t - \theta\|_{V_t} \leq r_t(\delta)\right\}. \tag{9}$$

Here, $r_t(\delta)$ is defined as

$$r_t(\delta) = \sigma_\eta \sqrt{d \log\left(\frac{1+tL^2/\lambda}{\delta}\right)} + \sqrt{\lambda} S, \tag{10}$$

where $L = \max_{x\in\mathcal{X}} \|x\|$.

In the following section, we propose a method for safe exploration using the characterization of the confidence ellipsoids introduced in Theorem 1.

4.2 Safe Exploration

We now describe a novel approach to "safe exploration" that will be utilized in the design of the proposed learning algorithm. At each stage $t \geq 1$, given a risk level δ_t, the SEGE algorithm constructs a safe exploration arm (X_t^{SE}) as a convex combination of a (δ_t, b_0)-safe arm (X_t^{S}) and a random exploratory arm (U_t), i.e.,

$$X_t^{\mathsf{SE}} = (1-\rho) X_t^{\mathsf{S}} + \rho U_t. \tag{11}$$

Qualitatively, the user-specified parameter $\rho \in (0,1)$ controls the balance between safety and exploration. Figure 1 provides a graphical illustration of the set of all safe exploration arms induced by a given safe arm X_t^{S} according to (11).

The random exploratory arm process $\{U_t\}_{t=1}^{\infty}$ is generated according to

$$U_t = \bar{x} + H^{1/2}\zeta_t, \tag{12}$$

where the random process $\{\zeta_t\}_{t=1}^{\infty}$ is assumed to be a sequence of independent, zero-mean, and symmetric random vectors. For each element of the sequence, we require that $\|\zeta_t\| = 1$ almost surely and $\sigma_\zeta^2 = \lambda_{\min}(\mathrm{Cov}\,(\zeta_t)) > 0$. Additionally, we define $\sigma^2 = \lambda_{\min}(\mathrm{Cov}\,(U_t))$. The parameters σ and ρ both determine how aggressively the algorithm can explore the set of allowable arms. However, exploration that is too aggressive may result in a violation of the stagewise safety constraint. In the following Lemma, we establish an upper bound on ρ such that for all choices of $\rho \in (0, \bar{\rho}]$, the arm X_t^{SE} is guaranteed to be safe for any $\sigma \geq 0$. The proof of Lemma 1 can be found in the appendix to (Khezeli and Bitar 2019).

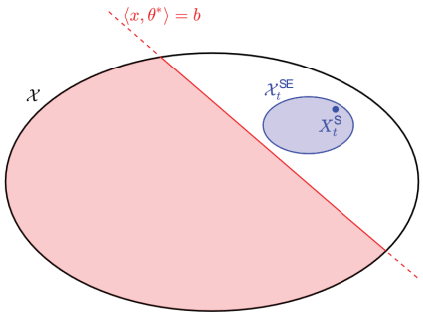


Figure 1: The figure illustrates the effect of the safety constraint on the learner's decision making ability. The shaded blue ellipse $\mathcal{X}_t^{\mathsf{SE}}$ depicts the set of all safe exploration arms constructed using the safe arm X_t^{S} under the SEGE algorithm, i.e., $\mathcal{X}_t^{\mathsf{SE}} = \{(1-\rho)X_t^{\mathsf{S}} + \rho x \mid \rho \in (0, \bar{\rho}], x \in \partial\mathcal{X}\}$. The red shaded area depicts the set of unsafe arms. The black ellipse (and its interior) depicts the entire set of allowable arms.

Lemma 1 *Let $\rho \in (0, \bar{\rho}]$ where $\bar{\rho} > 0$ is defined as*

$$\bar{\rho} = \min\left\{1, \frac{b_0 - b}{2S\sqrt{\lambda_{\max}(H)}}\right\}. \tag{13}$$

Then, for every stage $t \geq 1$, the safe exploration arm X_t^{SE} defined in Equation (11) is (δ, b)-safe for any $\delta \in [0,1]$.

As the SEGE algorithm expands its set of safe arms over time, it attempts to increase the stagewise efficiency with which it safely explores by exploring in the vicinity of the safe arm with the largest lower confidence bound on its expected reward. More specifically, at each stage t, the SEGE algorithm constructs a confidence set $\mathcal{C}_{t-1}(\delta_t)$ according to Equation (9). With this confidence set in hand, the proposed algorithm calculates a lower confidence bound (LCB) on the expected reward of each arm $x \in \mathcal{X}$ according to

$$\mathrm{LCB}_t(x) = \min_{\theta \in \mathcal{C}_{t-1}(\delta_t)} \langle x, \theta \rangle.$$

It is straightforward to show that the lower confidence bound defined above admits the closed-form expression:

$$\mathrm{LCB}_t(x) = \langle x, \widehat{\theta}_{t-1} \rangle - r_t(\delta_t)\|x\|_{V_{t-1}^{-1}}.$$

We define the LCB arm (X_t^{LCB}) to be the arm with the largest lower confidence bound on its expected reward among all allowable arms. It is given by:

$$X_t^{\mathsf{LCB}} = \mathrm{argmax}_{x\in\mathcal{X}}\ \mathrm{LCB}_t(x). \tag{14}$$

Clearly, the LCB arm is guaranteed to be (δ_t, b_0)-safe if $\mathrm{LCB}_t(X_t^{\mathsf{LCB}}) \geq b_0$. In this case, the SEGE algorithm relies on the LCB arm for safe exploration, as its expected

reward is *potentially* superior to the baseline arm's expected reward.[4] Putting everything together, the SEGE algorithm sets the safe arm (X_t^{S}) at each stage t according to:

$$X_t^{\mathsf{S}} = \begin{cases} X_t^{\mathsf{LCB}}, & \text{if } \mathrm{LCB}_t(X_t^{\mathsf{LCB}}) \geq b_0, \\ X_0, & \text{otherwise.} \end{cases} \tag{15}$$

Before closing this section, it is important to note that the LCB arm (14) can be calculated in polynomial time by solving a second-order cone program. This is in stark contrast to the non-convex optimization problem that needs to be solved when computing the UCB arm (i.e., the arm with the largest upper confidence bound on the expected reward)—a problem that has been shown to be NP-hard in general (Dani, Hayes, and Kakade 2008).

4.3 Safe Greedy Exploitation

We now describe a novel approach to "safe exploration" that will be utilized in the design of the proposed learning algorithm. Exploitation under the SEGE algorithm relies on the certainty equivalence principle. That is, the algorithm first estimates the unknown reward parameter according to Equation (7). Then, the algorithm chooses an arm that is optimal for the given parameter estimate. Given the ellipsoidal structure of the set of allowable arms, the optimal arm X^* can be calculated as

$$X^* = \bar{x} + \frac{H\theta^*}{\|\theta^*\|_H}. \tag{16}$$

Similarly, the certainty equivalent (greedy) arm can be calculated as

$$X_t^{\mathsf{CE}} = \bar{x} + \frac{H\widehat{\theta}_{t-1}}{\|\widehat{\theta}_{t-1}\|_H}, \tag{17}$$

where $\widehat{\theta}_{t-1}$ is the regularized least-squares estimate of the unknown reward parameter, as defined in Equation (7).

It is important to note that the SEGE algorithm only plays the greedy arm (17) when the lower confidence bound on its expected reward is greater than or equal to the safety threshold b. This ensures that the greedy arm is only played when it is safe.

5 Theoretical Results

We now present our main theoretical results showing that the SEGE algorithm exhibits near optimal regret for a large class of risk levels (cf. Theorem 3), in addition to being safe at every stage of play (cf. Theorem 2). As an immediate corollary to Theorem 3, we establish sufficient conditions under which the SEGE algorithm is also guaranteed to satisfy the conservative bandit constraint (6), while preserving the upper bound on regret in Theorem 3 (cf. Corollary 1).

Theorem 2 (Stagewise Safety Guarantee) *The SEGE algorithm is (δ_t, b)-safe at each stage, i.e.,*

$$\mathbb{P}\left(\langle X_t, \theta^* \rangle \geq b\right) \geq 1 - \delta_t$$

for all $t \geq 1$.

[4] It is important to note that the condition $\mathrm{LCB}_t(X_t^{\mathsf{LCB}}) \geq b_0$ does not guarantee superiority of the LCB arm to the baseline arm, as b_0 is only assumed to be a lower bound on the baseline arm's expected reward.

Algorithm 1 SEGE Algorithm

Input: X_0, b_0, $\mathcal{X}$, $S > 0$, $c > 0$, $\lambda > 0$, $b < b_0$, $\rho \in (0, \bar{\rho}]$, $\delta_t \in [0, 1]\ \forall t \geq 1$
for $t = 1, 2, 3, \ldots$ **do**
{*Parameter Estimation*}
  Set $\widehat{\theta}_{t-1}$ according to Eq. (7)
  Set $\mathcal{C}_{t-1}(\delta_t)$ according to Eq. (9)
{*Safe Greedy Exploitation*}
  if $\mathrm{LCB}_t(X_t^{\mathsf{CE}}) \geq b$ **and** $\lambda_{\min}(V_t) \geq c\sqrt{t}$
    Set $X_t = X_t^{\mathsf{CE}}$ according to Eq. (17)
{*Safe Exploration*}
  else
    Set $X_t = X_t^{\mathsf{SE}}$ according to Eq. (11)
  end if
  Observe $Y_t = \langle X_t, \theta^* \rangle + \eta_t$
end for

The ability to enforce safety in the sequence of arms played is not surprising given the assumption of a known baseline arm that is guaranteed to be safe at the outset. However, given the potential suboptimality of the baseline arm, a naïve policy that plays the baseline arm at every stage will likely incur an expected regret that grows linearly with the number of stages played T. In constrast, we show, in Theorem 3, that the SEGE algorithm exhibits an expected regret that is no greater than $O(\sqrt{T}\log(T))$ after T stages—a regret rate that is near optimal given existing $\Omega(\sqrt{T})$ lower bounds on regret (Dani, Hayes, and Kakade 2008; Rusmevichientong and Tsitsiklis 2010).

Theorem 3 (Upper Bound on Expected Regret) *Fix $\overline{\delta} \in (0, 1]$ and $K \geq 0$. Let $\{\delta_t\}_{t=1}^{\infty}$ be any sequence of risk levels satisfying*

$$\delta_t \geq \overline{\delta} e^{-K\sqrt{t}} \tag{18}$$

for all $t \geq 1$. Then, there exists a finite positive constant C such that the expected regret of the SEGE algorithm is upper bounded as

$$R_T \leq C\sqrt{T}\log(T) \tag{19}$$

for all $T \geq 1$.

In what follows, we provide a high-level sketch of the proof of Theorem 3. The complete proof can be found in the appendix to (Khezeli and Bitar 2019). We bound the expected regret incurred during the safe exploration and the greedy exploitation stages separately. First, we show that the stagewise expected regret incurred when playing the greedy arm is proportional to the mean squared parameter estimation error. We then employ Theorem 1 to show that, conditioned on the event $\{\lambda_{\min}(V_t) \geq c\sqrt{t}\}$, the mean squared parameter estimation error at each stage t is no greater than $O(\log(t)/\sqrt{t})$. It follows that the cumulative expected regret incurred during the exploitation stages is no more than $O(\sqrt{T}\log(T))$ after T stages of play. Now, in order to upper bound the expected regret accumulated during the safe exploration stages, it suffices to upper bound the expected

number of safe exploration stages, since the stagewise regret can be upper bounded by a finite constant under any admissible policy. We show that the expected number of safe exploration stages is no more than $O(\sqrt{T})$ after T stages of play for any sequence of risk levels that does not decay faster than the rate specified in (18).

We close this section with a result establishing sufficient conditions under which the SEGE algorithm is guaranteed to satisfy the conservative performance constraint (6), in addition to being stagewise safe, while satisfying an upper bound on its expected regret that matches that of the CLUCB algorithm (Kazerouni et al. 2017, Theorem 5). Corollary 1 is stated without proof, as it is an immediate consequence of Theorems 2 and 3.

Corollary 1 (Conservative Performance Guarantee) *Let $\delta \in (0, 1)$. Assume, in addition to the standing assumptions of Theorem 3, that $\{\delta_t\}_{t=1}^{\infty}$ is a summable sequence satisfying $\sum_{t=1}^{\infty} \delta_t \leq \delta$. Then, the SEGE algorithm satisfies the conservative performance constraint* (6)*, and exhibits an expected regret that is upper bounded by $O(\sqrt{T}\log(T))$ for all $T \geq 1$.*

6 Simulation Results

In this section, we conduct a simple numerical study to illustrate the qualitative features of the SEGE algorithm and compare it with the CLUCB algorithm introduced by (Kazerouni et al. 2017).

6.1 Simulation Setup

Model Parameters. We consider a linear bandit with a two-dimensional input space ($d = 2$), and restrict the set of allowable arms $\mathcal{X}$ to be closed disk of radius $r = 1$ centered at $\bar{x} = (1, 1)$. The true reward parameter is taken to be $\theta^* = (0.6, 0.8)$, and the upper bound on its norm is set to $S = 1$. We select a baseline arm at random from the set of allowable arms as $X_0 = (1.2, 1.9)$, and set the baseline expected reward to $b_0 = \langle X_0, \theta^* \rangle = 2.24$. We set the safety threshold to $b = 0.8 \times b_0$. The observation noise process $\{\eta_t\}_{t=1}^{\infty}$ is assumed to be an IID sequence of zero-mean Normal random variables with standard deviation $\sigma_\eta = 1$.

SEGE Algorithm. We set the parameters of the SEGE algorithm to $c = 0.5$, $\lambda = 0.1$, and $\rho = \bar{\rho} = 0.224$. We generate the random exploration process according to $U_t = \bar{x} + \zeta_t$, where $\{\zeta_t\}_{t=1}^{\infty}$ is a sequence of IID random variables that are uniformly distributed on the unit circle. To enable a direct comparison between the SEGE and CLUCB algorithms, we restrict our attention to a summable sequence of risk levels that satisfy the conditions of Corollary 1. Specifically, we set the sequence of risk levels to $\delta_t = 6\bar{\delta}/(\pi^2 t^2)$ for all stages $t \geq 1$, where $\bar{\delta} = 0.1$.

CLUCB Algorithm. We note that the implementation of the CLUCB algorithm requires the repeated solution of a non-convex optimization problem in order to compute UCB arms. To circumvent this intractable calculation, we approximate the continuous set of arms $\mathcal{X}$ by a finite set of arms $\widehat{\mathcal{X}}$ that correspond to a uniform discretization of the boundary of $\mathcal{X}$. The error induced by this approximation is negligible, as $\max_{x \in \mathcal{X}} \langle x, \theta^* \rangle - \max_{x \in \widehat{\mathcal{X}}} \langle x, \theta^* \rangle \leq 3 \times 10^{-3}$.

6.2 Performance of the SEGE Algorithm

We first discuss the transient behavior and performance of the SEGE algorithm. As one might expect, the SEGE algorithm initially relies on the baseline arm for safe exploration as depicted in Figure 3(a). Over time, as the algorithm accumulates information, it is able to gradually expand the set of safe arms as shown in Figure 2. This expansion enables the algorithm to increase the stagewise efficiency with which it safely explores by selecting arms in the vicinity of the safe arm with the largest lower confidence bounds on their expected rewards. In turn, the SEGE algorithm is able to exploit the information gained to play the greedy with increasing frequency over time. As a result, the growth rate of regret diminishes over time as depicted in Figure 3(c). Critically, Figure 3(a) also shows that the SEGE algorithm maintains stagewise safety throughout each of the 250 independent experiments.

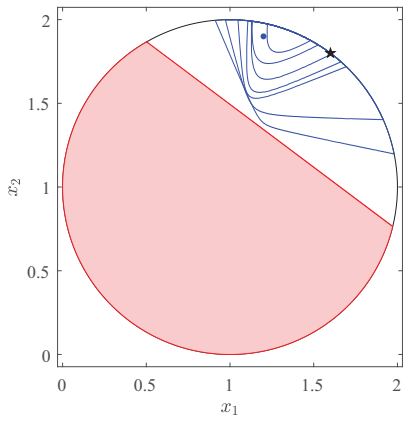
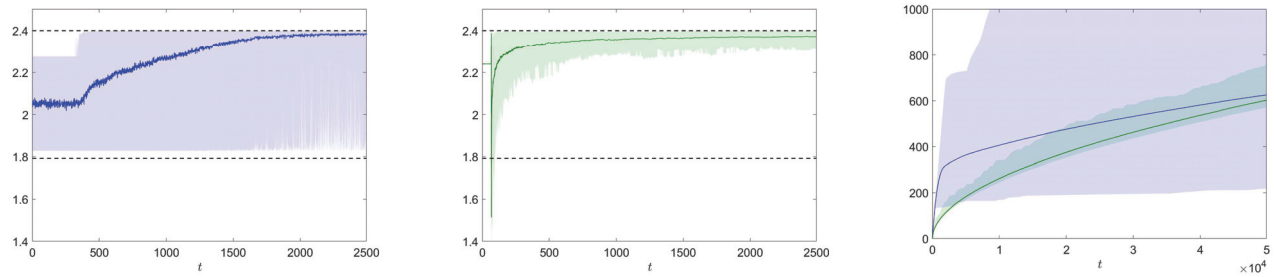


Figure 2: The blue curves depict the gradual expansion of the set of safe arms $\{x \in \mathcal{X} \mid \text{LCB}_t(x) \geq b\}$ over time under the SEGE algorithm for $t = 250$, 500, 1000, 2000, 5000, 10000, and 50000. The blue dot depicts the baseline arm X_0, the black star depicts the optimal arm X^*, and the red shaded area depicts the set of unsafe arms.

6.3 Comparison with the CLUCB Algorithm

Unlike the SEGE algorithm, the CLUCB algorithm is seen to violate the stagewise safety constraint at an early stage in the learning process as depicted in Figure 3(b). The violation of the stagewise safety constraint by the CLUCB algorithm is not surprising as it is only guaranteed to respect the conservative performance constraint (6). The SEGE algorithm, on the other hand, is guaranteed to satisfy the conservative performance constraint, in addition to being stagewise safe (cf. Corollary 1). However, as one might expect, the more stringent safety guarantee of the SEGE algorithm comes at a cost. Specifically, the regret under the SEGE algorithm initially grows more rapidly than the regret incurred



(a) Stagewise expected reward under the SEGE algorithm.

(b) Stagewise expected reward under the CLUCB algorithm.

(c) Cumulative regret of the SEGE algorithm (blue) and the CLUCB algorithm (green).

Figure 3: These figures illustrate the empirical performance of the SEGE and CLUCB algorithms. The solid lines depict empirical means and the shaded regions depict empirical ranges computed from 250 independent simulations.

by the CLUCB algorithm, as shown in Figure 3(c). However, over time the growth rate of regret of the SEGE algorithm slows down as information accumulates and the need for safe exploration diminishes enabling the algorithm to play the greedy arm more frequently.

Acknowledgments

This material is based upon work supported by the Holland Sustainability Project Trust, and the National Science Foundation under grant no. ECCS-135162 and IIP-1632124.

References

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

Agrawal, S., and Goyal, N. 2013. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, 127–135.

Berry, D. A., and Pearson, L. M. 1985. Optimal designs for clinical trials with dichotomous responses. *Statistics in Medicine* 4(4):497–508.

Cassel, A.; Mannor, S.; and Zeevi, A. 2018. A general approach to multi-armed bandits under risk criteria. In *Conference On Learning Theory*, 1295–1306.

Dani, V.; Hayes, T. P.; and Kakade, S. M. 2008. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, 355–366.

David, Y.; Szörényi, B.; Ghavamzadeh, M.; Mannor, S.; and Shimkin, N. 2018. PAC bandits with risk constraints. In *ISAIM*.

den Boer, A. V., and Zwart, B. 2013. Simultaneously learning and optimizing using controlled variance pricing. *Management science* 60(3):770–783.

Galichet, N.; Sebag, M.; and Teytaud, O. 2013. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, 245–260.

Kazerouni, A.; Ghavamzadeh, M.; Abbasi, Y.; and Van Roy, B. 2017. Conservative contextual linear bandits. In *Advances in Neural Information Processing Systems*, 3910–3919.

Keskin, N. B., and Zeevi, A. 2014. Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research* 62(5):1142–1167.

Khezeli, K., and Bitar, E. 2017. Risk-sensitive learning and pricing for demand response. *IEEE Transactions on Smart Grid* 9(6):6000–6007.

Khezeli, K., and Bitar, E. 2019. Safe linear stochastic bandits. *arXiv preprint*.

Li, C.; Kveton, B.; Lattimore, T.; Markov, I.; de Rijke, M.; Szepesvári, C.; and Zoghi, M. 2019. Bubblerank: Safe online learning to re-rank via implicit click feedback. In *The Conference on Uncertainty in Artificial Intelligence*.

Rusmevichientong, P., and Tsitsiklis, J. N. 2010. Linearly parameterized bandits. *Mathematics of Operations Research* 35(2):395–411.

Sani, A.; Lazaric, A.; and Munos, R. 2012. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, 3275–3283.

Sui, Y.; Gotovos, A.; Burdick, J.; and Krause, A. 2015. Safe exploration for optimization with gaussian processes. In *International Conference on Machine Learning*, 997–1005.

Sui, Y.; Burdick, J.; Yue, Y.; et al. 2018. Stagewise safe bayesian optimization with gaussian processes. In *International Conference on Machine Learning*, 4788–4796.

Sun, W.; Dey, D.; and Kapoor, A. 2017. Safety-aware algorithms for adversarial contextual bandit. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3280–3288. JMLR. org.

Usmanova, I.; Krause, A.; and Kamgarpour, M. 2019. Safe convex learning under uncertain constraints. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2106–2114.

Vakili, S., and Zhao, Q. 2015. Mean-variance and value at risk in multi-armed bandit problems. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 1330–1335. IEEE.

Vakili, S., and Zhao, Q. 2016. Risk-averse multi-armed bandit problems under mean-variance measure. *IEEE Journal of Selected Topics in Signal Processing* 10(6):1093–1111.

Villar, S. S.; Bowden, J.; and Wason, J. 2015. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics* 30(2):199.

Wu, Y.; Shariff, R.; Lattimore, T.; and Szepesvári, C. 2016. Conservative bandits. In *International Conference on Machine Learning*, 1254–1262.