A Proof of Lemma 1

In order to bound the minimum eigenvalue of the Gram matrix at round $T'+1$, we use the Matrix Chernoff Inequality (Tropp et al., 2015, Thm. 5.1.1).

Theorem 4 (Matrix Chernoff Inequality, Tropp et al. (2015))**.** *Consider a finite sequence $\{X_k\}$ of independent, random, symmetric matrices in $\mathbb{R}^d$. Assume that $\lambda_{\min}(X_k) \geq 0$ and $\lambda_{\max}(X_k) \leq L$ for each index k. Introduce the random matrix $Y = \sum_k X_k$. Let μ_{min} denote the minimum eigenvalue of the expectation $\mathbb{E}[Y]$,*

$$\mu_{\min} = \lambda_{\min}\left(\mathbb{E}[Y]\right) = \lambda_{\min}\left(\sum_k E[X_k]\right).$$

Then, for any $\epsilon \in (0,1)$, it holds,

$$\Pr\left(\lambda_{\min}(Y) \leq \epsilon\mu_{\min}\right) \leq d \cdot \exp\left(-(1-\epsilon)^2 \frac{\mu_{\min}}{2L}\right).$$

Proof of Lemma 1. Let $X_t = x_t x_t^\dagger$ for $t \in [T']$, such that each X_t is a symmetric matrix with $\lambda_{\min}(X_t) \geq 0$ and $\lambda_{\max}(X_t) \leq L^2$. In this notation, $A_{T'+1} = \lambda I + \sum_{t=1}^{T'} X_t$. In order to apply Theorem 4, we compute:

$$\mu_{\min} := \lambda_{\min}\left(\sum_{t=1}^{T'} \mathbb{E}[X_t]\right) = \lambda_{\min}\left(\sum_{t=1}^{T'} \mathbb{E}[x_t x_t^\dagger]\right) = \lambda_{\min}\left(T'\Sigma\right) = \lambda_- T'.$$

Thus, the theorem implies the following for any $\epsilon \in [0,1)$:

$$\Pr\left[\lambda_{\min}(\sum_{t=1}^{T'} X_t) \leq \epsilon\lambda_- T'\right] \leq d \cdot \exp\left(-(1-\epsilon)^2 \frac{\lambda_- T'}{2L^2}\right). \tag{16}$$

To complete the proof of the lemma, simply choose $\epsilon = 0.5$ (say) and $T' \geq \frac{8L^2}{\lambda_-}\log(\frac{d}{\delta})$ in (16). This gives $\Pr\left[\lambda_{\min}(A_{T'+1}) \geq \lambda + \frac{\lambda_- T'}{2}\right] \geq 1-\delta$, as desired. □

B Proof of Theorems 2 and 3

In this section, we present the proofs of Theorems 2 and 3.

B.1 Preliminaries

Conditioning on $\mu \in \mathcal{C}_t,\ \forall t > 0$. Consider the event

$$\mathcal{E} := \{\mu \in \mathcal{C}_t,\ \forall t > 0\}, \tag{17}$$

that μ is inside the confidence region for all times t. By Theorem 1 the event holds with probability $1-\delta$. Onwards, we condition on this event, and we make repeated use of the fact that $\mu \in \mathcal{C}_t$ for all $t > 0$, without further explicit reference.

Decomposing the regret in two terms. Recall the decomposition of the instantaneous regret in two terms in (10) as follows:

$$r_t = \mu^\dagger x_t - \mu^\dagger x^* = \underbrace{\mu^\dagger x_t - \tilde{\mu}_t^\dagger x_t}_{\text{Term I}} + \underbrace{\tilde{\mu}_t^\dagger x_t - \mu^\dagger x^*}_{\text{Term II}}. \tag{18}$$

As discussed in Section 3.1, we control the two terms separately.

B.2 Bounding Term I

The results in this subsection are by now rather standard in the literature (see for example (Abbasi-Yadkori et al., 2011)). We provide the necessary details for completeness.

We start with the following chain of inequalities, that hold for all $t \geq T' + 1$:

$$\begin{aligned}\text{Term I} &:= \mu^\dagger x_t - \tilde{\mu}_t^\dagger x_t = (\mu^\dagger x_t - \hat{\mu}_t^\dagger x_t) + (\hat{\mu}_t^\dagger x_t - \tilde{\mu}_t^\dagger x_t) \\ &\leq \|\mu - \hat{\mu}_t\|_{A_t} \|x_t\|_{A_t^{-1}} + \|\hat{\mu}_t - \tilde{\mu}_t\|_{A_t} \|x_t\|_{A_t^{-1}} \\ &\leq 2\beta_t \|x_t\|_{A_t^{-1}}. \end{aligned} \tag{19}$$

The last inequality (19) follows from Theorem 1 and the fact that μ and $\tilde{\mu}_t \in \mathcal{C}_t$. Recall, from Assumption 2, the trivial bound on the instantaneous regret

$$r_t = \mu^\dagger x_t - \mu^\dagger x^* \leq 2.$$

Thus, we conclude with the following

$$\text{Term I} \leq 2 \min\left(\beta_t \|x_t\|_{A_t^{-1}}, 1\right). \tag{20}$$

The next lemma bounds the total contribution of the (squared) terms in (19) across the entire horizon $t = T' + 1, \ldots, T$.

Lemma 3 (Term I). *Let Assumptions 1 and 2 hold. Fix any $\delta \in (0, 0.5)$ and assume that T' is such that $T' \geq \frac{8L^2}{\lambda_-} \log\left(\frac{d}{\delta}\right)$. Then, with probability at least $1 - \delta$, it holds*

$$\sum_{t=T'+1}^{T} \min\left(\|x_t\|^2_{A_t^{-1}}, 1\right) \leq 2d \log\left(\frac{2TL^2}{d(2\lambda + \lambda_- T')}\right).$$

Thus, with probability at least $1 - 2\delta$, it holds

$$\sum_{t=T'+1}^{T} (\mu^\dagger x_t - \tilde{\mu}_t^\dagger x_t) \leq 2\beta_T \sqrt{2d\,(T - T')\, \log\left(\frac{2TL^2}{d\,(2\lambda + \lambda_- T')}\right)}. \tag{21}$$

Proof. The proof is mostly adapted from (Dani et al., 2008, Lem. 9) but we also exploit the bound on $\lambda_{\min}(A_{T'+1})$ thanks to Lemma 1. We present the details for the reader's convenience.

With probability at least $1 - \delta$, we find that for all $t \geq T' + 1$:

$$\begin{aligned}\det(A_{t+1}) &= \det(A_t + x_t x_t^\dagger) = \det(A_t) \det(I + (A_t^{-\frac{1}{2}} x_t)(A_t^{-\frac{1}{2}} x_t)^\dagger) = \det(A_t)(1 + \|x_t\|^2_{A_t^{-1}}) \\ &= \ldots = \det(A_{T'+1}) \prod_{\tau=T'+1}^{t} (1 + \|x_\tau\|^2_{A_\tau^{-1}}) \\ &\geq \left(\lambda + \frac{\lambda_- T'}{2}\right)^d \prod_{\tau=T'+1}^{t} (1 + \|x_\tau\|^2_{A_\tau^{-1}}),\end{aligned}$$

where the last inequality follows form Lemma 1 and the fact that $\det(A) = \prod_{i=1}^d \lambda_i(A) \geq (\lambda_{\min}(A))^d$. Furthermore, by the AM-GM inequality applied to the eigenvalues of A_{t+1}, if holds

$$\det(A_{t+1}) = \prod_{i=1}^{d} \lambda_i(A_{t+1}) \leq \left(\frac{tL^2}{d}\right)^d,$$

where we also used the fact that $\|x_t\|_2 \leq L$ for all t. These combined yield,

$$\prod_{\tau=T'+1}^{t} (1 + \|x_\tau\|^2_{A_\tau^{-1}}) \leq \left(\frac{2tL^2}{d(2\lambda + \lambda_- T')}\right)^d.$$

Next, using the fact that for any $0 \leq y \leq 1$, $\log(1+y) \geq y/2$, we have

$$\sum_{t=T'+1}^{T} \min\left(\|x_t\|^2_{A_t^{-1}}, 1\right) \leq 2 \sum_{t=T'+1}^{T} \log\left(\|x_t\|^2_{A_t^{-1}}+1\right) = 2\log\Big(\prod_{t=T'+1}^{T}(\|x_t\|^2_{A_t^{-1}}+1)\Big)$$
$$\leq 2d\log\left(\frac{2TL^2}{d(2\lambda+\lambda\_T')}\right).$$

It remains to prove (21). Recall from (20) that for any $T' < t \leq T$, with probability at least $1-\delta$ (note that we have conditioned in the event $\mathcal{E}$ in (17)),

$$(\mu^\dagger x_t - \tilde{\mu}_t^\dagger x_t) \leq 2\min\left(\beta_t\|x_t\|_{A_t^{-1}}, 1\right) \leq 2\beta_T \min\left(\|x_t\|_{A_t^{-1}}, 1\right),$$

where for the inequality we have used the fact that $\beta_t \leq \beta_T$ (and assumed for simplicity that T large enough such that $\beta_T > 1$). Thus, the desired bound in (21) follows from applying Cauchy-Schwartz inequality to the above. □

B.3 Bounding Term II

As discussed in Section 3.2, the challenge in bounding Term II in (10) is that , in general, $\mathcal{D}_t^{\mathsf{S}} \neq \mathcal{D}_0^{\mathsf{S}}$, so x^* might not belong in $\mathcal{D}_t^{\mathsf{S}}$. Bounding Term II amounts to bounding a certain "distance" of the set $\mathcal{D}_t^{\mathsf{S}}$ from the set $\mathcal{D}_0$. In order to accomplish this task, we proceed as follows. First, we define a shrunk version $\widetilde{\mathcal{D}}_t^{\mathsf{S}}$ of $\mathcal{D}_t^{\mathsf{S}}$, for which we have a more convenient characterization, compared to the original $\widetilde{\mathcal{D}}_t^{\mathsf{S}}$. Then, we select the point z_t in $\widetilde{\mathcal{D}}_t^{\mathsf{S}}$ that is in the direction of x^* and is as close to it as possible. Finally, we are able to bound the distance of z_t to x^*.

A shrunk safe region $\widetilde{\mathcal{D}}_t^{\mathsf{S}}$. Consider an enlarged confidence region $\tilde{\mathcal{C}}_t$ centered at μ defined as follows:

$$\tilde{\mathcal{C}}_t := \{v \in \mathbb{R}^d : \|v-\mu\|_{A_t} \leq 2\beta_t\} \supseteq \mathcal{C}_t. \tag{22}$$

The inclusion property above holds since $\mu \in \mathcal{C}_t$, and, by triangle inequality, for all $v \in \mathcal{C}_t$, one has that $\|v-\mu\|_{A_t} \leq \|v-\hat{\mu}_t\|_{A_t} + \|\hat{\mu}_t - \mu\|_{A_t} \leq 2\beta_t$.

The definition of the enlarged confidence region in (22) naturally leads to the definition of a corresponding shrunk safe decision set $\widetilde{\mathcal{D}}_t^{\mathsf{S}}$. Namely, let

$$\widetilde{\mathcal{D}}_t^{\mathsf{S}} := \{x \in \mathcal{D}_0 : v^\dagger Bx \leq c,\ \forall v \in \tilde{\mathcal{C}}_t\} = \{x \in \mathcal{D}_0 : \max_{v \in \tilde{\mathcal{C}}_t} v^\dagger Bx \leq c\}$$
$$= \{x \in \mathcal{D}_0 : \mu^\dagger Bx + 2\beta_t\|Bx\|_{A_t^{-1}} \leq c\}, \tag{23}$$

and observe that $\widetilde{\mathcal{D}}_t^{\mathsf{S}} \subseteq \mathcal{D}_t^{\mathsf{S}}$. Note here that since by Assumption 3 zero is in the interior of $\mathcal{D}_0$, the sets $\widetilde{\mathcal{D}}_t^{\mathsf{S}}$ and $\mathcal{D}_t^{\mathsf{S}}$ have a nonempty interior.

A point $z_t \in \widetilde{\mathcal{D}}_t^{\mathsf{S}}$ close to x^*. Let z_t be a vector in the direction of x^* that belongs in $\widetilde{\mathcal{D}}_t^{\mathsf{S}}$ and is closest to x^*. Formally, $z_t := \alpha_t x^*$, where

$$\alpha_t := \max\left\{\alpha \in [0,1] \,|\, z_t = \alpha x^* \in \widetilde{\mathcal{D}}_t^{\mathsf{S}}\right\}.$$

Since both 0 and $x^* \in \mathcal{D}_0$, and, $\mathcal{D}_0$ is convex by assumption, it follows in view of (23) that

$$\alpha_t := \max\left\{\alpha \in [0,1] \,|\, \alpha \cdot \left(\mu^\dagger Bx^* + 2\beta_t\|Bx^*\|_{A_t^{-1}}\right) \leq c\right\}. \tag{24}$$

Recall that $C > 0$, thus (24) can be simplified to the following:

$$\alpha_t = \begin{cases} 1 & \text{, if} \quad \mu^\dagger Bx^* + 2\beta_t\|Bx^*\|_{A_t^{-1}} \leq c, \\ \min\left(\frac{c}{\mu^\dagger Bx^* + 2\beta_t\|Bx^*\|_{A_t^{-1}}}, 1\right) & \text{, otherwise.} \end{cases} \tag{25}$$

Bounding Term II **in terms of** α_t**.** Due to the fact that $\widetilde{\mathcal{D}}_t^{\mathrm{S}} \subseteq \mathcal{D}_t^{\mathrm{S}}$, it holds that $z_t \in \mathcal{D}_t^{\mathrm{S}}$. Using this, and optimality of $(\tilde{\mu}, x_t)$ in the minimization in Step 10 of Algorithm 1, we can bound Term II as follows:

$$\begin{aligned}\text{Term II} &:= \tilde{\mu}_t^\dagger x_t - \mu^\dagger x^* \\ &\le \mu^\dagger z_t - \mu^\dagger x^* = \alpha_t \mu^\dagger x^* - \mu^\dagger x^* \\ &\le |\alpha_t - 1|\, |\mu^\dagger x^*| \\ &\le |\alpha_t - 1| = (1 - \alpha_t). \end{aligned} \tag{26}$$

The inequality in the last line uses Assumption 2. For the last equality recall that $\alpha_t \in [0, 1]$

To proceed further from (26) we consider separately the two cases $\Delta > 0$ and $\Delta = 0$ that lead to Theorems 2 and 3, respectively.

B.3.1 Bound for the case $\Delta > 0$

Here, assuming that $\Delta > 0$, we prove that if the duration T' of the pure exploration phase of Safe-LUCB is chosen appropriately, then $\alpha_t = 1$, and equivalently, $x^* \in \mathcal{D}_t^{\mathrm{S}}$. The precise statement is given in Lemma 4 below, which is a restatement of Lemma 2, given here for the reader's convenience.

Lemma 4 ($\Delta > 0 \implies x^* \in \mathcal{D}_t^{\mathrm{S}}$)**.** *Let Assumptions 1, 2 and 3 hold for all $t \in [T]$. Fix any $\delta \in (0, 0.5)$ and assume a positive safety gap $\Delta > 0$. Initialize Safe-LUCB with*

$$T' \ge \left(\frac{8L^2 \|B\|^2 \beta_T^2}{\lambda_- \Delta^2} - \frac{2\lambda}{\lambda_-} \right) \vee t_\delta. \tag{27}$$

Then, with probability at least $1 - 2\delta$, for all $t = T' + 1, \dots, T$ it holds that

$$\text{Term II} := \tilde{\mu}_t^\dagger x_t - \mu^\dagger x^* \le 0.$$

Thus, with the same probability

$$\sum_{t=T'+1}^{T} (\tilde{\mu}_t^\dagger x_t - \mu^\dagger x^*) \le 0. \tag{28}$$

Proof. Recall from (26), that for any $T' < t \le T$, with probability at least $1 - \delta$ (note that we have conditioned in the event $\mathcal{E}$ in (17)), Term II $= 1 - \alpha_t$. Thus, in view of (25), it suffices to prove that for any $T' < t \le T$, with probability at least $1 - \delta$, it holds $\alpha_t = 1$, or equivalently,

$$\mu^\dagger B x^* + 2\beta_t \|Bx^*\|_{A_t^{-1}} \le c \;\;\Leftrightarrow\;\; \beta_t \|Bx^*\|_{A_t^{-1}} \le \Delta/2. \tag{29}$$

For any $T' < t \le T$, we have

$$\beta_t \|Bx^*\|_{A_t^{-1}} \le \frac{\beta_t \|Bx^*\|_2}{\sqrt{\lambda_{\min}(A_t)}} \le \frac{\beta_T \|Bx^*\|_2}{\sqrt{\lambda_{\min}(A_{T'+1})}} \le \frac{\beta_T \|B\| L}{\sqrt{\lambda_{\min}(A_{T'+1})}}, \tag{30}$$

where, in the second inequality we used $\beta_t \le \beta_T$ and $\lambda_{\min}(A_t) \ge \lambda_{\min}(A_{T'+1})$, and in the last inequality we used Assumption 2. Next, since $t_\delta \le T'$, we may apply Lemma 1 to find from (30), that for all $T' + 1 \le t \le T$, with probability at least $1 - \delta$:

$$\beta_t \|Bx^*\|_{A_t^{-1}} \le \frac{\sqrt{2}\|B\| L \beta_T}{\sqrt{2\lambda + \lambda_- T'}}. \tag{31}$$

To complete the proof of the lemma note that the assumption $T' \ge \frac{8\|B\|^2 L^2 \beta_T^2}{\lambda_- \Delta^2} - \frac{2\lambda}{\lambda_-}$ when combined with (31), it guarantees (29), as desired. □

Remark 2. *We remark on a simple tweak in the algorithm that results in a constant T' (i.e., independent of T) in Lemma 4. However, this does* not *change the final order of regret bound in Theorem 2. In particular, we modify Safe-LUCB to use the nested (as is called in Kazerouni et al. (2017)) confidence region $\mathcal{B}_t = \cap_{\tau=1}^t \mathcal{C}_\tau$ at round t such that $\dots \subseteq \mathcal{B}_{t+1} \subseteq \mathcal{B}_t \subseteq \mathcal{B}_{t-1} \subseteq \dots$. According to Theorem 1, it is guaranteed that for all $t > 0$, $\mu \in \mathcal{B}_t$, with high probability. Applying these nested confidence regions in creating safe sets, results in $\dots \subseteq \mathcal{D}_{t-1}^s \subseteq \mathcal{D}_t^s \subseteq \mathcal{D}_{t+1}^s \subseteq \dots$. Thanks to this, it is now guaranteed that once $x^* \in \mathcal{D}_t^s$, the optimal action x^* will remain inside the safe decision sets for all rounds after t. Thus, it is sufficient to find the first round t, such that $x^* \in \mathcal{D}_t^s$. This leads to a shorter duration T' for the pure exploration phase. In particular, following the arguments in Lemma 4, it can be shown that T' becomes the smallest value satisfying $2\sqrt{2}\|B\| L \beta_{T'} \le \Delta \sqrt{2\lambda + \lambda_- T'}$, which is now a constant independent of T.*

B.3.2 Bound for the case $\Delta = 0$

Lemma 5 (Term II for $\Delta = 0$)**.** *Let Assumptions 1, 2 and 3 hold. Fix any $\delta \in (0, 0.5)$ and assume that T' is such that $T' \geq t_\delta$. Then, with probability at least $1-\delta$, it holds*

$$\sum_{t=T'+1}^{T} 1 - \alpha_t \leq \frac{2\sqrt{2}\|B\|L\beta_T(T-T')}{c\sqrt{2\lambda + \lambda_- T'}}. \tag{32}$$

Therefore, with probability at least $1-2\delta$, it holds

$$\sum_{t=T'+1}^{T} (\tilde{\mu}_t^\dagger x_t - \mu^\dagger x^*) \leq \frac{2\sqrt{2}\|B\|L\beta_T(T-T')}{c\sqrt{2\lambda + \lambda_- T'}}. \tag{33}$$

Proof. Recall from (26), that for any $T' < t \leq T$, with probability at least $1-\delta$ (note that we have conditioned in the event $\mathcal{E}$ in (17)), Term II $= 1 - \alpha_t$. Thus, (33) directly follows once we show (32). In what follows, we prove (32).

The definition of α_t in (25) and the fact that $\mu^\dagger B x^* \leq c$ imply that

$$\alpha_t = \begin{cases} 1 & \text{, if } \mu^\dagger B x^* + 2\beta_t \|Bx^*\|_{A_t^{-1}} \leq c, \\ \frac{c}{\mu^\dagger B x^* + 2\beta_t \|Bx^*\|_{A_t^{-1}}} \geq \frac{c}{c + 2\beta_t \|Bx^*\|_{A_t^{-1}}} & \text{, otherwise.} \end{cases}$$

Thus, for all $t \geq T' + 1$:

$$\alpha_t \geq \frac{c}{c + 2\beta_t \|Bx^*\|_{A_t^{-1}}},$$

from which it follows,

$$1 - \alpha_t \leq \frac{2\beta_t \|Bx^*\|_{A_t^{-1}}}{c + 2\beta_t \|Bx^*\|_{A_t^{-1}}} \leq \frac{2\beta_t}{c} \|Bx^*\|_{A_t^{-1}} \leq \frac{2\beta_t \|Bx^*\|_2}{c\sqrt{\lambda_{\min}(A_t)}} \leq \frac{2\beta_t \|B\| L}{c\sqrt{\lambda_{\min}(A_{T'+1})}}.$$

The last two inequalities follow as in (30). To complete the proof, note that since $T' \geq t_\delta$, we can apply Lemma 1. Thus, with probability at least $1-\delta$ it holds,

$$\sum_{t=T'+1}^{T} 1 - \alpha_t \leq \frac{2\beta_T \|B\| L(T-T')}{c\sqrt{\lambda_{\min}(A_{T'+1})}} \leq \frac{2\sqrt{2}\|B\|L\beta_T(T-T')}{c\sqrt{2\lambda + \lambda_- T'}},$$

as desired. □

B.4 Completing the proof of Theorem 2

We are now ready to complete the proof of Theorem 2. Let T sufficiently large such that

$$T > T' \geq \left(\frac{8L^2\|B\|^2\beta_T^2}{\lambda_- \Delta^2} - \frac{2\lambda}{\lambda_-} \right) \vee t_\delta. \tag{34}$$

We combine Lemma 3 (specifically, Eqn. (21)), Lemma 4 (specifically, Eqn. (28)), and, the decomposition in (18), to conclude that

$$R_T = \sum_{t=1}^{T'} r_t + \sum_{t=T'+1}^{T} r_t \leq 2T' + 2\beta_T \sqrt{2d(T-T') \log\left(\frac{2TL^2}{d(2\lambda + \lambda_- T')} \right)}.$$

Specifically, choosing $T' = \left(\frac{8L^2\|B\|^2\beta_T^2}{\lambda_- \Delta^2} - \frac{2\lambda}{\lambda_-} \right) \vee t_\delta$ in the above, results in

$$R_T = \mathcal{O}\left(\frac{\|B\|^2}{\lambda_- \Delta^2} d\sqrt{T} \log T \right), \tag{35}$$

where the constant in the Big-O notation may only depend on L, S, R, λ and δ.

B.5 Completing the proof of Theorem 3

We are now ready to complete the proof of Theorem 3. Let T sufficiently large such that

$$T > T' \geq t_\delta.$$

We combine Lemma 3 (specifically, Eqn. (21)), Lemma 5 (specifically, Eqn. (33)), and, the decomposition in (18), to conclude that

$$R_T = \sum_{t=1}^{T'} r_t + \sum_{t=T'+1}^{T} r_t \leq 2T' + 2\beta_T \sqrt{2d(T-T')\log\left(\frac{2TL^2}{d(2\lambda+\lambda_- T')}\right)} + \frac{2\sqrt{2}\|B\|L\beta_T(T-T')}{c\sqrt{2\lambda+\lambda_- T'}}.$$

Specifically, choosing $T' = \left(\frac{\|B\|L\beta_T T}{c\sqrt{2\lambda_-}}\right)^{\frac{2}{3}} \vee t_\delta$ in the above, results in

$$R_T = \mathcal{O}\left(\left(\frac{\|B\|}{c}\right)^{\frac{2}{3}} \lambda_-^{-1/3} d\, T^{2/3} \log T\right), \tag{36}$$

where as in (35) the constant in the Big-O notation may only depend on L, S, R, λ and δ.

C Extension to linear contextual bandits

In this section, we present an extension to the setting of K-armed contextual bandit. At each round $t \in [T]$, the learner observes a context consisting of K action vectors, $\{y_{t,a} : a \in [K]\} \subset \mathbb{R}^d$ and chooses one action denoted by a_t and observes its associated loss, $\ell_t = \mu^\dagger y_{t,a_t} + \eta_t$. We consider the same constraint (1) which results in a *safe* set of actions at each round $\{y_{t,a} \,|\, a \in [K], \mu^\dagger B y_{t,a} \leq c\}$. The optimal action at round t is denoted by y_{t,a_t^*} where

$$a_t^* \in \underset{a\in[K], \mu^\dagger B y_{t,a} \leq c}{\arg\min} \mu^\dagger y_{t,a}. \tag{37}$$

If the chosen action at round t is denoted by $x_t := y_{t,a_t}$ and the optimal one by $x_t^* := y_{t,a_t^*}$, the cumulative regret over total T rounds will be

$$R_T = \sum_{t=1}^{T} \mu^\dagger x_t - \mu^\dagger x_t^*.$$

We briefly discuss how Safe-LUCB extends to the K-armed contextual setting with provable regret guarantees under the following assumptions.

First, we need the standard Assumptions 1 and 2 that naturally extend to the linear contextual bandit setting. Beyond these, in order for the safe-bandit problem to be well-defined, we assume that safe actions exist at each round. Equivalently, the feasible set in (37) is nonempty and x_t^* is well-defined. Moreover, in order to be able to run the pure-exploration phase of Safe-LUCB with random actions (that guarantee Lemma 1 holds) we further require that at least one of these safe actions is randomly sampled at each round t (technically, we need this assumption to hold only for rounds $1, \ldots, T'$). These two assumptions are both implied by Assumption 4 below.

Assumption 4 (Nonempty safe sets)**.** *Consider the set $\mathcal{D}^w = \{x \in \mathbb{R}^d : \|Bx\|_2 \leq \frac{c}{S}\}$. Then, at each round t, $N_t \geq 1$ number of K action vectors lie within $\mathcal{D}^w$.*

Finally, in order to guarantee that Safe-LUCB has sub-linear regret for the K-armed linear setting we need that the safety gap at each round is strictly positive.

Assumption 5 (Nonzero Δ)**.** *The safety gap $\Delta_t = c - \mu^\dagger B x_t^*$ at each round t is positive.*

Under these assumptions, Safe-LUCB naturally extends to the K-armed linear bandit setting. Specifically, at rounds $t \leq T'$, Safe-LUCB randomly selects x_t to be one of the available N_t action vectors that belong to the set $\mathcal{D}^w$. Assume that $\lambda_{\min}(\mathbb{E}[x_t x_t^\dagger]) \geq \lambda_- > 0$ for all $t \in [T']$.

After round T', Safe-LUCB implements the safe exploration-exploitation phase by choosing safe actions based on OFU principle as in (9). Therefore line 10 of Safe-LUCB changes to

$$a_t = \arg\min_{a \in \mathcal{A}_t^{\mathrm{S}}} \min_{v \in \mathcal{C}_t} v^\dagger y_{t,a}, \tag{38}$$

where the safe set at rounds $t \geq T' + 1$ is defined by

$$\mathcal{A}_t^{\mathrm{S}} = \{a \in [K] : v^\dagger B y_{t,a} \leq c, \forall v \in \mathcal{C}_t\}. \tag{39}$$

With these and subject to Assumptions 1, 2, 4 and 5, it is straightforward to extend the results of Theorem 2 to the setting considered here. Namely, under these assumptions, Safe-LUCB achieves regret $\widetilde{\mathcal{O}}(\sqrt{T})$ when T' is set to T_Δ as in (13) for $\Delta = \min_{t \in [T]} \Delta_t$.

D Safe-LUCB with ℓ_1-confidence region

In this section we briefly discussed a modified ℓ_1-confidence region (as in Dani et al. (2008)), which is used in our numerical experiments.

Motivation. The minimization in (9) involves solving a bilinear optimization problem. In view of (6) and (8) it is not hard to show that (9) can be equivalently expressed as follows:

$$\tilde{\mu}_t^\dagger x_t = \min_x \ \hat{\mu}_t^\dagger x - \beta_t \|x\|_{A_t^{-1}} \quad \text{sub.to} \quad \hat{\mu}_t^\dagger B x + \beta_t \|Bx\|_{A_t^{-1}} \leq c, \ \ x \in \mathcal{D}_0 \ .$$

This is a non-convex optimization problem. Thus, we present a variant of Safe-LUCB (and its analysis) and we show that it can be efficiently implemented (particularly so, when the decision set is a polytope) Dani et al. (2008). We use this variant in our simulation results (see Appendix F).

Algorithm and guarantees. We adapt the procedure first presented in Dani et al. (2008) to our new Safe-LUCB algorithm. The pure-exploration phase of the algorithm remains unaltered. In the safe exploration-exploitation phase, the only thing that changes is the definition of the confidence region in Line 8 in Algorithm 1. Specifically, we define the modified ℓ_1-confidence region as follows:

$$\mathcal{C}_t^{\ell_1} := \{v \in \mathbb{R}^d : \|v - \hat{\mu}_t\|_{A_{t,1}} \leq \beta_t \sqrt{d}\}. \tag{40}$$

Note that for any $v \in \mathcal{C}_t$ and all $t > 0$, $\|A_t^{1/2}(v - \hat{\mu}_t)\|_1 \leq \sqrt{d}\|A_t^{1/2}(v - \hat{\mu}_t)\|_2 \leq \sqrt{d}\beta_t$. Thus, $\mathcal{C}_t \subseteq \mathcal{C}_t^{\ell_1}, \quad \forall t > 0$. From this and Theorem 1, we conclude $\Pr(\mu \in \mathcal{C}_t^{\ell_1}, \forall t > 0) \geq 1 - \delta$. Then, the natural modification of (9) becomes

$$\tilde{\mu}_t^\dagger x_t = \min_{x \in \mathcal{D}_t^{\mathrm{S}}, v \in \mathcal{C}_t^{\ell_1}} v^\dagger x = \min_{v \in \mathcal{C}_t^{\ell_1}} \ f(v), \tag{41}$$

where

$$f(v) := \min_{\substack{x \in \mathcal{D}_0 \\ \hat{\mu}_t^\dagger B x + \sqrt{d}\beta_t \|Bx\|_{A_t^{-1}} \leq C}} \nu^\dagger x. \tag{42}$$

From these, it is clear that all the results and theorems can be directly applied to the modified algorithm which uses ℓ_1-confidence region in (40), with $\beta_t\sqrt{d}$ instead of β_t. As noted in Dani et al. (2008) the regret of the modified algorithm does not optimally scale with the dimension d (since there is an extra factor of $\sqrt{d}$ introduced by the substitution $\beta_t \leftarrow \beta_t\sqrt{d}$). However, as explained next, solving (41) is now computationally tractable.

On computational efficiency. Note that the minimization in (42) is a convex program that can be efficiently solved for fixed ν. In particular, if $\mathcal{D}_0$ is a polytope then the minimization in (42) is a quadratic program. Moreover, note that $f(v)$ is positive homogeneous of degree one, i.e., $f(\theta v) = \theta f(v)$ for any $\theta \geq 0$. Therefore, in order to solve (41) it suffices to evaluate the function $f(v)$ at the $2d$ vertices $v_1, \ldots, v_{2d}$ of $\mathcal{C}_t^{\ell_1}$ in (40) and choose the minimum $f_{\min} := \min_{v_i, \ i \in [2d]} f(v_i)$.

In order to see this, let $v^* \in \arg\min_{v \in \mathcal{C}_t^{\ell_1}} f(v)$ and $\theta_1, \ldots, \theta_{2d} \geq 0, \sum_{i=1}^{d} \theta_i = 1$ such that $v^* = \sum_{i=1}^{2d} \theta_i v_i$. Then,

$$\min_{v \in \mathcal{C}_t^{\ell_1}} f(v) = f(v^*) = \sum_{i=1}^{2d} \theta_i f(v_i) \geq f_{\min} \sum_{i=1}^{2d} \theta_i = f_{\min} \geq \min_{v \in \mathcal{C}_t^{\ell_1}} f(v).$$

Thus,

$$\min_{v \in \mathcal{C}_t^{\ell_1}} f(v) = \min_{v_i,\ i \in [2d]} f(v_i). \tag{43}$$

To sum up, we see from (43) that solving (41) amounts to solving $2d$ quadratic programs (when $\mathcal{D}_0$ is a polytope).

E On GSLUCB

Having no knowledge of the safety gap Δ, GSLUCB starts conservatively by setting the length of the pure exploration phase to its largest possible value, which is equal to T_0 defined in Theorem 3 (corresponding to $\Delta = 0$). The idea behind GSLUB is to generate at each round t of the pure-exploration phase a certain value Δ_t that serves as a lower bound for the unknown safety gap Δ. We discuss possible ways to do so next, but for now let us describe how these lower estimates of Δ can be useful. Owing to the result of Theorem 2, at each round t, GSLUCB computes a pure exploration duration $T_t' = T_{\Delta_t}$, which is associated with the lower confidence bound Δ_t (Eqn. (13) for $\Delta = \Delta_t$). If at some round t, the computed T_t' becomes less than t, then Theorem 2 guarantees that $x^* \in \mathcal{D}_t^{\mathrm{S}}$ and the algorithm switches to the exploration-exploitation phase.

One way to compute the Δ_t's that guarantees $\Delta_t \leq \Delta$ is as follows. For each vector $v \in \mathcal{C}_t$ denote $x_v^* \in \arg\min_{x \in \mathcal{D}_0^{\mathrm{s}}(v)} v^\dagger x$, where $\mathcal{D}_0^{\mathrm{S}}(v) := \{x \in \mathcal{D}_0 : v^\dagger B x \leq c\}$ and define

$$\Delta_t := \min_{v \in \mathcal{C}_t} \Delta_v, \tag{44}$$

where $\Delta_v := c - v^\dagger B x_v^*$. Since $\mu \in \mathcal{C}_t$ with high probability (cf. Theorem 1) and by definition of Δ, it can be seen that $\Delta_t \leq \Delta$. Unfortunately, solving (44) can be challenging and, in general, one has to resort to relaxed versions of the optimization involved, but ones that guarantee $\Delta_t \leq \Delta$ (at least after a few rounds). We leave the study of this general case to future work and we discuss here a special case in which this is possible. We have implemented this special case in the simulation results presented in Figure 1a (see Appendix F). Specifically, we consider a finite K-armed linear bandit setting with feature vectors denoted by $y_1, \ldots, y_K$. We produce lower estimates Δ_t as follows. For all $i \in [K]$, we form the following two sets. (i) The set $\mathcal{C}_t^i = \{v \in \mathcal{C}_t \mid v^\dagger B y_i \leq c\}$ of all vectors in the confidence region for which the action y_i is deemed safe; (ii) The set $\mathcal{Y}_t^i = \{y_j, j \in [K] \mid \max_{v \in \mathcal{C}_t^i} v^\dagger B y_j \leq c\}$ of all actions that are considered safe with respect to all $v \in \mathcal{C}_t^i$. Then, we define

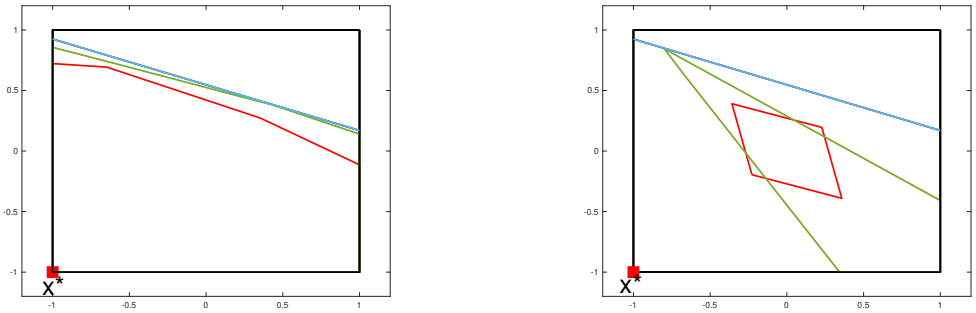
$$\Delta_t^i := \min_{\substack{v \in \mathcal{C}_t^i \\ v^\dagger y_i \leq v^\dagger y, \text{ for all } y \in \mathcal{Y}_t^i}} c - v^\dagger B y_i. \tag{45}$$

It can be checked that $\min_{i \in [K]} \Delta_t^i \leq \Delta$. Thus we rely on $\min_{i \in [K]} \Delta_t^i$ as our lower confidence bound on Δ. Note that computing $\min_{i \in [K]} \Delta_t^i$ is computationally tractable for finite K and an ℓ_1 confidence region.

F Simulation Results

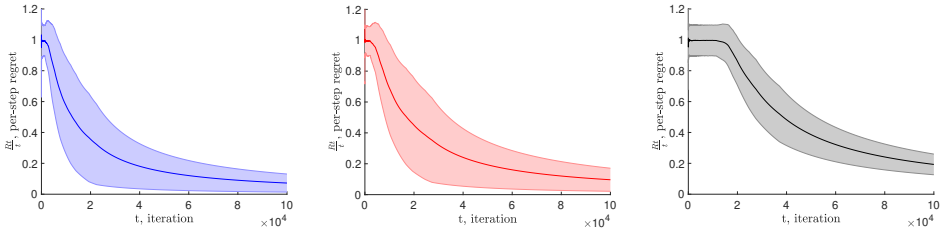
In this section, we provide the details of our numerical experiments. In view of our discussion in Appendix D, we implement a modified version of Safe-LUCB which uses 1-norms instead of 2-norms (as in Dani et al. (2008); see also Appendix D for details). We have taken $\delta = 0.01$, $\lambda = 1$, and $R = 0.1$ in all cases.

Figure 1a compares the average per-step regret of 1) Safe-LUCB with knowledge of Δ; 2) Safe-LUCB without knowledge of Δ (hence, assuming $\Delta = 0$); 3) GSLUCB without knowledge of Δ (the algorithm creates a lower confidence bound for Δ as the pure exploration phase runs). Figure 3



(a) Safe-LUCB with pure exploration phase. (b) Safe-LUCB without pure exploration phase

Figure 2: Growth of $\mathcal{D}_t^{\mathrm{S}}$ with and without pure exploration phase. In both figures: $\mathcal{D}_0$ (in black) $\mathcal{D}_0^{\mathrm{S}}$ (in blue), $\mathcal{D}_{T'+1}^{S}$ (in red), $\mathcal{D}_{5e4}^{S}$ (in green). Also, shown the optimal action x^*. Note that $x^* \in \mathcal{D}_{T'+1}^{S}$ when pure exploration phase is used as suggested by Lemma 2.



(a) Safe-LUCB, $T' = T_\Delta$ (b) GSLUCB (c) Safe-LUCB, $T' = T_0$

Figure 3: Comparison of mean per-step regret for Safe-LUCB($T' = T_\Delta$), GSLUCB, and Safe-LUCB($T' = T_0$). The shaded regions show one standard deviation around the mean. The results are averages over 20 problem realizations.

highlights the sample standard deviation of regret around the average per-step regret for each of the above-mentioned cases. We considered a time independent decision set of 15 arms in $\mathbb{R}^4$ such that 5 of the feature vectors are drawn uniformly from $\mathcal{D}^w$ and the other 10 are drawn uniformly from unit ball in $\mathbb{R}^4$. Moreover, μ is drawn from $\mathcal{N}(0, I_4)$ and then normalized to unit norm. B and c are drawn uniformly from $[0, 0.5]^{4\times 4}$ and [0,1] respectively. The results shown depict averages over 20 realizations. It can be seen from the figure that GSLUCB performs significantly better than the worst case suggested by Theorem 3 (aka Safe-LUCB assuming $\Delta = 0$). In fact, it appears that it approaches the improved regret performance suggested by Theorem 2 of Safe-LUCB with knowledge of Δ.

Our second numerical experiment serves to showcase the value of the safe exploration phase as discussed in Section 3.3. We focus on an instance with positive safety gap $\Delta > 0$ to verify the validity of Lemma 2, namely that $x^* \in \mathcal{D}_t^{\mathrm{S}}$ for $t \geq T' + 1$, when T' is appropriately chosen. Furthermore, we compare the performance with a "naive" variation of Safe-LUCB that only implements the safe exploration-exploitation phase (aka, no pure exploration phase). The regret plots of the two algorithms (with and without pure exploration phase) shown in Figure 1b clearly demonstrate the value of the pure exploration phase for the simulated example. Specifically, for the simulation, we consider a horizon $T = 100000$ with decision set $\mathcal{D}_0$ the unit ℓ_∞-ball in $\mathbb{R}^2$, and, the following parameters: $\mu = \begin{bmatrix} 0.9 \\ 0.044 \end{bmatrix}$, $B = \begin{bmatrix} 0.6 & 1.8 \\ 1.8 & 0.4 \end{bmatrix}$, $c = 0.9$. We have chosen a low-dimensional instance, because we find it instructive to also depict the the growth of the safe sets for the two algorithms. This is done in Figures 2a and 3c, where we illustrate the safe sets of Safe-LUCB with and without pure exploration phase, respectively. Black lines denote the (border of) the polytope $\mathcal{D}_0$; blue lines denote the linear constraint in (1); red lines denote the (border of) $\mathcal{D}_{T'+1}^{\mathrm{s}}$, where $T' = T_\Delta = 1054$ and $T' = 0$ for Figures 2a and 3c, respectively; and, green lines denote (the border of) safe sets $\mathcal{D}_{50000}^{\mathrm{s}}$ at round 50000. Also depicted the optimal action x^* with coordinates $\{-1, -1\}$. As expected,

Safe-LUCB starts the exploration-exploitation phase with a safe set that includes x^* while, without the pure exploration phase, the algorithm starts the exploration-exploitation phase with a smaller safe set which does not include x^* and as a results, fails in expanding the safe set to include x^* even after $T = 50000$ rounds. This results in the bad regret performance in Figure 1b.