Proceedings of Machine Learning Research vol 291:1–25, 2025 38th Annual Conference on Learning Theory

Linear Bandits on Ellipsoids: Minimax Optimal Algorithms

Raymond Zhang* RAYMOND.ZHANG@CENTRALESUPELEC.FR
Hédi Hadiji* HEDI.HADIJI@CENTRALESUPELEC.FR
Richard Combes RICHARD.COMBES@CENTRALESUPELEC
Laboratoire des signaux et systèmes, Université Paris-Saclay, CNRS, CentraleSupélec, France

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We consider linear stochastic bandits where the set of actions is an ellipsoid. We provide the first known minimax optimal algorithm for this problem. We first derive a novel information-theoretic lower bound on the regret of any algorithm, which must be at least $\Omega(\min(d\sigma\sqrt{T}+d\|\theta\|_A, \|\theta\|_A T))$ where d is the dimension, T the time horizon, σ^2 the noise variance, A a matrix defining the set of actions and θ the vector of unknown parameters. We then provide an algorithm whose regret matches this bound to a multiplicative universal constant. The algorithm is non-classical in the sense that it is not optimistic, and it is not a sampling algorithm. The main idea is to combine a novel sequential procedure to estimate $\|\theta\|$, followed by an explore-and-commit strategy informed by this estimate. The algorithm is highly computationally efficient, and a run requires only time $O(dT+d^2\log(T/d)+d^3)$ and memory $O(d^2)$, in contrast with known optimistic algorithms, which are not implementable in polynomial time. We go beyond minimax optimality and show that our algorithm is locally asymptotically minimax optimal, a much stronger notion of optimality. We further provide numerical experiments to illustrate our theoretical findings. The code to reproduce the experiments is available at https://github.com/RaymZhang/LinearBanditsEllipsoidsMinimaxCOLT.

1. Introduction

1.1. Model

We consider the problem of stochastic linear bandits over ellipsoids with independent subgaussian rewards. Time is discrete with a horizon T, and at time $t \in [T] = \{1, .., T\}$ a learner selects an action $x_t \in \mathcal{X}$ where action set $\mathcal{X}$ is a subset of $\mathbb{R}^d$ initially known to the learner. The action set $\mathcal{X}$ is an ellipsoid described by

$$\mathcal{X} := \{x \in \mathbb{R}^d : \|x - c\|_{A^{-1}} \leqslant 1\}$$

with vector $c \in \mathbb{R}^d$, positive definite matrix $A \in \mathbb{R}^{d\times d}$ and $\|u\|_M := \sqrt{u^\top M u}$ for $u \in \mathbb{R}^d$ is the norm associated to positive definite matrix M. After choosing action x_t the learner observes a scalar reward $y_t = x_t^\top \theta + z_t$ where the vector $\theta \in \mathbb{R}^d$ is initially unknown to the learner, and $z_t \in \mathbb{R}$ is additive noise. Random variables $(z_t)_{t\in\mathbb{N}}$ are assumed independent, centered and subgaussian with variance proxy σ^2 so that conditionally on past observations

$$\mathbb{E}\left(\exp\left(\lambda z_t\right)\right) \leqslant \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \text{ for all } t \in \mathbb{N} \text{ and } \lambda \in \mathbb{R}\,.$$

* The first two authors contributed equally to this work.

The learner selects action x_t solely based on the observed rewards up to time t, that is, $y_1, ..., y_{t-1}$. The goal of the learner is to minimize the (expected) regret

$$R_T(\theta) := \mathbb{E}\left[\sum_{t=1}^{T} x^\star(\theta)^\top\theta - x_t^\top\theta\right] \text{ where } x^\star(\theta) \in \arg\max_{x\in\mathcal{X}} x^\top\theta$$

which is the expectation of the difference between the rewards obtained by the learner and the reward obtained by an oracle who knows θ and hence always select the decision with the largest expected reward. From the Karush-Kush-Tucker conditions, $A^{-1}(x^\star(\theta) - c)$ should be proportional to θ, so that the decision chosen by the oracle is $x^\star(\theta) = c + A\theta/\|\theta\|_A$.

For two functions f, g of $\phi = (T, \theta, \mathcal{X}, d, \sigma)$, we write $f(\phi) = O(g(\phi))$ and $f(\phi) = \Omega(g(\phi))$ if there exists c_1, c_2 strictly positive *universal constants* (not allowed to depend on any parameter such as ϕ) such that $f(\phi) \leqslant c_1 g(\phi)$ and $f(\phi) \geqslant c_2 g(\phi)$ respectively, for all ϕ. We do not use the $\widetilde{O}$ notation in which logarithmic terms are hidden, to avoid confusion.

1.2. Our contribution

We make two main contributions to this problem. First we derive a novel information-theoretic lower bound on the regret of any algorithm, stating that for any algorithm and any $B \geqslant 0$

$$\max_{\theta\in\mathcal{E}_A(B)} R_T(\theta) = \Omega\left(\min(BT, d\sigma\sqrt{T})\right) \text{ with } \mathcal{E}_A(B) := \{\theta \in \mathbb{R}^d : \|\theta\|_A = B\}\,.$$

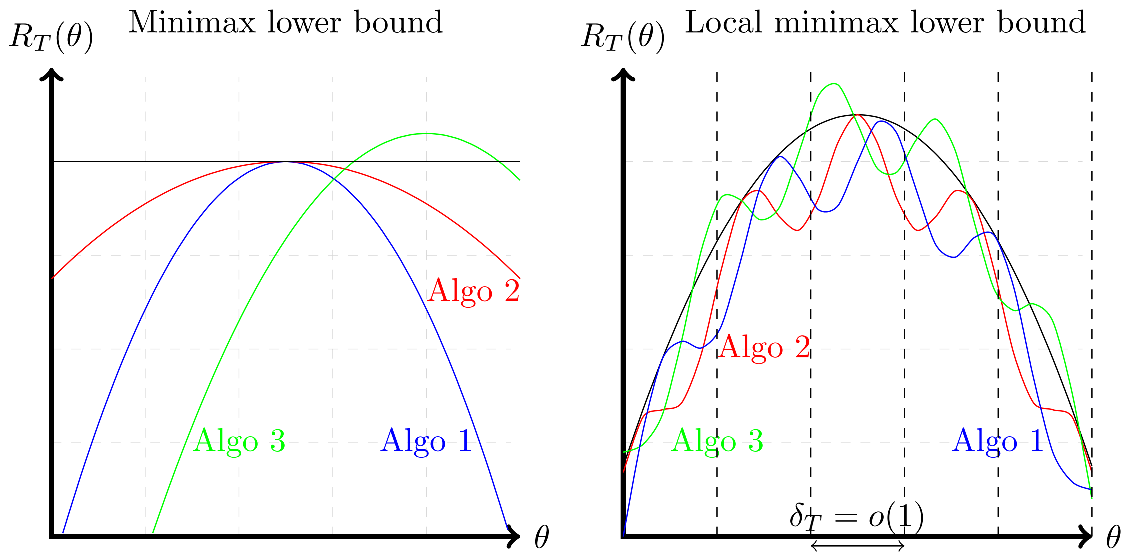
In fact we prove an even stronger result showing that the same hold true when $\mathcal{E}_A(B)$ is replaced by a neighborhood of θ whose size vanishes with T. Second, we propose a novel algorithm called `E2TC` (`Explore-Explore-Then-Commit`) matching this upper bound so that under `E2TC`

$$\max_{\theta\in\mathcal{E}_A(B)} R_T(\theta) = O\left(\min(BT, d\sigma\sqrt{T} + dB)\right)\,.$$

Algorithm `E2TC` has low computational complexity (time $O(dT + d^2\log(T/d) + d^3)$ and memory $O(d^2)$) so it can be applied to high dimensions, and is not based on optimism or sampling. Rather, it combines a non-trivial, adaptive initialization procedure to estimate the norm of θ followed by an Explore-Then-Commit (`ETC`) strategy. We emphasize that initially the learner has no information about the norm of θ. In fact not only do our results show that `E2TC` is minimax optimal with low computational complexity, but they also show that `E2TC` is locally asymptotically minimax optimal, a much stronger notion of optimality than minimax optimality, further discussed next section.

1.3. Locally asymptotic minimax optimality

We use a refined notion of optimality called "locally asymptotic minimax optimality" proposed by Hájek (1972) for estimation. Denote by $R_T^{\mathcal{A}}(\theta)$ the regret of algorithm $\mathcal{A}$ under parameter $\theta \in \mathbb{R}^d$ and let $\mathcal{P} \subset \mathbb{R}^d$ the set of allowed θ. Algorithm $\mathcal{A}$ is minimax optimal if for any algorithm $\mathcal{A}'$ there exists $\theta \in \mathcal{P}$ such that with $R_T^{\mathcal{A}'}(\theta) = \Omega(\max_{\theta\in\mathcal{P}} R_T^{\mathcal{A}}(\theta))$. Algorithm $\mathcal{A}$ is locally asymptotically minimax optimal if for any algorithm $\mathcal{A}'$ there exists θ'_T such that $R_T^{\mathcal{A}'}(\theta'_T) = \Omega(R_T^{\mathcal{A}}(\theta))$ and $\lim_{T\to\infty}\theta'_T = \theta$. Local asymptotic minimax optimality, as illustrated in Figure 1, is stronger than minimax optimality.



(*a*) Minimax optimality (*b*) Locally asymptotic minimax optimality

Figure 1: Schematic illustration of the differences between minimax and locally asymptotically minimax lower bounds. Minimax lower bounds require that the regret curve of any algorithm goes above the bound at least once, whereas locally minimax lower bounds require that the curve of any algorithm goes above the bound *in any region of size* $\delta_T = o(1)$. In both figures, Algorithms 1 and 2 match the lower bound, and Algorithm 3 does not.

Locally optimal algorithms perform well in the worst case, but also adapt to the information-theoretic "local difficulty" of the problem.

For instance when $\mathcal{X}$ is the unit ball, and the set of allowed θ is $\mathcal{P} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leqslant 1\}$ Lattimore and Szepesvári (2020) prove the minimax lower bound $\max_{\theta \in \mathcal{P}} R_T(\theta) = \Omega(d\sqrt{T})$, but their analysis only involves values of θ very close to 0. Our analysis paints a more complex picture, where the difficulty depends on $\|\theta\|_A$; locally minimax optimal algorithms like `E2TC` adapt to this.

1.4. Related Work

We now summarize the current state of the art on the problem of stochastic linear bandits with i.i.d. subgaussian rewards. For fair comparison, we first highlight the difference in assumptions made by authors, and whenever those assumptions differ from ours, we specialize their results to our setting.

Different settings Dani et al. (2008) consider an arbitrary action set $\mathcal{X}$ and assume that $\max_{x \in \mathcal{X}} |\theta^\top x| \leqslant 1$, which in our setting is equivalent to $|c^\top \theta + \|\theta\|_A| \leqslant 1$. Rusmevichientong and Tsitsiklis (2010) consider an arbitrary compact action set $\mathcal{X}$, but they assume intricate conditions on its curvature, see below. Abbasi-Yadkori et al. (2011) consider an arbitrary, possibly time-dependent, action set $\mathcal{X}_t$ and assume $\max_{x \in \mathcal{X}_t} |\theta^\top x| \leqslant 1$, which in our setting is equivalent to $|c^\top \theta + \|\theta\|_A| \leqslant 1$. They further assume that an upper bound for $\|\theta\|_2$ is known and provided as an input to the algorithm. Abeille and Lazaric (2017) consider an action set $\mathcal{X}$ which is an arbitrary closed subset of the unit ball, and also assume that an upper bound for $\|\theta\|_2$ is known and provided as an input to the algorithm.

Regret upper bounds The two versions of the `CB` (`ConfidenceBall`) algorithm of Dani et al. (2008) achieve regret upper bounds of $O(d\sqrt{T}\log(T)^3)$ and $O(d^{3/2}\sqrt{T}\log(T)^3)$ in our setting. The

PEGE (Phased Exploration and Greedy Exploitation) algorithm of Rusmevichientong and Tsitsiklis (2010) achieves a regret of $O((\|\theta\|_2 + 1/\|\theta\|_2)d\sqrt{T})$ when $\mathcal{X}$ is a unit ball. The UE (Uncertainty Ellipsoid) algorithm of Rusmevichientong and Tsitsiklis (2010) achieves a regret of $O(f(\mathcal{X})d\sqrt{T\log(T)^3})$ when $\mathcal{X}$ is an ellipsoid where f has an intricate dependency on the curvature of $\mathcal{X}$, and can only be bounded by a universal constant if the maximal ratio of eigenvalues of A, $|\lambda_{\max}(A)/\sqrt{\lambda_{\min}(A)}|$ is bounded by another universal constant, so that $\mathcal{X}$ is "close" to the unit ball. The OFUL (Optimism in the Face of Uncertainty Linear bandit) algorithm of Abbasi-Yadkori et al. (2011) achieves a regret $O((\|\theta\|_2 + \sigma\sqrt{d}\log(T))\sqrt{dT\log(T)})$ assuming that the algorithm is given an upper bound on $\|\theta\|_2$ as prior information in order to set the algorithm parameters correctly. The TS (Thompson Sampling) algorithm of Abeille and Lazaric (2017) achieves a regret of $O(\sigma d^{3/2}\sqrt{T\log(dT)^3})$ assuming that the algorithm is given an upper bound on $\|\theta\|_2$ as prior information in order to set the algorithm parameters correctly. The OLSOFUL (Ordinary Least Squares OFUL) algorithm of Gales et al. (2022) achieves a regret of $O(\sigma d\sqrt{T\log(T/d)^2} + \|\theta\|_2 d\log(d))$ without prior information on θ. Therefore, the two versions of CB, PEGE, UE, OFUL, OLSOFUL, TS are not provably minimax optimal, in contrast with E2TC.

Regret lower bounds Rusmevichientong and Tsitsiklis (2010) show that the minimax regret of any algorithm is at least $\Omega(d\sqrt{T})$ when the set of actions is the unit ball, using a Bayesian approach. Lattimore and Szepesvári (2020) show the same lower bound in a more general setting, using an Assouad-style construction. Our locally asymptotically minimax lower bound is a stronger result. with the correct dependency on d, σ, T and $\|\theta\|_A$.

Computational complexity To the best of our knowledge, even when $\mathcal{X}$ is a unit ball, none of the optimistic algorithms such as CB, UE, OFUL and OLSOFUL cannot be implemented in polynomial time, because they must repeatedly maximize the bilinear function $x^\top\theta$ over the couple $(x, \theta) \in \mathcal{X} \times \mathcal{C}_t$ where $\mathcal{C}_t$ is the confidence ellipsoid computed at time t. To the best of our knowledge, no efficient algorithm is known to solve this bilinear optimitzation problem when $\mathcal{X}$ is an ellipsoid. The only case in which this is feasible in polynomial time is when $\mathcal{X}$ is finite with polynomial size in d. This is corroborated by our numerical experiments. Both PEGE, TS can be implemented in polynomial time, as they require to minimize a linear function over $\mathcal{X}$, and/or to compute the least-squares estimate of θ. Our results show that there is no regret/complexity trade-off in linear bandits, in the sense that E2TC achieves both minimax regret, and low computational complexity.

Related problems Other authors have considered different but related problems in linear bandits, for instance: Zhu and Nowak (2022) consider the case where θ is sparse, Banerjee et al. (2022) lower bound asymptotically the eigenvalues of the expected design matrix, Jun and Kim (2024) consider the case where the variance of the reward from the optimal decision is small, and Lattimore and Szepesvári (2020) consider the adversarial case (we refer the reader to their discussion of the major differences between stochastic and adversarial linear bandits).

2. Regret lower bound

Our first main result is an information-theoretic lower bound on the regret of any algorithm on ellipsoids, presented in Theorem 1. Our result is a locally asymptotically minimax lower bound (stronger than a minimax lower bound) indeed, since given any θ, there must exist a θ' in the vicinity of θ on which regret is larger than the lower bound, with $\|\theta - \theta'\| \to 0$ when T is large. The lower bound is a minimum of two terms, the first proportional to $d\sigma\sqrt{T}$, and the second to $T\|\theta\|_A$, which depends on the norm of θ. Those two terms identify two "regimes" depending on $\|\theta\|_A$, suggesting that the correct measure of problem complexity is $\|\theta\|_A$ which captures both the magnitude and the orientation of θ, as well as the curvature of the action set $\mathcal{X}$.

Similarly to the minimax lower bound for the ball (Thm 24.2 in Lattimore and Szepesvári (2020)), our proof is built around an Assouad-like construction where we define 2^d alternative mean-reward vectors $\theta(\xi)$, with $\xi \in \{-1,+1\}^d$, and argue that no algorithm can have low regret on all such parameters. We show our stronger result by enforcing that the alternative vectors be in the vicinity of the given θ, and of the same norm. This comes with technical difficulty, as we need to lower bound the regret of playing actions that are close to optimal. We also took extra care to exhibit clearly the different regimes involving T, σ, B and d.

Theorem 1 *For any algorithm, any $B > 0$, any $T \geqslant 1$ and any $\theta \in \mathcal{E}_A(B) = \left\{\theta \in \mathbb{R}^{d+2}, \|\theta\|_A = B\right\}$ there exists a bandit problem with mean-rewards vector $\theta' \in \mathcal{E}_A(B)$ in a neighbourhood of θ such that $\|\theta - \theta'\|_A^2 \leqslant \min(\sigma d B/\sqrt{T}\,, 4B^2)$ and :*

$$R_T(\theta') \geqslant \min\left(\frac{\sigma d\sqrt{T}}{16}, \frac{BT}{4}\right).$$

Proof Let us set $B > 0$ and $\theta \in \mathbb{R}^{d+2}$ such that $\|\theta\|_A = B$, and fix any algorihm. We shall consider problems with gaussian rewards of variance σ^2 which are in particular subgaussian with proxy variance σ^2. For all t, define $w_t := A^{-1/2}(x_t - c)$, so that $x_t = A^{1/2}w_t + c$ and $w_t^\top w_t \leqslant 1$. The expected reward of action x_t is then $\theta^\top x_t = (A^{1/2}\theta)^\top w_t + \theta^\top c$ and the reward of the optimal action is $B + \theta^\top c$ with $B = \sqrt{\theta^\top A\theta}$. The difference between the two equals $B - (A^{1/2}\theta)^\top w_t$. We work with $(e_i)_{1\leqslant i\leqslant d+2}$ an orthonormal basis of $\mathbb{R}^{d+2}$ such that $e_{d+1} \propto A^{-1/2}\theta$ and $\mathrm{Span}(e_{d+1}, e_{d+2}) = \mathrm{Span}(A^{-1/2}\theta, A^{-1/2}c)$. We also set $C := \min({}^1/_{2\sqrt{2}}, {}^{B\sqrt{2T}}/_{\sigma d})$ and $\varepsilon^2 := \min(1, {}^{d\sigma C}/_{B\sqrt{2T}})$. Given $\xi \in \{-1,+1\}^d$ consider the family of problems defined by the mean-reward vectors

$$\theta(\xi) := A^{-\frac{1}{2}}\phi(\xi) \text{ and } \phi(\xi) := B\left[e_{d+1}\sqrt{1-\varepsilon^2} + \frac{\varepsilon}{\sqrt{d}}\sum_{i=1}^d \xi_i e_i\right].$$

Moreover, $\theta(\xi)^\top A\theta(\xi) = \phi(\xi)^\top\phi(\xi) = B^2$. By inspection, the regret at time step t satisfies

$$\begin{aligned} B - (A^{\frac{1}{2}}\theta(\xi))^\top w_t = B - \phi(\xi)^\top w_t &\overset{(1)}{=} \frac{B}{2}\left[1 - w_t^\top w_t + \left\|\frac{\phi(\xi)}{B} - w_t\right\|_2^2\right] \\ &\overset{(2)}{\geqslant} \frac{B}{2}\left\|\frac{\phi(\xi)}{B} - w_t\right\|_2^2 = \frac{B}{2}\left((w_{t,d+1} - \sqrt{1-\varepsilon^2})^2 + \sum_{i=1}^d\left(w_{t,i} - \xi_i\frac{\varepsilon}{\sqrt{d}}\right)^2 + w_{t,d+2}^2\right), \end{aligned}$$

using the facts that (1) $-a^\top b = \frac{1}{2}(\|a-b\|^2 - a^\top a - b^\top b)$ and (2) $w_t^\top w_t \leqslant 1$. Summing over t we get a lower bound for the regret:

$$\begin{aligned} R_T(\theta) = \sum_{t=1}^T\left[B - \phi(\xi)^\top w_t\right] &\geqslant \frac{B}{2}\sum_{t=1}^T\left[\left(w_{t,d+1} - \sqrt{1-\varepsilon^2}\right)^2 + \sum_{i=1}^d\left(w_{t,i} - \xi_i\frac{\varepsilon}{\sqrt{d}}\right)^2 + w_{t,d+2}^2\right] \\ &\geqslant \frac{B}{2}\sum_{i=1}^d\sum_{t=1}^T\left(w_{t,i} - \xi_i\frac{\varepsilon}{\sqrt{d}}\right)^2 \geqslant \frac{B}{2}\sum_{i=1}^d U_i(\xi_i) \qquad \text{where} \end{aligned}$$

$$\tau_i(T) := \min\left(\min\left\{t : \sum_{s=1}^t w_{s,i}^2 \geqslant \frac{d\sigma^2C^2}{2\varepsilon^2B^2}\right\}, T\right) \text{ and } U_i(\xi) := \sum_{t=1}^{\tau_i(T)}\left(w_{t,i} - \xi_i\frac{\varepsilon}{\sqrt{d}}\right)^2.$$

This is upper bounded with probability 1 by:

$$U_i(\xi) \leqslant 2 \sum_{t=1}^{\tau_i(T)} w_{ti}^2 + \frac{\varepsilon^2}{d} \leqslant 2 \left[\frac{d\sigma^2 C^2}{2\varepsilon^2 B^2} + \frac{T\varepsilon^2}{d} \right],$$

using the Cauchy-Schwarz inequality, $(a-b)^2 \leqslant 2(a^2+b^2)$ and the definition of $\tau_i(T)$. For all $i \in [d]$ define $\xi^{(-i)}$ with coordinates $\xi_j^{(-i)} := (-1)^{\mathbb{1}(i=j)} \xi_j$ and define $P_{\theta(\xi),i}$ the probability distribution of $(w_t)_{t \leqslant \tau_i(T)}$ under parameter $\theta(\xi)$. From the choice of $\theta(\xi)$ we have

$$\begin{aligned} \theta(\xi)^\top x_t - \theta(\xi^{(-i)})^\top x_t &= (\phi(\xi) - \phi(\xi^{(-i)}))^\top w_t + (\theta(\xi) - \theta(\xi^{(-i)}))^\top c \\ &= (\phi(\xi) - \phi(\xi^{(-i)}))^\top w_t + 2\frac{\varepsilon}{\sqrt{d}} e_i^\top A^{-\frac{1}{2}} c = (\phi(\xi) - \phi(\xi^{(-i)}))^\top w_t = \frac{2\varepsilon B}{\sqrt{d}} w_{t,i}, \end{aligned}$$

as $e_i^\top A^{-1/2} c = 0$ for all $i \leqslant d$. By Lattimore and Szepesvári (2020)[Exercises 14.7 and 15.8], the relative entropy between the distributions of the observations in both problems bounded by

$$\begin{aligned} D\big(P_{\theta(\xi),i} | P_{\theta(\xi^{(-i)}),i}\big) &= \frac{1}{2\sigma^2} \mathbb{E}\left(\sum_{s=1}^{\tau_i(T)} \left(\theta(\xi)^\top x_t - \theta(\xi^{(-i)})^\top x_t \right)^2 \right) \\ &= \frac{1}{2\sigma^2} \mathbb{E}\left(\sum_{s=1}^{\tau_i(T)} \left(\phi(\xi)^\top w_t - \phi(\xi^{(-i)})^\top w_t \right)^2 \right) = \mathbb{E}\left(\sum_{s=1}^{\tau_i(T)} w_{si}^2 \right) \frac{2\varepsilon^2 B^2}{d\sigma^2} \leqslant C^2 . \end{aligned}$$

Then because $U_i(\xi)$ is a function of $(w_t)_{t \leqslant \tau_i(T)}$ and is bounded by $2\big(\frac{d\sigma^2 C^2}{2\varepsilon^2 B^2} + \frac{T\varepsilon^2}{d}\big)$, we can use a total variation and Pinsker's inequalities to get that:

$$\begin{aligned} \mathbb{E}_{\theta(\xi)}(U_i(\xi)) &\geqslant \mathbb{E}_{\theta(\xi^{(-i)})}(U_i(\xi)) - 2 \left[\frac{d\sigma^2 C^2}{2\varepsilon^2 B^2} + \frac{T\varepsilon^2}{d} \right] \sqrt{\frac{1}{2} D(P_{\theta(\xi),i} | P_{\theta(\xi^{(-i)}),i})} \\ &\geqslant \mathbb{E}_{\theta(\xi^{(-i)})}(U_i(\xi)) - \sqrt{2} \left[\frac{d\sigma^2 C^2}{2\varepsilon^2 B^2} + \frac{T\varepsilon^2}{d} \right] C . \end{aligned}$$

Adding on both sides and using the definition of U:

$$\begin{aligned} \mathbb{E}_{\theta(\xi)}(U_i(\xi_i)) + \mathbb{E}_{\theta(\xi^{(-i)})}(U_i(\xi^{(-i)})) &\geqslant \mathbb{E}_{\theta(\xi)}(U_i(\xi) + U_i(\xi^{(-i)})) - \sqrt{2} \left[\frac{d\sigma^2 C^2}{2\varepsilon^2 B^2} + \frac{T\varepsilon^2}{d} \right] C \\ &= 2\mathbb{E}_{\theta(\xi)} \left(\sum_{t=1}^{\tau_i(T)} \frac{\varepsilon^2}{d} + w_{ti}^2 \right) - \sqrt{2} \left[\frac{d\sigma^2 C^2}{2\varepsilon^2 B^2} + \frac{T\varepsilon^2}{d} \right] C . \end{aligned}$$

Either $\tau_i(T) = T$ so $\sum_{t=1}^{\tau_i(T)} \frac{\varepsilon^2}{d} = \frac{T\varepsilon^2}{d}$, or $\tau_i(T) < T$ and then $\sum_{t=1}^{\tau_i(T)} w_{ti}^2 \geqslant \frac{d\sigma^2 C^2}{2\varepsilon^2 B^2}$. Therefore

$$\mathbb{E}_{\theta(\xi)}(U_i(\xi)) + \mathbb{E}_{\theta(\xi^{(-i)})}(U_i(\xi^{(-i)})) \geqslant \min \left(2T\frac{\varepsilon^2}{d}, \frac{2d\sigma^2 C^2}{2\varepsilon^2 B^2} \right) - \sqrt{2} \left[\frac{d\sigma^2 C^2}{2\varepsilon^2 B^2} + \frac{T\varepsilon^2}{d} \right] C .$$

If $d < 4B\sqrt{T}/\sigma$, then $C = 2\sqrt{2}/2$ and $\varepsilon^2 = d\sigma/(4B\sqrt{T}) < 1$ so

$$\mathbb{E}_{\theta(\xi)}(U_i(\xi)) + \mathbb{E}_{\theta(\xi^{-i})}(U_i(\xi^{(-i)})) \geqslant \frac{\sigma\sqrt{T}}{4B} .$$

If $d \geqslant 4B\sqrt{T}/\sigma$, then $C = \sqrt{2T}B/(\sigma d)$ and $\varepsilon^2 = 1$. So $2T/d \leqslant d\sigma^2 C^2/B^2$ and

$$\mathbb{E}_{\theta(\xi)}(U_i(\xi)) + \mathbb{E}_{\theta(\xi^{(-i)})}(U_i(\xi^{(-i)})) \geqslant \frac{2T}{d} - \sqrt{2}\frac{d\sigma^2 C^3}{2B^2} \geqslant \frac{2T}{d} - \sqrt{2}\frac{(2T)^{3/2}B}{2\sigma d^2} \geqslant \frac{T}{d}\,.$$

Putting both cases together and summing the regret over ξ:

$$\begin{aligned}\sum_{\xi \in \{-1,+1\}^d} \mathbb{E}_{\theta(\xi)}(R_T(\theta)) &\geqslant \frac{B}{2} \sum_{\xi \in \{-1,+1\}^d} \sum_{i=1}^{d} \mathbb{E}_{\theta(\xi)}(U_i(\xi)) \\ &\geqslant \frac{B}{4} \sum_{i=1}^{d} \sum_{\xi \in \{-1,+1\}^d} \mathbb{E}_{\theta(\xi)}(U_i(\xi)) + \mathbb{E}_{\theta(\xi^{(-i)})}(U_i(\xi^{(-i)})) \geqslant 2^d \min\left(\frac{\sigma d\sqrt{T}}{16}, \frac{TB}{4}\right).\end{aligned}$$

We conclude the proof by saying that there must exist at least a ξ such that

$$\mathbb{E}_{\theta(\xi)}(R_T(\theta)) \geqslant \min\left(\frac{\sigma d\sqrt{T}}{16}, \frac{TB}{4}\right) \qquad \blacksquare$$

To complete the picture, we also show that there is a minimal regret of $\Omega(d\|\theta\|_A)$ as soon as $T \geqslant d$. While this seems intuitively straightforward, as the learner should have no choice but to explore all dimensions at least once, the proof (in Appendix C) requires some technical work.

Proposition 2 *For any algorithm, for any $B > 0$, if $T \geqslant d$, there exists a noiseless bandit problem, with mean-rewards parameter $\theta \in \mathbb{R}^d$, such that $\|\theta\|_A \leqslant 4B$ and*

$$R_T(\theta) \geqslant 0.017\, d\|\theta\|_A\,.$$

3. The `E2TC` algorithm

We now describe the `E2TC` algorithm and prove its optimality. We first describe the algorithm and bound its regret when $\mathcal{X}$ is a centered ellipsoid i.e., $c = 0$, and then extend our results to $c \neq 0$ using a reduction. To ease notation, we define $\overline{\log}(x) := 1 + \log(\max(x, 1))$, $(a)^- := \min(a, 0)$, and given a set of instants $\mathcal{T} \subset \mathbb{N}$, we define vectors $Y_\mathcal{T} := (y_t)_{t\in\mathcal{T}} \in \mathbb{R}^{|\mathcal{T}|}$ and $Z_\mathcal{T} := (z_t)_{t\in\mathcal{T}} \in \mathbb{R}^{|\mathcal{T}|}$ and matrix $X_\mathcal{T} := (x_t)_{t\in\mathcal{T}} \in \mathbb{R}^{|\mathcal{T}|\times d}$ so that $Y_\mathcal{T} = X_\mathcal{T}\theta + Z_\mathcal{T}$. When $X_\mathcal{T}$ has full rank we define $\mathrm{LS}(X_\mathcal{T}, Y_\mathcal{T}) = (X_\mathcal{T}^\top X_\mathcal{T})^{-1} X_\mathcal{T}^\top Y_\mathcal{T}$ the (unregularized) ordinary least-squares estimate of θ when using only observations at times $t \in \mathcal{T}$. For $i \in \mathbb{N}$ we define $n_i = d2^{i-1}$, $T_i = d(2^i - 1)$ and $\delta_i = \min(dn_i/T, 1)$.

3.1. Algorithm description

Consider the centered case $c = 0$. The pseudocode of `E2TC` is presented as Algorithm 1 and is composed of three phases, whose names are inspired by the "explore-then-commit" strategy of Garivier et al. (2016). The main addition is the first phase that allows to determine how many exploration rounds are necessary, as those must depend on $\|\theta\|_A$, as shown by our analysis.

Phase 1: ("warmup phase", lines 1-5): We sample actions in $(A^{1/2}e_j)_{j\in[d]}$ in a round-robin fashion. At times $t \in (T_i)_{i\in\mathbb{N}}$ with we compute $\hat{\theta}_i := \mathrm{LS}(X_i, Y_i)$ the least-squares estimate of θ

using the observations during time interval $[T_i, T_{i+1}-1]$ and if $\|\hat{\theta}_i\|_A$ exceeds a threshold equal to $\alpha U(\delta_i, n_i)$ where $\alpha \in \mathbb{R}^+$ is a positive parameter of the algorithm and with

$$U(\delta_i, n_i)^2 = \frac{\sigma^2 d^2}{n_i}\left(1 + 2\sqrt{\frac{1}{d}\log\frac{1}{\delta_i}} + \frac{2}{d}\log\frac{1}{\delta_i}\right). \tag{1}$$

We exit this phase that we number $\hat{\imath}$, otherwise we keep going. When exiting, we compute $\hat{B} := \|\hat{\theta}_{\hat{\imath}}\|_A$ the norm estimate. The threshold U is designed to estimate the magnitude of θ correctly, so that with high probability $c_1\|\theta\|_A \leqslant \hat{B} \leqslant c_2\|\theta\|_A$, with c_1, c_2 well-chosen universal constants.

Phase 2: ("exploration phase", lines 6-9): We sample actions in $(A^{1/2}e_j)_{j\in[d]}$ in a round-robin fashion during an amount of time $N_e := d\sigma\lceil\sqrt{T}/\hat{B}\rceil$. The goal of this phase is simply to gather enough samples to estimate θ accurately, to be able to find a good action by the end of the phase with high probability. The analysis shows that the number of samples needs to depend on $\|\theta\|_A$, which is why we use the estimate $\hat{B}$ from the previous phase.

Phase 3: ("commit phase", lines 10-12): We compute $\hat{\theta} := \mathrm{LS}(X, Y)$ the least-squares estimate of θ using the past observations gathered during phase 2, and play greedily by selecting decision $\arg\max_{x\in\mathcal{X}} x^\top\hat{\theta}$ until the time horizon runs out. The goal of this phase is simply to maximize reward, as we are certain that enough samples have been gathered and this decision should be close to the optimal one $x^\star(\theta)$ with high probability.

Algorithm 1: `E2TC (Explore-Explore-Then-Commit)`

Input : Ellipsoid matrix A, noise level σ, time horizon T, parameter $\alpha \geqslant 0$.
Init: $i = 0$.

1 **while** *warm-up* **do**
2   **for** $t = T_i$ *to* $T_{i+1}-1$ **do**
3   Play $x_t = A^{1/2}e_{(t \bmod d)+1}$. Receive y_t.
4   Set $X_i = x_{[T_i, T_{i+1}-1]}$ and $Y_i = y_{[T_i, T_{i+1}-1]}$. Set $\hat{\theta}_i = \mathrm{LS}(X_i, Y_i)$. Update $i \leftarrow i+1$.
5   Exit warm-up if $\|\hat{\theta}_i\|_A > \alpha U(\delta_i, n_i)$, defined in (1)
6 Set $\hat{B} = \|\hat{\theta}_i\|_A$ and $N_e = d\sigma\lceil\sqrt{T}/\hat{B}\rceil$
7 **for** $t = T_i$ *to* $T_i + N_e - 1$ **do**
8   Play $x_t = A^{1/2}e_{(t \bmod d)+1}$. Receive y_t.
9 Set $X = x_{[T_i, T_i+N_e-1]}$ and $Y = y_{[T_i, T_i+N_e-1]}$.
10 Set $\hat{\theta} = LS(X, Y)$.
11 **for** $t = T_i + N_e(\hat{B})$ *to* T **do**
12   Play $x_t = x^\star(\hat{\theta})$. Receive y_t.

Remark 3 (Computational complexity) *A run of* `E2TC` *takes time* $O(dT + d^2\log(T/d) + d^3)$ *and memory* $O(d^2)$*, since it involves at most* $\log_2(T/d) + 1$ *least-squares estimates of* θ*, and all those estimations require inverting* $d \times d$ *matrices which are all proportional to* A*, so it suffices to invert* A *once.*

3.2. Regret upper bound

We now upper bound the regret of E2TC, in the centered case $c = 0$. We present the main elements of analysis here, and present complete proofs in appendix.

Regret of phases 2 and 3 Assume that phase 1 was successful. Denote by $\mathbb{E}_{\text{expl}}$ the expectation conditionally on all rounds before phase 3. The regret caused by a round of phase 2 is at most $2\|\theta\|_A$ by the Cauchy-Schwarz inequality. The regret caused by a round of phase 3 is $\theta(x^\star(\theta) - x^\star(\hat{\theta}))$ we always select the same action. The regret caused by phases 2 and 3 is hence upper bounded by

$$2N_e\|\theta\|_A + T\,\mathbb{E}_{\text{expl}}\left[\theta^\top\left(x^\star(\theta) - x^\star(\hat{\theta})\right)\right].$$

We note that, during phases 1 and 2, we play uniformly along the axes of the ellipsoid, which ensures that the design matrix $X^\top X$ stays proportional to A (after a number of exploration rounds multiple of d, which is always the case at the end of exploration in our algorithm).

As observed by Rusmevichientong and Tsitsiklis (2010), when the action set has positive curvature, the commit error scales *quadratically* with the mean-squared error of estimation (see Lemma 10 in Appendix A), which itself scales with $1/N_e$; this is crucial to ensure that the regret grows proportionally to $\sqrt{T}$.

Lemma 4 *Let $\hat{\theta}$ the least-squares estimator after N_e rounds of exploration, then conditionally on the exploration rounds,*

$$\mathbb{E}_{\text{expl}}\left[\theta^\top\left(x^\star(\theta) - x^\star(\hat{\theta})\right)\right] \leqslant \frac{d^2\sigma^2}{\|\theta\|_A N_e} + 2\|\theta\|_A e^{(2d/3 - N_e\|\theta\|^2/(3\sigma^2 d))^-}.$$

See Appendix A.1 for a proof. The regret incurred by phases 2 and 3 is hence upper bounded as

$$2N_e\|\theta\|_A + T\frac{d^2\sigma^2}{\|\theta\|_A N_e} + T\left(\frac{d^2\sigma^2}{\|\theta\|_A N_e} + 2\|\theta\|_A e^{(2d/3 - N_e\|\theta\|^2/(3\sigma^2 d))^-}\right).$$

Our analysis therefore suggests selecting $N_e \approx \sigma d\sqrt{T}/\|\theta\|_A$ to balance the first two terms in the right-hand side. Doing this would yield a regret upper bound of $O(d\sqrt{T})$ which is the correct minimax rate. Of course, since $\|\theta\|_A$ is initially unknown, we must estimate it during phase 1.

Regret of phase 1 Lemma 5 shows that the regret caused by phase 1 is small enough to ensure minimax optimality, and that the value of $\hat{B}$ output at the end of phase 1 is a good enough estimate of $\|\theta\|_A$ with high probability. This lemma also justifies the choice of the stopping criterion for phase 1.

Lemma 5 (Regret of Phase 1) *Set $\alpha = 3$ and let $\hat{B}$ be the output of phase 1. We have*

$$\mathbb{P}\left[\hat{B} \notin \left[(1/2)\|\theta\|_A, (3/2)\|\theta\|_A\right]\right] \leqslant \frac{164\,\sigma^2 d^2}{T\|\theta\|_A^2}\overline{\log}\left(\frac{T\|\theta\|_A^2}{d^2\sigma^2}\right) + 48\frac{d}{T},$$

and the expected length of phase 1 is at most

$$\mathbb{E}[N_{\text{warm-up}}] \leqslant \frac{164\,\sigma^2 d^2}{\|\theta\|_A^2}\overline{\log}\left(\frac{T\|\theta\|_A^2}{\sigma^2 d^2}\right) + 48d.$$

On the design of phase 1 The threshold $U(\delta_i, n_i)$ is chosen such that the probability of stopping too early at time T_i is less than δ_i. So $U(\delta_i, n_i)$ and δ_i can be interpreted as an upper confidence bound, and a failure probability respectively. The exploration design is chosen so that $X_i^\top X_i = (n_i/d)A$ after a phase of length n_i, which guarantees at the end of subphase i of the warm-up, the probability of failure is less than δ_i:

$$\mathbb{P}\big[\|\hat{\theta}_i - \theta\|_A^2 \geqslant U(n_i, \delta_i)\|\big] \leqslant \delta_i \,.$$

Interestingly, our analysis shows that δ_i must (counterintuitively) be an increasing function of i to yield a minimax optimal algorithm. The cornerstone of both the design of phase 1 and its analysis is a concentration inequality for least-squares regression, by Hsu et al. (2011), which is an extension of the tail bounds for chi-squared variables by Laurent and Massart (2000) to sub-gaussian noise.

Lemma 6 (Concentration of error in linear regression) *In linear regression with $Y = X\theta + Z$ and $\hat{\theta} = \mathrm{LS}(X, Y)$, assume that all entries of Z are σ^2-subgaussian and independent, and that X is full rank. Then for all $x \geqslant 0$*

$$\mathbb{P}\big[\|\hat{\theta} - \theta\|_{X^\top X}^2 \geqslant \sigma^2(d + 2\sqrt{dx} + 2x)\big] \leqslant e^{-x} \,.$$

Putting the three phases together, we can now state the regret upper bound of E2TCin Theorem 7, and the detailed proof is presented in appendix. Corollary 8 shows that the regret upper bound matches our regret lower bound of Theorem 1 up to a universal multiplicative constant, so that E2TC is locally asymptotically minimax optimal as announced.

Theorem 7 *Consider $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_{A^{-1}} \leqslant 1\}$ a centered ellipsoid. The regret of E2TC tuned with $\alpha = 3$ admits the upper bound for any $\theta \in \mathbb{R}^d$ and any $T \geqslant 0$:*

$$R_T(\theta) \leqslant 6d\sigma\sqrt{T} + \frac{984\,\sigma^2 d^2}{\|\theta\|_A}\overline{\log}\left(\frac{T\|\theta\|_A^2}{\sigma^2 d^2}\right) + 290\,d\|\theta\|_A + 2T\|\theta\|_A e^{(2d/3-(2/9)\sqrt{T}\|\theta\|_A/\sigma)^-} \,.$$

Corollary 8 *E2TC is locally asymptotically minimax optimal: for any $\theta \in \mathbb{R}^d$,*

$$R_T(\theta) = \mathcal{O}\Big(\min\big(\sigma d\sqrt{T} + d\|\theta\|_A, T\|\theta\|_A\big)\Big) \,.$$

Proof of Corollary 8 We proceed by considering multiple cases, depending on the value of T. If $T \leqslant 36d^2\sigma^2/\|\theta\|_A^2$, then $2T\|\theta\|_A \leqslant 12\sigma d\sqrt{T}$. So by the Cauchy-Schwarz bound, the regret is less than $R_T(\theta) \leqslant \min\big(2T\|\theta\|_A, 12\sigma d\sqrt{T}\big)$. Otherwise,

$$2d/3 - (2/9)\sqrt{T}\|\theta\|_A/\sigma \leqslant -(1/9)\sqrt{T}\|\theta\|_A/\sigma \,,$$

and the final term in the regret bound is upper bounded by

$$\begin{aligned} T\|\theta\|_A e^{(2d/3-(2/9)\sqrt{T}\|\theta\|_A/\sigma)^-} &\leqslant T\|\theta\|_A e^{-(1/9)\sqrt{T}\|\theta\|_A/\sigma} = \sigma\sqrt{T}\left(\frac{\sqrt{T}\|\theta\|_A}{\sigma}e^{-(1/9)\sqrt{T}\|\theta\|_A/\sigma}\right) \\ &\leqslant 9\sigma\sqrt{T}\sup_{X \geqslant 2d/3}\big(Xe^{-X}\big) \leqslant 6e^{-2d/3}\sigma d\sqrt{T} \,. \end{aligned}$$

Similarly, for the first logarithmic term,

$$\frac{\sigma^2 d^2}{\|\theta\|_A}\overline{\log}\left(\frac{T\|\theta\|_A^2}{\sigma^2 d^2}\right) \leqslant \sigma d\sqrt{T}\frac{\sigma d}{\|\theta\|_A\sqrt{T}}\overline{\log}\left(\frac{T\|\theta\|_A^2}{\sigma^2 d^2}\right) \leqslant \sigma d\sqrt{T}\sup_{X \geqslant 1/6}\frac{1}{X}\overline{\log}(X^2) = 6\sigma d\sqrt{T} \,.$$

Therefore, the total bound is at most $O(d\sigma\sqrt{T} + d\|\theta\|)$ as announced. ■

3.3. Non-centered Action Sets

Now consider the general, non-centered case $c \neq 0$. In Appendix B, we describe a general procedure to extend E2TC (and any other algorithm with similar structure) designed for the centered case to the non-centered case. Theorem 9 shows that E2TC is also locally asymptotically minimax optimal in the non-centered case; it is a direct consequence of Theorem 13 in Appendix B

Theorem 9 *Consider $\mathcal{X} = \{x \in \mathbb{R}^d : \|x - c\|_{A^{-1}} \leqslant 1\}$ an arbitrary ellipsoid. Set $\alpha = 3$. The regret of E2TC with Reduction 2 admits the upper bound for any $\theta \in \mathbb{R}^d$ and any $T \geqslant 0$:*

$$R_T(\theta) \leqslant 7d\sigma\sqrt{T} + \frac{2622\,\sigma^2 d^2}{\|\theta\|_A}\overline{\log}\left(\frac{T\|\theta\|_A^2}{4\sigma^2 d^2}\right) + 2T\|\theta\|_A e^{(2d/3-(1/9)\sqrt{T}\|\theta\|_A/\sigma)^-} + 392\,d\|\theta\|_A\,.$$

4. Numerical Experiments

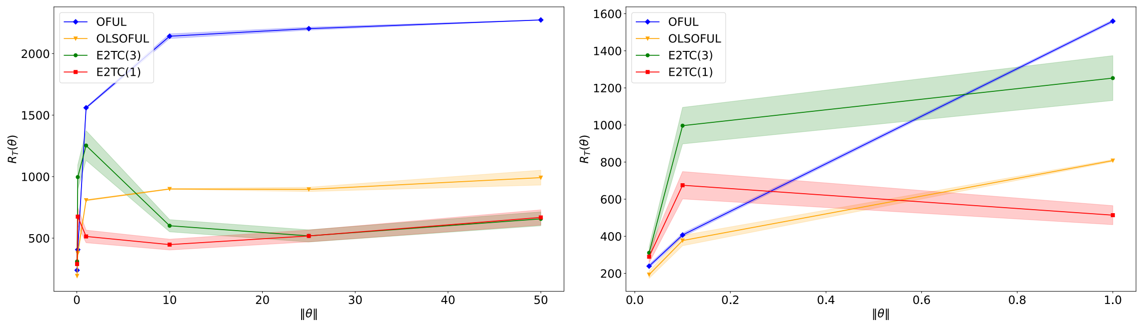
We now compare E2TC to state-of-the-art algorithms OFUL and OLSOFUL using numerical experiments. Regret of algorithms is averaged over 20 independent runs, and both averages and confidence intervals are presented. The source code for those numerical experiments will be released publicly after review. We denote by E2TC(α) the E2TC algorithm with parameter α, and we consider both E2TC(1) and E2TC(3). We consider action set $\mathcal{X}$ as the unit ball and a time horizon of $T = 10^4$. The confidence ellipsoids in OFUL and OLSOFUL require a confidence parameter δ, which can be interpreted as a probability of failure, and we set it to $\delta = 1/T$, to ensure that the regret caused by failure does not scale linearly in T.

We use a commercially available optimization solver Gurobi Optimization, LLC (2024) to handle the bilinear maximization problem $\max_{(x,\theta)} x^\top\theta$ subject to $(x,\theta) \in \mathcal{X} \times \mathcal{C}_t$ with $\mathcal{C}_t$ the confidence ellipsoid at time t, which must be solved at each step to implement OFUL and OLSOFUL. We limit Gurobi's available time to 1 second per time step, which already amounts to 2.7 hours per run, that is 2.5 days to average the regret over 20 independent runs. Gurobi is allowed to use all 20 cores of an i9-129000H processor, and outputs the best solution found within the allotted time, with an early stop if the estimated relative gap to optimality is less than 1%.

Regret $R_T(\theta)$ vs norm of θ We compare in Figure 2 the final regret of the algorithms for different values of $\|\theta\|_2$ chosen as $\|\theta\|_2 \in \{d/\sqrt{T}, 0.1, 1, 10, 25, 50\}$. OFUL requires an input parameter S which should be an upper bound for $\|\theta\|$. We set S as half the range norms so that $S = 25$. Due to the enormous computation complexity of optimistic algorithms, we limit ourselves to a small dimension $d = 3$: a larger dimension would already require over a week of computation to generate meaningful results.

Regret $R_T(\theta)$ vs dimension d We compare in Figure 3 the final regret of the algorithms for different values of d, and we choose θ with a fixed norm $\|\theta\|_2 = 10$. We consider dimensions $2 \leqslant d \leqslant 90$. We could not make the optimistic algorithms work in reasonable time for $d > 4$, so for high dimensions, solely E2TC is presented. As suggested by our theoretical results, the regret of E2TC seems to scale linearly with the dimension.

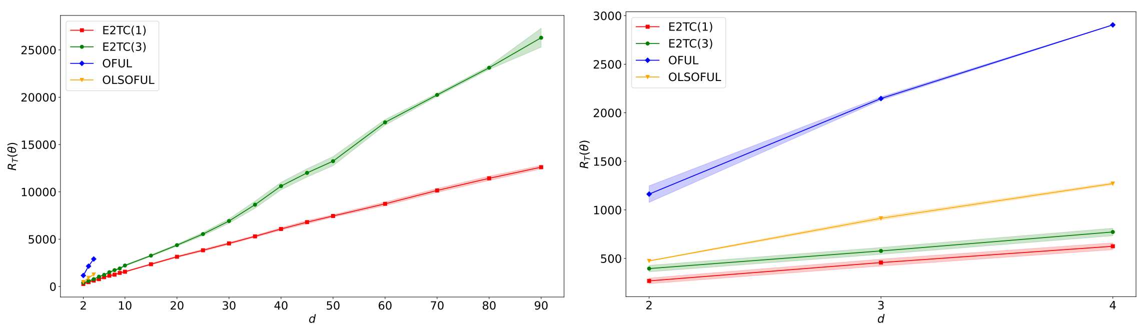
Runtime We compare the runtime (CPU time) of E2TC and OFUL for different norms $\|\theta\|$, where dimension is $d = 3$ in Fig.4(a) and for different dimensions d and fixed norm 10 in Fig.4(b). Runtime includes the time to compute the action to be played, the time used to solve the bilinear optimization problem, and the time to update the estimated parameters $\hat{\theta}_t$ using the Sherman Morrison algorithm for the inversion of the design matrix. We present the ratio between the runtime of OFUL and that of E2TC , and observe that OFUL is slower by about 5 orders of magnitude, an enormous



(a) Large norms (b) Small norms

Figure 2: Comparison of the final regret of E2TC, OFUL and OLSOFUL for different norms.

difference. This clearly illustrates the lower computational complexity of E2TC with respect to optimistic algorithms. Also, for small norms in dimension 3 and for dimension greater than 4, it is harder for Gurobi to find a good solution, and it maxes out its allotted time.



(a) Large dimensions (b) Small dimensions

Figure 3: Final regret of E2TC, OFUL and OLSOFUL as a function of the dimension d.

$\|\theta\|$	0.1	1	10	50
Runtime Ratio	6.1×10^4	1.0×10^5	1.7×10^5	8.7×10^4

(a) Norm runtime comparison

d	2	3	4
Runtime Ratio	1.2×10^4	1.7×10^5	6.7×10^5

(b) Dimension runtime comparison

Figure 4: Runtime comparison of E2TC and OFUL for different norms and dimensions.

5. Conclusion

We have proposed the first known provably minimax optimal algorithm for linear stochastic bandits over ellipsoids called E2TC. We derived a new information-theoretic lower bound on the regret of any algorithm equal to $\Omega(\min(d\sigma\sqrt{T}, \|\theta\|_A T))$, and showed that the regret of E2TC matches this bound so that it is minimax optimal. The E2TC algorithm is locally asymptotically minimax optimal, a stronger optimality notion than classical minimax optimality. Also E2TC is highly computationally efficient, requiring time $O(dT + d^2 \log(T/d) + d^3)$ and memory $O(d^2)$, unlike optimistic algorithms for which, to the best of our knowledge, no polynomial time implementation exists, even when the action set is a unit ball. Numerical experiments confirm our theoretical findings both in terms of regret as well as running time. An interesting direction for future research is whether or not the general structure of E2TC can be adapted to solve other structured bandit problems.

Acknowledgements

This paper is supported by the CHIST-ERA Wireless AI 2022 call MLDR project (ANR-23-CHR4-0005), partially funded by AEI and NCN under projects PCI2023-145958-2 and 2023/05/Y/ST7/00004, respectively.

References

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

Marc Abeille and Alessandro Lazaric. Linear Thompson Sampling Revisited. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 176–184. PMLR, 20–22 Apr 2017.

Debangshu Banerjee, Avishek Ghosh, Sayak Ray Chowdhury, and Aditya Gopalan. Exploration in linear bandits with rich action sets and its implications for inference. *arXiv preprint arXiv:2207.11597*, 2022.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic Linear Optimization under Bandit Feedback. In *Conference on Learning Theory*, 2008.

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

Aurelien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2024. URL https://www.gurobi.com.

Jaroslav Hájek. Local asymptotic minimax and admissibility in estimation. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 175–194, 1972.

Peter D Hoff. *A first course in Bayesian statistical methods*, volume 580. Springer, 2009.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors, October 2011. arXiv:1110.2842.

Kwang-Sung Jun and Jungtaek Kim. Noise-adaptive confidence sets for linear bandits and application to bayesian optimization. *arXiv preprint arXiv:2402.07341*, 2024.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 1 edition, July 2020. ISBN 978-1-108-57140-1 978-1-108-48682-8. doi: 10.1017/9781108571401.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000. doi: 10.1214/aos/1015957395.

Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010. ISSN 0364765X, 15265471.

Yinglun Zhu and Robert Nowak. Pareto optimal model selection in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 6793–6813. PMLR, 2022.

Appendix A. Proofs for Section 3

A.1. Proof of Theorem 7

Proof We use the natural template for the analysis of Explore-Then-Commit (ETC) algorithms. The first step is to use the Cauchy-Schwarz inequality to bound the instantaneous regret $\theta^\top(x^\star(\theta)-x_t) \leqslant 2\|\theta\|_A$ on all the exploration rounds:

$$R_T(\theta) \leqslant 2\|\theta\|_A \mathbb{E}[N_{\text{warm-up}} + N_e(\hat{B})] + T\,\mathbb{E}\big[\theta^\top(x^\star(\theta) - x^\star(\hat{\theta}))\big]\,.$$

We shall then control separately the expected total number of exploration rounds,

$$\mathbb{E}[N_{\text{warm-up}} + N_e(\hat{B})] \leqslant \frac{2d\sigma\sqrt{T}}{\|\theta\|_A} + \frac{328\,\sigma^2 d^2}{\|\theta\|_A^2}\overline{\log}\left(\frac{T\|\theta\|_A^2}{\sigma^2 d^2}\right) + 97d\,, \tag{2}$$

and the expected cost of the commit rounds,

$$\mathbb{E}\big[\theta^\top(x^\star(\theta)-x^\star(\hat{\theta}))\big] \leqslant \frac{3d\sigma}{2\sqrt{T}} + 2\|\theta\|_A e^{(2d/3-(2/9)\sqrt{T}\|\theta\|_A/\sigma)^-} + \frac{328\,\sigma^2 d^2}{T\|\theta\|_A}\overline{\log}\left(\frac{T\|\theta\|_A^2}{\sigma^2 d^2}\right) + 96\frac{d\|\theta\|_A}{T}\,. \tag{3}$$

The three bounds combined yield the claimed result. The first step in the proof of both of these bounds is to control the warm-up behavior. This is done via Lemma 5 stated above and proved below. It then suffices to analyze the algorithm on the event that the warm-up was successful. We thus define the corresponding good event $\mathcal{G}$ as

$$\mathcal{G} = \left\{ \frac{\|\theta\|_A}{2} \leqslant \hat{B} \leqslant \frac{3\|\theta\|_A}{2} \right\}.$$

Exploration rounds On the good event $\mathcal{G}$, the estimate $\hat{B}$ is at least $\|\theta\|/2$, so

$$N_e(\hat{B})\mathbb{1}\{\mathcal{G}\} = d\left\lceil \frac{\sigma\sqrt{T}}{\hat{B}} \right\rceil \mathbb{1}\{\mathcal{G}\} \leqslant d\left\lceil \frac{2\sigma\sqrt{T}}{\|\theta\|_A} \right\rceil \leqslant d\frac{2\sigma\sqrt{T}}{\|\theta\|_A} + d\,.$$

And therefore, by Lemma 5, in expectation, bounding $N_e(\hat{B})$ by the worst-case T on $\bar{\mathcal{G}}$,

$$\mathbb{E}[N_e(\hat{B})] \leqslant d\frac{2\sigma\sqrt{T}}{\|\theta\|_A} + d + T\mathbb{P}\big[\,\overline{\mathcal{G}}\,\big] \leqslant d\frac{2\sigma\sqrt{T}}{\|\theta\|_A} + d + \frac{164\,\sigma^2 d^2}{\|\theta\|_A^2}\overline{\log}\left(\frac{T\|\theta\|_A^2}{\sigma^2 d^2}\right) + 48d\,.$$

Commit error We isolate the guarantees for the least-squares estimator in Lemma 4, which we restate and prove below. Given this lemma, by the tower rule, and since $\mathcal{G}$ only depends on the exploration rounds (thus is measurable with respect to the σ-algebra generated by actions and observations in the exploration rounds),

$$\begin{aligned}
&\mathbb{E}[\theta^\top(x^\star(\theta) - x^\star(u))\mathbb{1}\{\mathcal{G}\}] = \mathbb{E}[\mathbb{E}_{\text{expl}}[\theta^\top(x^\star(\theta) - x^\star(u))]\mathbb{1}\{\mathcal{G}\}] \\
&\leqslant \mathbb{E}\left[\left(\frac{d^2\sigma^2}{\|\theta\|_A N_e} + 2\|\theta\|_A e^{(2d/3 - N_e\|\theta\|^2/(3\sigma^2 d))^-}\right)\mathbb{1}\{\mathcal{G}\}\right] \leqslant \frac{3d\sigma}{2\sqrt{T}} + 2\|\theta\|_A e^{(2d/3-(2/9)\sqrt{T}\|\theta\|_A/\sigma)^-}\,,
\end{aligned}$$

where we used the fact that $\hat{B} \leqslant 3\|\theta\|_A/2$ on the event $\mathcal{G}$, and thus that $N_e \geqslant d\sigma\sqrt{T}/(3\|\theta\|_A/2)$

$$\mathbb{E}[\theta^\top(x^\star(\theta) - x^\star(u))] \leqslant \mathbb{E}[\theta^\top(x^\star(\theta) - x^\star(u))\mathbb{1}\{\mathcal{G}\}] + 2\|\theta\|_A\mathbb{P}\big[\,\overline{\mathcal{G}}\,\big]\,.$$

The claimed bound follows by applying the previously shown inequalities. ∎

Lemma 4 *If $\hat{\theta}$ is the least-squares estimator after N_{e} rounds of exploration, then conditionally on the exploration rounds,*

$$\mathbb{E}_{\text{expl}}\big[\theta^\top\big(x^\star(\theta) - x^\star(\hat{\theta})\big)\big] \leqslant \frac{d^2\sigma^2}{\|\theta\|_A N_e} + 2\|\theta\|_A e^{(2d/3 - N_{\mathrm{e}}\|\theta\|^2/(3\sigma^2 d))^-} .$$

Proof To ease notation, we remove the subscript on $\mathbb{E}_{\text{expl}}$ inside this proof; all expectations and probability are taken conditionally on the randomness of the exploration rounds. At the end of the exploration phase, consider the matrices $Y \in \mathbb{R}^{N_{\mathrm{e}}}$, and $X \in \mathbb{R}^{N_{\mathrm{e}}\times d}$, defined by stacking the exploratory actions and responses so that

$$Y = X\theta + Z \quad \text{and} \quad \hat{\theta} = (X^\top X)^{-1} X^\top Y .$$

Let us start by evacuating very bad cases, in which the estimator is completely off. Depending on whether $\theta^\top A\hat{\theta} \leqslant 0$, we upper bound the instantaneous regret by either $2\|\theta\|_A$, or using Lemma 10,

$$\begin{aligned}\mathbb{E}\big[\theta^\top\big(x^\star(\theta) - x^\star(\hat{\theta})\big)\big] &\leqslant 2\|\theta\|_A \mathbb{P}\big[\theta^\top A\hat{\theta} \leqslant 0\big] + \frac{1}{\|\theta\|_A}\mathbb{E}\big[\|\theta - \hat{\theta}\|_A^2\big] \\ &\leqslant 2\|\theta\|_A \mathbb{P}[\|\hat{\theta} - \theta\|_A \geqslant \|\theta\|_A] + \frac{1}{\|\theta\|_A}\mathbb{E}\big[\|\theta - \hat{\theta}\|_A^2\big] . \quad (4)\end{aligned}$$

The second term is essentially the mean-squared error of the least-squares estimator

$$\mathbb{E}[\|\theta - \hat{\theta}\|_A^2] = \mathbb{E}[\|(X^\top X)^{-1}X^\top Z\|_A^2] = \mathbb{E}[\|A^{1/2}(X^\top X)^{-1}X^\top Z\|^2]$$

Now note that for any matrix $M \in \mathbb{R}^{N_{\mathrm{e}}\times d}$, since the covariance matrix of Z is diagonal with entries in $[0, \sigma^2]$, we have

$$\mathbb{E}[\|MZ\|^2] = \mathbb{E}[\mathrm{Tr}(ZZ^\top M^\top M)] = \sum_{t\in\mathcal{T}_{\text{expl}}} \mathbb{E}[z_t^2](M^\top M)_{t,t} \leqslant \sigma^2\,\mathrm{Tr}(M^\top M) .$$

In particular, taking $M = A^{1/2}(X^\top X)^{-1}X^\top$,

$$\mathbb{E}[\|A(X^\top X)^{-1}X^\top Z\|^2] \leqslant \sigma^2\,\mathrm{Tr}\left(A(X^\top X)^{-1}\right) = \frac{d^2\sigma^2}{N_e} ,$$

where we used the fact that $X^\top X = (N_e/d)A$ to obtain the final equality.

To treat the first term in (4), we use again the sub-gaussian concentration, Lemma 6, to obtain

$$\mathbb{P}[\|\hat{\theta} - \theta\|_A \geqslant \|\theta\|_A] \leqslant e^{2d/3 - N_{\mathrm{e}}\|\theta\|^2/(3\sigma^2 d)} . \quad (5)$$

To see this, note that for any $x \geqslant 0$, applying the inequality $2\sqrt{xd} \leqslant x + d$, we have

$$\frac{\sigma^2 d}{N_e}\big(d + 2\sqrt{xd} + 2x\big) \leqslant \frac{\sigma^2 d}{N_e}\big(2d + 3x\big) .$$

Then consider $x = N_{\mathrm{e}}\|\theta\|_A^2/(3d\sigma^2) - 2d/3$, so that $\frac{\sigma^2 d}{N_e}\big(2d + 3x\big) = \|\theta\|_A^2$. If $x \leqslant 0$, then the claim (5) holds trivially as a probability is less than 1. Otherwise, we apply subgaussian concentration (Lemma 6) to get

$$\begin{aligned}\mathbb{P}\big[\|\hat{\theta} - \theta\|_A^2 \geqslant \|\theta\|_A^2\big] &= \mathbb{P}\big[\|\hat{\theta} - \theta\|_{X^\top X}^2 \geqslant (\sigma^2 d/N_e)\big(2d + 3x\big)\big] \\ &\leqslant \mathbb{P}\big[\|\hat{\theta} - \theta\|_{X^\top X}^2 \geqslant (\sigma^2 d^2/N_e)\big(d + 2\sqrt{dx} + 2x\big)\big] \\ &\leqslant e^{-x} = e^{2d/3 - N_{\mathrm{e}}\|\theta\|^2/(3\sigma^2 d)} .\end{aligned}$$

■

Lemma 5 *Let $\hat{B}$ be the output of the warm-up, then,*

$$\mathbb{P}\Big[\hat{B} \notin \big[(1/2)\|\theta\|_A, (3/2)\|\theta\|_A\big]\Big] \leqslant \frac{164\,\sigma^2 d^2}{T\|\theta\|_A^2}\overline{\log}\bigg(\frac{T\|\theta\|_A^2}{d^2\sigma^2}\bigg) + 48\frac{d}{T}\,,$$

and the expected number of time-steps until the end the warm-up is

$$\mathbb{E}[N_{\mathrm{warm-up}}] \leqslant \frac{164\,\sigma^2 d^2}{\|\theta\|_A^2}\overline{\log}\bigg(\frac{T\|\theta\|_A^2}{\sigma^2 d^2}\bigg) + 48d\,.$$

Proof Define the threshold index

$$i^\star = \min\big\{i:\ U^2(\delta_i, n_i) \leqslant \|\theta\|_A^2\big\}\,,$$

around which the warm-up should terminate. We first prove an upper bound on the length of the phase at that threshold index,

$$2^{i^\star} \leqslant \frac{(2+\sqrt{2})\sigma^2 d}{\|\theta\|_A^2}\,\overline{\log}\bigg(\frac{T\|\theta\|_A^2}{\sigma^2 d^2}\bigg) + 1\,. \tag{6}$$

Proof of (6) This essentially consists in inverting the formula of the definition of U. To this end define,

$$\bar{n} = \frac{(1+\sqrt{2}/2)\sigma^2 d^2}{\|\theta\|^2}\overline{\log}\bigg(\frac{T\|\theta\|^2}{\sigma^2 d^2}\bigg)\,.$$

Define $\log_+(x) = \log(\max(x,1))$. If $n_i \geqslant \bar{n}$, then, using the inequality $2\sqrt{u} \leqslant \sqrt{2}/2 + (\sqrt{2})u$ for $u \geqslant 0$, we get

$$\begin{aligned}
U(\delta_i, n_i) &= \frac{\sigma^2 d^2}{n_i}\Bigg(1 + 2\sqrt{\frac{1}{d}\log_+\Big(\frac{d}{Tn_i}\Big)} + \frac{2}{d}\log_+\Big(\frac{d}{Tn_i}\Big)\Bigg)\\
&\leqslant \frac{\sigma^2 d^2}{\bar{n}}\Bigg(1 + 2\sqrt{\frac{1}{d}\log_+\Big(\frac{d}{T\bar{n}}\Big)} + \frac{2}{d}\log_+\Big(\frac{d}{T\bar{n}}\Big)\Bigg)\\
&\leqslant \frac{\sigma^2 d^2}{\bar{n}}\bigg(2 + \frac{3}{d}\log_+\Big(\frac{d}{T\bar{n}}\Big)\bigg) = \|\theta\|^2\frac{1+\sqrt{2}/2 + \frac{2+\sqrt{2}}{d}\log_+\Big(\frac{d}{T\bar{n}}\Big)}{(1+\sqrt{2}/2)\overline{\log}\Big(\frac{T\|\theta\|^2}{\sigma^2 d^2}\Big)}\\
&\leqslant \|\theta\|^2\frac{1+\sqrt{2}/2 + \frac{2+\sqrt{2}}{d}\log_+\Big(\frac{T\|\theta\|^2}{2\sigma^2 d}\Big)}{(1+\sqrt{2}/2)\overline{\log}\Big(\frac{T\|\theta\|^2}{\sigma^2 d^2}\Big)}\\
&\leqslant \|\theta\|^2\frac{1+\sqrt{2}/2 + (1+\sqrt{2}/2)\log_+\Big(\frac{T\|\theta\|^2}{\sigma^2 d^2}\Big)}{(1+\sqrt{2}/2)\overline{\log}\Big(\frac{T\|\theta\|^2}{\sigma^2 d^2}\Big)} = \|\theta\|^2\,,
\end{aligned}$$

where we also used the facts that $d \geqslant 2$ and $1/\bar{n} \leqslant \|\theta\|^2/(\sigma^2 d^2)$. Therefore, we have shown that if $n_i \geqslant \bar{n}$, then $i \geqslant i^\star$, hence

$$2^{i^\star} \leqslant \min\{2^i \mid n_i \geqslant \bar{n}\} \leqslant \max\Big(\frac{2\bar{n}}{d}, 1\Big) \leqslant \frac{2\bar{n}}{d} + 1\,,$$

which is the claimed inequality.

Now that (6) is proved, let us proceed with the next steps. Let $\hat{\imath}$ be the index of the warm-up round in which the algorithm terminates. Then by design, $\|\hat{\theta}_{\hat{\imath}}\| > 3U(\delta_{\hat{\imath}}, n_{\hat{\imath}})$. Therefore, by Lemma 11 and the triangle inequality,

$$\begin{aligned}\mathbb{P}\Big[\big|\|\hat{\theta}_{\hat{\imath}}\| - \|\theta\|\big| > \frac{1}{2}\|\theta\|\Big] &\leqslant \mathbb{P}\Big[\big|\|\hat{\theta}_{\hat{\imath}}\| - \|\theta\|\big| > \frac{1}{3}\|\hat{\theta}_{\hat{\imath}}\|\Big] \\ &\leqslant \mathbb{P}\Big[\big|\|\hat{\theta}_{\hat{\imath}}\| - \|\theta\|\big| > U(\delta, n_{\hat{\imath}})\Big] \leqslant \mathbb{P}\Big[\|\hat{\theta}_{\hat{\imath}} - \theta\| > U(\delta, n_{\hat{\imath}})\Big]\,.\end{aligned}$$

With a union bound on the possible values of $\hat{\imath}$, and applications of Lemma 6, the probability above is less than

$$\begin{aligned}\sum_{i=1}^{i^\star+4} \mathbb{P}\Big[\|\hat{\theta}_i - \theta\| > U(\delta_i, n_i)\Big] + \mathbb{P}\big[\hat{\imath} > i^\star + 4\big] &\leqslant \sum_{i=1}^{i^\star+4} \delta_i + \mathbb{P}\big[\hat{\imath} > i^\star + 4\big] \\ &\leqslant \frac{d}{T} 2^{i^\star+5} + \mathbb{P}\big[\hat{\imath} > i^\star + 4\big]\,. \end{aligned} \tag{7}$$

The number of warm-up rounds can be written as

$$N_{\text{warm}-\text{up}} = \min\left(T, \sum_{i=1}^{\hat{\imath}} n_i\right)$$

so that, doing a union bound over the value of $\hat{\imath}$ as above,

$$\mathbb{E}[N_{\text{warm}-\text{up}}] \leqslant \sum_{i=1}^{i^\star+4} n_i + T\mathbb{P}\big[\hat{\imath} > i^\star + 4\big] \leqslant d2^{i^\star+5} + T\mathbb{P}\big[\hat{\imath} > i^\star + 4\big]\,. \tag{8}$$

Now observe that on the event $\{\hat{\imath} > i^\star + 4\}$, we have

$$\|\hat{\theta}_{i^\star+4}\| \leqslant 3U(\delta_{i^\star+4}, n_{i^\star+4}) \quad \text{and} \quad U(\delta_{i^\star+4}, n_{i^\star+4}) < \frac{1}{4} U(\delta_{i^\star}, n_{i^\star}) \leqslant \frac{1}{4}\|\theta\|\,.$$

Therefore, by combining both inequalities, on this event,

$$\|\theta - \hat{\theta}_{i^\star+4}\| \geqslant \|\theta\| - \|\hat{\theta}_{i^\star+4}\| > 4U(\delta_{i^\star+4}, n_{i^\star+4}) - 3U(\delta_{i^\star+4}, n_{i^\star+4}) = U(\delta_{i^\star+4}, n_{i^\star+4})\,.$$

We have thus shown that

$$\mathbb{P}\big[\hat{\imath} > i^\star + 4\big] \leqslant \mathbb{P}\big[\|\theta - \hat{\theta}_{i^\star+4}\| > U(\delta_{i^\star+4}, n_{i^\star+4})\big] \leqslant \delta_{i^\star+4} = \frac{d}{T} 2^{i^\star+4}\,.$$

Plug this in (7) and (8) to conclude.

$$\mathbb{P}\Big[\hat{B} \notin \big[(1/2)\|\theta\|_A, (3/2)\|\theta\|_A\big]\Big] \leqslant \frac{48d}{T} 2^{i^\star} \quad \text{and} \quad \mathbb{E}[N_{\text{warm}-\text{up}}] \leqslant 48d2^{i^\star}\,. \qquad \blacksquare$$

A.2. Technical lemmas

Lemma 10 *For any $\theta, u \in \mathbb{R}^d$, if $\theta^\top A u \geqslant 0$ then,*

$$\theta^\top (x^\star(\theta) - x^\star(u)) \leqslant \frac{1}{\|\theta\|_A} \|\theta - u\|_A^2 \,.$$

Proof Without loss of generality $A = I_d$, since $\|A^{1/2}\theta\| = \|\theta\|_A$ and

$$\theta^\top \big(x_A^\star(\theta) - x_A^\star(u)\big) = A^{1/2}\theta^\top \big(x_I^\star(A^{1/2}\theta) - x_I^\star(A^{1/2}u)\big) \,.$$

Then $x^\star(\theta) = (1/\|\theta\|)\theta$ and $x^\star(u) = (1/\|u\|)u$, so

$$\theta^\top (x^\star(\theta) - x^\star(u)) = \frac{\|\theta\|}{2} \Big\| \frac{1}{\|\theta\|}\theta - \frac{1}{\|u\|}u \Big\|^2 = \frac{1}{2\|\theta\|} \Big\| \theta - \frac{\|\theta\|}{\|u\|}u \Big\|^2 \,.$$

Furthermore, using the fact that $0 \leqslant \theta^\top u \leqslant \|\theta\|\|u\|$,

$$\Big\| \theta - \frac{\|\theta\|}{\|u\|}u \Big\|^2 = 2\|\theta\|^2 - 2\frac{\|\theta\|}{\|u\|}\theta^\top u \leqslant 2\|\theta\|^2 - 2\frac{1}{\|u\|^2}(\theta^\top u)^2 \,.$$

Recognizing that the right-most term is twice the distance from θ to its projection on the line with direction u,

$$\Big\| \theta - \frac{\|\theta\|}{\|u\|}u \Big\|^2 \leqslant 2\Big(\|\theta\|^2 - \frac{1}{\|u\|^2}(\theta^\top u)^2\Big) = 2 \inf_{\lambda \in \mathbb{R}} \|\theta - \lambda u\|^2 \leqslant 2\|\theta - u\|^2 \,.$$ ■

Lemma 11 *For any $a, b, \lambda \geqslant 0$ such that $|a - b| \geqslant \lambda a$, we have $|a - b| \geqslant \lambda/(1+\lambda)b$.*

Proof We consider two cases. If $a \geqslant b$, then $|a - b| \geqslant \lambda a \geqslant \lambda b \geqslant \lambda/(1+\lambda)b$ proving the claim. Otherwise, if $a < b$, then $(1+\lambda)|a - b| = |a - b| + \lambda(b - a) \geqslant \lambda a + \lambda(b - a) = \lambda b$. ■

Appendix B. Action Sets That Do Not Contain Zero

B.1. Non-centered Action Sets

We now describe how to modify our algorithm to handle non-centered ellipsoids, i.e., sets of the form

$$\mathcal{X} = \{x \in \mathbb{R}^d : \|x - c\|_{A^{-1}} \leqslant 1\}$$

for $c \neq 0$.

ETC-type Algorithms The modification applies more generally to any ETC-type algorithm, going from action sets $\mathcal{X}$ that contain 0 to the translated $c + \mathcal{X}$ for any $c \in \mathbb{R}^d$. Let us give some definitions.

Definition 12 *An algorithm is an ETC-type algorithm if there exists a stopping time T_{commit} such that for any $t > T_{\mathrm{commit}}$, the algorithm commits to the action $x_t = x_{T_{\mathrm{commit}}+1}$.*

In the case of Algorithm 1, the warm-up rounds should be counted as exploration rounds. We denote the commit action of an ETC-type algorithm by $x_{\mathrm{commit}} := x_{T_{\mathrm{commit}}+1}$. An ETC-type algorithm comes with a natural upper bound on its regret, as

$$R_T(\mathcal{A}, \mathcal{X}, \theta) \leqslant \Big(\max_{x,y \in \mathcal{X}} \|x - y\| \Big) \|\theta\|_\star \, \mathbb{E}[T_{\mathrm{commit}}] + T \, \mathbb{E}[\theta^\top (x^\star(\theta) - x_{\mathrm{commit}})] \,, \tag{9}$$

where $\|\cdot\|$ denotes any norm on $\mathbb{R}^d$ and $\|\cdot\|_\star$ denotes its dual norm.

Consider a parametric ETC-type algorithm Alg, that takes a variance proxy and time-horizon as inputs. Assume that for any linear bandit problem, the algorithm $\mathcal{A} = \mathrm{Alg}(\sigma^2, T)$ tuned with σ^2 and T satisfies

$$\mathbb{E}\big[T_{\mathrm{commit}}(\mathcal{A})\big] \leqslant B(\theta, \sigma^2, T) \quad \text{and} \quad \mathbb{E}\big[\theta^\top \big(x^\star_{\mathcal{X}}(\theta) - x_{\mathrm{commit}}(\mathcal{A})\big)\big] \leqslant C(\theta, \sigma^2, T) \,, \tag{10}$$

for some B and C, called the expected exploration time and expected commit error, respectively.

A Reduction We transform any ETC-type algorithm for action sets $\mathcal{X} \ni 0$ into an algorithm for the translated set $c + \mathcal{X}$ which retains essentially the same bound as above up to constant factors.

Consider a parametric algorithm Alg for $\mathcal{X}$ that takes the variance proxy and horizon as inputs, and let $\mathcal{A} = \mathrm{Alg}(2\sigma^2, T/2)$. The recipe consists in replacing the plays $\tilde{x}_t \in \mathcal{X}$ in each round of exploration made by $\mathcal{A}$ by two rounds of exploration, playing first the reference point c and then $c + \tilde{x}_t$, at times $2t - 1$ and $2t$. Now observe that

$$y_{2t-1} - y_{2t} = \theta^\top \tilde{x}_t + (z_{2t-1} - z_{2t}) \,.$$

Therefore, by providing $y_{2t-1} - y_{2t}$ as inputs to $\mathcal{A}$, the observations received by $\mathcal{A}$ are consistent with the linear bandit model with action set $\mathcal{X}$, except that the sub-gaussian variance proxy is doubled. Finally, since the optimal action is equivariant by translation of the action set, the commit error is the same in both cases.

B.2. Analysis of the Reduction

The next theorem states that the algorithm obtained by applying the reduction enjoys essentially the same guarantees as the initial algorithm.

Algorithm 2: Reduction for non-centered action sets for ETC algorithms

Input : ETC-style algorithm $\mathcal{A}$ for action set $\mathcal{X}$ containing 0. Translation vector $c \in \mathbb{R}^d$.
Init: $t = 1$
while $t \leqslant T_{\text{commit}}(\mathcal{A})$ **do**
  Receive $\tilde{x}_t \in \mathcal{X}$ from $\mathcal{A}$
  Play $x_{2t-1} = c$, receive y_{2t-1}. Play $x_{2t} = \tilde{x}_t$, receive y_{2t}.
  Give $y_{2t-1} - y_{2t}$ as input to $\mathcal{A}$
Get x_{commit} from $\mathcal{A}$
while $t \leqslant T$ **do**
  Play $c + x_{\text{commit}}$

Theorem 13 *Let $\mathcal{X}$ denote an action set containing $0 \in \mathbb{R}^d$. Let Alg be an ETC type algorithm with bounds B and C (see (10)). Denoting by $\mathcal{A}_c$ the algorithm obtained by applying the reduction Algorithm 2 to $\mathrm{Alg}(2\sigma^2, T/2)$ for translation vector $c \in \mathbb{R}^d$, for any linear bandit problem with parameters θ and σ^2,*

$$R_T(\mathcal{A}_c, c + \mathcal{X}, \theta) \leqslant 2\Big(\max_{x,y \in \mathcal{X}} \|x - y\|\Big) \|\theta\|_\star B(\theta, 2\sigma^2, T/2) + T\, C(\theta, 2\sigma^2, T/2)\,.$$

This implies in particular that if there is a good ETC-type algorithm for an action set containing 0, then applying a translation to the action set can only make the linear bandit problem easier. Note also that the translation vector c does not appear in the upper bound, perhaps surprisingly.

Proof Let $T_{\text{commit}}(\mathcal{A})$ denote the commit time, and $x_{\text{commit}}(\mathcal{A}) \in \mathcal{X}$ the commit action of $\mathcal{A}$ on its observations. The sequence of observations given to $\mathcal{A}$ fit exactly a linear bandit model with subvariance proxy $2\sigma^2$ and tuned for horizon $T/2$, so the guarantees on $\mathcal{A}$ hold:

$$\mathbb{E}\big[T_{\text{commit}}(\mathcal{A})\big] \leqslant B(\theta, 2\sigma^2, T/2) \quad \text{and} \quad \mathbb{E}\big[\theta^\top\big(x^\star_{\mathcal{X}}(\theta) - x_{\text{commit}}(\mathcal{A})\big)\big] \leqslant C(\theta, 2\sigma^2, T/2)\,.$$

By construction $\mathcal{A}_c$ is an ETC-type algorithm with $T_{\text{commit}}(\mathcal{A}_c) = 2T_{\text{commit}}(\mathcal{A})$, and the commit action is $x_{\text{commit}}(\mathcal{A}_c) = c + x_{\text{commit}}(\mathcal{A})$. So the regret decomposition (9) applies and

$$\begin{aligned} R_T(\theta) &\leqslant \Big(\max_{x,y \in \mathcal{X}} \|x - y\|\Big) \|\theta\|_\star\, \mathbb{E}[T_{\text{commit}}(\mathcal{A}_c)] + T\,\mathbb{E}[\theta^\top(x^\star_{c+\mathcal{X}}(\theta) - x_{\text{commit}}(\mathcal{A}_c))] \\ &\leqslant 2\Big(\max_{x,y \in \mathcal{X}} \|x - y\|\Big) \|\theta\|_\star\, \mathbb{E}[T_{\text{commit}}(\mathcal{A})] + T\,\mathbb{E}[\theta^\top(x^\star_{\mathcal{X}}(\theta) - x_{\text{commit}}(\mathcal{A}))]\,. \end{aligned}$$ ■

Appendix C. Proof details for Section 2

Proof of Proposition 2 We consider a noiseless linear bandit problem with mean parameter θ sampled from a Gaussian distribution $\mathcal{N}(0,(B^2/d)A^{-1})$. The idea of the proof is to lower bound the expected regret of any algorithm when averaged over this distribution.

After $t \leqslant d$ rounds of playing, let $\mathcal{F}_t$ denote the σ-algebra generated by the observations $y_1,\dots,y_t$, and the chosen actions $x_1,\dots,x_{t+1}$ (we include x_{t+1} to account for the possible internal randomization of the algorithm). At time $t+1$, if we integrate both over the initial randomization on θ and on the internal randomization, by the tower rule and since x_{t+1} is $\mathcal{F}_t$-measurable,

$$\mathbb{E}[(x_{t+1}-c)^\top\theta] = \mathbb{E}[(x_{t+1}-c)^\top\mathbb{E}[\theta|\mathcal{F}_t]] \leqslant \mathbb{E}\big[\|\mathbb{E}\big[\theta|\mathcal{F}_t\big]\|_A\big]\,.$$

By standard results on Bayesian linear regression, (see, e.g., Chapter 9.2 in Hoff (2009)), the distribution of $A^{1/2}\theta$ conditional on $\mathcal{F}_t$ is $\mathcal{N}\big(P_t\theta,(B^2/d)(I_d-P_t)\big)$ where P_t denotes the orthogonal projection on the span of $x_1,\dots,x_t$. Orthonormalizing the family (x_t) yields orthonormal vectors $e_1,\dots,e_t$ such that e_i is $\mathcal{F}_{i-1}$-measurable and

$$\|A^{-1/2}P_tA^{1/2}\theta\|_A^2 = \|P_tA^{1/2}\theta\|^2 = \sum_{i=1}^t((A^{1/2}\theta)^\top e_i)^2\,.$$

Then, since e_i is $\mathcal{F}_{i-1}$-measurable, the vector $(A^{1/2}\theta)^\top e_i$ is normally distributed with distribution $\mathcal{N}\big(0,(B^2/d)e_i^\top(I_d-P_{i-1})e_i\big) = \mathcal{N}\big(0,B^2/d\big)$. So for all $i\leqslant d$, we have the equality $\mathbb{E}\big[((A^{1/2}\theta)^\top e_i)^2|\mathcal{F}_{i-1}\big] = B^2/d$. Applying this repeatedly, we see that

$$\mathbb{E}[\|A^{-1/2}P_tA^{1/2}\theta\|_A] \leqslant \sqrt{\mathbb{E}\big[\|A^{-1/2}P_tA^{1/2}\theta\|_A^2\big]} = B\sqrt{\frac{t}{d}}\,.$$

This implies that the averaged regret up to time d is at least

$$\begin{aligned}\mathbb{E}[R_d(\theta)] &= \sum_{t=1}^d \mathbb{E}[(x^\star(\theta)-c)^\top\theta] - \mathbb{E}[(x_t-c)^\top\theta]\\ &\geqslant \sum_{t=1}^d \mathbb{E}[\|\theta\|_A] - \mathbb{E}[\|A^{-1/2}P_{t-1}A^{1/2}\|_A] \geqslant \sum_{t=1}^d \mathbb{E}[\|\theta\|_A] - B\sqrt{\frac{t-1}{d}}\,.\end{aligned}$$

Note also that since $A^{1/2}$ is a standard gaussian vector,

$$\mathbb{E}_\theta\big[\|A^{1/2}\theta\|\big] = \frac{B}{\sqrt{d}}\frac{\sqrt{2}\Gamma((d+1)/2)}{\Gamma(d/2)} \geqslant B\sqrt{(d-1)/d}\,,$$

where we used Gautschi's inequality,

$$x^{1-s} \leqslant \frac{\Gamma(x+1)}{\Gamma(x+s)},$$

with $x=(n-1)/2$ and $s=1/2$. All in all, this implies that

$$\mathbb{E}[R_T(\theta)] \geqslant B\sum_{t=1}^d\sqrt{\frac{d-1}{d}} - \sqrt{\frac{t-1}{d}} \geqslant \frac{1}{3}dB\,.$$

To complete the averaging argument, we also need to ensure that the norm of θ is large enough. Let c be some numerical parameter, of which we will set the value later on. By a case disjunction,

$$\mathbb{E}\big[R_d(\theta)\mathbb{1}\{\|\theta\|_A \leqslant cB\}\big] = \mathbb{E}[R_d(\theta)] - \mathbb{E}[R_d(\theta)\mathbb{1}\{\|\theta\|_A > cB\}]\,.$$

We apply a version of the peeling trick decomposing over the values of $\|\theta\|_A$ in a grid to bound

$$\begin{aligned}\mathbb{E}[R_d(\theta)\mathbb{1}\{\|\theta\|_A > cB\}] &\leqslant \sum_{i=1}^{+\infty} 2t\sqrt{i+1}cB\mathbb{P}[\|\theta\|_A > \sqrt{i}cB] \\ &\leqslant 2\sqrt{2}dB\sum_{i=1}^{+\infty}\sqrt{c^2 i}\mathbb{P}[\|\theta\|_A > \sqrt{i}cB] = \frac{2\sqrt{2}dB}{\sqrt{d}}\sum_{i=1}^{+\infty}\sqrt{dic^2}\mathbb{P}[\|\theta\|_A > \sqrt{i}cB]\,.\end{aligned}$$

Use the concentration inequality Lemma 6, which applies in particular to the Gaussian vector $A^{1/2}\theta$,

$$\mathbb{P}\Big[\|A^{1/2}\theta\|^2 \geqslant B^2\Big(2+\frac{3x}{d}\Big)\Big] \leqslant e^{-x} \quad \text{with} \quad x = \frac{d}{3}(c^2\, i - 2)\,.$$

If $c^2 \geqslant 8/7$, then $c^2 i - 2 \geqslant 3c^2 i/4$ as $7ci^2/4 \geqslant 2$ (remember $i \geqslant 1$) so

$$\mathbb{P}\big[\|A^{1/2}\theta\|^2 \geqslant ic^2B^2\big] \leqslant \exp\Big(-\frac{idc^2}{4}\Big)$$

Note furthermore that since the function $xe^{-x^2/8}$ is decreasing on $[4,+\infty)$, so for any $x \geqslant 4$, $xe^{-x^2/4} \leqslant 4e^{-2}e^{-x^2/8}$. Consequently, if $c \geqslant 2\sqrt{2}$, then $\sqrt{idc^2} \geqslant 4$ since $d \geqslant 2$, and

$$\sum_{i=1}^{+\infty}\sqrt{idc^2}e^{-idc^2/4} \leqslant \sum_{i=1}^{+\infty}e^{-idc^2/8} = \frac{e^{-dc^2/8}}{1-e^{-dc^2/8}}\,.$$

We have thus lower bounded the regret on the event that $\|\theta\|_A > cB$ as

$$\begin{aligned}\mathbb{E}[R_t(\theta)\mathbb{1}\{\|\theta\|_A > cB\}] &\geqslant \frac{dB}{2} - \frac{2\sqrt{2}dB}{\sqrt{d}}\frac{e^{-dc^2/8}}{1-e^{-dc^2/8}} = \Big(\frac{1}{2} - \frac{2\sqrt{2}}{\sqrt{d}}\frac{e^{-dc^2/8}}{1-e^{-dc^2/8}}\Big)dB \\ &\geqslant \Big(\frac{1}{2} - \frac{2e^{-c^2/4}}{1-e^{-c^2/4}}\Big)dB \geqslant \Big(\frac{1}{2} - \frac{2e^{-2}}{1-e^{-2}}\Big)dB \geqslant 0.18dB\,,\end{aligned}$$

we also used the facts that $d \geqslant 2$ and $c \geqslant 2\sqrt{2}$. Applying our result exactly with $c = 4$, we deduce the existence of some θ such that $\|\theta\|_A \leqslant 4B$ and

$$R_T(\theta) \geqslant R_d(\theta) \geqslant 0.07dB \geqslant \frac{0.07}{4}d\|\theta\|_A \geqslant 0.017d\|\theta\|_A\,. \qquad \blacksquare$$

Appendix D. Glossary

Notation related to problem setting

Notation	Description	Page List
T	Time horizon	1
$[T]$	Set of time step until $[T] = \{1, 2, ..., T\}$	1
d	Dimension of the problem considered	1
A	Definite positive matrix defining a norm and an ellipsoid action set	1
$\lambda_{\max}$	Maximal eigenvalue of the matrix	4
$\lambda_{\min}$	Minimal eigenvalue of the matrix	4
$\|u\|_M$	Norm defined by the matrix M, $\|u\|_M = \sqrt{u^\top M u}$	1
$\mathcal{X}$	The action set $\mathcal{X} \subset \mathbb{R}^d$	1
c	Center of the ellipsoid that define the action set	1
θ	Unkown parameter of the model $\theta \in \mathbb{R}^d$	1
$x^\star(\theta)$	Optimal action as a function θ and A	2
x_t	Action taken at time t	1
y_t	Reward received at time t	1
z_t	Noise at time t	1
σ	Subgaussian parameter of the noise	1
$R_T(\theta)$	Average regret at time T on the environment parametrized by θ	2

Notation related to the algorithm and its analysis

Notation	Description	Page List
Ω	Lowerbound Notation, "$f(\phi) = \Omega(g(\phi))$" iff $\exists c > 0, \forall \phi, f(\phi) \geqslant cg(\phi)$	2
O	Upperbound Notation, "$f(\phi) = O(g(\phi))$" iff $\exists c > 0, \forall \phi, f(\phi) \leqslant cg(\phi)$	2
$\widetilde{O}$	Upperbound Notation with hidden polylog terms	2
B	Norm A of the parameter θ	2
$\mathcal{E}_A(B)$	Elliptical ensemble of vectors for which the norm is B in $\|.\|_A$ norm	2
$\mathcal{A}$	Generic notation for an algorithm	2
w_t	Reparametrization of the action x_t centered and in base A, $x_t = A^{\frac{1}{2}} w_t + c$	5
ε^2	Parameter in the proof to set the distance between two confusing problems' parameter	5
ϕ	Function to create confusing parameters in the lowerbound	5
ξ	Vector in $\xi \in \{-1, 1\}^d$ to create confusing parameters in the lowerbound	5
$\theta(\xi)$	Confusing parameters in the lowerbound	5
τ_i	Stopping time to control relative entropy in the i-th direction	5
U_i	Random variable representing directional regret in the i-th direction	5
$\xi^{(-i)}$	Flip the ith coordinate of the vector ξ. $\forall j, \xi_j^{(-i)} := (-1)^{\mathbb{1}(i=j)} \xi_j$	6
$P_{\theta(\xi),i}$	the probability distribution of $(w_t)_{t \leqslant \tau_i(T)}$ under parameter $\theta(\xi)$	6
C	Parameter used to set a constant relative entropy between problems	5
D	Relative entropy between two distributions	6

Notation	Description	Page List
LS	Least-squares estimator with observed features X and rewards Y. $\mathrm{LS}(X, Y) = (X^\top X)^{-1} X^\top Y$	7
$X^\top X$	The design matrix with the oberving features $X \in \mathbb{R}^{n \times d}$	9
$\mathcal{T}$	Set of instant $\mathcal{T} \subset \mathbb{N}$	7
$X_\mathcal{T}$	Submatrix of the features of the reward $X_\mathcal{T} := (x_t)_{t \in \mathcal{T}} \in \mathbb{R}^{\lvert\mathcal{T}\rvert \times d}$	7
$Y_\mathcal{T}$	Subvector of the reward $Y_\mathcal{T} := (y_t)_{t \in \mathcal{T}} \in \mathbb{R}^{\lvert\mathcal{T}\rvert}$	7
$Z_\mathcal{T}$	Subvector of the noise $Z_\mathcal{T} := (z_t)_{t \in \mathcal{T}} \in \mathbb{R}^{\lvert\mathcal{T}\rvert}$	7
$\hat{\theta}_i$	The Least square estimation of the parameter θ at the end of phase i	7
α	Parameter used in the warm-up phase	8
$\hat{\imath}$	Phase when the warm-up phase ends	8
$\hat{B}$	The estimation of $\lVert\theta\rVert_A$ at the end of the warm-up phase, $\hat{B} = \lVert\hat{\theta}_{\hat{\imath}}\rVert_A$	8
δ_i	Probability of failure of the i-th phase of the warm-up procedure	7
n_i	Duration of the i-th phase of the warm-up procedure	7
T_i	Number of rounds after the i-th phase of the warm-up procedure	7
$U(\delta_i, n_i)$	Confidence bound of the estimated norm distribution with probability $1 - \delta_i$ and n_i samples. Related to the χ^2 law.	8
N_e	Number of exploration round during the exploration phase	8
$\hat{\theta}$	The least-squares estimation of the parameter θ at the beginning of the commit phase	8
$\mathbb{E}_{\text{expl}}$	Expectation conditionally on all rounds before phase 3	9
$\mathcal{C}_t$	Confidence ellipsoid at time t of the parameter θ in optimistic algorithms	4
$\overline{\log}$	$\overline{\log}(x) = 1 + \log(\max(1, x))$	7
$-$	$a^- := \min(a, 0)$	7

Abbreviation used in the paper

Notation	Description	Page List
`ETC`	`Explore-Then-Commit`, Type of algorithm that first explores the environment then commits to the best actions	2
`E2TC`	`Explore-Explore-Then-Commit`, The locally asymptotic minimax algorithm introduced in this paper	2
`CB`	`ConfidenceBall`, Optimistic based algorithm of Dani et al. (2008)	3
`PEGE`	`Phased Exploration and Greedy Exploitation` Nearly ETC like algorithm of Rusmevichientong and Tsitsiklis (2010)	4
`UE`	`Uncertainty Ellipsoid`, Optimistic algorithm of Rusmevichientong and Tsitsiklis (2010)	4
`OFUL`	`Optimism in the Face of Uncertainty Linear bandit`, Optimistic algorithm of Abbasi-Yadkori et al. (2011)	4
`OLSOFUL`	`Ordinary Least Squares OFUL`, Optimistic algorithm of Gales et al. (2022)	4
`TS`	`Thompson Sampling`, Sampling based algorithm of Abeille and Lazaric (2017)	4