OPT2020: 12th Annual Workshop on Optimization for Machine Learning

Employing No Regret Learners for Pure Exploration in Linear Bandits

Mohammadi Zaki MOHAMMADI@IISC.AC.IN
Electrical Communication Engineering,
Indian Institute of Science,
Bangalore 560012.

Avinash Mohan AVINASHMOHAN@CAMPUS.TECHNION.AC.IL
Faculty of Electrical Engineering,
Technion, Israel Institute of Technology,
Haifa 3200003.

Aditya Gopalan ADITYA@IISC.AC.IN
Electrical Communication Engineering,
Indian Institute of Science,
Bangalore 560012.

Abstract

We study the best arm identification problem in linear multi-armed bandits (LMAB) in the fixed confidence (δ-PAC) setting; this is also the problem of optimizing an unknown linear function over a discrete ground set with noisy, zeroth-order access. We propose an explicitly implementable and provably order-optimal sample-complexity algorithm to solve this problem. Most previous approaches rely on access to a minimax optimization oracle which is at the heart of the complexity of the problem. We propose a method to solve this optimization problem (upto suitable accuracy) by interpreting the problem as a two-player zero-sum game, and attempting to sequentially converge to its saddle point using low-regret learners to compute the players' strategies in each round which yields a concrete querying algorithm. The algorithm, which we call the *Phased Elimination Linear Exploration Game* (PELEG), maintains a high-probability confidence ellipsoid containing θ^* in each round and uses it to eliminate suboptimal arms in phases. We analyze the sample complexity of PELEG and show that it matches, up to order, an instance-dependent lower bound on sample complexity in the linear bandit setting without requiring boundedness assumptions on the parameter space. PELEG is, thus, the first algorithm to achieve both order-optimal sample complexity and explicit implementability for this setting. We also provide numerical results for the proposed algorithm consistent with its theoretical guarantees.

1. Introduction

We study the problem of best arm identification (BAI) in linearly parameterised multi-armed bandits. Given a (finite) set of feature vectors $\mathcal{X} = \{x_1, x_2, \ldots, x_K\}$, a confidence parameter δ and an unknown vector θ^*, the goal is to identify $\mathrm{argmax}_{x\in\mathcal{X}}\, x^T\theta^*$, with probability at least $1-\delta$, using noisy measurements of the form $x^T\theta^*$ (fixed-confidence setting) as quickly as possible. Formally, the agent plays sequentialy and in every round $t = 1, 2, \ldots$ the agent chooses an arm $x_t \in \mathcal{X}$, and receives a reward $y(x_t) = \theta^{*T} x_t + \eta_t$. Recently, Degenne et al.[4] use an approach similar to Degenne et al. [3] to design an algorithm called LinGame for pure exploration in LMAB. Their algorithm achieves the information-theoretic lower bound for sample complexity in the limit as $\delta \downarrow 0$.

A closer look, however, indicates various interesting directions for improvement. Firstly, being a fully adaptive algorithm, the stopping criteria used in Degenne et al. [4] rely on potentially weaker concentration results, which aim at controlling deviations in all (i.e., 2^d) directions. We address this by proposing a phased algorithm which relies on tighter deviation bounds, along only the difference directions obtained from the surviving arms in each phase. Second, LinGame requires knowledge of an explicit upper bound on $\|\theta^*\|_2$ – an unknown model parameter. In fact, Degenne et.al. [4] leave it as an open problem to remove the requirement of a bound on $\|\theta^*\|_2$. We instead make use of an action-set property, namely, $C := \lambda_{min}\left(\sum_{x\in\mathcal{X}_B} xx^T\right)$, which is computable in advance, indicating that this is possible, but with the additional dependence on C. Here $\mathcal{X}_B$ denotes a barycentric spanner of the armset which consists of atmost d arms. We note that the lower bound (i.e., with constant factor 1) in Degenne et al. [4] is achieved only in the limit as $\delta \downarrow 0$; its non-asymptotic (in $1/\delta$) version has additive (second-order) terms (see Table 1) which can be large and effect the sample complexity of their algorithm.

We try to follow, at a high level, the template of Fiez et al. [5], who give an algorithm with information-theoretically optimal (instance-dependent) PAC sample complexity. Their algorithm however, requires repeated oracle access to a minimax optimization problem; it is not clear, from a performance standpoint, in what manner (and to what accuracy) this optimization problem should be *practically* solved (for its experiments, the paper implements a (approximate) minimax oracle using the Frank-Wolfe algorithm and a heuristic stopping rule, but this is not rigorously justifiable for nonsmooth optimization, see Sec. 3) to enjoy the claimed sample complexity. In this paper, we give an explicit linear bandit best-arm identification algorithm with instance-optimal PAC sample complexity and, more importantly, a clearly quantified computational effort by using new techniques: the main ingredient in the proposed algorithm is a game-theoretic interpretation of the minimax optimization problem that is at the heart of the instance-based sample complexity lower bound. This in turn yields an adaptive, sample-based approach using carefully constructed confidence sets for the unknown parameter θ^*. The adaptive sampling strategy is driven by the interaction of 2 no-regret online learning subroutines that attempt to solve the minimax problem approximately.

Assumptions and Notation. The noise η_t is zero-mean assumed to be conditionally $1-$ sub-Gaussian. We denote by $\nu^k_{\theta^*}$ the distribution of the reward obtained by pulling arm $k \in [K]$, i.e., $\forall t \geqslant 1, y(x_t) \sim \nu^k_{\theta^*}$, whenever $x_t = x_k$. Given two probability distributions μ, ν over $\mathbb{R}$, $KL(\mu,\nu)$ denotes the KL Divergence of μ and ν (assuming $\mu \ll \nu$). Given $\theta \in \mathbb{R}^d$, let $a^* \equiv a^*(\theta) = \underset{a\in[K]}{\operatorname{argmax}}\, \theta^T x_a$, where we assume that θ is such that the argmax is unique. We assume that $\|x_k\|_2 \leqslant 1, \forall x_k \in \mathcal{X}$. Given a positive definite matrix A, $\|x\|_A := \sqrt{x^T A x}$ denotes the matrix norm induced by A. For any $i \in [K], i \neq a^*$, let $\Delta_i := \theta^{*T}(x_{a^*} - x_i)$ be the gap between the largest expected reward and the expected reward of (suboptimal) arm x_i. Let $\Delta_{\min} := \min_{i\in[K]} \Delta_i$. We denote $B(z,r)$ as the *closed* ball with center z and radius r. We define $\mathcal{P}_K$ to be the set of all probability mass functions on an alphabet of size K.For the benefit of the reader, we provide a glossary of commonly used symbols in Sec. A.

2. The Minimax Optimization Problem and Pure-exploration games

We first note that a lower bound on the sample complexity of any δ- PAC algorithm for the canonical (i.e., unstructured) bandit setting [6] was generalized by Fiez et al [5] to the linear bandit setting. This result states that any δ-PAC algorithm in the linear setting must satisfy $\mathbb{E}_{\theta^*}[\tau] \geqslant$

$(\log 1/2.4\delta)\,\frac{1}{T_{\theta^*}} = (\log 1/2.4\delta)\,\frac{1}{D_{\theta^*}}$, where $T_{\theta^*} := \max_{w\in\mathcal{P}(\mathcal{X})}\min_{\theta:a^*(\theta)\neq a^*(\theta^*)}\sum_{k\in[K]} w_k KL\left(\nu_\theta^k,\nu_{\theta^*}^k\right)$ and $D_{\theta^*} := \max_{w\in\mathcal{P}(\mathcal{X})}\min_{\substack{x\in\mathcal{X}\\ x\neq x^*}}\frac{\left(\theta^{*T}(x^*-x)\right)^2}{\|x^*-x\|^2_{\left(\sum_{x\in\mathcal{X}} w_x xx^T\right)^{-1}}}$, where $x^* = x_{a^*}$. The bound suggests a natural δ-PAC strategy, namely, to sample arms according to the distribution

$$w^* = \operatorname*{argmin}_{w\in\mathcal{P}(\mathcal{X})}\max_{x\in\mathcal{X}\setminus\{x^*\}}\frac{\|\,x^*-x\,\|^2_{\left(\sum_{x\in\mathcal{X}} w_x xx^T\right)^{-1}}}{\left(\left(x^*-x\right)^T\theta^*\right)^2}. \tag{1}$$

However, x^* is unknown. Fiez et al [5] design a nontrivial strategy (RAGE) that attempts to mimic the optimal allocation w^* in phases. Specifically, in phase m, it tries to eliminate arms that are about 2^{-m}-suboptimal (in their gaps), by solving (1) with a plugin estimate of θ^*. This approach, however, is based crucially on solving minimax problems of the form (1). Though the inner (max) function is convex as a function of w on the probability simplex (see e.g., Lemma 1 in [12]), it is *non-smooth*, and it is not clear how, and to what extent, it must be solved in [5]. We are able to circumvent this obstacle by using ideas from games between no-regret online learners with optimism, as introduced by the work of Degenne et al [3] for unstructured bandits.

We consider the following related geometrical optimization problem of fitting an ellipsoid inside a polygon, both centered at origin, namely $\max_{w\in\mathcal{P}_K}\min_{\lambda\in\cup_{x\in\mathcal{X}_m}\mathcal{C}_m(x)}\|\lambda\|^2_{\sum_{x\in\mathcal{X}} w_x xx^T}$, where $\mathcal{C}_m(x)$ is a union of halfspaces, defined in 1. By subsequently reducing the size of the polygon (i.e., value of ε_m in alg 1) we get arbitary reduction in size of the confidence-ellipsoid. Consider the two-player, zero-sum *Pure-exploration Game* in which the MAX player (or column player) plays an arm $k_t \in [K]$ while the MIN (or row) player chooses an $\lambda_t \in \cup_{x\in\mathcal{X}_m}\mathcal{C}_m(x)$. MAX then receives a payoff of $\|\lambda_t\|^2_{x_k x_k^T}$ from MIN. With MAX moving first and playing a mixed strategy $w \in \mathcal{P}(\mathcal{X})$, the value of the game becomes $B_m = \max_{w\in\mathcal{P}_K}\min_{\lambda\in\cup_{x\in\mathcal{X}_m}\mathcal{C}_m(x)}\|\lambda\|^2_{\sum_{x\in\mathcal{X}} w_x xx^T}$. As is shown in Appendix G Prop. 14 this quantity is directly related to $1/D_{\theta^*}$.

We crucially employ no-regret online learners to solve this Pure Exploration Game. More precisely, no-regret learning with the well-known Exponential Weights rule/Negative-entropy mirror descent algorithm [9] on one hand, and a best-response convex programming subroutine on the other, provides a *direct* sampling strategy that obviates the need for separate allocation optimization and rounding for sampling as in [5]. One crucial advantage of our approach (inspired by [3]) is that we only use a best response oracle to solve for $T_{\theta^*}(w)$, which gives us a computational edge over [5] who employ the computationally more costly max-min oracle to solve $T_{\theta^*}(w)$, or, its linear bandit equivalent, D_{θ^*}.

3. Algorithm and Sample Complexity Bound

Our algorithm, that we call "Phased Elimination Linear Exploration Game" (PELEG), is presented in detail in Appendix B as Algorithm 1. PELEG proceeds in phases with each phase consisting of multiple rounds, maintaining a set of *active* arms $\mathcal{X}_m$ for testing during Phase m. An OLS estimate $\widehat{\theta}_m$ of θ^* is used to estimate the mean reward of active arms and, at the end of phase m, every active arm with a plausible reward more than $\approx 2^{-m}$ below that of some arm in $\mathcal{X}_m$ is eliminated. Suppose $\mathcal{S}_m := \left\{x \in \mathcal{X}\setminus\{x^*\} : \theta^{*T}\left(x^*-x\right) < \frac{1}{2^m}\right\}$. If we can ensure that $\mathcal{X}_m \subset \mathcal{S}_m$

in every Phase $m \geqslant 1$, then PELEG will terminate within $\lceil \log_2(1/\Delta_{\min}) \rceil$ phases, where $\Delta_{\min} = \min_{x \neq x^*} \theta^{*T}(x^* - x)$. This statement is proved in Corollary 13 in the Supplementary Material.

Approximating the minimax problem using no regret learners. We formulate the minimax problem discussed in Sec. 2 as a two player, zero-sum game. We solve the game sequentially, converging to its Nash equilibrium by invoking the use of the EXP-WTS algorithm [1]. Specifically, in each round t in a phase, PELEG supplies EXP-WTS (MAX player) with an appropriate loss function l_{t-1}^{MAX} and receives the requisite sampling distribution w_t . This w_t is then fed to the second no-regret learner (MIN player) – a best response subroutine – that finds the 'most confusing' plausible model λ to focus on. This is a minimization of a quadratic function over a union of finitely many convex sets (halfspaces intersecting a ball) which can be transparently implemented in polynomial time. Once the sampling distribution is found, we use an efficient *tracking* procedure to ensure that for every $t \geqslant 1$, $\sum_{s=1}^{t} w_s^k - (\log K) \leqslant n_t^k \leqslant \sum_{s=1}^{t} w_s^k + 1$ (see [4] for a proof). This procedure avoids the use of explicit rounding techniques.

Finally, in each phase m, we need to sample arms often enough to (i) construct confidence intervals of size at most $2^{-(m+1)}$ around $(x - x')^T\theta^*$, $\forall x, x' \in \mathcal{X}_m$, (ii) ensure that $\mathcal{X}_m \subset \mathcal{S}_m$ and (iii) that $x^* \in \mathcal{X}_m$. In Sec. F, we prove a Key Lemma (whose argument is discussed in Sec. 4) to show that our novel *Phase Stopping Criterion* ensures this with high probability. It is worth remarking that the use of phased elimination template of Fiez etal [5] eliminates the need to use more complex, self-tuning online learners like AdaHedge [2], as used in [3] and more recently [4], in favour of the simpler Exponential Weights (Hedge). The main theoretical result of this paper is the following performance guarantee. A more detailed version is in Appendix Sec. G.

Theorem 1 (Sample Complexity of Algorithm 1) *With probability at least $1 - \delta$, the worst-case sample complexity of PELEG is bounded as*

$$\tau \leqslant \tilde{\mathcal{O}}\left(\frac{\log^2(K^2/\delta)}{C^2 D_{\theta^*}}\right). \tag{2}$$

In Sec. 4, we sketch the arguments behind the result with the full proof in Sec. G.

Remark 2 *As explained in Sec. 2, the optimal (oracle) allocation requires $\mathcal{O}\left(\frac{1}{D_{\theta^*}} \log \frac{K}{\delta}\right)$ samples. Comparing this with (2), we see that our algorithm is instance optimal up to logarithmic factors, barring the $\frac{\log(K/\delta)}{C^2}$ term. In most applications, feature vectors (i.e., $x_1, \cdots, x_K$) are chosen to represent the feature space well which translates to a high value of C(i.e., $C = \Omega(1)$) .*

Remark 3 *We conjecture that the extra $\log(1/\delta)$ factor arises only because of the way the analysis is carried out. As is shown in the proof of lemma 9, in the definition of ε_m (line 1), whenever $\varepsilon_m == 1.\left(\frac{1}{2}\right)^{m+1}$, the phase length can be bounded without the extra $\log(1/\delta)$ term. On the other hand, if an upperbound $\|\theta^*\|_2 \leqslant S$ is known, then we can to use this information into our algorithm. More details can be found in appendix (Sec. H), where we also sketch a sample complexity bound for this new version of PELEG (we call PELEG-S, alg 2) which does not depend on the parameter C. This brings out an intrinsic trade-off between the knowledge of S and C.*

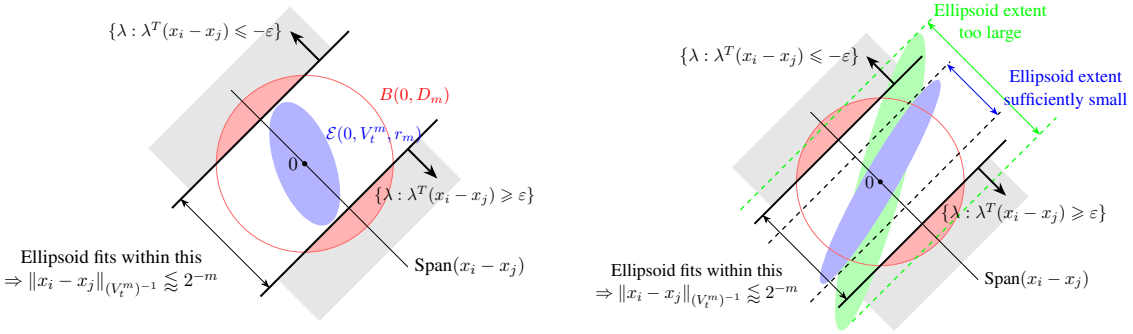
4. Sketch of Sample Complexity Analysis

At a high level the proof of Theorem 1 involves two main parts: (1) a correctness argument for the central **while** loop that eliminates arms, and (2) a bound for its length, which, when added across all phases, gives the overall sample complexity bound.

1. Ensuring progress (arm elimination) in each phase. At the heart of the analysis is the following result which guarantees that upon termination of the central while loop, the uncertainty in estimating all differences of means among the surviving (i.e., non-eliminated) arms remains bounded.

Lemma 4 (Key Lemma) *After each phase* $m \geqslant 1$, $\max_{x,x' \in \mathcal{X}_m, x \neq x'} \|x - x'\|^2_{(V^m_{N_m})^{-1}} \leqslant \frac{\left(\left(\frac{1}{2}\right)^{m+1}\right)^2}{8 \log K^2/\delta_m}$.

Proof sketch. Phase m ends at round t when the ellipsoid $\mathcal{E}(0, V^m_t, r_m)$, with center 0 and shape according to the arms played in the phase so far, becomes small enough to avoid intersecting the half spaces $\mathcal{C}_m(x)$, for all surviving arms x, within the ball $\cap B(0, D_m)$ (Phase Stopping Criteria of the algorithm) which is required to keep loss functions bounded for no-regret properties. Consider the simpler situation when there are only two arms remaining: x_i, x_j. When a phase ends we have one of two possibilities. Fig. 1 (a) shows a situation when the ellipsoid V^m_t, shaded in blue, has just broken away from the red regions *in the interior of the ball*. Because its extent in the direction $x_i - x_j$ lies within the strip between the two hyperplanes bounding $\mathcal{C}_m(i), \mathcal{C}_m(j)$, it can be shown (see proof of lemma in appendix) that $\|x_i - x_j\|_{(V^m_t)^{-1}}$ is small enough to not exceed roughly 2^{-m}. The more challenging situation is when the ellipsoid V^m_t breaks away from the red regions by *breaching the boundary of the ball* $B(0, D_m)$, as in Figure 1 (b). The **while** loop terminating at this time would not satisfy the objective of controlling $\|x_i - x_j\|_{(V^m_t)^{-1}}$ to within 2^{-m}, since the extent of the ellipsoid in the direction $x_i - x_j$ is larger than the gap between the halfspaces $\mathcal{C}_m(x_i)$ and $\mathcal{C}_m(x_j)$.



(a) 'Easy' case: The blue ellipsoid separates from the halfspaces intersecting the ball (red) by *staying within*.

(b) 'Difficult' case: The green ellipsoid separates from the halfspaces intersecting the ball (red) by *breaching*.

Figure 1: Phase stopping condition in Algorithm 1 ensures $\|x_i - x_j\|_{(V^m_t)^{-1}} \lessapprox 2^{-m}$ after phase m.

2. Bounding the number of arm pulls in a phase. The main bound on the length of the central **while** loop is the following result.

Lemma 5 (Phase length bound) *Let* $B_m := \min_{w \in \mathcal{P}_K} \max_{x,x' \in \mathcal{X}_m, x \neq x'} \|x - x'\|^2_{W^{-1}}$. *There exists* δ_0 *such that* $\forall \delta < \delta_0$, *the phase length* N_m *for any* m *is bounded as :*

$$N_m \leqslant \begin{cases} \max\left\{8B_m \left(2^m\right)^2 \left[\frac{r_m^4 \log K}{(\sqrt{2}-1)^2 C^2}\right] + 1, d\right\}, & \varepsilon_m = \frac{D_m \sqrt{C}}{r_m} \left(\frac{1}{2}\right)^{m+1}, \\ \max\left\{8B_m \left(2^m\right)^2 r_m^2 + 1, d\right\}, & \varepsilon_m = \left(\frac{1}{2}\right)^{m+1}. \end{cases}$$

To prove this we use the no-regret property of both the best-response MIN and the EXP-WTS MAX learner (the full proof appears in the appendix). A key novelty here is the introduction of the ball $B(0, D_m)$ as a technical device to control the 2-norm radius of the final stopped ellipsoid $\mathcal{E}(0, V_t^m, r_m)$ (inequality (i) in the proof).

5. Experiments

We numerically evaluate PELEG, against the algorithms $\mathcal{XY}$-static ([10]), LUCB ([7]), ALBA ([11]), LinGapE ([8]) and RAGE ([5]), for 3 common benchmark settings. The oracle lower bound is also calculated. Note: In our implementation, we ignore the term $B(0, D_m)$ in the phase stopping criterion; this has the advantage of making the criterion check-able in closed form. We simulate independent, $\mathcal{N}(0, 1)$ observation noise in each round. All results reported are averaged over 50 trials. We also empirically observe a 100% success rate in identifying the best arm, although a confidence value of $\delta = 0.1$ is passed in all cases.

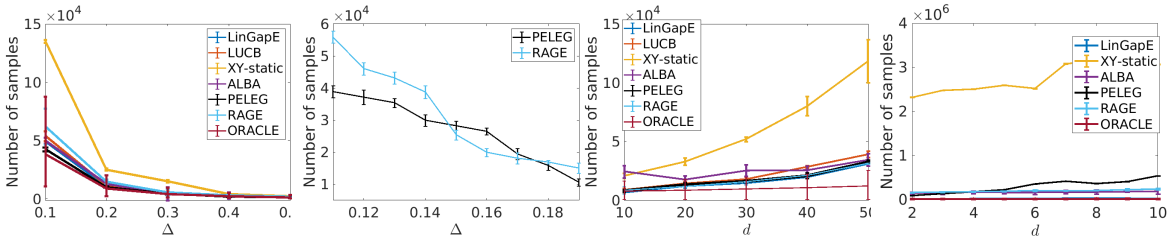
Setting 1: Standard bandit. The arm set is the standard basis $\{e_1, e_2, \ldots, e_5\}$ in 5 dimensions. The unknown parameter θ^* is set to $(\Delta, 0, \ldots, 0)$, where $\Delta > 0$, with Δ swept across $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. As noted in [8], for Δ close to 0, $\mathcal{XY}$-static's essentially uniform allocation is optimal, since we have to estimate all directions equally accurately. However, PELEG performs better (Fig. 2(a)) due to being able to eliminate suboptimal arms earlier instead of uniformly across all arms. Fig. 2(b) compares PELEG and RAGE in the smaller window $\Delta \in [0.11, 0.19]$, where PELEG is found to be competitive (and often better than) RAGE.

Setting 2: Unit sphere. The arms set comprises of 100 vectors sampled uniformly from the surface of the unit sphere $\mathbb{S}^{d-1}$. We pick the two closest arms, say u and v, and then set $\theta^* = u + \gamma(v - u)$ for $\gamma = 0.01$, making u the best arm. We simulate all algorithms over dimensions $d = 10, 20, \ldots, 50$. This setting was first introduced in [11], and PELEG is uniformly competitive with the other algorithms (Fig. 2(c)).

Setting 3: Standard bandit with a confounding arm [10]. We instantiate d canonical basis arms $\{e_1, e_2, \ldots, e_d\}$ and an additional arm $x_{d+1} = (\cos(\omega), \sin(\omega), 0, \ldots, 0)$, $d = 2, \ldots, 10$, with $\theta^* = e_1$ so that the first arm is the best arm. By setting $0 < \omega << 1$, the $d + 1$th arm becomes the closest competitor. Here, the performance critically depends on how much an agent focuses on comparing arm 1 and arm $d + 1$. LinGapE performs very well in this setting, and PELEG and RAGE are competitive with it (Fig. 2(d)).

6. Concluding Remarks

We proposed a new, explicitly described algorithm for BAI in linear bandits, using tools from game theory and no-regret learning to solve minimax games. The algorithm proposed is the first attempt towards instance-optimality without the explicit knowledge of a bound on the parameter space available to the learner. Several interesting directions remain unexplored. Removing the extra logarithmic dependence on $\log(1/\delta)$ is perhaps the most interesting technical question. It is also of great interest to see whether machinery can be extended to solve for best policies in general MDPs.



(a) (b) (c) (d)

Figure 2: Sample complexity performance of LMAB algorithms for 3 different settings: Standard bandit (Figs. a, b), Unit Sphere (Fig. c) and Standard bandit with confounding arm (Fig. d).

Acknowledgement

This work was supported by the Science and Engineering Research Board, Department of Science and Technology [grant no. EMR/2016/002503] and Govt. of Israel's PBC Fellowship and ISF grant number 1380/16." ?

References

[1] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games.* Cambridge University Press, 2006.

[2] Steven de Rooij, Tim van Erven, Peter Grunwald, and Wouter Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15:1281–1316, 2014.

[3] Rémy Degenne, Wouter M. Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games. In *NeurIPS*, 2019.

[4] Rèmy Degenne, Pierre Mènard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020.

[5] Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. In *Advances in Neural Information Processing Systems*, pages 10666–10676, 2019.

[6] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference On Learning Theory*, pages 998–1027, Jun. 2016.

[7] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. In *ICML*, 2012.

[8] Masashi Sugiyama Liyuan Xu, Junya Honda. Fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, 2017.

[9] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011.

[10] Marta Soare, Alessandro Lazaric, and Remi Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems 27*, pages 828–836, 2014.

[11] Chao Tao, Saúl Blanco, and Yuan Zhou. Best arm identification in linear bandits with linear dimension dependency. volume 80 of *Proceedings of Machine Learning Research*, pages 4877–4886. PMLR, 2018.

[12] Mohammadi Zaki, Avinash Mohan, and Aditya Gopalan. Towards optimal and efficient best arm identification in linear bandits. *arXiv preprint arXiv:1911.01695*, 2019.

Appendix A. Glossary of symbols

1. $\mathcal{A}_m^{MAX}$: the EXP-WTS algorithm, used to compute the mixed strategy of the MAX player in each round of PELEG.
2. a^* : the index of the best arm, i.e., $a^* := \text{argmax}_{i\in[K]}\, x_i^T\theta^*$.
3. $B(0, D_m)$: the *closed* ball of radius D_m in $\mathbb{R}^d$, centered at 0.
4. $C = \lambda_{\min}\left(\sum_{x\in\mathcal{X}} xx^T\right)$.
5. $\mathcal{C}_m(x) := \left\{\lambda \in \mathbb{R}^d : \exists x' \in \mathcal{X}_m, x' \neq x | \lambda^T x' \geqslant \lambda^T x + \varepsilon_m\right\}$ is the union of all hyperplanes $\{\lambda \in \mathbb{R}^d | \lambda^T(x' - x) \geqslant \varepsilon_m\}$.
6. $D_m := 2(\sqrt{2}-1)\sqrt{\frac{C}{\max\limits_{x,x'\in\mathcal{X}_m, x\neq x'} \|x-x'\|_2^2 \log K}}$.
7. d : dimension of space in which the feature vectors $x_1, \cdots, x_K$ reside.
8. $\Delta_i = (x^* - x_i)^T\theta^*,\ i \neq a^*$.
9. $\Delta_{\min} = \min_{i\neq a^*} \Delta_i$.
10. δ : maximum allowable probability of erroneous arm selection (a.k.a confidence parameter).
11. $\delta_m = \frac{\delta}{m^2}$.
12. $\mathcal{E}(0, V, r) := \{\lambda \in \mathbb{R}^d \mid \lambda^T V \lambda \leqslant r^2\}$, is the confidence ellipsoid with center 0, shaped by V and r.
13. $K = |\mathcal{X}|$ number of feature vectors.
14. N_m : the length of Phase m.
15. ν_k : rewards from Arm k are all drawn IID from ν_k.
16. $\mathcal{P}(\Omega) := \{p \in [0,1]^{|\Omega|} :\| p \|_1 = 1\}$, the set of all probability measures on some given set Ω.
17. $r_m = \sqrt{8 \log \frac{K^2}{\delta_m}}$.
18. θ^* : fixed but unknown vector in $\mathbb{R}^d$ that parameterizes the means of ν_k, i.e., the mean of ν_k is $x_k^T\theta^*$.
19. n_t^k : number of times Arm k has been sampled up to Round t of PELEG.
20. $\widehat{\theta}_m$: OLS estimate of θ^* at the end of Phase m of PELEG.
21. $V_t^m = \sum_{s\leqslant t} x_s x_s^T$ the design matrix in Round t of Phase m.
22. $W_t = \sum_{x\in\mathcal{X}} w_x xx^T$ the design matrix formed by sampling arms$\sim w \in \mathcal{P}(\mathcal{X})$.
23. $\mathcal{X} = \{x_1, \cdots, x_K\}$, the feature set.
24. $\mathcal{X}_m$ the set of features that survive Phase m of PELEG.

A.1. Sample complexity comparison of BAI algorithms for LMAB in literature

Algorithm	Sample Complexity	Remarks
$\mathcal{XY}$-static [10]	$\mathcal{O}\Big(\frac{d}{\Delta_{\min}}(\ln\frac{1}{\delta}+\ln K+\ln\frac{1}{\Delta_{min}})+d^2\Big)$	Static allocation, worst-case optimal Dependence on d cannot be removed
LinGapE[1] [8]	$\mathcal{O}\Big(dH_0\log\Big(dH_0\log\frac{1}{\delta}\Big)\Big)$	Fully adaptive, sub-optimal in general.
ALBA [11]	$\mathcal{O}\Big(\sum_{i=1}^{d}\frac{1}{\Delta_{(i)}^2}\ln\Big(\frac{K}{\delta}+\ln\frac{1}{\Delta_{min}}\Big)\Big)$	Fully adaptive, sub-optimal in general (see [5])
RAGE [5]	$\mathcal{O}\left(\frac{1}{D_{\theta^*}}\log 1/\Delta_{\min}\log\left(\frac{K^2\log^2 1/\Delta_{min}}{\delta}\right)\right)$	Instance-optimal, but **Minimax oracle required**
LinGame [4]	$\mathcal{O}\left(\frac{\log 1/\delta+K^2d\sqrt{\log 1/\delta}+S^2(\log 1/\delta)^{3/4}}{D_{\theta^*}}\right)$	**Requires knowledge of a upperbound on** $\|\theta^*\|$ Optimal lower bound is achieved only in the limit as $\delta\downarrow 0$
PELEG (this paper)	$\mathcal{O}\left(\frac{\log_2(1/\Delta_{min})}{D_{\theta^*}}\left[\frac{\log^2\big((\log_2(1/\Delta_{min}))^2K^2/\delta\big)}{C^2}\right]\right)$	Instance-optimal (upto a factor of $\log(K/\delta)/C^2$), **Explicitly implementable** **Only requires knowledge of** C **which can be computed in advance**

Table 1: Comparison of Sample complexities achieved by various algorithms for LMAB in the literature. Here S is a bound on 2-norm of θ^* and H_0 is a complicated term defined in terms of a solution to an offline optimization problem in [8].

Appendix B. Details of Algorithm 1

Appendix C. Technical lemmas

C.0.1. DETAILS OF $\mathcal{A}_m^{MAX}$ (EXP-WTS)

We employ the EXP-WTS algorithm to recommend to the MAX player, the arm to be played in round $t > K$. At the start of every phase $m \geqslant 1$, an EXP-WTS subroutine is instantiated afresh, with initial weight vectors to be 1 for each of the K experts. The K experts are taken to be standard unit vectors $(0, 0, \ldots, 0, 1, 0, \ldots, 0)$ with 1 at the k^{th} position, $k \in [K]$. The EXP-WTS subroutine recommends an exponentially-weighted probability distribution over the number of arms, depending upon the weights on each expert. The loss function supplied to update the weights of each expert, is indicated in Step 1 of Algorithm 1.

EXP-WTS requires a bound on the losses (rewards) in order to set its learning parameter optimally. This is ensured by passing an upper-bound of D_m^2 ($\because$ in any Phase m, $\|\lambda\|_2 \leqslant D_m$, see Step **??** of Algorithm 1).

Lemma 6 *In any phase m, at any round $t > K$, $\mathcal{A}_m^{MAX}$ has a regret bounded as*

$$R_t \leqslant \frac{D_m^2}{\sqrt{2}-1}\sqrt{t\log K}.$$

Proof The proof involves a simple modification of the proof of the regret analysis of the EXP-WTS algorithm (see for example, [1]), with loss scaled by $[0, D_m^2]$ followed by the well-known *doubling trick.* ■

Algorithm 1 Phased Elimination Linear Exploration Game (PELEG)

Input: $\mathcal{X}$, a barycentric spanner of $\mathcal{X}$: $\mathcal{X}_B, \delta$.
Init: $m \leftarrow 1, \mathcal{X}_m \leftarrow \mathcal{X}$.
while $\{|\mathcal{X}_m| > 1\}$ **do**
  $\delta_m \leftarrow \frac{\delta}{m^2}$.
  $D_m \leftarrow 2(\sqrt{2}-1)\sqrt{C\left(\max_{x,x' \in \mathcal{X}_m, x \neq x'} \|x - x'\|_2^2 \log K\right)^{-1}}$
  $r_m \leftarrow \sqrt{8 \log\left(K^2/\delta_m\right)}$
  $\varepsilon_m \leftarrow \min\left\{1, D_m\sqrt{C} r_m^{-1}\right\}\left(\frac{1}{2}\right)^{m+1}$.
  Let $\mathcal{C}_m(x) := \left\{\lambda \in \mathbb{R}^d : \exists x' \in \mathcal{X}_m, x' \neq x | \lambda^T x' \geqslant \lambda^T x + \varepsilon_m\right\}$, for $x \in \mathcal{X}_m$.
  Play each arm in $\mathcal{X}_B$ once and collect rewards. Burn-in period
  $\forall k \in [K], n_k^d = \mathbb{1}\{x_k \in \mathcal{X}_B\}, V_d^m \leftarrow \sum_{x \in \mathcal{X}_B} x x^T, t \leftarrow d$.
  Initialize $\mathcal{A}_m^{MAX} \equiv EXP - WTS$ with expert set $\{\widehat{e}_1, \cdots, \widehat{e}_K\} \subset \mathbb{R}^K$ and loss function $l_{t-1}^{MAX}()$. *MAX* player: EXP-WTS
  Phase Stopping Criterion
  while $\left\{\min_{\lambda \in \cup_{x \in \mathcal{X}_m} \mathcal{C}_m(x) \cap B(0, D_m)} \|\lambda\|_{V_t^m}^2 \leqslant r_m^2\right\}$ **do**
    Get w_t from $\mathcal{A}_m^{MAX}$ and form the matrix $W_t = \sum_{k=1}^K w_t^k x_k x_k^T$.
    $\lambda_t \leftarrow \operatorname{argmin}_{\lambda \in \cup_{x \in \mathcal{X}_m} \mathcal{C}_m(x) \cap B(0, D_m)} \|\lambda\|_{W_t}^2$. *MIN* player: Best response
    For $k \in [K], U_t^k := \left(\lambda_t^T x_k\right)^2$.
    Construct loss function $l_t^{MAX}(w) = -w^T U_t$
    Play arm $k_t = \operatorname{argmax}_{k \in [K]} \sum_{s=1}^t w_s^k - n_{t-1}^k$ Tracking
    $n_t^{k_t} \leftarrow n_t^{k_t} + 1$
    Collect sample $Y_t = \theta^{*T} x_{k_t} + \eta_t$
    $V_t^m = V_{t-1}^m + x_{k_t} {x_{k_t}}^T$.
  end
  $N_m \leftarrow t$
  $\widehat{\theta}_m \leftarrow \left(V_{N_m}^m\right)^{-1}\left(\sum_{s=1}^{N_m} Y_s x_{k_s}\right)$ LSE of θ^*
  $\widehat{x}_{m+1} \leftarrow \operatorname{argmax}_{x \in \mathcal{X}_m} \widehat{\theta}_m^T x$. $\mathcal{X}_{m+1} \leftarrow \left\{x \in \mathcal{X}_m | \widehat{\theta}_m^T(\widehat{x}_{m+1} - x) \leqslant 2^{-(m+2)}\right\}$.
  $m \leftarrow m + 1$.
end
Return $\mathcal{X}_m$ Output surviving arm

Appendix D. Proof of Key Lemma

Lemma 7 (Key Lemma) *At the end of each phase $m \geqslant 1$,*

$$\max_{x,x' \in \mathcal{X}_m, x \neq x'} \left\|x - x'\right\|^2_{(V^m_{N_m})^{-1}} \leqslant \frac{\left(\left(\frac{1}{2}\right)^{m+1}\right)^2}{8 \log K^2/\delta_m}.$$

Proof Let $r_m := \sqrt{8 \log K^2/\delta_m}$, for ease of notation. The phase stopping criterion is

$$\text{STOP at round } t \geqslant K \text{ if: } \min_{\lambda \in \bigcup\limits_{x \in \mathcal{X}_m} \mathcal{C}_m(x) \cap B(0, D_m)} \|\lambda\|^2_{(V^m_{N_m})} > r_m^2. \tag{3}$$

Note that the set $\mathcal{C}_m(x)$ depends on the value that ε_m takes in phase m. Depending on the value of ε_m, we divide the analysis into the following two cases.

Case 1. $\varepsilon_m = (1/2)^{m+1}$**.**

In this case $\frac{D_m\sqrt{C}}{r_m} \geqslant 1$. For any phase $m \geqslant 1$, and $t \geqslant 1$, let us define the ellipsoid $\mathcal{E}\left(0, V_t^m, r_m\right) := \left\{\theta \in \mathbb{R}^d : \|\theta\|^2_{V_t^m} \leqslant r_m^2\right\}$. The phase stopping rule at round $t \geqslant K$ is equivalent to :

$$\begin{aligned} \text{STOP if } : \qquad & \mathcal{E}(0, V_t^m, r_m) \bigcap \left\{ \bigcup_{x \in \mathcal{X}_m} \mathcal{C}_m(x) \cap B(0, D_m) \right\} = \emptyset \text{ (empty set)} \\ \Leftrightarrow \qquad & \{\mathcal{E}(0, V_t^m, r_m) \cap B(0, D_m)\} \bigcap \left\{ \bigcup_{x \in \mathcal{X}_m} \mathcal{C}_m(x) \right\} = \emptyset. \end{aligned}$$

However by Rayleigh' inequality[2] followed by the fact that $\frac{D_m\sqrt{C}}{r_m} \geqslant 1$, we have for any $\theta \in \mathcal{E}(0, V_t^m, r_m)$,

$$\|\theta\|_2^2 \leqslant \frac{\|\theta\|^2_{V_t^m}}{\lambda_{min}(V_t^m)} \overset{(*)}{\leqslant} \frac{\|\theta\|^2_{V_t^m}}{\lambda_{min}(\sum\limits_{k=1}^{K} x_k x_k^T)} \leqslant \frac{r_m^2}{C} \leqslant D_m^2.$$

The inequality (*) follows from the following fact: for $t \geqslant K$, $V_t^m = \sum\limits_{k=1}^{K} x_k x_k^T + \sum\limits_{s=K+1}^{t} x_s x_s^T \succcurlyeq \sum\limits_{k=1}^{K} x_k x_k^T$.

$\therefore \mathcal{E}(0, V_t^m, r_m) \subseteq B(0, D_m), \forall t \geqslant K$. Hence the phase stopping rule reduces to,

$$\begin{aligned} \text{STOP if: } \mathcal{E}(0, V_t^m, r_m) \bigcap \left\{ \bigcup_{x \in \mathcal{X}_m} \mathcal{C}_m(x) \right\} = \emptyset \Leftrightarrow & \min_{\lambda \in \cup_{x \in \mathcal{X}_m} \mathcal{C}_m(x)} \|\lambda\|^2_{V_t^m} > r_m^2 \\ \Leftrightarrow & \min_{\lambda \in \bigcup\limits_{(x,x') \in \mathcal{X}_m^2} \{\lambda' : \lambda'^T x' \geqslant \lambda'^T x + (1/2)^{m+1}\}} \|\lambda\|^2_{V_t^m} > r_m^2. \end{aligned}$$

2. for any PSD matrix A and $x \in \mathbb{R}^d$, $\lambda_{min}(A) \leqslant \frac{x^T A x}{x^T x} \leqslant \lambda_{max}(A)$

The above reduction is a minimization problem over union of halfspaces. For any fixed pair $(x, x') \in \mathcal{X}_m^2, x \neq x'$, this is a quadratic optimization problem with linear constraints, which can be explicitly solved using standard Lagrange method.

Lemma 8 (Supporting Lemma for Lem. 7) *For any two arms x and x', we have that*

$$\min_{\lambda \in \{\lambda' : \lambda'^T x' \geqslant \lambda'^T x + \left(\frac{1}{2}\right)^{m+1}\}} \|\lambda\|^2_{V_t^m} = \frac{\left(\left(\frac{1}{2}\right)^{m+1}\right)^2}{\|x - x'\|^2_{(V_t^m)^{-1}}}.$$

Proof The result follows by solving the optimization problem explicitly using the Lagrange multiplier method. ■

By using the above lemma we obtain:

$$\text{STOP if:} \forall x, x' \in \mathcal{X}_m, x \neq x', \left\|x - x'\right\|^2_{(V_t^m)^{-1}} < \frac{\left(\left(\frac{1}{2}\right)^{m+1}\right)^2}{8 \log K^2/\delta_m}.$$

Hence, at round $t = N_m$ we have, $\forall x, x' \in \mathcal{X}_m, x \neq x', \|x - x'\|^2_{(V_{N_m}^m)^{-1}} < \frac{\left(\left(\frac{1}{2}\right)^{m+1}\right)^2}{8 \log K^2/\delta_m}$.

Case 2. $\varepsilon_m = \frac{D_m \sqrt{C}}{r_m} \left(\frac{1}{2}\right)^{m+1}$.

In this case, we have $\frac{D_m \sqrt{C}}{r_m} < 1$.
The phase ends when $\forall (x, x') \in \mathcal{X}_m^2, \min_{\lambda \in \left\{\lambda \in \mathbb{R}^d : \lambda^T x' \geqslant \lambda^T x + \varepsilon_m\right\} \cap B(0, D_m)} \|\lambda\|^2_{V_t^m} > r_m^2$. Let us decompose the optimization problem defining the phase stopping criteria into smaller sub-problems, depending on pair of arms (x, x') in $\mathcal{X}_m^2$. That is, we split the set $\cup_{x \in \mathcal{X}_m} \mathcal{C}_m(x)$ in equation (3), and consider the following problem: for any pair of distinct arms $(x, x') \in \mathcal{X}_m$, consider

$$P(x, x') : \min_{\lambda \in \left\{\lambda \in \mathbb{R}^d : \lambda^T x' \geqslant \lambda^T x + \varepsilon_m\right\} \cap B(0, D_m)} \|\lambda\|^2_{V_t^m}.$$

let $t_{x,x'}$ be the first round when $\min_{\lambda \in \left\{\lambda \in \mathbb{R}^d : \lambda^T x' \geqslant \lambda^T x + \varepsilon_m\right\} \cap B(0, D_m)} \|\lambda\|^2_{V_t^m} > r_m^2$. Clearly, we have $N_m = \max_{(x,x') \in \mathcal{X}_m^2, x \neq x'} t_{x,x'}$. In addition, for any $t \geqslant t_{x,x'}$, $\|\lambda\|^2_{V_t^m} = \lambda^T \left(V^m_{t_{x,x'}} + \sum_{s=t_{x,x'}+1}^t x_s x_s^t\right) \lambda = \|\lambda\|^2_{V^m_{t_{x,x'}}} + \sum_{s=t_{x,x'}+1}^t (x_s^T \lambda)^2 \geqslant \|\lambda\|^2_{V^m_{t_{x,x'}}} > r_m^2$. Hence, once the inequality for a given pair of arms (x, x') is fulfilled it is satisfied for all subsequent rounds. We will now analyze the problem $P(x, x')$ for each pair of arms $(x, x') \in \mathcal{X}_m^2$ individually.

For any $t \geqslant 1$, define $\lambda_t^* \in \underset{\substack{\lambda \in \left\{\lambda \in \mathbb{R}^d : \lambda^T x' \geqslant \lambda^T x + \varepsilon_m\right\} \\ \cap B(0, D_m)}}{\operatorname{argmin}} \|\lambda\|^2_{V_t^m}$. Note that λ_t^* is specific to the pair (x, x').

CLAIM 1. $\lambda_t^{*T}(x' - x) = \varepsilon_m, \forall t \geqslant 1$.

Proof [Proof of Claim 1] For the proof, let's denote $\lambda^* \equiv \lambda_t^*$. Now, suppose that the claim was not true, i.e., $\lambda^{*T}(x' - x) = \varepsilon_m + a$ for some $a > 0$. Let $b = \frac{a}{\lambda^{*T}(x'-x)}$. Then $0 < b < 1$. Define $\lambda' := (1-b)\lambda^*$. By construction, $\lambda'^T(x' - x) = \varepsilon_m$, and $\|\lambda'\|_2 = (1-b)\|\lambda^*\|_2 < \|\lambda^*\|_2$. Hence $\lambda' \in \{\lambda \in \mathbb{R}^d : \lambda^T x' \geqslant \lambda^T x + \varepsilon_m\} \cap B(0, D_m)$. However, $\|\lambda'\|_{V_t^m} = (1-b)\|\lambda^*\|_{V_t^m} < \|\lambda^*\|_{V_t^m}$, which is a contradiction. ■

At $t = t_{x,x'}$, we have $\min_{\substack{\lambda \in \{\lambda \in \mathbb{R}^d : \lambda^T x' \geqslant \lambda^T x + \varepsilon_m\} \\ \cap B(0, D_m)}} \|\lambda\|_{V_t^m}^2 > r_m^2$. We have two sub-cases depending on the 2-norm of λ_t^*.

SUB-CASE 1. $\|\lambda_t^*\|_2 < D_m$.

In this case, we have the following equivalence:

$$\min_{\substack{\lambda \in \{\lambda \in \mathbb{R}^d : \lambda^T x' \geqslant \lambda^T x + \varepsilon_m\} \\ \cap B(0, D_m)}} \|\lambda\|_{V_t^m}^2 \equiv \min_{\lambda \in \{\lambda \in \mathbb{R}^d : \lambda^T x' \geqslant \lambda^T x + \varepsilon_m\}} \|\lambda\|_{V_t^m}^2 .$$

This can be seen by noting that if $\|\lambda_t^*\|_2 < D_m$, then the corresponding Lagrange multiplier is zero. Hence at round $t = t_{x,x'}$, by solving a standard Lagrange optimization problem, we get $\|x - x'\|_{(V_t^m)^{-1}}^2 < \frac{\varepsilon_m^2}{8 \log K^2/\delta_m} = \frac{D_m^2 C}{r_m^2} \frac{\left(\frac{1}{2}\right)^{2(m+1)}}{8 \log K^2/\delta_m} < \frac{\left(\frac{1}{2}\right)^{2(m+1)}}{8 \log K^2/\delta_m}$. The last inequality follows from the hypothesis of Case 2. Since $N_m \geqslant t_{x,x'}$, we get $\|x - x'\|_{(V_{N_m}^m)^{-1}}^2 \leqslant \|x - x'\|_{\left(V_{t_{x,x'}}^m\right)^{-1}}^2 < \frac{\left(\left(\frac{1}{2}\right)^{m+1}\right)^2}{8 \log K^2/\delta_m}$.

SUB-CASE 2. $\|\lambda_t^*\|_2 = D_m$.

The sub-case when $\|\lambda_t^*\|_2 = D_m$, is more involved. Let's enumerate the properties of λ_t^* at $t = t_{x,x'}$ that we have.

- $\|\lambda_t^*\|_{V_t^m}^2 > r_m^2$.
- $\|\lambda_t^*\|_2 = D_m$.
- $\lambda_t^{*T}(x - x') = \varepsilon_m$.

We divide the analysis of this sub-case into two further sub-cases.

SUB-SUB-CASE 1. $r_m^2 \|x - x'\|_{(V_t^m)^{-1}}^2 > \varepsilon_m^2$.

Let $\theta_t^* := \operatorname{argmax}_{\theta \in \mathcal{E}(0, V_t^m, r_m)} \theta^T(x' - x)$. Then, one can verify by solving the maximization problem explicitly that $\theta_t^{*T}(x' - x) = r_m \|x' - x\|_{(V_t^m)^{-1}}$. Let $\theta_1 := \frac{\theta_t^{*T}(x'-x)}{\|x'-x\|_2^2}(x' - x)$. We have the following properties of θ_1 by construction, which are straight-forward to verify.

- $\|\theta_1\|_2 = \frac{r_m \|x'-x\|_{(V_t^m)^{-1}}}{\|x'-x\|_2}$.

- $\theta_1^T(\theta_t^* - \theta_1) = 0$.

Let $\lambda_1 := \frac{\lambda_t^{*T}(x'-x)}{\|x'-x\|_2^2}(x'-x)$. It follows that, $\|\lambda_1\|_2 = \frac{\left|\lambda_t^{*T}(x'-x)\right|}{\|x'-x\|_2} = \frac{\varepsilon_m}{\|x'-x\|_2}$.

Finally, let us define two more quantities. Let $\lambda_2 := \frac{r_m\|x'-x\|_{(V_t^m)^{-1}}}{\varepsilon_m}\lambda_t^*$ and $\theta_2 := \frac{\varepsilon_m}{r_m\|x'-x\|_{(V_t^m)^{-1}}}\theta_t^*$.

We have by the hypothesis of sub-sub-case 1, that $\|\theta_2\|_2^2 < \|\theta_t^*\|_2^2$. This implies that $\theta_2 \in \mathcal{E}(0, V_t^m, r_m)$.

Next, we make the following two claims on the 2-norms of θ_2 and $\theta_t^* - \theta_1$.

CLAIM. $\|\theta_2\|_2 > D_m$.

Proof [Proof of Claim.] Suppose that $\theta_2 \in B(0, D_m)$. By construction, $\theta_2^T(x'-x) = \varepsilon_m$. Hence, $\theta_2 \in \left\{\lambda \in \mathbb{R}^d : \lambda^T x' \geqslant \lambda^T x + \varepsilon_m\right\} \cap B(0, D_m)$. Since, $\theta_2 \in \mathcal{E}(0, V_t^m, r_m)$, this implies that $\|\theta_2\|_{V_t^m} \leqslant r_m$. However, this is a contradiction since at round t, $\min_{\substack{\lambda \in \left\{\lambda^T x' \geqslant \lambda^T x + \varepsilon_m\right\} \\ \cap B(0, D_m)}} > r_m^2$. ■

Hence, we have the following,

$$D_m^2 < \|\theta_2\|_2^2 = \frac{\varepsilon_m^2}{r_m^2\|x'-x\|_{(V_t^m)^{-1}}^2}\|\theta_t^*\|_2^2 = \frac{D_m^2}{\|\lambda_2\|_2^2}\|\theta_t^*\|_2^2 \Rightarrow \|\theta_t^*\|_2^2 > \|\lambda_2\|_2^2 .$$

CLAIM. $\|\theta_t^* - \theta_1\|_2^2 > \|\lambda_2 - \theta_1\|_2^2$.

Proof [Proof of Claim.] First we note that,

$$\begin{aligned}\theta_1^T(\theta_t^* - \lambda_2) &= \frac{\theta_t^{*T}(x'-x)}{\|x'-x\|_2^2}(x'-x)^T\left(\theta_t^* - \frac{r_m\|x'-x\|_{(V_t^m)^{-1}}}{\varepsilon_m}\lambda_t^*\right) \\ &= \frac{r_m^2\|x'-x\|_{(V_t^m)^{-1}}^2}{\|x'-x\|_2^2} - \frac{r_m^2\|x'-x\|_{(V_t^m)^{-1}}^2}{\|x'-x\|_2^2} = 0.\end{aligned}$$

Next observe that,

$$\begin{aligned}\|\theta_t^* - \theta_1\|_2^2 &= \|\theta_t^*\|_2^2 + \|\theta_1\|_2^2 - 2\theta_t^{*T}\theta_1 \\ &= \|\theta_t^*\|_2^2 + \|\theta_1\|_2^2 - 2(\theta_t^* - \lambda_2)^T\theta_1 - 2\theta_1^T\lambda_2 \\ &= \|\theta_t^*\|_2^2 + \|\theta_1\|_2^2 - 2\theta_1^T\lambda_2 \\ &> \|\lambda_2\|_2^2 + \|\theta_1\|_2^2 - 2\theta_1^T\lambda_2 = \|\lambda_2 - \theta_1\|_2^2 .\end{aligned}$$

■

Putting things together we have,

$$\begin{aligned}&\|\theta_t^*\|_2^2 = \|\theta_t^* - \theta_1\|_2^2 + \|\theta_1\|_2^2 \\ \Rightarrow\ &\|\theta_1\|_2^2 = \|\theta_t^*\|_2^2 - \|\theta_t^* - \theta_1\|_2^2 \\ \Rightarrow\ &\|\theta_1\|_2^2 < \|\theta_t^*\|_2^2 - \|\lambda_2 - \theta_1\|_2^2\end{aligned}$$

$$\Rightarrow \frac{r_m^2 \left\|x'-x\right\|^2_{(V_t^m)^{-1}}}{\left\|x'-x\right\|_2^2} < \frac{r_m^2}{C} - r_m^2 \left\|x'-x\right\|^2_{(V_t^m)^{-1}} \left(\frac{D_m^2}{\varepsilon_m^2} - \frac{1}{\left\|x'-x\right\|_2^2}\right)$$

$$\Rightarrow \frac{\left\|x'-x\right\|^2_{(V_t^m)^{-1}}}{\left\|x'-x\right\|_2^2} < \frac{1}{C} - \left\|x'-x\right\|^2_{(V_t^m)^{-1}} \left(\frac{D_m^2}{\varepsilon_m^2} - \frac{1}{\left\|x'-x\right\|_2^2}\right)$$

$$\Rightarrow \frac{\left\|x'-x\right\|^2_{(V_t^m)^{-1}}}{\left\|x'-x\right\|_2^2} < \frac{1}{C} - \left\|x'-x\right\|^2_{(V_t^m)^{-1}} \frac{D_m^2}{\varepsilon_m^2} + \frac{\left\|x'-x\right\|^2_{(V_t^m)^{-1}}}{\left\|x'-x\right\|_2^2}$$

$$\Rightarrow \left\|x'-x\right\|^2_{(V_t^m)^{-1}} < \frac{\varepsilon_m^2}{D_m^2 C} = \frac{D_m^2 C}{r_m^2 D_m^2 C}\left(\frac{1}{2}\right)^{2(m+1)} = \frac{\left(\left(\frac{1}{2}\right)^{m+1}\right)^2}{8\log K^2/\delta_m}.$$

SUB-SUB-CASE 2. $r_m^2 \left\|x-x'\right\|^2_{(V_t^m)^{-1}} \leqslant \varepsilon_m^2$.

This case is trivial as by the hypothesis,

$$\left\|x-x'\right\|^2_{(V_t^m)^{-1}} \leqslant \frac{\varepsilon_m^2}{r_m^2} = \frac{D_m^2 C}{r_m^2}\frac{1}{r_m^2}\left(\left(\frac{1}{2}\right)^{m+1}\right)^2 < \frac{\left(\left(\frac{1}{2}\right)^{m+1}\right)^2}{8\log K^2/\delta_m}.$$

This completes the proof of the key lemma. ∎

Appendix E. Proofs of bounds on phase length

In this section we will provide an upper-bound on the length of any phase $m \geqslant 1$. Clearly, the length of any phase m is governed by the value of ε_m in that phase. Towards this, we have the following lemma.

Lemma 9 (Phase length bound) *Let* $B_m := \min_{w\in\mathcal{P}_K} \max_{x,x'\in\mathcal{X}_m, x\neq x'} \left\|x-x'\right\|^2_{W^{-1}}$. *There exists* δ_0 *such that* $\forall \delta < \delta_0$*, the length* N_m *of any phase* m *is bounded as :*

$$N_m \leqslant \begin{cases} \max\left\{2B_m\left(2^{m+1}\right)^2\left[\frac{r_m^4 \log K}{(\sqrt{2}-1)^2 C^2}\right]+1, d\right\} & \text{if } \varepsilon_m = \frac{D_m\sqrt{C}}{r_m}\left(\frac{1}{2}\right)^{m+1}, \\ \max\left\{2B_m\left(2^{m+1}\right)^2 r_m^2 + 1, d\right\} & \text{if } \varepsilon_m = \left(\frac{1}{2}\right)^{m+1}. \end{cases}$$

Proof Clearly by the design of the algorithm, every phase has a minimum of d phases as $—\mathcal{X}_B| = d$. Recall that $r_m = \sqrt{8\log K^2/\delta_m}$. Let t be the last round in phase m, *before* the phase ends. Then by definition of phase stopping rule (Step 12 of the algorithm),

$$r_m^2 \geqslant \min_{\lambda \in \bigcup_{x\in\mathcal{X}_m} \mathcal{C}_m(x)\cap B(0,D_m)} \left\|\lambda\right\|^2_{V_t^m}$$

$$\overset{(i)}{\geqslant} \min_{\lambda \in \bigcup_{x\in\mathcal{X}_m} \mathcal{C}_m(x)\cap B(0,D_m)} \sum_{s=1}^{t} \left\|\lambda\right\|^2_{W_s} - D_m^2 K\log K$$

$$
\begin{aligned}
&\overset{(ii)}{\geqslant} \sum_{s=1}^{t} \|\lambda_s\|^2_{W_s} - D_m^2 K \log K \\
&\overset{(iii)}{=} \sum_{s=1}^{t}\sum_{k=1}^{K} w_s^k \left(\lambda_s^T x_k\right)^2 - D_m^2 K \log K \\
&\overset{(iv)}{\geqslant} \max_{w\in\mathcal{P}_K} \sum_{s=1}^{t}\sum_{k=1}^{K} w^k \left(\lambda_s^T x_k\right)^2 - \frac{D_m^2}{\sqrt{2}-1}\sqrt{t\log K} - D_m^2 K\log K \\
&= \max_{w\in\mathcal{P}_K} \sum_{s=1}^{t} \|\lambda_s\|^2_W - \frac{D_m^2}{\sqrt{2}-1}\sqrt{t\log K} - D_m^2 K\log K \\
&= t.\max_{w\in\mathcal{P}_K} \sum_{s=1}^{t} \frac{1}{t}\|\lambda_s\|^2_W - \frac{D_m^2}{\sqrt{2}-1}\sqrt{t\log K} - D_m^2 K\log K \\
&\overset{(v)}{\geqslant} t.\max_{w\in\mathcal{P}_K} \min_{q\in\mathcal{P}\left(\bigcup_{x\in\mathcal{X}_m}\mathcal{C}_m(x)\cap B(0,D_m)\right)} \mathbb{E}_{\lambda\sim q}\left[\|\lambda\|^2_W\right] - \frac{D_m^2}{\sqrt{2}-1}\sqrt{t\log K} - D_m^2 K\log K \\
&\overset{(vi)}{\geqslant} t.\max_{w\in\mathcal{P}_K} \min_{q\in\mathcal{P}\left(\bigcup_{x\in\mathcal{X}_m}\mathcal{C}_m(x)\right)} \mathbb{E}_{\lambda\sim q}\left[\|\lambda\|^2_W\right] - \frac{D_m^2}{\sqrt{2}-1}\sqrt{t\log K} - D_m^2 K\log K \\
&\overset{(vii)}{=} t\frac{\varepsilon_m^2}{B_m} - \frac{D_m^2}{\sqrt{2}-1}\sqrt{t\log K} - D_m^2 K\log K.
\end{aligned}
$$

Here the inequalities follow because of (i) lemma **??**, (ii) best-response of MIN player as given in Step 15 of the algorithm, (iii) by definition of W_s in Step 14, (iv) regret property of MAX player (see lemma 6), (v) $\sum_{s=1}^{t}\frac{1}{t}\mathbb{1}\{\lambda=\lambda_s\} \in \mathcal{P}\left(\bigcup_{x\in\mathcal{X}_m}\mathcal{C}_m(x)\cap B(0,D_m)\right)$, (vi) taking minimum over a larger set, and (vii) follows by explicitly solving the minimization problem and recalling the definition of B_m We have that,

$$
t - \frac{B_m}{(\sqrt{2}-1)\varepsilon_m^2} D_m^2\sqrt{\log K}\sqrt{t} \leqslant \frac{B_m}{\varepsilon_m^2} r_m^2 + \frac{B_m}{\varepsilon_m^2} D_m^2 K\log K. \tag{4}
$$

We will do the analysis depending on the value that ε_m takes in phase m.

Case 1. $\varepsilon_m = \frac{D_m\sqrt{C}}{r_m}\left(\frac{1}{2}\right)^{m+1}$**.**

In this case we have, $\frac{D_m\sqrt{C}}{r_m} < 1$. Applying the value of ε_m in eq. (4), we have

$$
\begin{aligned}
t - \frac{B_m}{(\sqrt{2}-1)\varepsilon_m^2} D_m^2\sqrt{\log K}\sqrt{t} &\leqslant \frac{B_m}{\varepsilon_m^2} r_m^2 + \frac{B_m}{\varepsilon_m^2} D_m^2 K\log K \\
\Rightarrow t - \frac{B_m}{(\sqrt{2}-1)C} r_m^2\left(2^{m+1}\right)^2\sqrt{\log K}\sqrt{t} &\leqslant \frac{B_m}{D_m^2 C} r_m^4\left(2^{m+1}\right)^2 + \frac{B_m}{C} r_m^2\left(2^{m+1}\right)^2 K\log K.
\end{aligned} \tag{5}
$$

Let $T_m := \frac{B_m}{D_m^2 C} r_m^4 \left(2^{m+1}\right)^2 + \frac{B_m}{C} r_m^2 \left(2^{m+1}\right)^2 K \log K$. The function $t \mapsto \sqrt{t}$ is a differentiable concave function, meaning for any $t_1, t_2 > 0$, $\sqrt{t_2} \leqslant \sqrt{t_1} + \frac{1}{2\sqrt{t_1}}(t_2 - t_1)$. We therefore have

$$\sqrt{t} \leqslant \sqrt{T_m} + \frac{1}{2\sqrt{T_m}} \left(t - T_m\right).$$

Applying both these to (5) and rearranging, we get

$$t \leqslant T_m \left(1 + \frac{2B_m r_m^2 \left(2^{m+1}\right)^2 \sqrt{\log K}}{2(\sqrt{2}-1)C\sqrt{T_m} - B_m r_m^2 \left(2^{m+1}\right)^2 \sqrt{\log K}}\right).$$

Note that for small enough δ, the first term in the definition of T_m dominates the second term, i.e., there exists $\delta_0^{(1)} > 0$ such that $\forall \delta < \delta_0^{(1)}$,

$$\begin{aligned} \frac{B_m}{C} r_m^2 \left(2^{m+1}\right)^2 K \log K &\leqslant \frac{B_m}{D_m^2 C} r_m^4 \left(2^{m+1}\right)^2, \\ \Rightarrow r_m^2 &\geqslant K \log K D_m^2. \end{aligned} \tag{6}$$

This means that $T_m \leqslant 2\frac{B_m}{D_m^2 C} r_m^4 \left(2^{m+1}\right)^2$, and hence,

$$\begin{aligned} t &\leqslant 2\frac{B_m r_m^4 \left(2^{m+1}\right)^2}{D_m^2 C} \left(1 + \frac{2B_m r_m^2 \left(2^{m+1}\right)^2 \sqrt{\log K}}{2(\sqrt{2}-1)C\sqrt{\frac{B_m r_m^4 (2^{m+1})^2}{D_m^2 C}} - B_m r_m^2 \left(2^{m+1}\right)^2 \sqrt{\log K}}\right) \\ &= 2\frac{B_m r_m^4 \left(2^{m+1}\right)^2}{D_m^2 C} \left(1 + \frac{2D_m \sqrt{B_m \left(2^{m+1}\right)^2 \log K}}{2(\sqrt{2}-1)\sqrt{C} - D_m \sqrt{B_m \left(2^{m+1}\right)^2 \log K}}\right). \end{aligned}$$

We note here the following lower bound on B_m.

$$\begin{aligned} B_m &= \min_{w \in \mathcal{P}_K} \max_{x,x' \in \mathcal{X}_m, x \neq x'} \left\|x - x'\right\|_{W^{-1}}^2 \\ &\geqslant \min_{w \in \mathcal{P}_K} \max_{x,x' \in \mathcal{X}_m, x \neq x'} \lambda_{min}(W^{-1}) \left\|x - x'\right\|_2^2 \\ &= \min_{w \in \mathcal{P}_K} \max_{x,x' \in \mathcal{X}_m, x \neq x'} \frac{1}{\lambda_{max}(W)} \left\|x - x'\right\|_2^2 \\ &\geqslant \min_{w \in \mathcal{P}_K} \max_{x,x' \in \mathcal{X}_m, x \neq x'} \left\|x - x'\right\|_2^2 \\ &= \max_{x,x' \in \mathcal{X}_m, x \neq x'} \left\|x - x'\right\|_2^2. \end{aligned}$$

By using the value of D_m as given in Step 6 of the algorithm, we note that

$$D_m \sqrt{B_m \left(2^{m+1}\right)^2 \log K} = 2(\sqrt{2}-1) \sqrt{\frac{C}{\max_{x,x' \in \mathcal{X}_m, x \neq x'} \|x - x'\|_2^2 \log K}} \sqrt{B_m \left(2^{m+1}\right)^2 \log K}$$

$$\geqslant 2(\sqrt{2}-1)\sqrt{\frac{C}{\max\limits_{x,x'\in\mathcal{X}_m,x\neq x'}\|x-x'\|_2^2\log K}}\sqrt{\max_{x,x'\in\mathcal{X}_m,x\neq x'}\|x-x'\|_2^2\left(2^{m+1}\right)^2\log K}$$
$$=\left(2^{m+1}\right).2(\sqrt{2}-1)\sqrt{C}>2(\sqrt{2}-1)\sqrt{C}.$$

Using this we get a bound on t as:

$$t\leqslant 2\frac{B_m r_m^4\left(2^{m+1}\right)^2}{D_m^2 C}=2\frac{B_m r_m^4\left(2^{m+1}\right)^2}{4(\sqrt{2}-1)^2C^2}\left(\max_{x,x'\in\mathcal{X}_m,x\neq x'}\left\|x-x'\right\|_2^2\log K\right)$$
$$\leqslant 2B_m\left(2^{m+1}\right)^2\left[\frac{r_m^4\log K}{(\sqrt{2}-1)^2C^2}\right].$$

Since, by assumption, $C\equiv\lambda_{min}\left(\sum\limits_{k=1}^{K}x_k x_k^T\right)=\Theta(1)$, we have $t\leqslant\lim O\left(B_m\left(2^{m+1}\right)^2 r_m^4\log K\right),\ \forall\delta<\delta_0^{(1)}$.

Case 2. $\varepsilon_m=\left(\frac{1}{2}\right)^{m+1}$**.**

We have in this case that, $\frac{D_m\sqrt{C}}{r_m}\geqslant 1$. Applying the value of ε_m in eq. (4), we obtain

$$t-\frac{B_m}{(\sqrt{2}-1)\varepsilon_m^2}D_m^2\sqrt{\log K}\sqrt{t}\leqslant\frac{B_m}{\varepsilon_m^2}r_m^2+\frac{B_m}{\varepsilon_m^2}D_m^2K\log K \tag{7}$$
$$\Rightarrow t-\frac{B_m}{(\sqrt{2}-1)}D_m^2\left(2^{m+1}\right)^2\sqrt{\log K}\sqrt{t}\leqslant B_m r_m^2\left(2^{m+1}\right)^2+B_m\left(2^{m+1}\right)^2D_m^2K\log K. \tag{8}$$

Let $T_m:=B_m r_m^2\left(2^{m+1}\right)^2+B_m\left(2^{m+1}\right)^2D_m^2K\log K..$ As before, noting that $t\mapsto\sqrt{t}$ is a concave, differentiable function, we have

$$\sqrt{t}\leqslant\sqrt{T_m}+\frac{1}{2\sqrt{T_m}}\left(t-T_m\right).$$

Applying this to (8) and rearranging, we get

$$t\leqslant T_m\left(1+\frac{2B_m r_m^2\left(2^{m+1}\right)^2\sqrt{\log K}}{2(\sqrt{2}-1)C\sqrt{T_m}-B_m r_m^2\left(2^{m+1}\right)^2\sqrt{\log K}}\right).$$

Going along the same lines as Case 1, we see that there exists $\delta_0^{(2)}>0$ such that $\forall\ \delta<\delta_0^{(2)}$, $T_m\leqslant 2B_m r_m^2\left(2^{m+1}\right)^2$, whence

$$t\leqslant 2B_m\left(2^{m+1}\right)^2 r_m^2.$$

We now set $\delta_0=\min\{\delta_0^{(1)},\delta_0^{(2)}\}$.

■

Appendix F. Justification of elimination criteria

In this section, we argue that progress is made after every phase of the algorithm. We will also show the correctness of the algorithm. Let us define a few terms which will be useful for analysis.

Let $\mathcal{S}_m := \left\{x \in \mathcal{X} : \theta^{*T}(x^* - x) < \frac{1}{2^m}\right\}$. Let $B_m^* := \min_{w \in \mathcal{P}_K} \max_{(x,x') \in \mathcal{S}_m^2, x \neq x'} \|x - x'\|_{W^{-1}}^2$, where $W = \sum_{k=1}^{K} w^k x_k x_k$. Finally, define $T_m^* := \frac{B_m^*}{D_m^2 C} r_m^4 \left(2^{m+1}\right)^2 + \frac{B_m^*}{C} r_m^2 \left(2^{m+1}\right)^2 D_m^2 K \log K$.

Define a sequence of favorable events $\{\mathcal{G}_m\}_{m \geqslant 1}$ as,

$$\mathcal{G}_m := \left\{ N_m \leqslant T_m^* \left(1 + \frac{2B_m^* r_m^2 \left(2^{m+1}\right)^2 \sqrt{\log K}}{2(\sqrt{2}-1)C\sqrt{T_m^*} - B_m^* r_m^2 \left(2^{m+1}\right)^2 \sqrt{\log K}}\right)\right\} \bigcap \{x^* \in \mathcal{X}_{m+1}\} \bigcap \{\mathcal{X}_{m+1} \subseteq \mathcal{S}_{m+1}\}.$$

Remark 10 *Conditioned on the event $\mathcal{G}_{m-1}$, $x^* \in \mathcal{X}_m$ and $\mathcal{X}_m \subseteq \mathcal{S}_m$. Hence, $B_m \leqslant B_m^*$ and $T_m \leqslant T_m^*$. Hence, under the event $\mathcal{G}_{m-1}$,*

$$N_m \leqslant T_m^* \left(1 + \frac{2B_m^* r_m^2 \left(2^{m+1}\right)^2 \sqrt{\log K}}{2(\sqrt{2}-1)C\sqrt{T_m^*} - B_m^* r_m^2 \left(2^{m+1}\right)^2 \sqrt{\log K}}\right) a.s.$$

Note here that the right hand side is a non-random quantity.

Lemma 11 $\mathbb{P}\left[\mathcal{G}_m \mid \mathcal{G}_{m-1}, \ldots, \mathcal{G}_1\right] \geqslant 1 - \delta_m$.

Proof [Proof of lemma 11] Let $y = x_i - x_j$ for some $x_i, x_j \in \mathcal{X}_m, x_i \neq x_j$. Since $\widehat{\theta}_m$ is a least squares estimate of θ^*, conditioned on the realization of the set $\mathcal{X}_m$, $y^T\left(\widehat{\theta}_m - \theta^*\right)$ is a $\|y\|^2_{(V^m_{N_m})^{-1}}$ −sub-Gaussian random variable.

By the key lemma 7 we have that $\|y\|^2_{(V^m_{N_m})^{-1}} \leqslant \frac{1}{8(2^{m+1})^2 \log(K^2/\delta_m)}$. Using property of sub-Gaussian random variables, we write for any $\eta \in (0,1)$,

$$\mathbb{P}\left[\left|y^T\left(\widehat{\theta}_m - \theta^*\right)\right| > \sqrt{2\|y\|^2_{(V^m_{N_m})^{-1}} \log(2/\eta)} \mid \mathcal{G}_{m-1}, \ldots, \mathcal{G}_1\right] \leqslant \eta,$$

which implies that

$$\mathbb{P}\left[\left|y^T\left(\widehat{\theta}_m - \theta^*\right)\right| > \sqrt{\frac{2\log(2/\eta)}{8\left(2^{m+1}\right)^2 \log(K^2/\delta_m)}} \mid \mathcal{G}_{m-1}, \ldots, \mathcal{G}_1\right] \leqslant \eta.$$

Taking intersection over all possible $y \in \mathcal{Y}(\mathcal{X}_m)$, and setting $\eta = 2\delta_m/K^2$, gives

$$\mathbb{P}\left[\forall y \in \mathcal{Y}(\mathcal{X}_m) : \left|y^T\left(\theta^* - \widehat{\theta}_m\right)\right| \leqslant 2^{-(m+2)} \mid \mathcal{G}_{m-1}, \ldots, \mathcal{G}_1\right] > 1 - \delta_m. \tag{9}$$

Conditioned on $\mathcal{G}_{m-1}$, $x^* \in \mathcal{X}_m$. Let $x' \in \mathcal{X}_m$ be such that $x' \notin \mathcal{S}_{m+1}$. Let $y = (x^* - x')$. Then $y \in \mathcal{Y}(\mathcal{X}_m)$. By eq. (9) we have with probability $\geqslant 1 - \delta_m$:

$$\left(x^* - x'\right)^T\left(\theta^* - \widehat{\theta}_m\right) \leqslant 2^{-(m+2)} \Rightarrow \widehat{\theta}_m^T\left(x^* - x'\right) > 2^{-(m+1)} - 2^{-(m+2)} = 2^{-(m+2)}.$$

Thus arm x' will get eliminated after phase m by the elimination criteria of algorithm 1(see step 25 of algorithm 1). Hence $\mathcal{X}_{m+1} \subseteq \mathcal{S}_{m+1}$ w.p. $\geqslant 1-\delta_m$.
Next, we show that conditioned on $\mathcal{G}_{m-1}$, $x^* \in \mathcal{X}_{m+1}$, w.p. $\geqslant 1-\delta_m$. Recall that $\widehat{x}_m$ is the empirically best arm at the end of phase m. Hence $\widehat{x}_m \in \mathcal{X}_{m+1}$. Suppose that x^* gets eliminated at the end of phase m. This means that $\widehat{\theta}_m^T\left(\widehat{x}_m - x^*\right) > 2^{-(m+2)}$. However, by eq. 9,

$$\left(\widehat{x}_m - x^*\right)^T \left(\widehat{\theta}_m - \theta^*\right) \leqslant 2^{-(m+2)} \Rightarrow \theta^{*T}\left(x^* - \widehat{x}_m\right) < 0$$

which is a contradiction, since x^* is a best arm. This, along with note 10 shows that $\mathbb{P}\left[\mathcal{G}_m \;\middle|\; \mathcal{G}_{m-1}, \ldots, \mathcal{G}_1\right] \geqslant 1-\delta_m$.

■

Corollary 12 *Let* $\mathcal{G} := \bigcap_{m\geqslant 1} \mathcal{G}_m$.

$$\mathbb{P}\left[\bigcap_{m\geqslant 1} \mathcal{G}_m\right] \geqslant \prod_{m=1}^{\infty}\left(1-\frac{\delta}{m^2}\right) \geqslant 1-\delta.$$

Corollary 13 *Under the event* $\mathcal{G}$*, the maximum number of phases of Algorithm 1 is bounded by* $\log_2 \frac{1}{\Delta_{min}}$.

Proof Recall that $\Delta_{min} = \min_{x\in\mathcal{X}: x\neq x^*} \theta^{*T}\left(x^* - x\right)$. The proof follows by observing that after any phase m, under the favorable event $\mathcal{G}_{m-1}$, $\mathcal{X}_m \subseteq \mathcal{S}_m$. Since the size $\mathcal{S}_m$ shrinks exponentially with the number of phases $\left(\text{because } \mathcal{S}_m = \left\{x \in \mathcal{X} : \theta^{*T}\left(x^* - x\right) < \frac{1}{2^m}\right\}\right)$, we have the result. ■

Appendix G. Proof of bound on sample complexity

We begin by observing the following useful result from [5]. Recall that

$$D_{\theta^*} = \max_{w\in\Delta_K} \min_{x\in\mathcal{X}, x\neq x^*} \frac{\left(\theta^{*T}\left(x^* - x\right)\right)^2}{\|x^* - x\|^2_{W^{-1}}}$$

Proposition 14 ([5])

$$\sum_{m=1}^{\log_2 \frac{1}{\Delta_{min}}} (2^m)^2 B_m^* \leqslant \frac{4\log_2\left(1/\Delta_{min}\right)}{D_{\theta^*}}.$$

Using proposition 14 we now give a bound on the asymptotic sample complexity of algorithm 1.

Theorem 15 *With probability at least* $1-\delta$*, PEPEG returns the optimal arm after* τ *rounds, with*

$$\tau \leqslant \left(2048\frac{\log_2\left(1/\Delta_{min}\right)}{D_{\theta^*}}\left[\frac{\left(\log\left(\left(\log_2\left(1/\Delta_{min}\right)\right)^2 K^2/\delta\right)\right)^2 \log K}{(\sqrt{2}-1)^2 C^2}\right]\right) + \left(256\frac{\log_2\left(1/\Delta_{min}\right)}{D_{\theta^*}} \log\left(\left(\log_2\left(1/\Delta_{min}\right)\right)^2 K^2/\delta\right)\right).$$

Proof The proof follows from Lemma 9 (phase length bound), Corollary 13 (bound on number of phases), Prop. 14 above and the fact that the sum of several non negative quantities is bigger than their max.

To begin with, the discussion in Sec. F shows that in every phase, $B_m \leqslant B_m^*$. Next, Lemma 9 gives us (w.h.p),

$$\begin{aligned}
\tau &= \sum_{m=1}^{\log_2(1/\Delta_{min})} N_m \\
&\leqslant \sum_{m=1}^{\log_2(1/\Delta_{min})} \max\left\{2B_m^*\left(2^{m+1}\right)^2\left[\frac{r_m^4\log K}{(\sqrt{2}-1)^2C^2}\right], 2B_m^*\left(2^{m+1}\right)^2 r_m^2\right\} + d\log_2\left(1/\Delta_{min}\right) \\
&\leqslant \sum_{m=1}^{\log_2(1/\Delta_{min})} 2B_m^*\left(2^{m+1}\right)^2\left[\frac{r_m^4\log K}{(\sqrt{2}-1)^2C^2}\right] + \sum_{m=1}^{\log_2(1/\Delta_{min})} 2B_m^*\left(2^{m+1}\right)^2 r_m^2 + d\log_2\left(1/\Delta_{min}\right)
\end{aligned}$$

Hence, using the fact that $r_m = \sqrt{8\log K^2/\delta_m}$ and invoking Prop. 14 we get

$$\begin{aligned}
\tau &\leqslant \sum_{m=1}^{\log_2(1/\Delta_{min})} 512B_m^*\left(2^m\right)^2\left[\frac{\left(\log\left(K^2/\delta_m\right)\right)^2\log K}{(\sqrt{2}-1)^2C^2}\right] \\
&\quad + \sum_{m=1}^{\log_2(1/\Delta_{min})} 64B_m^*\left(2^{m+1}\right)^2\log\left(K^2/\delta_m\right) + d\log_2\left(1/\Delta_{min}\right) \\
&\overset{(*)}{\leqslant} 2048\frac{\log_2\left(1/\Delta_{min}\right)}{D_{\theta^*}}\left[\frac{\left(\log\left(\left(\log_2\left(1/\Delta_{min}\right)\right)^2K^2/\delta\right)\right)^2\log K}{(\sqrt{2}-1)^2C^2}\right] \\
&\quad + 256\frac{\log_2\left(1/\Delta_{min}\right)}{D_{\theta^*}}\log\left(\left(\log_2\left(1/\Delta_{min}\right)\right)^2K^2/\delta\right) + d\log_2\left(1/\Delta_{min}\right),
\end{aligned}$$

where $(*)$ follows from the fact that $\frac{K^2}{\delta_m} = \frac{m^2K^2}{\delta} \leqslant \frac{(\log_2(1/\Delta_{min}))^2K^2}{\delta}$. ■

Appendix H. Knowledge of a bound on $\|\theta^*\|$

In this section we give a sketch for the sample complexity in the case when an upperbound on $\|\theta^*\|$ is known. The algorithm is shown here (Alg. 2).

Let S be an upperbound on $\|\theta^*\|_2$. We use a regularized version for the grammian matrix. Note a separate burn-in phase for each phase is not reqiured in this case. For $\lambda > 0$ and any phase $m \geqslant 1$, let $V_t^m = \lambda I + \sum_{s=1}^{t} x_s x_s^T$. The ridge estimate $\widehat{\theta}_m$ of θ^* is given by $\widehat{\theta}_m = {V_{N_m}^m}^{-1}\sum_{s=1}^{N_m} Y_s x_{k_s}$. Let $S_t := \sum_{s=1}^{t} Y_s x_{k_s}$. We sketch the sample complexity analysis of algorithm 2.

- **Concentration.** We first observe the following useful lemma.

Algorithm 2 Phased Elimination Linear Exploration Game-known S (PELEG-S)

Input: $\mathcal{X}, S, \delta$.
Init: $m \leftarrow 1, \mathcal{X}_m \leftarrow \mathcal{X}$.
while $\{|\mathcal{X}_m| > 1\}$ **do**
  $\delta_m \leftarrow \frac{\delta}{m^2}$.
  $(\varepsilon_m, D_m, r_m) =$ **SetParams-S**$(\mathcal{X}, S, m, \delta_m)$
  Let $\mathcal{C}_m(x) := \left\{\lambda \in \mathbb{R}^d : \exists x' \in \mathcal{X}_m, x' \neq x | \lambda^T x' \geqslant \lambda^T x + \varepsilon_m\right\}$, for $x \in \mathcal{X}_m$.
  $\forall k \in [K], n_k^d = 0, V_0^m \leftarrow \lambda I, t \leftarrow 0$.
  Initialize $\mathcal{A}_m^{MAX} \equiv EXP - WTS$ with expert set $\{\widehat{e}_1, \cdots, \widehat{e}_K\} \subset \mathbb{R}^K$ and loss function $l_{t-1}^{MAX}()$. MAX player: EXP-WTS
  Phase Stopping Criterion
  while $\left\{ \min_{\lambda \in \bigcup_{x \in \mathcal{X}_m} \mathcal{C}_m(x) \cap B(0, D_m)} \|\lambda\|_{V_t^m}^2 \leqslant r_m^2 \right\}$ **do**
    $t \leftarrow t + 1$
    Get w_t from $\mathcal{A}_m^{MAX}$ and form the matrix $W_t = \sum_{k=1}^{K} w_t^k x_k x_k^T$.
    $\lambda_t \leftarrow \underset{\lambda \in \cup_{x \in \mathcal{X}_m} \mathcal{C}_m(x) \cap B(0, D_m)}{\operatorname{argmin}} \|\lambda\|_{W_t}^2$. MIN player: Best response
    For $k \in [K], U_t^k := \left(\lambda_t^T x_k\right)^2$.
    Construct loss function $l_t^{MAX}(w) = -w^T U_t$
    Play arm $k_t = \underset{k \in [K]}{\operatorname{argmax}} \sum_{s=1}^{t} w_s^k - n_{t-1}^k$ Tracking
    $n_t^{k_t} \leftarrow n_t^{k_t} + 1$
    Collect sample $Y_t = \theta^{*T} x_{k_t} + \eta_t$
    $V_t^m = V_{t-1}^m + x_{k_t} {x_{k_t}}^T$.
  end
  $N_m \leftarrow t$
  $\widehat{\theta}_m \leftarrow \left(V_{N_m}^m\right)^{-1} \left(\sum_{s=1}^{N_m} Y_s x_{k_s}\right)$ LSE of θ^*
  $\widehat{x}_{m+1} \leftarrow \underset{x \in \mathcal{X}_m}{\operatorname{argmax}} \widehat{\theta}_m^T x$.
  $\mathcal{X}_{m+1} \leftarrow \left\{x \in \mathcal{X}_m | \widehat{\theta}_m^T (\widehat{x}_{m+1} - x) \leqslant 2^{-(m+2)}\right\}$.
  $m \leftarrow m + 1$.
end
Return $\mathcal{X}_m$ Output surviving arm

Lemma 16

$$\left\|\widehat{\theta}_m - \theta^*\right\|_{V_{N_m}^m} \leqslant \|S_{N_m}\|_{V_{N_m}^{m\,-1}} + \sqrt{\lambda} S.$$

Proof Let $V(\lambda) := \lambda I + \sum_{s=1}^{N_m} x_s x_s^T$ and $V_0 := \sum_{s=1}^{N_m} x_s x_s^T$.

$$\left\|\widehat{\theta}_m - \theta^*\right\|_{V(\lambda)} = \left\|V(\lambda)^{-1} S_{N_m} + V(\lambda)^{-1} V_0 \theta^* - \theta^*\right\|_{V(\lambda)}$$

Algorithm 3 SetParams-S$(\mathcal{X}, S, m, \delta_m)$

$D_m \leftarrow 2(\sqrt{2}-1)\sqrt{\frac{\lambda}{\max\limits_{x,x' \in \mathcal{X}_m, x \neq x'} \|x-x'\|_2^2 \log K}}$

$r_m \leftarrow 4\sqrt{d} + 2\sqrt{(\log(K^2/\delta_m))} + \sqrt{\lambda}S.$

$\varepsilon_m \leftarrow \min\left\{1, \frac{D_m\sqrt{(\lambda)}}{r_m}\right\}\left(\frac{1}{2}\right)^{m+1}.$

Return $\varepsilon_m, D_m, r_m.$

$$
\begin{aligned}
&= \left\| V(\lambda)^{-1} S_{N_m} + \left(V(\lambda)^{-1} V_0 - I \right) \theta^* \right\|_{V(\lambda)} \\
&\leqslant \left\| V(\lambda)^{-1} S_{N_m} \right\|_{V(\lambda)} + \left\| \left(V(\lambda)^{-1} V_0 - I \right) \theta^* \right\|_{V(\lambda)} \\
&= \|S_{N_m}\|_{V(\lambda)^{-1}} + \sqrt{\theta^{*T} \left(V(\lambda)^{-1} V_0 - I \right) V(\lambda) \left(V(\lambda)^{-1} V_0 - I \right) \theta^*} \\
&= \|S_{N_m}\|_{V(\lambda)^{-1}} + \sqrt{\theta^{*T} \left(V(\lambda)^{-1} V_0 - I \right) (V_0 - V(\lambda)) \, \theta^*} \\
&= \|S_{N_m}\|_{V(\lambda)^{-1}} + \sqrt{\lambda}\sqrt{\theta^{*T} \left(I - V(\lambda)^{-1} V_0 \right) \theta^*} \leqslant \|S_{N_m}\|_{V_{N_m}^{m}{}^{-1}} + \sqrt{\lambda} S.
\end{aligned}
$$

■

By standard sub-Gaussianity bounds and by observing that $V_t^{-1/2} S_t \sim \mathcal{N}\left(0, V_t^{-1/2} V_t V_t^{-1/2}\right)$, we get the following result.

Lemma 17

$$
\mathbb{P}\left[\left\| \widehat{\theta}_t - \theta^* \right\|_{V_t} \leqslant 4\sqrt{d} + 2\sqrt{(\log(1/\delta_m))} + \sqrt{\lambda} S \right] \geqslant 1 - \delta_m.
$$

Hence for any fixed $y \in \mathbb{R}^d$, with probability $\geqslant 1 - \delta_m$,

$$
\left| \left(\widehat{\theta}_m - \theta^* \right)^T y \right| \leqslant \|y\|_{V_{N_m}^{m}{}^{-1}} \left(4\sqrt{d} + 2\sqrt{(\log(1/\delta_m))} + \sqrt{\lambda} S \right). \tag{10}
$$

Hence we have,

$$
\mathbb{P}\left[\forall y \in \mathcal{Y}(\mathcal{X}_m) : \left| \left(\widehat{\theta}_m - \theta^* \right)^T y \right| \leqslant \|y\|_{V_{N_m}^{m}{}^{-1}} \left(4\sqrt{d} + 2\sqrt{(\log(K^2/\delta_m))} + \sqrt{\lambda} S \right) \right] \geqslant 1 - \delta_m. \tag{11}
$$

- **Uncertainity control.** After every phase $m \geqslant 1$,

$$
\max_{\substack{x,x', x \neq x' \\ (x,x') \in \mathcal{X}_m^2}} \left\| x - x' \right\|_{V_{N_m}^{m}{}^{-1}}^2 \geqslant \left(\frac{(1/2)^{m+1}}{r_m} \right)^2.
$$

The proof is same as D.

- **Phase length bound.** With $r_m := 4\sqrt{d} + 2\sqrt{(\log(K^2/\delta_m))} + \sqrt{\lambda}S$, as defined in Alg 3 we follow similar steps as in E to obtain the following bound on phaselength.

  Lemma 18 (Phase length bound for alg. 2) *Let* $B_m := \min_{w \in \mathcal{P}_K} \max_{x,x' \in \mathcal{X}_m, x \neq x'} \|x - x'\|^2_{W^{-1}}$. *There exists* δ_0 *such that* $\forall \delta < \delta_0$, *the length* N_m *of any phase* m *is bounded as :*

$$N_m \leqslant \begin{cases} 2B_m \left(2^{m+1}\right)^2 \left[\frac{r_m^4 \log K}{(\sqrt{2}-1)^2\lambda^2}\right] + 1 & \text{if } \varepsilon_m = \frac{D_m\sqrt{\lambda}}{r_m}\left(\frac{1}{2}\right)^{m+1}, \\ 2B_m \left(2^{m+1}\right)^2 r_m^2 + 1 & \text{if } \varepsilon_m = \left(\frac{1}{2}\right)^{m+1}. \end{cases}$$

- Finally, putting things together we get the following bound for high probability sample complexity bound.

  Theorem 19 *With probability at least* $1-\delta$, *PEPEG-S returns the optimal arm after* τ *rounds, with*

$$\tau \leqslant \left(C_1 \frac{\log_2(1/\Delta_{min})}{D_{\theta^*}} \left[\frac{\left(\log\left(\left(\log_2(1/\Delta_{min})\right)^2 K^2/\delta\right)\right)^2 \log K + \lambda^2 S^4}{\lambda^2} \right] \right) + \left(C_2 \frac{\log_2(1/\Delta_{min})}{D_{\theta^*}} \log\left(\left(\log_2(1/\Delta_{min})\right)^2 K^2/\delta\right) \right) + C_3 d^2 \log_2(1/\Delta_{min}).$$

Appendix I. Experiment Details

In this section, we provide some details on the implementation of each algorithm. Each experiment was repeated 50 times and the errorbar plots show the mean sample complexity with 1-standard deviations.

- For implementation of PELEG, as mentioned in Sec. 5, we ignore the intersection with the ball $B(0, D_m)$ in the phase stopping criterion. This helps in implementing a closed form expression for the stopping rule. The learning rate parameter in the EXP-WTS subroutine is set to be equal to $(1/D_m^2)\sqrt{8\log K/t}$.
- LinGapE: In the paper of [8] LinGapE was simulated using a greedy arm selection strategy that deviates from the algorithm that is analyzed. We instead implement the LinGapE algorithm in the form that it is analyzed.
- For implementation of RAGE, ALBA and $\mathcal{XY}-$ORACLE, we have used the code provided in the Supplementary material of Fiez et al [5]. We refer the readers to Appendix Sec. F of [5] for further details of their implementations.