Electronic Journal of Statistics
Vol. 15 (2021) 5652–5695
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1909

Regret lower bound and optimal algorithm for high-dimensional contextual linear bandit

Ke Li

Department of Statistics, University of Illinois at Urbana-Champaign
e-mail: kel6@illinois.edu

Yun Yang

Department of Statistics, University of Illinois at Urbana-Champaign
e-mail: yy84@illinois.edu
and

Naveen N. Narisetty

Department of Statistics, University of Illinois at Urbana-Champaign
e-mail: naveen@illinois.edu

Abstract: In this paper, we consider the multi-armed bandit problem with high-dimensional features. First, we prove a minimax lower bound, $\mathcal{O}\big((\log d)^{\frac{\alpha+1}{2}} T^{\frac{1-\alpha}{2}} + \log T\big)$, for the cumulative regret, in terms of horizon T, dimension d and a margin parameter $\alpha \in [0,1]$, which controls the separation between the optimal and the sub-optimal arms. This new lower bound unifies existing regret bound results that have different dependencies on T due to the use of different values of margin parameter α explicitly implied by their assumptions. Second, we propose a simple and computationally efficient algorithm inspired by the general Upper Confidence Bound (UCB) strategy that achieves a regret upper bound matching the lower bound. The proposed algorithm uses a properly centered ℓ_1-ball as the confidence set in contrast to the commonly used ellipsoid confidence set. In addition, the algorithm does not require any forced sampling step and is thereby adaptive to the practically unknown margin parameter. Simulations and a real data analysis are conducted to compare the proposed method with existing ones in the literature.

MSC2020 subject classifications: Primary 62L05.
Keywords and phrases: Contextual linear bandit, high-dimension, minimax regret, sparsity, upper confidence bound.

Received December 2020.

Contents

1. Introduction

In the big data era, the abundance of personalized information enables decision-makers to make individualized decisions for improving the long term reward by incorporating this contextual information. For example, internet marketing companies may utilize users' searching history and demographics to display personalized online advertisements to improve the conversion rate (Abe et al., 2003). In personalized medicine, doctors assign treatments tailored to the individual patient based on the context of the patient's own medical records and genetic information such as biomarkers (Bastani and Bayati, 2019). In these examples, data are often collected sequentially and decision-makers need to pick the best action to maximize the long term reward based on the current predict response in a sequential fashion. Mathematically, this problem can be formulated as a contextual bandit problem (Abe et al., 2003; Chu et al., 2011), where an agent sees a d-dimensional feature vector and has to choose among K actions (arms) at each of the T rounds to maximize the cumulative reward. When the expected reward is a linear function of the features, this problem is known as the linear bandit problem, or the multi-armed bandit problem with linear payoff functions (Abe et al., 2003; Auer, 2003). Under this setting, Dani et al. (2008), Chu et al. (2011) and Abbasi-Yadkori et al. (2011) showed a polynomial dependence of the cumulative regret on dimension d and time horizon T. Precisely, Chu et al. (2011) proved a regret upper bound scaling as $\mathcal{O}(\sqrt{dT})$, while Dani et al. (2008) and Abbasi-Yadkori et al. (2011) proved regret upper bounds scaling as $\mathcal{O}(d\sqrt{T})$.

We focus on the high-dimensional regime where the dimension d of the feature vector can be comparable, or even much larger than the total number of rounds T that plays the role of "sample size" in the statistical perspective. The high-dimensionality of the features and the dependence between the samples induced by the bandit policy make the high dimensional linear bandit problem very challenging both for methodological development and theoretical analysis. In particular, the ordinary least squares (OLS), a traditional statistical method for linear regression and its variants serve as the cornerstones for most linear bandit algorithms, they require a substantial number of samples to accurately estimate the parameters, incurring the polynomial-d dependence of the regret. In the statistical literature, it is well-known that imposing extra lower-dimensional structures such as sparsity on the model improves the minimax risk, or the sample complexity of the best learning algorithm. For example, under the sparsity assumption that the response, or reward, only depends on a small subset of all features with size $s_0 \ll d$, the minimax risk reduces from $\sqrt{d/T}$ to $\sqrt{s_0 \log d/T}$, where s_0 now plays the role of the "effective dimension" and the extra $\log d$ term is due to the uncertainty in feature selection. Unfortunately, sparsity generally does not help improve the regret in linear bandit. Roughly speaking, this failure of sparsity adaptation is due to the exploitation nature of a good bandit problem that tends to choose an optimal arm as often as possible, which prevents the exploration of different directions in the feature space and results in an ill-conditioned design matrix. Interestingly, we find that sparsity adaptation is still possible in the high-dimensional linear bandit when features associated with different arms have a sufficient separation. In particular, we show that under the stochastic assumption that feature vectors are random whose population-level covariance matrices are well-conditioned (see Assumption 2 for a precise definition), the randomness in the features prevents the selected arm to collapse and leads to a regret bound that scales logarithmically in the feature dimension d. Moreover, this randomness in the feature implicitly encourages explorations in the feature space, due to which no *forced exploration* is needed. A simple and computationally efficient algorithm combining the Upper Confidence Bound (UCB) technique (Auer et al., 2002a; Auer, 2003) with the LASSO (Tibshirani, 1996) estimator induces sparsity in the estimated regression parameter and enjoys the improved regret bound.

In this paper, we make a number of contributions to the multi-armed bandit literature. Theoretically, we introduce a relaxation of the margin condition (Assumption 1(b)) from Wang et al. (2018) and Bastani and Bayati (2019), which precisely captures the hardness of the linear bandit problem via a margin parameter $\alpha \in [0, +\infty]$ that controls the "separation" between optimal and sub-optimal arms. With this additional assumption, we prove a regret lower bound of $\mathcal{O}(\log d + \log T)$ for $\alpha = 1$ and $\mathcal{O}((\log d)^{\frac{\alpha+1}{2}} T^{\frac{1-\alpha}{2}})$ for $\alpha \in [0, 1)$. The dependence of our lower bound on the margin parameter α unifies both the polynomial and the logarithmic regret bound dependencies on T in the literature. More details are provided in the discussion of related works in Section 3.4. Methodologically, we propose the use of a properly centered ℓ_1-ball as the confidence set in contrast to the commonly used ellipsoid confidence sets in the UCB algorithm. We show that the ℓ_1-ball confidence set captures the uncertainty and implicitly manages the trade-off between exploration versus exploitation. We also prove that the algorithm is optimal up to a logarithmic factor in T.

Technically, our proof of the upper bound is based on non-asymptotic analysis for the LASSO estimator for dependent data based on a novel proof strategy for verifying the *restricted eigenvalue condition* for the sample covariance matrix under dependent data. In particular, our analysis uses an anti-concentration technique to show that the randomness in the feature vectors facilitates an automatic feature space exploration and prevents feature collapse. Computationally, optimizing the expected reward jointly over the ℓ_1-ball confidence set and the covariate set is equivalent to finding the arm that maximizes the estimated reward plus a term proportional to the ℓ_∞-norm (dual norm to the ℓ_1-norm) of the corresponding feature, which is easy to implement and computationally efficient.

The rest of the paper is organized as follows. In Section 2, we review the background and discuss our assumptions. After that, we present the regret lower bound for policies. In section 3, we propose ℓ_1-confidence ball based algorithm for the high-dimensional linear bandit problem. We present a novel non-asymptotic analysis of the LASSO estimator under the bandit setting, and provide an regret upper bound for the proposed algorithm. Section 4 reports experimental results comparing with some existing bandit algorithms. Detailed proofs of theorems are deferred to the appendix section.

2. Problem formulation

Let T be the total time number of steps, which is allowed to be unknown, and K be the number of possible actions (arms). At each round $t = 1, \dots, T$, for each arm $a \in [K]$, the learner observes one of K feature vectors $X_{a,t} \in \mathbb{R}^d$. We consider the high-dimensional regime where the dimension d of the feature space can be comparable or even much larger than the time horizon T. We adopt a standard random-design framework by assuming that for each arm $a \in [K]$, the observed sequence of feature vectors $\{X_{a,t}\}_{t\geq 0}$ are i.i.d. sub-Gaussian random vectors drawn from an unknown distribution $\mathbb{P}_a$ in $\mathbb{R}^d$. The distributions $\mathbb{P}_a$'s for $a \in [K]$ are allowed to have dependence across different arms. More precisely, recall the following notion of sub-Gaussian random variables.

Definition 1 (Sub-Gaussian random variable)**.** *A real-valued random variable z is σ-sub-Gaussian if $\mathbb{E}[e^{tz}] \leq e^{\sigma^2 t^2/2}$ for every $t \in \mathbb{R}$. This definition implies $\mathbb{E}[z] = 0$ and $Var[z] \leq \sigma^2$.*

Let $\mathcal{D}_t = \{X_{a,t}\}_{a=1}^K$ denote the covariate set consisting of feature vectors corresponding to all the arms at time $t \in [T]$. The feature vectors are allowed to be dependent across arms but are assumed to be independent across different time points t. The decision-maker has access to K arms and each arm yields a random reward, the expected value of which follows a linear function of the associated feature. All arms share the same unknown parameter $\beta_* \in \mathbb{R}^d$ in the linear reward function $\langle \cdot\,, \beta_*\rangle$. In our setup, if the learner pulls arm $a \in [K]$ at time t, then the following random reward is observed $Y_{a,t} := \langle X_{a,t}, \beta_*\rangle + \epsilon_{a,t}$, where $\epsilon_{a,t}$ are independent σ-sub-Gaussian random variables that are also independent of the sequence $\{X_{a,t}\}_{t\geq 0}$ for all $a \in [K]$. Define σ-algebra $\mathcal{F}_t = \sigma(X_{a,1}, \epsilon_{a,1}, \ldots, X_{a,t-1}, \epsilon_{a,t-1}, X_{a,t} : \forall a \in [K])$. Then the sequence of $\{\mathcal{F}_t\}_{t=0}^{\infty}$ is a *filtration*, and the errors $\{\epsilon_{a,t}\}_{t=1}^{\infty}$ for each arm $a \in [K]$ forms a *martingale difference sequence* relative to this filtration.

Our goal is to design a sequential decision-making policy π that learns the parameter β_* over time in order to maximize the expected cumulative reward over the time horizon. Let $\hat{a}_t \in [K]$ denote the arm chosen by the policy π at time $t \in [T]$. Then, an admissible policy π is a sequence of random variables $\hat{a}_1, \hat{a}_2, \cdots$ taking values in the set $\{1, \cdots, K\}$ such that $\hat{a}_t$ is measurable with respect to the σ-algebra $\mathcal{F}_t^+$ generated by the previous feature vectors from each arm and observed rewards of the chosen arms $\{X_{a,s}, Y_{\hat{a}_s,s} : s = 1, \cdots, t-1; a = 1, \cdots, K\}$ and by the current feature vectors $\{X_{1,t}, \cdots, X_{K,t}\}$,

$$\mathcal{F}_t^+ = \sigma\left(X_{a,1}, Y_{\hat{a}_1,1}, \cdots, X_{a,t-1}, Y_{\hat{a}_{t-1},t-1}, X_{a,t} : \forall a \in [K]\right).$$

To characterize the quality of the policy π, we compare it with the oracle policy π^* that uses the knowledge of β_* to choose the best arm $a_t^* = \arg\max_a \langle X_{a,t}, \beta_*\rangle$ maximizing the expected reward at each round $t \in [T]$. Note that under our random-design assumption, a_t^* is also random. To summarize, any policy incurs at each time $t \in [T]$ an expected regret

$$r_t = \mathbb{E}\left[\max_{a\in[K]} \langle X_{a,t}, \beta_*\rangle - \langle X_{\hat{a}_t,t}, \beta_*\rangle\right], \tag{1}$$

where the expectation is taken with respect to the randomness in the feature vector $\{X_{a,t}\}$ and the stochastic reward through $\{\epsilon_{a,t}\}$. Our goal is to seek a policy that minimizes the expected cumulative regret $R_T = \sum_{t=1}^{T} r_t$.

In the high-dimensional regime, the regression coefficient β_*, which is a high-dimensional parameter vector, is assumed to admit a sparse structure, namely, the number of nonzero entries in β_* is much smaller than the ambient dimension d. We denote $S = \{j : \beta_{*,j} \neq 0\}$ as the unknown index set of the nonzero entries of true parameter β_* that collects all influential feature components. Let $s_0 = |S|$ be the number of the nonzero entries in β_*, which satisfies $s_0 \ll d$. For each $t \in [T]$, let $X_t \in \mathbb{R}^{t\times d}$ be the time t-observed design matrix whose rows correspond to all the selected feature vectors $X_{\hat{a}_s,s}$ for $s = 1, \ldots, t$ after $\hat{a}_s$ being pulled, and vector $Y_t = (Y_1, \ldots, Y_t)^T \in \mathbb{R}^t$ collects the observed rewards.

2.1. Assumptions

Let $\hat{\Sigma}_t = X_t^T X_t / t$ denote the sample covariance matrix at the end of round $t \in [T]$. For each matrix $A \in \mathbb{R}^{n\times d}$ and index set $I \subset [d]$, we use A_I to denote the submatrix of A that collects all columns whose indices are in I. Let $\mathbb{S}^{d\times d}$

denote the cone of all $d \times d$ positive semidefinite matrices. We now define the *sparse eigenvalue condition* for a matrix in $\mathbb{S}^{d\times d}$.

Definition 2 (Sparse eigenvalues). *For a $d\times d$ matrix $A \in \mathbb{S}^{d\times d}$ and $1 \le m \le d$, define $\rho_{\min}(m, A)$ and $\rho_{\max}(m, A)$ as the minimum and maximum m-sparse eigenvalues of A if*

$$\rho_{\min}(m, A) \le \frac{\Delta^T A \Delta}{\|\Delta\|_2^2} \le \rho_{\max}(m, A) \textit{ for any } \Delta \neq 0 \textit{ with } \|\Delta\|_0 \le m.$$

Sparse eigenvalue condition is an important requirement in high-dimensional estimation problems (Oliveira, 2016; Rudelson and Zhou, 2013). In particular, it often serves as a bridge for proving the *restricted eigenvalue condition* (c.f. Definition 3 In Section 3.5.2), which is nearly necessary for accurately learning the unknown parameter β_* according to high-dimensional statistics literature Bühlmann and Geer (2011).

We now state two key assumptions for our theoretical analysis.

Assumption 1 (On the true reward). *(a) Sparsity condition: There exist positive constants b and s_0 such that $\|\beta_*\|_1 \le b$ and $\|\beta_*\|_0 \le s_0$.*

(b) Margin condition: There exist positive constants Δ_*, C_1 and $\alpha \in [0, +\infty]$, such that for $h \in \left[C_1\sqrt{\frac{\log d}{T}}, \Delta_*\right]$, $P(\langle X_{a_t^*,t}, \beta_*\rangle \le \max_{b\neq a_t^*}\langle X_{b,t}, \beta_*\rangle + h) \le \frac{1}{2}(\frac{h}{\Delta_*})^\alpha$.

The first part of the assumption requires boundedness and sparsity of the true parameter β_* for its estimability, which are standard requirements in the literature (Abbasi-Yadkori et al., 2011; Bastani and Bayati, 2019).

The second part of the assumption controls the probability of the features falling into an h-neighborhood of the decision boundary. As α increases, the margin condition becomes stronger since the feature vectors are less likely to fall close to the decision boundary. As a result, it will be easier for a bandit policy to distinguish the optimal arm from sub-optimal arms. In particular, in the most extreme case with $\alpha = +\infty$, there is a deterministic positive gap between the rewards of the optimal arm and sub-optimal arms, making the bandit problem easiest. On the contrary, the margin condition becomes weaker as α decreases as it is satisfied by a large class of distributions. For example, with $\alpha = 0$, the feature vectors can have arbitrary distributions around the decision boundary and do not need to exhibit any separation, making the bandit problem hardest. In particular, our condition under $\alpha = 0$ corresponds to the setup considered in Abbasi-Yadkori et al. (2011), where at each round t, the decision-maker is given an arbitrary set $\mathcal{D}_t$ of feature vectors with no separation assumptions on the feature vectors. As a consequence, the cumulative regret of their algorithm scales $\mathcal{O}(\sqrt{T})$ in horizon T.

The margin condition with $\alpha = 1$ has been assumed by Goldenshluger and Zeevi (2013), Wang et al. (2018) and Bastani and Bayati (2019) and will be satisfied when the density of $\langle X_{a,t}, \beta_*\rangle$ is upper bounded uniformly for $\forall a \in [K]$.

Assumption 1(b) is a probabilistic relaxation of the usual gap assumption made for problem dependent bounds in the multi-armed bandit literature. For example, Abbasi-Yadkori et al. (2011) assume that there exists some gap $\Delta > 0$ between the rewards of the best arm and the "second best" arm in the covariate set $\mathcal{D}_t$ such that $\min_{b \neq a_t^*} \langle X_{a_t^*,t} - X_{b,t}, \beta_* \rangle > \Delta$, which corresponds to our margin condition with $\alpha = +\infty$ and $\Delta_* = \Delta$. Assumption 1(b) relaxes this by allowing the gap $\langle X_{a_t^*,t}, \beta_* \rangle - \max_{b \neq a_t^*} \langle X_{b,t}, \beta_* \rangle$ to be arbitrarily close to zero. This assumption resembles the usual margin condition in the classification literature Audibert et al. (2007). The probability of features falling into an h-neighborhood of the decision boundary, $P(\{\langle X_{a_t^*,t}, \beta_* \rangle \leq \langle X_{b,t}, \beta_* \rangle + h,\ \forall b \neq a_t^*\})$, diminishes to zero as $h \to 0_+$, where a_t^* is the optimal arm at time t. In our regret analysis, we will choose a diminishing sequence of h values $\{h_t : t \in [T]\}$ to control the probability of pulling sub-optimal arms for deriving the regret bound in Theorem 2. The constant Δ_* in this assumption is related to the effective number of candidates for optimal arms. For example, if only m many arms have strictly positive probability to be optimal, then under the assumption that the probability density function of $\langle X_{a,t}, \beta_* \rangle$ is bounded for each $a \in [K]$, a simple union argument verifies Assumption (b) with $1/\Delta_* = O(m)$.

We provide a concrete example of distributions for which the margin condition holds with different values of α. We consider a 2-armed bandit problem with the covariate set $\mathcal{D}_t = \{X_t, X_t + Z_t\} \subseteq \mathbb{R}^d$ in each round t, where X_t can follow any distribution and $Z_t = (\zeta_t, 0, \cdots, 0)^T \in \mathbb{R}^d$ has a random variable ζ_t at the first entry and zero otherwise. In addition, we assume that the first entry of the parameter vector β_* is 1. The margin condition for the 2-armed bandit problem can be expressed as

$$P(|\langle Z_t, \beta_* \rangle| \leq h) = P(|\zeta_t| \leq h) \leq \frac{1}{2}(\frac{h}{\Delta_*})^\alpha.$$

Then, for any random variable ζ_t with distribution satisfying the above inequality around 0, the margin condition will hold with α. For example, ζ_t can be drawn from a random signed Beta distribution, i.e., $\zeta_t \sim \varepsilon\,\mathrm{Beta}(\alpha, 1)$, where ε equals to ± 1 with equal probabilities. In this example, parameter $\alpha > 0$ corresponds to the margin parameter. This 2-armed bandit problem is also applicable in real world applications, in which the competing two candidates are very similar and only a few features has marginal difference between them. In such cases, the margin condition can be viewed as an assumption on the difference, e.g., Z_t, between the feature vectors and the parameter α characterizes the level of similarity among different feature vectors.

Assumption 2 (On the features)**.** *(a) Boundedness: For some positive constant $x_{\max}$, $\|X_{a,t}\|_\infty \leq x_{\max}$ for all $a \in [K]$ and $t \in [T]$.*

(b) Anti-concentration: There exists a positive constant ζ such that for each $a \in [K]$, $t \in [T]$, $v \in \mathbb{R}^d$ and $h \in \mathbb{R}_+$, $P(\langle X_{a,t}, v \rangle^2 \leq h\|v\|_2^2) \leq \zeta h$.

(c) Sparse eigenvalues: There exist constants $\Lambda_0 \geq \lambda_0 > 0$ and $C^* > 0$ such that for each $a \in [K]$, the minimum and maximum $C^* s_0$-sparse eigenvalues of the matrix $\Sigma_a := \mathbb{E}[X_{a,t} X_{a,t}^T]$ are bounded below and above by

λ_0^2 and Λ_0^2, respectively. In addition, there exist positive constants $\Lambda_1^2 \geq \phi_0^2 > 0$ such that the minimum and maximum $C^* s_0$-sparse eigenvalues of $\mathbb{E}[X_{a_t^*} X_{a_t^*}^T | \Gamma_t]$ are bounded below and above by ϕ_0^2 and Λ_1^2, respectively. The event Γ_t is defined as $\Gamma_t := \left\{ \langle X_{a_t^*,t}, \beta_* \rangle \geq \max_{b \neq a_t^*} \langle X_{b,t}, \beta_* \rangle + \Delta_* \right\}$.

Assumption 2(a) together with Assumption 1(a) ensures that the maximum regret at each time is bounded, since $|\langle X_{a,t}, \beta_* \rangle| \leq x_{\max} b$ by Cauchy-Schwarz inequality. Assumption 2(b) is an anti-concentration condition that plays a critical role in controlling the estimation accuracy for the regression coefficient β_*. In particular, it ensures directions of random feature vectors $X_{a,t}$ from each arm to be uniformly scattered so that the sample covariance matrix $\hat{\Sigma}_t$ will not concentrated in a single direction such as the direction of β_*. This anti-concentration assumption is mild. For example, it is implied by the boundedness of the probability density function of $\langle X_{a,t}, v \rangle$ for each $v \in \mathbb{R}^d$. Moreover, the parameter ζ only appears in the regret upper bound via a burn-in term and does not affect the leading term. More details can be found in the discussion after Theorem 2.

Assumption 2(c) is made to ensure that the design matrix corresponding to each arm is well-behaved so that the sample covariance matrix $\hat{\Sigma}_t$ satisfies the restricted eigenvalue condition (Definition 3) with a high probability (c.f. Proposition 2 in Section 3.5). The restricted eigenvalue condition is required for the ℓ_1-error bound analysis of LASSO estimator (Proposition 1) and can be implied by the sparse eigenvalue condition (Definition 2) through the transfer principle (Lemma 5). Details of the proof can be found in Section 3.5.2. Here, C^* is a sufficiently large constant that could depend on K and can scale as $\mathcal{O}(K \log K)$ in the worst case. Similar assumptions are commonly adopted in existing work in the high dimensional bandit setting, e.g. Bastani and Bayati (2019) and more generally in the high-dimensional statistics literature, e.g. Candes and Tao (2005) and Wainwright (2019). This assumption is critical for controlling the estimation accuracy for the regression coefficient β_* and in turn the regret analysis.

When samples are independent, several existing results in the literature (e.g. Candes and Tao (2005) and Wainwright (2019)) show a similar condition, *Restricted Isometry Property* (RIP), on the sample covariance matrix $\hat{\Sigma}_t$. However, these results on RIP and the corresponding proofs are not applicable in our case as the independence between samples is violated in the bandit setting. In a nutshell, our proof shows that each vector in the restricted cone $\mathbb{C}(I) = \{v \in \mathbb{R}^d : \|v_{I^c}\|_1 \leq 3\|v_I\|_1\}$ with $I = S$ can be well approximated by a $C^* s_0$-sparse vector in $\mathbb{R}^d$ (c.f. Lemma 6 in the appendix), from which we can construct a covering set of this restricted cone with $C^* s_0$-sparse vectors of controlled cardinality. Then, we apply concentration inequalities after properly decoupling the dependence structure induced by the bandit policy (c.f. the proof of Proposition 2 in the appendix) to obtain a lower bound on the quadratic form $v^T \hat{\Sigma}_t v$ for each sparse vector $v \in \mathbb{R}^d$ in the covering set. Finally, we apply a union bound to prove a lower bound on $v^T \hat{\Sigma}_t v$ uniformly over all $v \in \mathbb{R}^d$ belonging to the restricted cone $\mathbb{C}(S)$, which implies the restricted eigenvalue condition for $\hat{\Sigma}_t$. Our actual proof is even more delicate than the proof outline described here and guarantees that the restricted eigenvalue does not depend on K. The description after Proposition 2 illustrates the high level idea of the actual proof.

To conclude this subsection, we provide a simple example where all our assumptions are satisfied with (ζ, Δ_*) not depending on the dimension d. For each arm, suppose the feature vector follows a truncated multivariate normal distribution truncated on the set $\|X_{a,t}\|_\infty \leq 1$ with different mean vectors μ_a and the same identity matrix as the covariance matrix. Under this setup, the inner product $\langle X_{a,t}, v\rangle$ follows a truncated normal distribution with bounded density function, so the first part of Assumption 2 is satisfied. Assumption 1(b) with $\alpha \in [0,1]$ can be verified similarly as the order statistics of $\{\langle X_{a,t}, \beta_*\rangle : a \in [K]\}$ have bounded density functions. For truncated normal distributions, the optimal region Γ_t is well-spread, so the eigenvalue condition (Assumption 2(b) and 2(c)) is also satisfied. In the same example, non-identity covariance matrix can also be allowed if the eigenvalues of the covariance matrix are lower and upper bounded by some absolute positive constants.

2.2. Regret lower bound

In this section we provide a regret lower bound for any policy π in the linear bandit environment under the assumptions made in Section 2.1. We also provide a sketched proof for Theorem 1 later in Section 3.5.1.

Theorem 1. *Suppose the margin parameter $\alpha \in [0,1]$. Let $\mathcal{P}_\alpha$ be the class of distributions of $(X_{a,t}, Y_{a,t} : a \in [d])$, where the feature vector $X_{a,t}$ is drawn from $\mathbb{P}_a$ and the reward $Y_{a,t} = \langle X_{a,t}, \beta_*\rangle + \epsilon_{a,t}$ for $a \in [K]$ satisfies Assumptions 1–2 with parameter α in Assumption 2(b). Then for large enough horizon T,*

$$\inf_{\pi} \sup_{P_{X,Y} \in \mathcal{P}_\alpha} R_T(\pi, \pi^*) \geq C_L \left((\log d)^{\frac{\alpha+1}{2}} T^{\frac{1-\alpha}{2}} + \log T \right), \tag{2}$$

where C_L is a constant independent of T and d.

Our Theorem 1 provides a more general result compared to existing results in the literature considering both the dimension d and horizon T by leveraging the margin condition (Assumption 1(b)). The theorem shows a poly-logarithmic dependence on dimension d. In addition, it unifies the polynomial and logarithmic dependence on horizon T through the margin parameter α, which describes the hardness of distinguishing the optimal arm from the sub-optimal arms. In particular, when $\alpha = 1$, the lower bound scales as $\Omega(\log d + \log T)$, which is logarithmic in both d and T; when $\alpha = 0$, the lower bound scales as $\Omega\left(\sqrt{T \log d}\right)$, which is poly-logarithmic in d and polynomial in T.

For comparison, Goldenshluger and Zeevi (2009) prove a lower bound of $\Omega(T^{\frac{1-\alpha}{2}})$ for one-dimensional one-armed linear bandit problem, where they assume a margin condition corresponding to Assumption 1(b) with $\alpha \in (0,1]$. In addition, they propose a policy with forced sampling, which achieves optimal regret in horizon T, i.e., $\mathcal{O}(T^{\frac{1-\alpha}{2}})$, when $\alpha \in (0,1)$. Later, Goldenshluger and

Zeevi (2013) prove a lower bound of $\Omega(\log T)$ for low-dimensional linear bandit problem, where they also assume a margin condition which corresponds to the $\alpha = 1$ case in Assumption 1(b). Meanwhile, they propose the OLS-bandit algorithm with $\mathcal{O}(d^3 \log T)$ regret upper bound, which is sub-optimal in d according to Theorem 1. Rusmevichientong and Tsitsiklis (2010) also prove a regret lower bound of $\Omega(d\sqrt{T})$ for low-dimensional linear bandit problem, where the covariate set $\mathcal{D}_t$ is a compact set of infinitely many feature vectors, e.g., the unit sphere $S^{d-1} \subseteq \mathbb{R}^d$. Later, Chu et al. (2011) prove a lower bound of $\Omega(\sqrt{Td})$ for low-dimensional contextual bandit problem, where no margin condition is assumed and the regret lower bound therein corresponds to the worst case when $\alpha = 0$, and propose the LinUCB algorithm with $\tilde{\mathcal{O}}(\sqrt{dT})$ regret upper bound, which is near-optimal according to Theorem 1. Our lower bound result unifies the existing results through the parameter α and applies to the high-dimensional linear bandit problems. More details of the comparison are deferred to Section 3.4.

Remark 1. *The problem formulation considered in Goldenshluger and Zeevi (2013) and Bastani and Bayati (2019) is slightly different from ours, where they assume that at each round, there is one common feature vector shared by all arms and each arm has its own parameter vector. Their formulation can be mathematically reparametrized into our formulation. In particular, they assume that each of the K arms has its own parameter vector $\beta_{*,a} \in \mathbb{R}^d$ for $a \in [K]$. The common feature vector X_t in each round is drawn i.i.d. from a distribution. Then, this formulation can be embedded into our formulation. Specifically, the feature vector $X_{a,t}$ for the a-th arm is a Kd-dimensional vector defined as $X_{a,t} := (0, \cdots, 0, X_t^T, 0, \cdots, 0)^T$, where the a-th block of $X_{a,t}$ is X_t and zero elsewhere. The common parameter $\beta_* := (\beta_{*,1}^T, \cdots, \beta_{*,K}^T)^T$ in our setting is a concatenation of all vectors $\beta_{*,a}$ for $a \in [K]$. Due to the reparametrization, the regret lower bound result, i.e., $\Omega(\log T)$, proved in Goldenshluger and Zeevi (2013) can also applied in our formulation.*

3. ℓ_1-Confidence ball algorithm for high-dimensional linear bandit

In this section we propose a conceptually simple and computationally efficient bandit algorithm. The proposed algorithm is nearly optimal—it matches the regret lower bound when margin parameter α belongs to $[0, 1)$ and is optimal up to a factor of $\log T$ when $\alpha = 1$.

3.1. Optimism in the face of uncertainty

In this section, we motivate our ℓ_1-confidence ball based method under the general framework of optimism in the face of uncertainty (OFU) introduced in Auer (2003) and Abbasi-Yadkori et al. (2011). More precisely, suppose that at the beginning of each round, say $t + 1$, we can construct a confidence set $\mathcal{C}_{t+1} \subseteq \mathbb{R}^d$ for the parameter β_* using the past selected features $X_{\hat{a}_1,1}, \ldots, X_{\hat{a}_t,t}$

and observed rewards $Y_1, \dots, Y_t$ up to round t, so that β_* belongs to $\mathcal{C}_{t+1}$ with high probability. Then a OFU-based algorithm chooses the action-estimate pair

$$(X_{\hat{a}_{t+1},t+1}, \tilde{\beta}_{t+1}) = \underset{(x,\beta)\in\mathcal{D}_{t+1}\times\mathcal{C}_{t+1}}{\arg\max} \langle x, \beta\rangle, \tag{3}$$

which jointly maximizes the expected reward within the covariate set $\mathcal{D}_{t+1} = \{X_{a,t+1}\}_{a=1}^{K}$ and the confidence set $\mathcal{C}_{t+1}$ of β_*. The shape and size of the confidence set $\mathcal{C}_{t+1}$ reflects our uncertainty on the unknown parameter β_*. In high-dimensional linear bandits, we consider to use the LASSO (Tibshirani, 1996) estimator as the center of $\mathcal{C}_{t+1}$, since it adapts the sparsity structure of the parameter. As a remark, other penalized estimators such as MCP (Zhang et al., 2010) and SCAD (Fan and Li, 2001) can also be considered as the center for the confidence set, though need more sophisticated regret analysis.

At the beginning of round $t+1$, we form the LASSO estimator $\hat{\beta}_t$ for β_* by viewing the observations as coming from the linear model $Y = X\beta_* + \epsilon$, with design matrix $X = X_t \in \mathbb{R}^{t\times d}$, response vector $Y = Y_t \in \mathbb{R}^t$, both defined in Section 2, and noise vector $\epsilon \in \mathbb{R}^t$ whose entries are independent σ-sub-Gaussian random variables. Given a regularization parameter $\lambda \geq 0$, the LASSO estimator is the solution of

$$\hat{\beta}_t \in \arg\min_{\beta}\{\mathcal{L}_t(\beta) + \lambda\|\beta\|_1\}, \tag{4}$$

where $\mathcal{L}_t$ denotes the empirical loss function at round t. In our regression framework, the loss function is the least squares objective $\mathcal{L}_t(\beta) = \frac{1}{2t}\|Y - X\beta\|_2^2$. Proposition 1 in Section 3.5.2 suggests a choice of the regularization parameter as $\lambda = \lambda_t = 2\sigma x_{\max}\sqrt{(2\log t + 2\log d)/t}$, where constant $x_{\max}$ (as well as the ϕ_0 in the following display) is defined in Section 2.1.

Having determined the center of the confidence set $\mathcal{C}_t$, it remains to determine its shape. Our proposal of the ℓ_1 ball as the shape of the confidence set is motivated by both theoretical and computational considerations: (i) Theoretically, the shape of the confidence set is important because the cumulative regret can be roughly characterized by the sum of the "volumes" of the confidence sets $\mathcal{C}_t$ over the time horizon. In the low-dimensional regime where $d \ll T$, since all norms in the Euclidean space are equivalent (up to some constant depending on the dimension), the shape of the ball induced by a norm has limited impact on the regret. In contrast, as the dimension d grows, unit balls under different norms have drastically different volumes making the choice of the shape crucial. For example, the volume ratio between unit ℓ_1-ball and unit ℓ_2-ball is $O(d^{-d/2})$, indicating the importance of using the right "norm" for constructing $\mathcal{C}_t$. As we will show in Corollary 1 in the appendix, the use of ℓ_1-ball improves the dependence of the regret on the feature dimension from d to s_0 modulo $\log d$ terms where recall that s_0 is the sparsity level of the model. (ii) Computationally, we want to maintain the convexity of the optimization problem (3) by requiring the confidence set $\mathcal{C}_t$ to be convex. These two factors together motivate us to use a properly centered ℓ_1-ball for $\mathcal{C}_t$, since $q = 1$ is the smallest number maintains the convexity of the ℓ_q-ball over $q \geq 0$.

Corollary 1 in the appendix provides a high probability upper bound to the ℓ_1-distance $\|\hat{\beta}_t - \beta_*\|_1$ between $\hat{\beta}_t$ and β_*, motivating us to use the following confidence set

$$\mathcal{C}_{t+1} = \left\{ \beta \in \mathbb{R}^d : \|\hat{\beta}_t - \beta\|_1 \leq \tau_t \right\}, \tag{5}$$

where $\tau_t = \frac{384 s_0 \sigma x_{\max}}{\phi_0^2} \sqrt{\frac{2\log t + 2\log d}{t}}$.

Note that the size of the confidence set scales poly-logarithmically in feature dimension d (i.e. $\mathcal{O}(\sqrt{\log d})$), which improves upon the $\mathcal{O}(\sqrt{d})$ size of the confidence ellipsoid in Abbasi-Yadkori et al. (2011) that has a polynomial dependence on dimension d. This is because we have leveraged on the sparsity structure of β_* captured by the LASSO estimator.

3.2. ℓ_1-Confidence ball based algorithm

In this subsection, we formally describe our algorithm for the high-dimensional linear bandit problem, which is summarized in Algorithm 1 below. The algorithm takes as inputs an initial regularization parameter λ_0 and an initial diameter of the confidence set τ_0. Recall from the previous subsection that the ℓ_1-confidence ball at time t is $\mathcal{C}_t = \{\beta : \|\beta - \hat{\beta}_{t-1}\|_1 \leq \tau_{t-1}\}$, where $\hat{\beta}_{t-1}$ is the LASSO estimator given by (4), and τ_{t-1} is given in (5). The action selection rule (3) of the algorithm can be reformulated as

$$\begin{aligned} X_{\hat{a}_t, t} &= \arg\max_{x \in \mathcal{D}_t} \left\{ \max_{\beta \in \mathcal{C}_t} \langle x, \beta \rangle \right\} = \arg\max_{x \in \mathcal{D}_t} \left\{ \max_{\|u\|_1 \leq \tau_{t-1}} \langle x, \hat{\beta}_{t-1} + u \rangle \right\} \\ &= \arg\max_{x \in \mathcal{D}_t} \left\{ \langle x, \hat{\beta}_{t-1} \rangle + \tau_{t-1} \|x\|_\infty \right\}. \end{aligned} \tag{6}$$

The criterion function in the second line in equation (6) is composed of a point estimate of the expected reward plus the confidence width of x through its ℓ_∞ norm, making the optimization simple and the algorithm computationally efficient. The confidence width, also known as the exploration bonus, encourages the learner to visit new areas in the feature space to attain an optimal exploration versus exploitation trade-off. Computation-wise, the OLS-bandit algorithm of Goldenshluger and Zeevi (2013) and the LASSO-bandit algorithm of Bastani and Bayati (2019) require a pre-processing step to select a subset of arms based on forced-sample estimation; the OFUL-LS algorithm of Abbasi-Yadkori et al. (2011) needs optimization over an ellipsoid that involves matrix inversion, which makes it computationally expensive.

Remark 2. *Our proposed algorithm reduces to a greedy algorithm when all $X_{a,t}$'s have the same ℓ_∞-norm. However, in cases when $X_{a,t}$'s have very different magnitudes across arms, their sup-norms are able to reflect the prediction uncertainty in the point estimates $\langle X_{a,t}, \hat{\beta}_{t-1} \rangle$'s. As a consequence, our algorithm encourages explorations on those arms with high uncertainty while the greedy algorithm completely ignores this information and tends to exhibit poor performance.*

Algorithm 1 ℓ_1-Confidence Ball Based Algorithm

```
Require: Initial regularization parameter $\lambda_0$, and initial diameter $\tau_0$ for the confidence set.
Initialize $\hat{\beta}_0$ by 0 in $\mathbb{R}^d$.
for $t = 1$ to $T$ do
    Observe $\mathcal{D}_t = \{X_{a,t}\}_{a=1}^K$.
    Obtain $\hat{\beta}_{t-1}$ by solving equation (4), i.e. $\hat{\beta}_{t-1} \in \arg\min_\beta\{\mathcal{L}_{t-1}(\beta) + \lambda_{t-1}\|\beta\|_1\}$.
    Construct confidence set $\mathcal{C}_t = \{\beta \in \mathbb{R}^d : \|\hat{\beta}_{t-1} - \beta\|_1 \leq \tau_{t-1}\}$.
    Select feature $X_{\hat{a}_t,t}$ from equation (6), and observe $Y_t = \langle X_{\hat{a}_t,t}, \beta_* \rangle + \epsilon_t$.
    Set $\lambda_t = \lambda_0\sqrt{\frac{\log d + \log t}{t}}$ and $\tau_t = \tau_0\sqrt{\frac{\log d + \log t}{t}}$.
end for
```

3.3. Regret analysis of proposed algorithm

In this subsection, we provide our theoretical regret analysis. The following theorem provides an upper bound on the cumulative regret for the proposed ℓ_1-confidence ball based algorithm.

Theorem 2. *Suppose that Assumptions 1–2 hold and $K \geq 2$, $d \geq 1$, $\lambda_0 = 2\sqrt{2}\sigma x_{\max}$ and $\tau_0 = \frac{384\sqrt{2}s_0\sigma x_{\max}}{\phi_0^2}$. The expected cumulative regret of our algorithm at time horizon T can be bounded as*

$$\begin{aligned} R_T \leq & I_1^\alpha + \overbrace{C_3 x_{\max} b \Lambda_0^2 (K \log K) s_0 \log d}^{I_2} \\ & + \overbrace{C_4 \sigma^2 x_{\max}^5 b s_0^2 \zeta^2 K^2 \log d / \Delta_*^2}^{I_3} + \overbrace{C_5 x_{\max} b \log T}^{I_4}, \end{aligned} \tag{7}$$

where

$$I_1^\alpha = \begin{cases} C_2 \frac{s_0^{\alpha+1}\sigma^{\alpha+1}x_{\max}^{2(\alpha+1)}}{\Delta_*^\alpha \phi_0^{2(\alpha+1)}} (\log d)^{\frac{\alpha+1}{2}} T^{\frac{1-\alpha}{2}}, & \textit{when} \quad \alpha \in [0,1), \\ C_2 \frac{s_0^2\sigma^2 x_{\max}^4}{\Delta_*\phi_0^4} [\log d + \log T] \log T, & \textit{when} \quad \alpha = 1, \\ C_2 \frac{s_0^2\sigma^2 x_{\max}^4}{(\alpha-1)\Delta_*\phi_0^4} \log d, & \textit{when} \quad \alpha \in (1, +\infty), \\ C_2 \frac{s_0^2\sigma^2 x_{\max}^4}{\Delta_*\phi_0^4} \log d, & \textit{when} \quad \alpha = +\infty, \end{cases}$$

which is the leading term of the regret bound that depends on the margin parameter α defined in Assumption 1(b), and C_2, C_3, C_4, C_5 are positive constants not depending on T, d or α.

The regret bound provided by Theorem 2 for our algorithm shows that the regret of the algorithm grows poly-logarithmically in d, i.e., $R_T = \mathcal{O}((\log d)^{\frac{\alpha+1}{2}})$, when $\alpha \in [0, 1)$; logarithmically in d, i.e., $\mathcal{O}(\log d)$, when $\alpha \in [1, +\infty]$. Meanwhile, the cumulative regret grows polynomially in T, i.e., $R_T = \mathcal{O}(T^{\frac{1-\alpha}{2}})$, when $\alpha \in [0, 1)$; poly-logarithmically in T, i.e., $R_T = \mathcal{O}((\log T)^2)$, when $\alpha = 1$; and logarithmically in T, i.e., $R_T = \mathcal{O}(\log T)$, when $\alpha \in (1, +\infty]$. In addition, for $\alpha \in (1, +\infty)$, the term I_1^α will increase as $\alpha \to 1_+$ and scale as $\mathcal{O}(\log d \log T)$ when $\alpha = 1$ in the extreme case, which fills the gap between

$\alpha = 1$ and $\alpha \in (1, +\infty)$ cases. Comparing the regret upper bound for our proposed method with the minimax regret lower bound established in Section 2.2, we can see that our method is optimal up to a factor of $\log T$ when $\alpha \in [0, 1]$.

Intuitively, $I_2 + I_3$ on the right-hand side of display (7) describes the regret caused by the "burn-in" period of exploring the space of the feature space which does not contribute to the asymptotic regret growth, while the last term I_4 is the cumulative regret when $\mathcal{C}_{t+1}$ does not contain β_* in Corollary 1. In the regret bound, constant Δ_* (defined in Assumption 1(b)) plays the role of the gap parameter as appeared in a typical problem-dependent regret bound (Abbasi-Yadkori et al., 2011). The first term is due to the risk of selecting sub-optimal arms when the features fall near the decision boundary, which is controlled by the margin condition (Assumption 1(b)). From the result of Theorem 2, we can see that when α is larger, the feature vectors are less likely to fall into the close neighborhood of the decision boundary and the bandit environment will become easier to learn. In particular, for the $\alpha = +\infty$ case when there is a deterministic positive gap Δ_* between the rewards of the optimal arm and sub-optimal arms, the regret bound for the algorithm scales polynomially in both dimension d and horizon T, i.e., $\mathcal{O}(\log d + \log T)$. We discuss how our regret bound compares with and provides a unified view of the existing results in Section 3.4.

3.4. Comparisons with existing literature

3.4.1. Problem formulation

The stochastic linear bandit problem was first introduced by Auer (2003), and was subsequently studied by Dani et al. (2008), Rusmevichientong and Tsitsiklis (2010) and Chu et al. (2011). Later, Abbasi-Yadkori et al. (2011) and Abbasi-Yadkori et al. (2012) proposed the OFUL algorithm for both low-dimensional and high-dimensional settings. The model parametrization considered in Abbasi-Yadkori et al. (2011), Dani et al. (2008) and Rusmevichientong and Tsitsiklis (2010) is the same as ours. However, there is one major difference between the two formulations that the covariate set in their problem consists of infinitely many feature vectors, e.g., the unit sphere $S^{d-1} \subseteq \mathbb{R}^d$, while the covariate set $\mathcal{D}_t$ in our paper consists of finitely many feature vectors drawn from some underlying distribution. Moreover, the covariate set in Dani et al. (2008) and Rusmevichientong and Tsitsiklis (2010) does not change over time, therefore the optimal arm will be the same in different rounds. In contrast, the optimal arm in our setting varies along the time due to the randomness in the feature vectors. In addition, the feature space can be explored implicitly in our setting and no explicit exploration phase is needed as in Rusmevichientong and Tsitsiklis (2010).

In another parallel strand of linear bandit research, Goldenshluger and Zeevi (2013) proposed an OLS-bandit algorithm for low-dimensional multi-armed bandit problems, where the formulation is different from ours. In particular, they consider the K-armed bandit problem with bounded linear response, where feature vector at each time is common to all arms and parameters are arm-specific.

After that, Bastani and Bayati (2019) proposed a Lasso-bandit algorithm for the high-dimensional linear bandit problem which requires forced sampling and prior knowledge on their gap parameter. Later, in the same setting as Bastani and Bayati (2019), Wang et al. (2018) introduced an MCP-bandit algorithm which also requires forced sampling and knowledge of gap parameter. As discussed in the remarks after Theorem 1, our problem formulation is also different from that in Bastani and Bayati (2019). The formulation in our paper is meaningful on its own in real world applications. In some real applications, many baseline features are available which have the same effect in the sense of sharing the same regression coefficient β_* across different arms. In such cases, our formulation can help estimate the parameters associated with the common features more accurately by using all the data. However, the formulation in Bastani and Bayati (2019) does not incorporate such structure, which will cause loss of information. For example, in clinical trials or advertisement problems, we can map the interactions between the features of patients (or customers) and treatments (or advertisements) as a feature vectors $X_{a,t}$ for each arm, with one common parameter vector β_* across arms. In these applications, the parameters corresponding to the main effect of patients (or customers) such as age and gender are the same for all arms. Our formulation facilitates sample-efficient estimation of these parameters by combining data from multiple arms. In comparison, the parameters under the formulation in Bastani and Bayati (2019) are estimated separately for different arms. In addition, the difference between the two formulations has been considered in Kim and Paik (2019), where they observe that our formulation is advantageous in such practical applications. In particular, when the number of arms is large, it is not practical to apply the formulation in Bastani and Bayati (2019) due to the large number of parameters. When the set of arms changes over time such as in online advertisement, it is also not feasible to assign a different parameter for every new incoming item as in Bastani and Bayati (2019).

Apart from the above benefits, the motivation for our setting comes from applications such as online advertisements, where there are finitely many available products (arms) for each incoming customer. In addition, available products are usually different for different customers and drawn from some underlying distribution, where each arm is a category of products that is associated with a distribution over the set of products.

3.4.2. Regret bound analysis

In terms of the regret bound analysis, Rusmevichientong and Tsitsiklis (2010) proved a lower bound of $\Omega(d\sqrt{T})$ for both cumulative regret and Bayes risk in low-dimensional linear bandit problem, where the covariate set $\mathcal{D}_t$ is a compact set of infinitely many feature vectors, e.g., the unit sphere $S^{d-1} \subseteq \mathbb{R}^d$. They also proposed an algorithm with matching regret upper bound that alternates between exploration and exploitation phases. Chu et al. (2011) proposed the LinUCB algorithm, which has the regret upper bound $\mathcal{O}(\sqrt{Td\log(T\log(T)/\delta)}$

with probability $1-\delta$, and proved a regret lower bound of $\Omega(\sqrt{dT})$ for the low-dimensional contextual bandit problem with linear payoff, which corresponds to our $\alpha = 0$ case. Abbasi-Yadkori et al. (2011) proved that the expected cumulative regret of the OFUL algorithm scales as $\mathcal{O}(d\sqrt{T})$ in both low-dimensional and high-dimensional settings. In the above literature, the feature vectors are allowed to be chosen arbitrarily by an adversary and the analysis is based on the worst case, leading to the polynomial dependence on dimension d. In contrast, Goldenshluger and Zeevi (2013) proved a cumulative regret of $\mathcal{O}(d^3 \log T)$ for their proposed OLS-bandit algorithm and a regret lower bound of $\Omega(\log T)$ corresponding to the margin condition (Assumption 1(b)) with $\alpha = 1$. However, the cumulative regret of the above algorithms scales polynomially in dimension d, which can be sub-optimal in the high-dimensional setting. The Lasso-bandit algorithm in Bastani and Bayati (2019) and the MCP-bandit algorithm in Wang et al. (2018) are shown to have an improved regret upper bound $\mathcal{O}(\log d)$ in dimension d under the margin condition with $\alpha = 1$, but require forced sampling, which could be costly in some practical cases such as medical and marketing applications. In comparison, our proposed algorithm (Algorithm 1) does not require any forced sampling or knowledge of the gap parameter. Meanwhile, our theoretical results on regret upper bound can be applied to any margin parameter $\alpha \in [0, +\infty]$. On the other hand, the setting of Wang et al. (2018) and Bastani and Bayati (2019) corresponds only to the case of $\alpha = 1$, for which our algorithm achieves the same regret bound dependence on d, i.e., $\mathcal{O}(\log d)$.

3.5. Proof sketch for the main results

In this section, we outline the proofs for Theorem 1 and Theorem 2.

3.5.1. Proof sketch for Theorem 1

In this section, we provide a sketch of the proof for Theorem 1. In Goldenshluger and Zeevi (2013), the author already proved a lower bound of $\Omega(\log T)$ under $\alpha = 1$ case. In this paper, we prove a lower bound on dimension d and also the polynomial dependence on horizon T when margin condition parameter $\alpha < 1$. In order to prove the lower bound result, we follow the standard proof technique by designing a class $\mathcal{P}_{\mathcal{M}}$ of distributions, where $\mathcal{M}$ denotes the parameter space for the regression coefficient vector β in the linear model $Y_{a,t} = \langle X_{a,t}, \beta \rangle + \epsilon_{a,t}$, and then reducing the worst case of cumulative regret in $\mathcal{P}_{\mathcal{M}}$ to the average case by introducing a prior distribution over the class $\mathcal{P}_{\mathcal{M}}$. The novelty of our proof comes from the design of the class $\mathcal{P}_{\mathcal{M}}$, which is a hard-to-learn subset of $\mathcal{P}_\alpha$ that yields the lower bound. Specifically, $\mathcal{P}_{\mathcal{M}}$ is composed of bandit problems with two arms, the first of which is $X^{(1)} = (X_0, X_1, \ldots, X_d)$ and the second is $X^{(2)} = (0, -X_1, \ldots, -X_d)$. Here X_0 is a random variable taking values in $\{-1, 0, 1\}$, while X_i for $i \in \{1, \cdots, d\}$ follows truncated normal distribution of $N(0, 1)$. We take $\mathcal{M} = \{\beta = (\beta_0, \beta_1, \cdots, \beta_d) \in \mathbb{R}^{d+1} : \beta_0 = 1, \beta_u = \theta, \beta_j = 0 \text{ for } j \neq u, u \in [d]\}$ as a set of true parameter β^*, where θ is a tuning parameter controlling the signal strength of the second largest component of β. Here we assume that $\beta_0 = 1$ is known and does not need to be estimated, and the other nonzero entry $\beta_u = c\sqrt{\frac{\log d}{T}}$ with a small constant c. The choice of $\mathcal{M}$ is to make sure that assumptions in Section 2.1 are satisfied and the magnitude of β_u is small so that it is hard for policy to distinguish it from 0. Moreover, we specify a uniform prior distribution on the parameter vector β_* over the set $\mathcal{M}$. According to Bayesian decision rule, we prove that the average cumulative regret can be lower bounded by the regret incurred by the Bayesian optimal policy. By following the common proof technique of reducing the problem of obtaining the estimation lower bound to a multiple hypothesis testing problem, we show that learning β from space $\mathcal{M}$ reduces to a variable selection problem that causes the poly-logarithmic term in d and the polynomial term in T in our new lower bound. Specifically, the multiple hypothesis testing problem is to determine the index u of the nonzero entry β_u from $\{1, \cdots, d\}$ and the optimal Bayes rule is $\hat{u}_{t-1} = \arg\max_{j\in[d]} |\hat{\beta}_{j,t-1}|$, where $\hat{\beta}_{t-1}$ is defined according to the Bayesian policy. Finally, by applying Fano's Lemma (Lemma 1), we derive the lower bound for the probability of error $P(\hat{u}_{t-1} \neq u)$ in multiple hypothesis testing problem that leads to the regret lower bound.

3.5.2. Proof sketch for Theorem 2

In this section, we provide the proof strategy for Theorem 2. In order to prove Theorem 2, we need to prove in Corollary 1 that the ℓ_1-confidence set $\mathcal{C}_t$ in (5) contains β_* with high probability for properly selected regularization parameters. Since we choose LASSO estimator as the center of the confidence set $\mathcal{C}_t$, we need to analyze the error bound of the LASSO estimator. This is challenging since the observed data are highly dependent from each other due to bandit policy, thus we cannot directly apply the standard error analysis of the LASSO estimator. In particular, the restricted eigenvalue condition for $\hat{\Sigma}_t$, which is required for the analysis of LASSO estimator, is hard to prove for dependent data. In this paper, we introduce a convergence result (Proposition 1) on LASSO estimator for dependent data and a novel method to prove the eigenvalue condition (Proposition 2). Finally, we derive the cumulative regret taking into account of the regret incurred under different circumstances.

Firstly, we provide an ℓ_1-error bound for the LASSO estimator. Specifically, Proposition 1 shows that the LASSO estimator falls into an $\mathcal{O}(\sqrt{\log d/t})$ neighborhood around the true parameter β_* with high probability. Before the statement of Proposition 1, we introduce the definition of *restricted eigenvalue condition* for positive semidefinite matrices.

Definition 3 (Restricted eigenvalue condition)**.** *For any set of indices $I \subseteq [d]$ and a positive constant ϕ, define the set*

$$\mathcal{C}(I,\phi) = \{M \in \mathbb{S}^{d\times d} \mid \forall v \in \mathbb{R}^d \ s.t.\ \|v_{I^c}\|_1 \leq 3\|v_I\|_1, \\ \text{we have } \|v\|_1^2 \leq |I|(v^T M v)/\phi^2\}.$$

A matrix that belongs to this set is said to satisfy the restricted eigenvalue condition over index set I with parameter ϕ.

On the event that the sample covariance matrix $\hat{\Sigma}_t$ at time t belongs to $\mathcal{C}(S, \phi)$ for some suitable constant ϕ, Proposition 1 extends the usual LASSO analysis for i.i.d. observations to our bandit setting where observations are dependent due to the sequential decision policy.

Proposition 1. *Let X_i denote the i^{th} row of X and $Y(i)$ denote the i^{th} entry of Y. The sequence $\{X_i, i = 1, \cdots, t\}$ forms an adapted sequence of observations, i.e., X_i depends on past features and their rewards $\{X_{i'}.Y(i')\}_{i'=1}^{i-1}$. Also, assume that all X_i satisfy $\|X_i\|_\infty \leq x_{\max}$ and the regularization parameter $\lambda = 2\sigma x_{\max}\sqrt{(2\log t + 2\log d)/t}$. Then for any $\phi > 0$, we have*

$$P\left[\|\hat{\beta} - \beta_*\|_1 \leq \frac{6s_0\sigma x_{\max}}{\phi^2}\sqrt{\frac{2\log t + 2\log d}{t}}\right]$$
$$\geq 1 - \frac{2}{t} - P[\hat{\Sigma}(X) \notin \mathcal{C}(supp(\beta_*), \phi)],$$

where $\hat{\beta}$ is the LASSO estimator by solving equation (4) with X, Y and regularization parameter λ.

Proposition 1 is a more general version of the LASSO oracle inequality. It is an adaption of Proposition 1 in Bastani and Bayati (2019), where we plug-in specific values for λ and the ℓ_1-error bound for $\hat{\beta}$. For self-containedness, we also include a proof for Proposition 1 in the appendix. It allows for the adapted sequence of observations and errors that are σ-sub-Guassian conditional on all past observations. Moreover, note that the performance of the LASSO estimator dependents on the choice of parameter ϕ and the structure of the sample covariance matrix. This is one of the reasons why multi-armed bandit algorithms such as Bastani and Bayati (2019) has forced sampling step to ensure that the sample covariance matrix $\hat{\Sigma}_t$ is positive definite in some directions with high probability.

Secondly, we prove the restricted eigenvalue condition for $\hat{\Sigma}_t$ based on dependent data. Proposition 2 shows that with the selection rule induced by our method and the construction of the confidence set, the sample covariance matrix will satisfy the restricted eigenvalue condition with high probability.

Proposition 2. *Suppose we construct the confidence set $\mathcal{C}_{t+1} = \{\beta \in \mathbb{R}^d : \|\hat{\beta}_t - \beta\|_1 \leq \frac{6s_0\sigma x_{\max}}{\phi^2}\sqrt{\frac{2\log t + 2\log d}{t}}\}$ with some constant $\phi > 0$ that is large enough. Then if the time horizon t exceeds a certain threshold (i.e. $t \geq \mathcal{O}(\log T + s_0 K \log K \log d + \frac{\sigma^2 x_{\max}^4 s_0^2 \zeta^2 K^2 \log d}{\Delta_*^2})$), the sample covariance matrix $\hat{\Sigma}_t$ are guaranteed to satisfy the compatibility condition with high probability, i.e.*

$$P[\hat{\Sigma}_t \in \mathcal{C}(supp(\beta_*), \phi_*)] \geq 1 - \mathcal{O}(1/t) - \mathcal{O}(e^{-t+s_0\log K \log d}),$$

where $\phi_* = \frac{\phi_0}{8}$.

The detailed proof of Proposition 2 is given in Section B.2. Here we only provide some intuition about how we prove Proposition 2. Firstly, we provide a crude result of the restricted eigenvalue condition for the sample covariance matrix in Proposition 3, i.e., the restricted eigenvalue condition constant is of order $1/K$ where K is the number of arms. The proof of Proposition 3 consists of three main steps:

(i) First, we discretize $\mathcal{H} = \{v \in S^{d-1} : \|v\|_0 \leq Cs_0\}$, where constant $C = \Theta(\zeta\Lambda_0^2 K \log K)$, with a ϵ-net $\mathcal{N}_\epsilon$ (Definition 5). Then, we prove the restricted eigenvalue condition for $\hat{\Sigma}_t$ over vectors in $\mathcal{N}_\epsilon$. In detail, We apply concentration inequalities after properly decoupling the dependence structure induced by the bandit policy to obtain a lower bound of order $1/K$ on the quadratic form $\langle v, \hat{\Sigma}_t v\rangle$ for all $v \in \mathcal{N}_\epsilon$ with high probability. This step is guaranteed by Assumption 2(b), which ensures that the probability of feature vectors locating at the original point is small.

(ii) We prove the restricted eigenvalue condition over $\mathcal{H}$ through approximation from $\mathcal{N}_\epsilon$.

(iii) We prove the eigenvalue condition over the cone $\mathbb{C}_3(S) = \{v \in \mathbb{R}^d : \|v_{S^c}\|_1 \leq 3\|v_S\|_1\}$ (Definition 4), which implies the restricted eigenvalue condition. Here we apply the Transfer Principle (Lemma 5) and carefully choose the constant C as discussed earlier in step (i).

Then in Proposition 2, we prove a refined restricted eigenvalue condition based on Proposition 3. The crucial part is to eliminate the dependence on K which is required in Proposition 3. This is challenging since the observed samples are high dependent and in Proposition 3, we need to consider the worst case across K arms. Instead of applying forced sampling to guarantee the eigenvalue condition of $\hat{\Sigma}_t$ as in Bastani and Bayati (2019), we apply a novel technique to prove the result in Proposition 2. In particular, with the crude restricted eigenvalue condition in Proposition 3, we can show that the optimal arm will be pulled sufficient times, i.e., a positive fraction of time horizon. As a result, the restricted eigenvalue condition in Proposition 2 is satisfied since Assumption 2(c), which ensures that the feature vectors are randomly diverse, i.e., the minimum eigenvalue of the population covariance matrix is positive. Intuitively, this can guarantee that the space of feature vectors can be automatically explored without forced sampling. The main proof of Proposition 2 can be summarized in two steps:

(i) In Lemma 10, we show that based on Proposition 3, the algorithm will pull the optimal arm at a positive fraction of times after some time point. In particular, Assumption 1(b) and 2(c) guarantee that $P(\Gamma_t) \geq 1/2$. Under event Γ_t, there is a positive gap Δ_* between the reward of optimal arm and sub-optimal arms. According to Proposition 1 and Proposition 3, for each arm $X_{a,t}$, the optimal reward of $\tilde{\beta}$ within $\mathcal{C}_t$ in (3) is close to the true reward of $X_{a,t}$. Thus, the algorithm will only select the optimal arm under Γ_t after some time point according to equation (3).

(ii) Assumption 2(c) guarantees the sparse eigenvalue condition with constant

ϕ_0 for optimal arm under event Γ_t. We have shown in Lemma 10 that the optimal arm will be pulled frequently. Then, combining the above results, we prove the restricted eigenvalue condition for $\hat{\Sigma}_t$ with constant $\phi_* = \phi_0/8$, which does not depend on K, even though the observed samples are dependent.

Thirdly, we provide the following corollary which shows that the true parameter β_* lies in the ℓ_1-confidence set $\mathcal{C}_{t+1}$ with high probability. Corollary 1 follows from a direct application of Proposition 1 and Proposition 2, and will play a crucial role in analyzing the cumulative regret of our algorithm.

Corollary 1. *Suppose that Assumptions 1–2 hold and the regularization parameter is chosen as $\lambda_t = 2\sigma x_{\max}\sqrt{(2\log t + 2\log d)/t}$. Then, there exist absolute constants $C, c_1, c_2 > 0$, such that for all $t \geq C\,(\log T + s_0 K \log K \log d + \frac{\sigma^2 x_{\max}^4 s_0^2 \zeta^2 K^2 \log d}{\Delta_*^2})$, with probability at least $1 - c_1 t^{-1} - c_2 e^{-t + s_0 \log K \log d}$, β_* lies in the confidence set $\mathcal{C}_{t+1}$ defined by Equation (5).*

The constants c_1, c_2 in Corollary 1 are absolute constants and can be chosen as $c_1 = 1/(2e)$ and $c_2 = 2e^2$. The size of the confidence set $\mathcal{C}_{t+1}$ scales as $\mathcal{O}(\sqrt{\frac{\log d + \log t}{t}})$ in dimension d and time t. In comparison, the size of the ellipsoid confidence set of Abbasi-Yadkori et al. (2011) centered at the ridge regression estimator scales as $\mathcal{O}(\sqrt{\frac{d \log t}{t}})$, which has a polynomial dependence on the dimension d, and is exponentially larger than the size of our ℓ_1 confidence set.

Finally, we apply the result in Corollary 1 to compute the cumulative regret of the proposed algorithm. As discussed after Theorem 2, for the instant regret incurred in "burn-in" period or when C_t does not contain β_*, we simply bound it with worst-case regret $2bx_{\max}$. When C_t contains β_*, we consider two complementary cases based on event $\Gamma_{\xi_t} = \{\omega \in \Omega : \langle X_{a_t^*,t}, \theta_* \rangle \geq \langle X_{b,t}, \theta_* \rangle + \xi_t,\ \forall b \neq a_t^*\}$ with $\xi_t = 2x_{\max}\tau_t$. We show that when Γ_{ξ_t} holds, the algorithm will only select optimal arm and incurs no regret. Then according to Assumption 1(b), we bound the regret incurred when Γ_t does not hold. In the end, we sum up the regret from different parts to obtain the cumulative regret.

4. Experimental results

We compare our ℓ_1-confidence ball based method with (i) the OFUL-LS method proposed in Abbasi-Yadkori et al. (2011), (ii) the OLS-bandit introduced in Goldenshluger and Zeevi (2013), and (iii) the LASSO-bandit algorithm in Bastani and Bayati (2019) in both synthetic data and real data experiments. The first two methods are not specifically designed for high-dimensional settings. We note that the parametrization of OLS-bandit and LASSO-bandit methods is slightly different from ours, because they treat different arms to have different parameter vectors and one common feature vector. We apply these methods after converting their parametrization into ours.

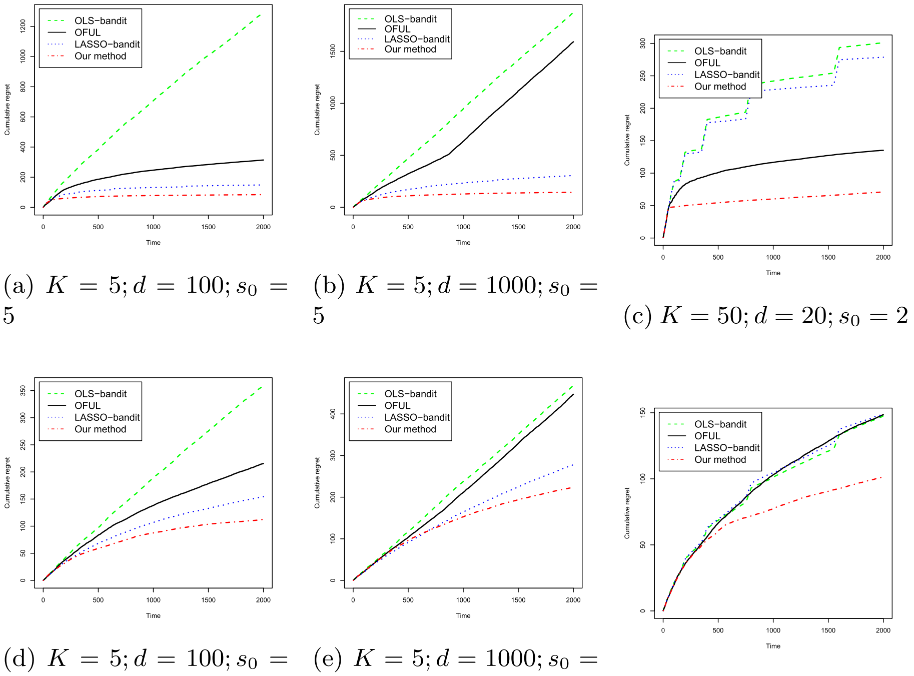
4.1. Synthetic data

In the synthetic data experiment, we consider three scenarios for K, d and s_0: (1) $K = 5$, $d = 100$, $s_0 = 5$; (2) $K = 5$, $d = 1000$, $s_0 = 5$; (3) $K = 50$, $d = 20$, $s_0 = 2$. In each case, a randomly chosen subset of s_0 features is predictive of the reward. At each time t, feature vectors for K arms are i.i.d. generated by truncating a Gaussian distribution $N(0, \Sigma)$ with $\Sigma_{ij} = 0.5^{|i-j|}$ so that $\|X_{a,t}\|_\infty \leq 1$. The error term $\epsilon_{a,t}$ follows a zero mean normal distribution with variance $\sigma^2 = 1$. For this setting, our margin condition holds with $\alpha = 1$ so that we expect a regret growing logarithmically in T. For our method, we choose the initial regularization parameter $\lambda_0 = 0.5$ and diameter $\tau_0 = 1$ and the method is robust to the choice of these tuning parameters. For LASSO and OLS, we choose the forced sampling parameter $q = 1$, the localization parameter $h = 5$ for LASSO-bandit and $h = 1$ for OLS-bandit. For LASSO-bandit, we further set the initial parameters $\lambda_1 = \lambda_{2,0} = 0.5$ and for OFUL-LS, we set $\lambda = 1$ and $\delta = 10^{-4}$.

Figure 1 compares our method with competitors on the aforementioned synthetic data with a time horizon of $T = 2000$. The curves are the average cumulative regrets over 5 trials. We observe that the proposed ℓ_1-confidence ball based algorithm outperforms all the three competing algorithms in all the cases. In cases (1) and (2) where dimension d is from moderately large to high, OLS-bandit and OFUL-LS algorithms do not perform well since they are not designed for high-dimensional settings and fail to capture the sparsity structure. In case (3) where K is large and the feature vector is low-dimensional, our ℓ_1-confidence ball algorithm still outperforms the competitors, while the performance of LASSO-bandit is no longer competitive. One possible reason is that since the number of arms is larger, it may need much more exploration in the feature space than it is allowed. LASSO-bandit does forced sampling only at a limited number of pre-specified times due to which it may not have sufficient exploration to accommodate the large number of arms. In contrast, our algorithm performs an implicit exploration and does not require forced sampling.

4.2. Warfarin dosing data

We now consider data from a real experiment from the healthcare context where a physician needs to determine the optimal warfarin dosage for each patient. The warfarin dataset Consortium (2009) has experimental data on 5528 patients, and contains information on patients' demographics, diagnosis and genetic information. The same dataset is also used by Bastani and Bayati (2019), where they demonstrated the benefits of applying LASSO-bandit algorithms over OFUL-LS, OLS-bandit algorithms, and other low-dimensional methods. Bastani and Bayati (2019) partitioned the dataset into three arms based on the optimal dosage for each patient. However, their partitions are highly unbalanced with one arm having only 13% and another arm having 54% of the patients. Therefore, in our analysis, we evenly partition the dataset into four arms based on the quantiles of the optimal warfarin dosages in the dataset: (1) Level 1: $(0.0, 19.5]$ mg/week,



(a) $K = 5; d = 100; s_0 = 5$

(b) $K = 5; d = 1000; s_0 = 5$

(c) $K = 50; d = 20; s_0 = 2$

(d) $K = 5; d = 100; s_0 = 5$

(e) $K = 5; d = 1000; s_0 = 5$

(f) $K = 50; d = 20; s_0 = 2$

FIG 1. *Comparison of the cumulative regret of the ℓ_1-confidence ball algorithm against competitors. Nonzero components in β_* are generated from Unif$(0, 1)$ in (a)-(c), and Unif$(0, 0.2)$ in (d)-(f). Lower value indicates better performance.*

(2) Level 2: $(19.5, 28.0]$ mg/week, (3): (28.0, 38.5] mg/week, and (4) Level 4: $(38.5, \infty)$ mg/week. The proportion of each arm is nearly 25% after this partition. Following Bastani and Bayati (2019), we construct 94 patient-specific covariates, including intercept and indicators for categorical variables.

We apply our ℓ_1-confidence ball based method to the dataset along with other methods. We include a Doctor's policy for comparison, which always assigns a level 2 dose that has the highest percentage (27.9%) of patients in the warfarin data. We also include oracle policies that assign the optimal dose given the true parameter β_*. Similar to Bastani and Bayati (2019), the "true" parameter vector is estimated using all patient outcomes. We consider two versions of optimal policy: Linear Oracle that estimates β_* using linear regression, and Logit Oracle that estimates β_* using multinomial logistic regression (with four categories).

For the real dataset, the expected regret (1) is not computable, since we do not know the true parameter vector β_*. Instead, we use a surrogate measure which calculates the misclassification rate of assigning optimal dosage to patients. The misclassification rate is calculated as the number of incorrect decisions divided by the number of patients. The lower the expected regret, the

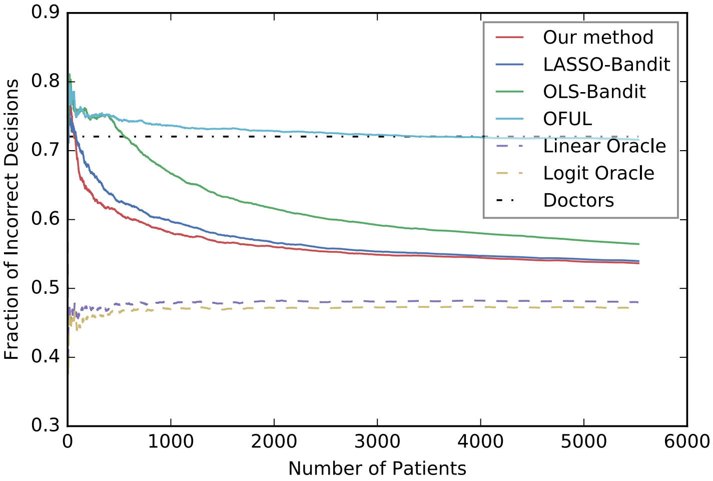


Fig 2. *Comparison of the fraction of incorrect dosing decisions under the oracle, LASSO Bandit, OLS Bandit, OFUL, the ℓ_1-confidence ball algorithm, and doctor policies in the warfarin data. Lower value indicates better performance.*

lower the misclassification rate. We consider 10 random permutations of patients and take the average of the misclassification rate of 10 permutations. Figure 2 illustrates the average fraction of incorrect dosing decisions under different policies. We observe that our method outperforms other competitors except for the oracle polices. Especially, when the number of observations is smaller than 1000, the performance of our method is clearly better than other algorithms, which indicates the benefits of our method when the sample size is relatively small compared to the dimension.

Appendix A: Proof of Theorem 1

We will apply the Bayesian decision theory to show the lower bound of the cumulative regret up to horizon time T. To prove the $\log T$ term in Theorem 1, we can follow the procedure in Goldenshluger and Zeevi (2013). For simplicity, we will focus on the terms in lower bound involving the dimension d.

Suppose we have two arms, the first of which is $X^{(1)} = (X_0, X_1, \dots, X_d)$ and the second is $X^{(2)} = (0, -X_1, \dots, -X_d)$. Here the two arms are not independent from each other. Entry X_0 in the arm vector follows a discrete distribution which will be specified later. The other entries $\{X_1, \dots, X_d\}$ are i.i.d. truncated normal distribution of $N(0,1)$. For simplicity, we first consider the standard normal case, later the proof can be easily applied to the truncated normal case. The common parameter vector is β is from a set $\mathcal{M} = \{\beta \in \mathbb{R}^{d+1} : \beta_0 = 1, \beta_u = \theta, \beta_j = 0 \text{ for } j \neq u, u \in [d]\}$. The sparsity of parameters in set $\mathcal{M}$ is then $s_0 = 2$. Within the set $\mathcal{M}$, the first entry β_0 of the parameter vector is assumed to be known, so we do not need to estimate it. The remaining d entries of the parameter vector has exactly one nonzero entry with value θ. We assume that the parameter vector β is uniformly distributed within the set $\mathcal{M}$. In order to guarantee that the assumptions in section 2.1 hold for the configuration of the parameters in $\mathcal{M}$ and data, we define $\beta_{\min} := \sqrt{\frac{\log d}{T}}$ and $\theta := c\beta_{\min}$, where the constant c is sufficiently small and will be defined momentarily, and

$$\begin{cases} P(X_0 = 0) = C\beta_{\min}^{\alpha} \\ P(X_0 = a) = \frac{1}{2}(1 - C\beta_{\min}^{\alpha}), \quad \text{when } a = \pm 1 \end{cases} \tag{8}$$

Define $\mathcal{P}_\alpha$ as the bandit environment where Assumptions 1–2 hold with constant α in Assumption 1(b), and $\mathcal{P}_{\mathcal{M}}$ be the set of environments defined as above with β from $\mathcal{M}$. We have $\mathcal{P}_{\mathcal{M}} \subseteq \mathcal{P}_\alpha$ according to the configuration of $\mathcal{P}_{\mathcal{M}}$. Moreover, we define $\mathcal{F}_{t-1}^{+} = \sigma(X_1^{(1)}, X_1^{(2)}, Y_1^{(\pi_1)}, \ldots, X_{t-1}^{(1)}, X_{t-1}^{(2)}, Y_{t-1}^{(\pi_t)}, X_t^{(1)}, X_t^{(2)})$ as a σ-algebra. Then for any policy π, the supreme of the cumulative regret of π at horizon T within $\mathcal{P}_\alpha$ can be lower bounded by:

$$\begin{aligned} &\sup\left\{R_T(\pi,\pi^*) : P_{X,Y} \in \mathcal{P}_\alpha\right\} \\ \geq &\sup_{P_{X,Y}\in\mathcal{P}_{\mathcal{M}}} \mathbb{E}\sum_{t=1}^{T} |\beta^T\Delta_{x,t}| \left[I\left\{\beta^T\Delta_{x,t} \geq 0, \pi_t = 2\right\} + I\left\{\beta^T\Delta_{x,t} < 0, \pi_t = 1\right\}\right] \\ \geq &\mathbb{E}\Bigg\{\sum_{t=1}^{T} \Big(\mathbb{E}_\beta\left[\beta^T\Delta_{x,t} I\left\{\beta^T\Delta_{x,t} \geq 0\right\} | \mathcal{F}_{t-1}^{+}\right] I\left\{\pi_t = 2\right\} \\ &\qquad - \mathbb{E}_\beta\left[\beta^T\Delta_{x,t} I\left\{\beta^T\Delta_{x,t} < 0\right\} | \mathcal{F}_{t-1}^{+}\right] I\left\{\pi_t = 1\right\}\Big)\Bigg\}, \end{aligned} \tag{9}$$

where $P_{X,Y} := \left(P_{X^{(1)},Y^{(1)}}, P_{X^{(2)},Y^{(2)}}\right)$ is the joint distribution of the feature vectors and the response variables, and $\Delta_{x,t} := X_t^{(1)} - X_t^{(2)} = (\Delta_{0,t}, \ldots, \Delta_{d,t})$. Here $\mathbb{E}_\beta\left[\cdot|\mathcal{F}_{t-1}^{+}\right]$ denotes the expectation of β in $\mathcal{M}$ conditioned on $\mathcal{F}_{t-1}^{+}$, which is the σ-algebra generated by feature vectors up to time t and observed response variables of the chosed arms up to time $t-1$. Then, according to the Bayesian decision rule in Goldenshluger and Zeevi (2013), the optimal policy is $\hat{\pi}_t = 1$ if

$$\begin{aligned} &\mathbb{E}_\beta\left[\beta^T\Delta_{x,t} I\left\{\beta^T\Delta_{x,t} \geq 0\right\} | \mathcal{F}_{t-1}^{+}\right] \geq -\mathbb{E}_\beta\left[\beta^T\Delta_{x,t} I\left\{\beta^T\Delta_{x,t} < 0\right\} | \mathcal{F}_{t-1}^{+}\right] \\ &\qquad \Longleftrightarrow \qquad \mathbb{E}_\beta\left[\beta^T\Delta_{x,t} | \mathcal{F}_{t-1}^{+}\right] \geq 0 \end{aligned} \tag{10}$$

and $\hat{\pi}_t = 2$ otherwise. Define $\mathcal{F}_{t-1} = \sigma(X_1^{(1)}, X_1^{(2)}, Y_1^{(\pi_1)}, \ldots, X_{t-1}^{(1)}, X_{t-1}^{(2)}, Y_{t-1}^{(\pi_t)})$ as the σ-algebra generated by feature vectors and observed responses up to time $t-1$. We set $\hat{\beta}_{t-1} = \mathbb{E}\left[\beta|\mathcal{F}_{t-1}^{+}\right]$. Since $\Delta_{x,t}$ is independent from $\mathcal{F}_{t-1}$ and β, we have $\mathbb{E}\left[\beta|\mathcal{F}_{t-1}^{+}\right] = \mathbb{E}\left[\beta|\mathcal{F}_{t-1}\right]$. The Bayesian policy is then $\hat{\pi} = \{\hat{\pi}_t : t \in [T]\}$, where $\hat{\pi}_t = I\left\{\hat{\beta}_{t-1}^T\Delta_{x,t} \geq 0\right\} + 2I\left\{\hat{\beta}_{t-1}^T\Delta_{x,t} < 0\right\}$. If we define $u \in [d]$ as the location of the nonzero entry θ in parameter vector β, then according to the distribution of β within $\mathcal{M}$, u is uniform distributed in $[d]$. Then, we have

$$\sup\left\{R_T(\pi,\pi^*) : P_{X,Y} \in \mathcal{P}_\alpha\right\}$$

$$
\begin{aligned}
&\geq \sup\left\{R_T\left(\hat{\pi}, \pi^*\right) : P_{X,Y} \in \mathcal{P}_{\mathcal{M}}\right\}\\
&\geq \sum_{t=1}^{T} \mathbb{E}_{\beta}\mathbb{E}_{X,Y|\beta}\left[|\beta^T\Delta_{x,t}| I\left\{\operatorname{sign}(\beta^T\Delta_{x,t}) \neq \operatorname{sign}(\hat{\beta}_{t-1}^T\Delta_{x,t})\right\}\right]\\
&= \sum_{t=1}^{T} \mathbb{E}_{\beta,u}\mathbb{E}\left[|X_{0,t} + 2\theta X_{u,t}| I\left\{\operatorname{sign}(\beta^T\Delta_{x,t}) \neq \operatorname{sign}(\hat{\beta}_{t-1}^T\Delta_{x,t})\right\}\right]\\
&\geq \sum_{t=1}^{T} \mathbb{E}_{\beta,u}P(X_0=0)\mathbb{E}_{X,Y|\beta}\left[|2\theta X_{u,t}| I\left\{\operatorname{sign}(\beta^T\Delta_{x,t}) \neq \operatorname{sign}(\hat{\beta}_{t-1}^T\Delta_{x,t})\right\} \,\middle|\, X_0=0\right]\\
&\geq \sum_{t=1}^{T} \mathbb{E}_{\beta,u}P(X_0=0)\frac{1}{4}\mathbb{E}_{X,Y|\beta}\left[|2\theta X_{u,t}| I\left\{\hat{\beta}_{t-1}^T\Delta_{x,t} < 0\right\} \,\middle|\, X_0=0, \beta^T\Delta_{x,t} \geq 0\right]\\
&\geq \sum_{t=1}^{T} \mathbb{E}_{\beta,u}\frac{1}{2}Cc\beta_{\min}^{\alpha+1}\mathbb{E}_{X,Y|\beta}\left[I\left\{1 < X_{u,t} < 2,\ \hat{\beta}_{t-1}^T\Delta_{x,t} < 0\right\} \,\middle|\, X_0=0, \beta^T\Delta_{x,t} \geq 0\right]\\
&= \sum_{t=1}^{T} \frac{1}{2}Cc\beta_{\min}^{\alpha+1}\mathbb{E}_{\beta,u}\left[P_{X,Y|\beta}\left(1 < X_{u,t} < 2,\ \hat{\beta}_{t-1}^T\Delta_{x,t} < 0\right)\right]. \qquad (11)
\end{aligned}
$$

The third inequality above is by the conditional expectation with respect to the event $\{X_0 = 0\}$, and the fourth inequality is by the symmetric distribution of $\beta^T\Delta_{x,t}$ when $X_0 = 0$. The last inequality is by the fact that when $X_{u,t} \in (1, 2)$, then $|\theta X_{u,t}| \geq c\beta_{\min}$.

Then, it suffices to prove the lower bound for the expectation term in the last equation, where

$$
\begin{aligned}
&\mathbb{E}_{\beta,u}\left[P_{X,Y|\beta}\left(1 < X_{u,t} < 2,\ \hat{\beta}_{t-1}^T\Delta_{x,t} < 0\right)\right]\\
=&\mathbb{E}_{\beta,u}\left[P\left(1 < X_{u,t} < 2\right)P_{X,Y|\beta}\left(\hat{\beta}_{u,t-1}X_{u,t} + \hat{\beta}_{-u,t-1}^T X_{-u,t} < 0 \mid 1 < X_{u,t} < 2\right)\right]\\
\geq&\tilde{C}P\left(\hat{\beta}_{u,t-1}X_{u,t} + \hat{\beta}_{-u,t-1}^T X_{-u,t} < 0 \mid 1 < X_{u,t} < 2\right). \qquad (12)
\end{aligned}
$$

Here $\hat{\beta}_{-u,t-1} \in \mathbb{R}^{d-1}$ is the subvector of $\hat{\beta}_{t-1}$ without the 0-th and u-th entries, and $X_{-u,t} := (X_{1,t}, \cdots, X_{u-1,t}, X_{u+1,t}, \cdots, X_d)^T \in \mathbb{R}^{d-1}$. The probability $\tilde{C} := P(1 < X_{u,t} < 2)$ is positive, since $X_{u,t}$ is standard normal by definition.

In order to lower bound the probability in (12), we use the Fano's inequality to reduce the probability to the false discovery rate in a multiple hypothesis testing problem. Specifically, we define a test function $\hat{u}_{t-1} := \arg\max_{j\in[d]} |\hat{\beta}_{j,t-1}|$. We then show that the probability in (12) is lower bounded by $P(\hat{u}_{t-1} \neq u)$. In fact, if event $\{\hat{u}_{t-1} \neq u\}$ holds, then we have $|\hat{\beta}_{u,t-1}|^2 \leq \|\hat{\beta}_{-u,t-1}\|_2^2$ by the definition of $\hat{u}_{t-1}$. In addition, we have

$$
\begin{aligned}
&P\left(\hat{\beta}_{u,t-1}X_{u,t} + \hat{\beta}_{-u,t-1}^T X_{-u,t} < 0 \mid 1 < X_{u,t} < 2, |\hat{\beta}_{u,t-1}|^2 \leq \|\hat{\beta}_{-u,t-1}\|_2^2\right)\\
\geq&P\left(|2\hat{\beta}_{u,t-1}| + \hat{\beta}_{-u,t-1}^T X_{-u,t} < 0 \mid 1 < X_{u,t} < 2, |\hat{\beta}_{u,t-1}|^2 \leq \|\hat{\beta}_{-u,t-1}\|_2^2\right)
\end{aligned}
$$

$$
\begin{aligned}
=&P\left(2+N\left(0,\frac{\|\hat{\beta}_{-u,t-1}\|_2^2}{|\hat{\beta}_{u,t-1}|^2}\right)<0 \mid 1<X_{u,t}<2, |\hat{\beta}_{u,t-1}|^2 \le \|\hat{\beta}_{-u,t-1}\|_2^2\right)\\
\ge&P(N(2,1)<0)=C' \qquad (13)
\end{aligned}
$$

The last equation is because that $X_{-u,t}$ is independent from $\hat{\beta}_{t-1}$ and $X_{u,t}$. In addition, the random variable $\hat{\beta}^T_{-u,t-1}X_{-u,t}$ conditioned on $\hat{\beta}_{t-1}$ follows normal distribution with mean 0 and variance $\|\hat{\beta}_{-u,t-1}\|_2^2$. Then, by applying the above inequalities to eq. (12), we have

$$
\mathbb{E}_{\beta,u}\left[P_{X,Y|\beta}\left(1<X_{u,t}<2,\ \hat{\beta}^T_{t-1}\Delta_{x,t}<0\right)\right] \ge \tilde{C}C'P(\hat{u}_{t-1}\neq u). \qquad (14)
$$

It suffices to lower bound $P(\hat{u}_{t-1}\neq u)$ for each time t. We first state a variant of Fano's lower bound in multiple hypothesis testing, the details of which can be found in Chapter 15 of Wainwright (2019).

Lemma 1. *Assume that U is uniform on $\mathcal{U}$. For any Markov chain $U \to (X,Y) \to \hat{U}$,*

$$
P(\hat{U}\neq U) \ge 1-\frac{I(X,Y;U)+\log 2}{\log(|\mathcal{U}|)}. \qquad (15)
$$

Here $I(X,Y;U)$ is the mutual information between (X,Y) and U, and $|\mathcal{U}|$ is the cardinality of set $\mathcal{U}$.

According to the definition of $\mathcal{P}_{\mathcal{M}}$, u is uniform in $[d]$, therefore $\log(|\mathcal{U}|) = \log d$. Then, we prove an upper bound for $I(X_{\{1:t-1\}}, Y_{\{1:t-1\}}; u)$ to guarantee that the probability $P(\hat{u}_{t-1}\neq u)$ in (14) is bounded away from 0. Here $X_{\{1:t-1\}}$ and $Y_{\{1:t-1\}}$ represent the set of feature vectors and response variables up to time $t-1$. Based on the chain rule of the mutual information, we have

$$
I(X,Y;u) = I(X;u) + I(Y;u|X) = I(Y;u|X), \qquad (16)
$$

where the last equality is due to the independence between X and the uniformly distributed u. Moreover, according to Chapter 15 of Wainwright (2019), the conditional mutual information of Y and u conditoned on X can be upper bounded by the Kullback-Leibler divergence, i.e.,

$$
I(Y;u|X) \le \frac{1}{|\mathcal{U}|^2}\sum_{j,k=1}^{|\mathcal{U}|} D(P_{\beta^j|X}\|P_{\beta^k|X}), \qquad (17)
$$

where $P_{\beta^j|X}$ is the distribution of Y conditioned on X and the parameter vector $\beta^j \in \mathcal{M}$ corresponding to $u^j \in \mathcal{U}$. The KL-divergence of between two normal distributions, i.e., $P_{\beta^j|X} = N(X\beta^j,\sigma^2)$ and $P_{\beta^k|X} = N(X\beta^k,\sigma^2)$, can be upper bounded as

$$
\begin{aligned}
&D(P_{\beta^j|X}\|P_{\beta^j|X})\\
=&\frac{1}{2\sigma^2}\|X(\beta^j-\beta^k)\|_2^2 = \frac{1}{2\sigma^2}\|(X_j-X_k)\theta\|_2^2
\end{aligned}
$$

$$\leq \frac{1}{2\sigma^2}\theta^2 4x_{\max}^2 = \frac{2x_{\min}^2}{\sigma^2} \cdot \frac{c^2 \log d}{T}. \tag{18}$$

Summing up the mutual information up to time $t-1$, and according to eq. (16), we have that

$$I(X_{\{1:t-1\}}, Y_{\{1:t-1\}}; u) = I(Y_{\{1:t-1\}}; u) \leq \sum_{s=1}^{t-1} \frac{2x_{\min}^2}{\sigma^2} \cdot \frac{c^2 \log d}{T} \leq \frac{2x_{\max}^2 c^2 \log d}{\sigma^2}. \tag{19}$$

Define constant $c := \frac{\sigma}{2x_{\max}}$. Then, if $\log d \geq 4 \log 2$, we have that

$$\begin{aligned} P(\hat{u}_{t-1} \neq u) &\geq 1 - \frac{I(X_{\{1:t-1\}}, Y_{\{1:t-1\}}; u) + \log 2}{\log |\mathcal{U}|} \\ &\geq 1 - \frac{\frac{1}{2}\log d + \log 2}{\log d} \geq \frac{1}{4}. \end{aligned} \tag{20}$$

By applying eq. (20) and eq. (14) to eq. (11), it can be derived that

$$\begin{aligned} &\sup \left\{ R_T\left(\pi, \pi^*\right) : P_{X,Y} \in \mathcal{P}_\alpha \right\} \\ \geq &\sum_{t=1}^{T} \frac{1}{2} C c \beta_{\min}^{\alpha+1} \mathbb{E}_{\beta,u} \left[P_{X,Y|\beta} \left(1 < X_{u,t} < 2,\ \hat{\beta}_{t-1}^T \Delta_{x,t} < 0 \right) \right] \\ \geq &\sum_{t=1}^{T} \frac{1}{2} C c \beta_{\min}^{\alpha+1} \tilde{C} C' P(\hat{u}_{t-1} \neq u) \\ \geq &\frac{\tilde{C} C' C c}{8} (\frac{\log d}{T})^{\frac{\alpha+1}{2}} T. \end{aligned}$$

Therefore, the supreme cumulative regret incurred by any policy π can be lower bounded as

$$\sup \left\{ R_T\left(\pi, \pi^*\right) : P_{X,Y} \in \mathcal{P}_\alpha \right\} \geq \begin{cases} \tilde{C}_L \left(\log d\right), & \alpha = 1 \\ \tilde{C}_L \left((\log d)^{\frac{\alpha+1}{2}} T^{\frac{1-\alpha}{2}} \right), & \alpha \in [0, 1), \end{cases} \tag{21}$$

where $\tilde{C}_L$ is some postive constant.

Combining the regret lower bound $\Omega(\log T)$ from Goldenshluger and Zeevi (2013) on the horizon T, we prove the regret lower bound in Theorem 1.

Appendix B: Proof of Theorem 2

In this section, we provide the proof for Theorem 2. As we discussed in Section 3.5.2, we first prove Proposition 1 and Proposition 2. Then Corollary 1 is an application of the above propositions. Lastly, we prove Theorem 2 by applying the result in Corollary 1.

B.1. Proof of Proposition 1

For consistency with Proposition 1, let X_i be the i^{th} row of X and $Y(i)$ be the i^{th} entry of Y. The sequence $\{X_i : i = 1, \dots, t\}$ forms an adapted sequence of observations, i.e. X_t may depend on the history $\{X_s, Y(s)\}_{s=1}^{t-1}$. And let $\epsilon \in \mathbb{R}^t$ be the σ-sub-Guassian errors.

Before proving Proposition 1, we first stat the following lemmas for adapted sequences.

Lemma 2 (Bernstein Concentration)**.** *Let $\{D_k, \mathcal{G}_k\}_{k=1}^{\infty}$ be a martingale difference sequence, and suppose that D_k is σ-sub-Guassian in an adapted sense, i.e,. for all $\alpha \in \mathbb{R}$, $\mathbb{E}[e^{\alpha D_k} | \mathcal{G}_{k-1}] \leq e^{\alpha^2\sigma^2/2}$ almost surely. Then, for all $t \geq 0$, $P[|\sum_{k=1}^{n} D_k| \geq t] \leq 2\exp[-t^2/2n\sigma^2]$.*

Proof of Lemma 2 follows from Theorem 2.3 of Wainwright (2019) when $\alpha_* = \alpha_k = 0$ and $\nu_k = \sigma$ for all k.

Lemma 3. *Define the event*

$$\mathcal{F}(\lambda_0(\gamma)) = \{\max_{r\in[d]}(2|\epsilon^T X^{(r)}|/t) \leq \lambda_0(\gamma)\},$$

where $X^{(r)}$ is the r^{th} column of X and $\lambda_0(\gamma) = 2\sigma x_{\max}\sqrt{(\gamma^2 + 2\log d)/t}$. Then we have $P[\mathcal{F}(\lambda_0(\gamma))] \geq 1 - 2\exp[-\gamma^2/2]$.

Proof of Lemma 3 can be found in Lemma EC.2. of Bastani and Bayati (2019).

Lemma 4. *For any $\lambda_0 \in \mathbb{R}^+$, when $\lambda \geq \lambda_0$, on event $\mathcal{F}(\lambda_0)$, we have*

$$\|X(\hat{\beta} - \beta_*)\|_2^2/t \leq 3\lambda\|\hat{\beta}_S - \beta_{*,S}\|_1 - \lambda\|\hat{\beta}_{S^c} - \beta_{*,S^c}\|_1.$$

Proof. According to the definition of the LASSO estimator (4), we have

$$\frac{1}{2t}\|Y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2t}\|Y - X\beta_*\|_2^2 + \lambda\|\beta_*\|_1. \tag{22}$$

Since $\lambda \geq \lambda_0$, then if event $\mathcal{F}(\lambda_0)$ holds, we have $\lambda \geq 2\|\epsilon^T X\|_\infty/t$. Thus,

$$\begin{aligned}\frac{1}{2t}\|X(\hat{\beta} - \beta_*)\|_2^2 &\leq \frac{1}{t}\|\epsilon^T X\|_\infty\|\hat{\beta} - \beta_*\|_1 + \lambda(\|\beta_*\|_1 - \|\hat{\beta}\|_1)\\ &\leq \frac{\lambda}{2}(\|\hat{\beta}_S - \beta_{*,S}\|_1 + \|\hat{\beta}_{S^c} - \beta_{*,S^c}\|_1) + \lambda(\|\beta_{*,S}\|_1 - \|\hat{\beta}\|_1)\\ &\leq \frac{\lambda}{2}(\|\hat{\beta}_S - \beta_{*,S}\|_1 + \|\hat{\beta}_{S^c} - \beta_{*,S^c}\|_1) \qquad (23)\\ &\quad + \lambda(\|\hat{\beta}_S - \beta_{*,S}\|_1 - \|\hat{\beta}_{S^c} - \beta_{*,S^c}\|_1)\\ &= \frac{3}{2}\lambda\|\hat{\beta}_S - \beta_{*,S}\|_1 - \frac{\lambda}{2}\|\hat{\beta}_{S^c} - \beta_{*,S^c}\|_1.\end{aligned}$$

□

Now we can prove Proposition 1. From Lemma 4, we have

$$\|\hat{\beta}_{S^c} - \beta_{*,S^c}\|_1 \leq 3\|\hat{\beta}_S - \beta_{*,S}\|_1. \tag{24}$$

Then we choose $\gamma = 2\log t$ for event $\mathcal{F}(\lambda_0(\gamma))$. By the definition of λ, we have $\lambda \geq \lambda_0$. Thus, if both events $\mathcal{F}(\lambda_0(\gamma))$ and $\{\hat{\Sigma}(X) \in \mathcal{C}(supp(\beta_*), \phi)\}$ hold, we have

$$\begin{aligned}\|\hat{\beta} - \beta_*\|_1^2 &\leq \frac{s_0}{\phi^2}\|X(\hat{\beta} - \beta_*)\|_2^2/t \leq \frac{s_0}{\phi^2}(\frac{2}{t}\|\epsilon^T X\|_\infty + 2\lambda)\|\hat{\beta} - \beta_*\|_1 \\ &\leq \frac{3s_0\lambda}{\phi^2}\|\hat{\beta} - \beta_*\|_1 = \frac{6s_0\sigma x_{\max}}{\phi^2}\sqrt{\frac{2\log t + 2\log d}{t}}\|\hat{\beta} - \beta_*\|_1. \end{aligned}\tag{25}$$

Thus,

$$\begin{aligned}&P\left[\|\hat{\beta} - \beta_*\|_1 \leq \frac{6s_0\sigma x_{\max}}{\phi^2}\sqrt{\frac{2\log t + 2\log d}{t}}\right] \\ \geq& P\left[\mathcal{F}(\lambda_0(\gamma))\right] - P\left[\{\hat{\Sigma}(X) \notin \mathcal{C}(supp(\beta_*), \phi)\}\right] \\ \geq& 1 - \frac{2}{t} - P\left[\{\hat{\Sigma}(X) \notin \mathcal{C}(supp(\beta_*), \phi)\}\right]. \end{aligned}\tag{26}$$

B.2. Proof of Proposition 2

Definition 4. *For a constant $\alpha \geq 1$ and index set I, define the set*

$$\mathbb{C}_\alpha(I) = \left\{\Delta \in \mathbb{R}^d : \|\Delta_{I^c}\|_1 \leq \alpha\|\Delta_I\|_1\right\}.$$

Our goal is to prove that with high probability, for $\forall v \in \mathbb{C}_3(S)$

$$v^T\hat{\Sigma}_t v = \frac{1}{t}\|X_t v\|_2^2 = \frac{1}{t}\sum_{s=1}^{t}\langle X_{\hat{a}_s,s}, v\rangle^2 \geq \kappa\|v\|_2^2,$$

for some constant $\kappa > 0$. To prove Proposition 2, we first state a weaker version about the compatibility condition of the sample covariance matrix.

Proposition 3. *For the adapted sequence $\{X_{\hat{a}_s,s} : s = 1, \ldots, t\}$ induced by the bandit policy, the sample covariance matrix $\hat{\Sigma}_t$ are guaranteed to satisfy the compatibility condition uniformly with high probability, i.e.*

$$P\left(\hat{\Sigma}_t \in \mathcal{C}(supp(\beta_*), \frac{1}{\sqrt{32\zeta K}})\right) \geq 1 - \mathcal{O}(e^{s_0 K\log K\log d}[e^{-c_0 t} + e^{-\tilde{C}\log(K)t}]),$$

where $c_0, \tilde{C}$ are constants. In addition, we can derive a uniform bound for the compatibility condition over t that exceeds a certain threshold (i.e. $t \geq \mathcal{O}(\log T + s_0 K\log K\log d)$) with high probability, i.e.

$$P\left(\forall t \geq T', \hat{\Sigma}_t \in \mathcal{C}(supp(\beta_*), \frac{1}{\sqrt{32\zeta K}})\right) \geq 1 - \frac{2}{T},$$

where $T' = \mathcal{O}(\log T + s_0 K\log K\log d)$.

We first provide the outline of the proof for Proposition 3.

(i) Discretize the unit sphere $S^{d-1} \in \mathbb{R}^d$, and show eigenvalue condition of $\hat{\Sigma}_t$ on a finite set $\mathcal{N}_0$.
(ii) Show eigenvalue condition for all m-sparse vectors within sphere S^{d-1}.
(iii) Transfer eigenvalue condition of $\hat{\Sigma}_t$ to vectors in $\mathbb{C}_3(S)$, which implies compatibility condition.

In the following, we state a result from Oliveira (2016), which is used for step (iii) in the proof.

Lemma 5 (Transfer Principle)**.** *Suppose $\hat{\Sigma}_t$ and Σ are matrix with non-negative diagonal entries, and assume $\eta \in (0,1)$, $m \in \{1,\ldots,d\}$ are such that*

$$\forall v \in \mathbb{R}^d \textit{ with } \|v\|_0 \le m, v^T\widehat{\Sigma}_t v \ge (1-\eta)v^T\Sigma v. \tag{27}$$

Assume D is a diagonal matrix whose elements $D_{j,j}$ are non-negative and satisfy $D_{j,j} \ge [\hat{\Sigma}_t]_{j,j} - (1-\eta)\Sigma_{j,j}$. Then

$$\forall x \in \mathbb{R}^d, x^T\widehat{\Sigma}_t x \ge (1-\eta)x^T\Sigma x - \frac{\|D^{1/2}x\|_1^2}{m-1}. \tag{28}$$

The proof of Lemma 5 can be found in Lemma 5.1 of Oliveira (2016).

According to Lemma 5, it suffices to prove the eigenvalue condition (27) for m-sparse vectors with sufficient large $m = Cs_0$, where constant C will be specified later. To prove this, we first introduce ϵ-net on unit sphere for approximation.

Definition 5 (Nets, covering numbers)**.** *Let (X,d) be a metric space and let $\epsilon > 0$. A subset $\mathcal{N}_\epsilon$ of X is called an ϵ-net of X if every point $x \in X$ can be approximated to within ϵ by some point $y \in \mathcal{N}_\epsilon$, i.e. so that $d(x,y) \le \epsilon$. The minimal cardinality of an ϵ-net of X, if finite, is denoted $\mathcal{N}(X,\epsilon)$ and is called the covering number of X (at scale of ϵ).*

The following provides a bound for the cardinality of the ϵ-net that can approximate the points on the unit Euclidean sphere within range of ϵ.

Lemma 6 (Covering numbers of the sphere)**.** *The unit Euclidean sphere S^{d-1} equipped with the Euclidean metric satisfies for every $\epsilon > 0$ that*

$$\mathcal{N}(S^{d-1},\epsilon) \le \left(1 + \frac{2}{\epsilon}\right)^d.$$

Denote $E_J = \text{span}\{e_j : j \in J\}$, then we define $\mathcal{N}_0$ as the ϵ-net for $S^{d-1} \cap (\cup_{|J|=Cs_0} E_J)$. For each subset $J \subseteq [d]$ with $|J| = Cs_0$, the set $S^{d-1} \cap E_J$ can be viewed as a unit sphere in $\mathbb{R}^{Cs_0}$. Then according to Lemma 6, we have a bound for the covering number N of $\mathcal{N}_0$ when $\epsilon \le 1$.

$$\log N = \log\left(\binom{d}{Cs_0}\left(1 + \frac{2}{\epsilon}\right)^{Cs_0}\right)$$

$$
\begin{aligned}
&\leq \log\left(\left(\frac{ed}{Cs_0}\right)^{Cs_0}\left(\frac{3}{\epsilon}\right)^{Cs_0}\right)\\
&\lesssim Cs_0[\log d + \log(\frac{1}{\epsilon})]. \qquad (29)
\end{aligned}
$$

According to Assumption (b), when $h > 0$ is small, i.e. $h = \mathcal{O}(\frac{1}{K})$, we have for $\forall v \in \mathbb{R}^d$ that:

$$P(\min_a \langle X_{a,t}, v\rangle^2 \leq h\|v\|_2^2) \leq \zeta K h.$$

From now on we fix a sparse vector v with $\|v\|_2 = 1$ and $\|v\|_0 \leq Cs_0$. Define

$$N_t = \sum_{s=1}^{t} \mathbb{1}_{\{\min_a \langle X_{a,s}, v\rangle^2 \leq h\|v\|_2^2\}}$$

Then N_t is the sum of i.i.d. indicator random variables which have expectation greater than $(1 - \zeta K h)$. Now we define $S_t = \sum_{s=1}^{t} Z_s$ where

$$Z_s = \mathbb{1}_{\{\min_a \langle X_{a,s}, v\rangle^2 \leq h\|v\|_2^2\}} - \mathbb{E}[\mathbb{1}_{\{\min_a \langle X_{a,s}, v\rangle^2 \leq h\|v\|_2^2\}}],$$

are centered random variables. Then by Hoeffding's Lemma, we have for $\forall t \in \mathbb{N}_+$ and $\delta \in \mathbb{R}$,

$$P(\frac{1}{t} S_t \leq \delta) \leq \exp(-2\delta^2 t).$$

Let $\delta = -\frac{1}{2}(1 - \zeta K h)$, we have

$$P\left(\frac{1}{t} N_t \leq \frac{1}{2}(1 - \zeta K h)\right) \leq \exp(-\frac{1}{2}(1 - \zeta K h)^2 t) = \exp(-c_0 t)$$

where $c_0 = \frac{1}{2}(1 - \zeta K h)^2$. Then taking the union of the event over all vectors in $\mathcal{N}_0$, we can obtain

$$
\begin{aligned}
&P\left(\forall v \in \mathcal{N}_0 : \frac{1}{t}\sum_{s=1}^{t} \langle X_{a_s,s}, v\rangle^2 \leq \frac{1}{2} h(1 - \zeta K h)\right)\\
\leq& N\exp(-c_0 t) \leq \exp(-c_0 t + Cs_0 \log d + Cs_0 \log(\frac{1}{\epsilon})). \qquad (30)
\end{aligned}
$$

In the above inequalities, we have proved the minimum eigenvalue condition for all vectors in ϵ-net $\mathcal{N}_0$. In the following, we will show that the restricted eigenvalue condition for Cs_0-sparse vectors can also be implied by the eigenvalue condition on $\mathcal{N}_0$. Before that, we state a result on the spectral norm of symmetric matrices.

Lemma 7 (Computing the spectral norm on a net)**.** *Let A be a symmetric $d \times d$ matrix, and let $\mathcal{N}_\epsilon$ be an ϵ-net of S^{d-1} for some $\epsilon \in [0, 1)$. The spectral norm of A can be computed via the associated quadratic form $\|A\| = \sup_{x \in S^{d-1}} |\langle Ax, x\rangle|$. Then*

$$\|A\| = \sup_{x \in S^{d-1}} |\langle Ax, x\rangle| \leq (1 - 2\epsilon)^{-1} \sup_{x \in \mathcal{N}_\epsilon} |\langle Ax, x\rangle|.$$

The proof of Lemma 7 can be found in Lemma 5.4 of Vershynin (2012).

According to the definition of $\mathcal{N}_0$, for any Cs_0-sparse vector v and symmetric $d \times d$ matrix A, there is a vector $u \in \mathcal{N}_0$, such that $\|u - v\|_2 \leq \epsilon$ and $\operatorname{supp}(v) \subseteq \operatorname{supp}(u)$. Then

$$\begin{aligned} v^T A v &= u^T A u + (u-v)^T A (u-v) + 2u^T A (v-u) \\ &\geq u^T A u - 2\|u-v\|_2 \|A\|_{op} \|u\|_2, \end{aligned} \tag{31}$$

where $\|A\|_{op}$ is the spectral norm of A constrained on sparse vectors with ℓ_0-norm not greater than Cs_0, i.e. $\|A\|_{op} = \max_{v \in S^{d-1}, \|v\|_0 \leq Cs_0} \|Av\|_2$.

We have proved a uniform lower bound of $u^T \hat{\Sigma}_t u$ for $u \in \mathcal{N}_0$ and $t \geq T_1$. Then to prove a lower bound of $v^T \hat{\Sigma}_t v$ for $v \in S^{d-1}$ and $\|v\|_0 \leq Cs_0$, based on (31) and the fact that there exit $u \in \mathcal{N}_0$ such that $\|u - v\|_2 \leq \epsilon$, it suffices to show that $\|\hat{\Sigma}_t\|_{op}$ is upper bounded. And by Lemma 7, we only need to bound the spectral norm of $\hat{\Sigma}_t$ on $\mathcal{N}_0$.

Here we have for a fixed vector $v \in \mathcal{N}_0$ and $\forall t \in \mathbb{N}_+$,

$$\frac{1}{t} \sum_{s=1}^{t} \langle X_{a_s,s}, v \rangle^2 \leq \frac{1}{t} \sum_{s=1}^{t} \max_{a \in [K]} \langle X_{a,s}, v \rangle^2.$$

Define $Z_s^a = \langle X_{a,s}, v \rangle^2 - v^T \Sigma_a v$, $S_t^a = \sum_{s=1}^t Z_s^a$ and $S_t = \sum_{a=1}^K S_t^a$ for $a \in [K]$. Then Z_s^a is centered sub-exponential random variables. Before proving the upper bound of $\|\hat{\Sigma}_t\|_{op}$, we will state some properties of sub-exponential random variables. Firstly, we introduce some parameters of sub-guassian and sub-exponential random variables.

Definition 6. *The sub-guassian norm of a sub-guassian random variable X, denoted $\|X\|_{\psi_2}$, is defined as*

$$\|X\|_{\psi_2} = \sup_{p \geq 1} p^{-1/2} \left(\mathbb{E}|X|^p\right)^{1/p}.$$

Similarly, the sub-exponential norm of a sub-exponential random variable X, denoted $\|X\|_{\psi_1}$, is defined as

$$\|X\|_{\psi_1} = \sup_{p \geq 1} p^{-1} \left(\mathbb{E}|X|^p\right)^{1/p}.$$

The details of the definition and related properties can be found in Section 5.2.3 and Section 5.2.4 in Vershynin (2012). The following is a relationship between sub-guassian random variables and sub-exponential random variables.

Lemma 8 (Sub-exponential is sub-guassian squared)**.** *A random variable X is sub-gaussian if and only if X^2 is sub-exponential. Moreover,*

$$\|X\|_{\psi_2}^2 \leq \|X^2\|_{\psi_1} \leq 2\|X\|_{\psi_2}^2.$$

Proof of Lemma 8 can be found in Lemma 5.14 of Vershynin (2012).

Lemma 9 (MGF of sub-exponential random variables)**.** *Let X be a centered sub-exponential random variable. Then for t such that $|t| \leq c_1/\|X\|_{\psi_1}$, one has*

$$\mathbb{E}\exp(tX) \leq \exp(c_2 t^2 \|X\|_{\psi_1}^2),$$

where $c_1, c_2 > 0$ are absolute constants.

Proof of Lemma 9 can be found in Lemma 5.15 of Vershynin (2012).

By the result in Lemma 8, we have for sub-exponential random variable Z_s^a,

$$\kappa_a = \|Z_s^a\|_{\psi_1} \leq 2\|\langle X_{a,s}, v\rangle^2\|_{\psi_1} \leq 4\|\langle X_{a,s}, v\rangle\|_{\psi_2}^2 \leq 4\lambda_{\max}^2(\Sigma_a).$$

Then for any $\mu > 0$ and $\delta > 0$, we have

$$\begin{aligned} P\left(\frac{1}{t}\sum_{s=1}^{t}\max_{a\in[K]}\langle X_{a,s}, v\rangle^2 - \Lambda_0^2 \geq \delta\right) &= P\left(\sum_{s=1}^{t}[\max_a\langle X_{a,s}, v\rangle^2 - \Lambda_0^2] \geq \delta t\right) \\ &\leq e^{-\mu\delta t}\mathbb{E}[\exp(\mu\sum_{s=1}^{t}[\max_a\langle X_{a,s}, v\rangle^2 - \Lambda_0^2])] \\ &= e^{-\mu\delta t}\prod_{s=1}^{t}\mathbb{E}[\exp(\mu[\max_a\langle X_{a,s}, v\rangle^2 - \Lambda_0^2])] \end{aligned} \tag{32}$$

According to Lemma 9, for μ such that $|\mu| \leq \frac{c_1}{\max_a \kappa_a}$, we have

$$\begin{aligned} \mathbb{E}[\exp(\mu[\max_a\langle X_{a,s}, v\rangle^2 - \Lambda^2])] &\leq \sum_{a=1}^{K}\mathbb{E}[\exp\{\mu(\langle X_{a,s}, v\rangle^2 - \mathbb{E}[\langle X_{a,s}, v\rangle^2])\}] \\ &\leq \sum_{a=1}^{K} e^{c_2\mu^2\kappa_a^2} \leq e^{c_2\mu^2\max_a\kappa_a^2 + \log(K)}. \end{aligned}$$

Applying the above inequality to inequality (32), we have for μ such that $|\mu| \leq \frac{c_1}{\max_a \kappa_a}$,

$$P\left(\frac{1}{t}\sum_{s=1}^{t}\max_{a\in[K]}\langle X_{a,s}, v\rangle^2 - \Lambda_0^2 \geq \delta\right) \leq \exp(-\mu\delta t + tc_2\mu^2\max_a\kappa_a^2 + t\log K). \tag{33}$$

The right hand side of the above inequality achieves the minimum value at $\mu = \frac{\delta}{2c_2\max_a\kappa_a^2}$, where the minimum value is $\exp(-\frac{\delta^2 t}{4c_2\max_a\kappa_a^2} + t\log K)$.

Now if $\frac{\delta}{2c_2\max_a\kappa_a^2} > \frac{c_1}{\max_a\kappa_a}$, the right hand side of the equation (33) obtain the minimum value at $\mu = \frac{c_1}{\max_a\kappa_a}$ and

$$P\left(\frac{1}{t}\sum_{s=1}^{t}\max_{a\in[K]}\langle X_{a,s}, v\rangle^2 - \Lambda_0^2 \geq \delta\right)$$

$$\leq \exp\left(-\frac{c_1 \delta t}{\max_a \kappa_a} + tc_2 \frac{c_1^2}{\max_a \kappa_a^2} \max_a \kappa_a^2 + t \log K\right).$$

Moreover, if $\frac{\delta}{2c_2 \max_a \kappa_a^2} > \frac{c_1}{\max_a \kappa_a}$, we apply this to the above inequality and obtain

$$P\left(\frac{1}{t}\sum_{s=1}^{t} \max_a \langle X_{a,s}, v\rangle^2 - \Lambda_0^2 \geq \delta\right) \leq \exp\left(-\frac{c_1 \delta t}{2 \max_a \kappa_a} + t \log K\right).$$

Now we have for $\forall \delta > 0$ and $t \in \mathbb{N}_+$

$$\begin{aligned} &P\left(\frac{1}{t}\sum_{s=1}^{t} \max_a \langle X_{a,s}, v\rangle^2 - \Lambda_0^2 \geq \delta\right) \\ \leq &\exp\left(-\min\{\frac{\delta^2}{4c_2 \max_a \kappa_a^2}, \frac{c_1 \delta}{2\max_a \kappa_a}\}t + t \log K\right). \end{aligned}$$

If we set $\delta = c_3 \log(K)\Lambda_0^2$ where c_3 is a sufficient large positive constant, then we have for a fixed vector $v \in \mathcal{N}_0$

$$P\left(\frac{1}{t}\sum_{s=1}^{t} \max_{a\in[K]} \langle X_{a,s}, v\rangle^2 - \Lambda_0^2 \geq \delta\right) \leq \exp\left(-\tilde{C}\log(K)t\right),$$

where $\tilde{C} = \min\{\frac{c_3^4 \log K}{64c_2}, \frac{c_1 c_3}{8}\} - 1 \leq \min\{\frac{c_3^2 \Lambda_0^4 \log K}{4c_2 \max_a \kappa_a^2}, \frac{c_1 c_3 \Lambda_0^2}{2\max_a \kappa_a}\} - 1$ is a positive constant independent on K and t. Taking the union of the probability over all vectors in $\mathcal{N}_0$, we have

$$\begin{aligned} &P\left(\forall v \in \mathcal{N}_0 : \langle v, \hat{\Sigma}_t v\rangle \geq \tilde{c}\log(K)\Lambda_0^2\right) \\ \leq &N \exp\left(-\tilde{C}\log(K)t\right) \\ \lesssim &\exp\left(-\tilde{C}\log(K)t + Cs_0 \log d + Cs_0 \log(\frac{1}{\epsilon})\right). \end{aligned} \tag{34}$$

Now we provide the proof of Proposition 3.

Proof. We define events $\mathcal{G}_1$ and $\mathcal{G}_2$ as

$$\begin{aligned} \mathcal{G}_1 &= \{\forall u \in \mathcal{N}_0 : \frac{1}{t}\sum_{s=1}^{t} \langle X_{\hat{a}_s,s}, u\rangle \geq \frac{1}{2}h(1-\zeta K h)\}; \\ \mathcal{G}_2 &= \{\forall u \in \mathcal{N}_0 : \langle u, \hat{\Sigma}_t u\rangle \leq \tilde{c}\log(K)\Lambda_0^2\}. \end{aligned}$$

If event $\mathcal{G}_2$ holds, then according to Lemma 7 and definition of $\|\cdot\|_{op}$ in (31), we have

$$\|\hat{\Sigma}_t\|_{op} \leq (1-2\epsilon)^{-1} \max_{v\in\mathcal{N}_0} \langle v, \hat{\Sigma}_t v\rangle \leq (1-2\epsilon)^{-1}\tilde{c}\log(K)\Lambda_0^2. \tag{35}$$

If both event $\mathcal{G}_1$ and $\mathcal{G}_2$ hold, based on (31), we can prove the restricted eigenvalue condition for all Cs_0-sparse vectors in unit sphere S^{d-1}, i.e. for $\forall v \in S^{d-1}$ and $\|v\|_0 \leq Cs_0$ that

$$\begin{aligned}\langle v, \hat{\Sigma}_t v\rangle &\geq \frac{1}{2}h(1-\zeta K h) - 2\epsilon\|\hat{\Sigma}_t\|_{op}\\ &\geq \frac{1}{2}h(1-\zeta K h) - \frac{2\epsilon}{1-2\epsilon}\tilde{c}\log(K)\Lambda_0^2. \end{aligned}\tag{36}$$

Now if we set $h = \frac{1}{2\zeta K} = \mathcal{O}(\frac{1}{K})$ and $\epsilon = \min\{\frac{1}{4}, \frac{1}{64\tilde{c}\log(K)\Lambda_0^2}\}$, then for $\forall v \in S^{d-1}$ and $\|v\|_0 \leq Cs_0$, we have

$$\langle v, \hat{\Sigma}_t v\rangle \geq \frac{1}{4}h(1-\zeta K h) = \frac{1}{16\zeta K}.$$

Now we proved part (i) and (ii) in Proposition 3, i.e. the minimum eigenvalue condition for all Cs_0-sparse vectors in S^{d-1}. To finished the proof, we need to apply Lemma 5 to prove part (iii).

To apply Lemma 5, we set $\Sigma = \frac{1}{8\zeta K}I_{d\times d}$ as a $d\times d$ diagonal matrix, $m = Cs_0$ and $\eta = \frac{1}{2}$. Moreover, we set $D_{j,j} = [\hat{\Sigma}_t]_{j,j}$ to satisfy the condition on diagonal matrix D. By greedy method for constructing $\mathcal{N}_0$, we can let $e_j \in \mathcal{N}_0$ for $\forall j \in [d]$, where e_j is unit vector in $\mathbb{R}^d$ having exactly one entry equal to 1 and 0 otherwise. Then under event $\mathcal{G}_2$, we have

$$D_{j,j} = [\hat{\Sigma}_t]_{j,j} = \langle e_j, \hat{\Sigma}_t e_j\rangle \leq \tilde{c}\log(K)\Lambda_0^2.$$

Now according to (28), if both events $\mathcal{G}_1$ and $\mathcal{G}_2$ hold, we have for $\forall x \in S^{d-1}$ and $x \in \mathbb{C}_3(S)$ that

$$\begin{aligned}x^T\hat{\Sigma}_t x &\geq \frac{1}{2}x^T\Sigma x - \frac{\|D^{1/2}x\|_1^2}{Cs_0-1}\\ &\geq \frac{1}{16\zeta K} - \frac{\tilde{c}\log(K)\Lambda_0^2\|x\|_1^2}{Cs_0-1}\\ &\geq \frac{1}{16\zeta K} - \frac{\tilde{c}\log(K)\Lambda_0^2\cdot 16s_0}{Cs_0-1}, \end{aligned}\tag{37}$$

where the last inequality above is because for $x \in S^{d-1}$ and $x \in \mathbb{C}_3(S)$, we have

$$\|x\|_1 \leq (1+3)\|x_S\|_1 \leq 4\sqrt{s_0}\|x_S\|_2 \leq 4\sqrt{s_0}.$$

Then, if we set $C = 2\cdot 16^2\tilde{c}\log(K)\Lambda_0^2\zeta K + \frac{1}{s_0} = \Theta(\zeta\Lambda_0^2 K\log K)$, we can have $x^T\hat{\Sigma}_t x \geq \frac{1}{32\zeta K}$ for unit vector $x \in \mathbb{C}_3(S)$, which implies compatibility condition.

Now, we prove the uniform bound on the compatibility condition in Proposition 3 by showing that events $\mathcal{G}_1$ and $\mathcal{G}_2$ hold with high probability.

At first, for any $T_1 \in \mathbb{N}_+$ we have

$$P\left(\forall t \geq T_1, \forall v \in \mathcal{N}_0 : \frac{1}{t}\sum_{s=1}^{t}\langle X_{a_s,s}, v\rangle^2 \leq \frac{1}{2}h(1-\zeta K h)\right)$$

$$\leq N \sum_{t=T_1}^{\infty} e^{-c_0 t} = \frac{1}{1-e^{-c_0}} \exp(-c_0 T_1 + \log N)$$
$$\lesssim \exp(-c_0 T_1 + C s_0 \log d + C s_0 \log(\frac{1}{\epsilon})) = \delta_2,$$

where $c_0 = \frac{1}{2}(1-\zeta K h)^2 = \frac{1}{8}$. If we set $\delta_2 = \frac{1}{T}$, then with the choice of $\epsilon = \min\{\frac{1}{4}, \frac{1}{64\tilde{c}\log(K)\Lambda_0^2}\}$ and C from above, we have

$$\begin{aligned} T_1 &= \mathcal{O}(\log T + C s_0[\log d + \log(\frac{1}{\epsilon})]) \\ &= \mathcal{O}(\log T + s_0 \zeta \Lambda_0^2 K(\log K)\log d + \log(\log K) + \log \Lambda_0^2) \\ &= \mathcal{O}(\log T + s_0 K(\log K)\log d). \end{aligned} \tag{38}$$

The above inequality provides a uniform lower bound for the eigenvalue of sample covariance matrix $\hat{\Sigma}_t$ over $\mathcal{N}_0$ for $t \geq T_1$.

Moreover, we can obtain an union probability bound for all vectors in $\mathcal{N}_0$ and some $T_2 \in \mathbb{N}_+$ in (34).

$$P\left(\forall t \geq T_2, \forall v \in \mathcal{N}_0 : \langle v, \hat{\Sigma}_t v\rangle \geq \tilde{c}\log(K)\Lambda_0^2\right)$$
$$\leq N \sum_{t=T_2}^{\infty} \exp(-\tilde{C}\log(K)t)$$
$$\lesssim \exp\left(-T_2 \tilde{C}\log(K) + C s_0 \log d + C s_0 \log(\frac{1}{\epsilon})\right) = \delta_3.$$

If we set $\delta_3 = \frac{1}{T}$, then we have

$$T_2 = \frac{\log(T) + C s_0 \log d + C s_0 \log(\frac{1}{\epsilon})}{\tilde{C}\log(K)} = \mathcal{O}\left(\log T + C s_0 \log d + C s_0 \log(\frac{1}{\epsilon})\right).$$

Now, we can show that compatibility condition holds on $\mathbb{C}_3(S)$ uniformly for all $t \geq \max(T_1, T_2) = \mathcal{O}(\log T + s_0 K \log K \log d)$ with high probability, i.e.

$$P\left(\forall t \geq \max(T_1, T_2), v \in \mathbb{C}_3(S) : v^T \hat{\Sigma}_t v \geq \frac{1}{32\zeta K}\right) \geq 1 - \delta_2 - \delta_3 = 1 - \frac{2}{T}. \quad \square$$

With the crude compatibility condition from Proposition 3, we can prove Proposition 2. Firstly, we prove that the optimal arm will be pulled a positive fraction of time by the ℓ_1-confidence ball based algorithm after some time point.

Lemma 10. *Suppose we construct the confidence set $\mathcal{C}_t$ in Proposition 2 with $\phi = \frac{1}{\sqrt{32\zeta K}}$, then when time horizon t exceeds a certain threshold (i.e. $t \geq \mathcal{O}(\log T + s_0 K \log K \log d + \frac{s_0^2 \zeta^2 K^2 \log d}{\Delta_*^2})$), decision-makers will pull the optimal arms a positive fraction of time.*

Proof. Before the proof, we first define the event in Proposition 1 with $\phi^2 = \frac{1}{32\zeta K}$,

$$\mathcal{G}_{3,t+1} = \left\{ \|\hat{\beta}_t - \beta_*\|_1 \leq \frac{6s_0\sigma x_{\max}}{1/(32\zeta K)} \sqrt{\frac{2\log t + 2\log d}{t}} \right\}.$$

According to Proposition 3, $\mathcal{G}_{3,t}$ will hold uniformly over $t \geq T'$ with high probability.

From Assumption 2(c), we have event Γ_t such that $P(\Gamma_t) \geq \frac{1}{2}$. Then with sufficient large $\phi \geq \frac{1}{\sqrt{32\zeta K}}$ in Proposition 2, if both Γ_t and $\mathcal{G}_{3,t}$ hold, we can only select the arm $b \in [K]$ instead of the optimal arm a_t^* when

$$\begin{aligned} &\max_{\beta \in \mathcal{C}_t} \langle X_{b,t}, \beta \rangle \geq \max_{\beta \in \mathcal{C}_t} \langle X_{a_t^*,t}, \beta \rangle \\ \Longrightarrow\ &\langle X_{b,t}, \beta_* \rangle + 2\|X_{b,t}\|_\infty \frac{6s_0\sigma x_{\max}}{1/(32\zeta K)} \sqrt{\frac{2\log(t-1) + 2\log d}{t-1}} \geq \langle X_{a_t^*,t}, \beta_* \rangle. \end{aligned}$$

The second inequality is because for all $\beta \in \mathcal{C}_t$, by triangle inequality we have

$$\begin{aligned} \|\beta - \beta_*\|_1 &\leq \|\beta - \hat{\beta}_t\|_1 + \|\hat{\beta}_t - \beta_*\|_1 \\ &\leq (\frac{1}{\phi^2} + 32\zeta K)(6s_0\sigma x_{\max}) \sqrt{\frac{2\log(t-1) + 2\log d}{t-1}} \\ &\leq 384\zeta K s_0 \sigma x_{\max} \sqrt{\frac{2\log(t-1) + 2\log d}{t-1}} \end{aligned}$$

Then we will always select the optimal arm a_t^* if

$$\begin{aligned} \Delta_* &\geq 384 s_0 \sigma x_{\max}^2 \zeta K \sqrt{\frac{2\log(t-1) + 2\log d}{t-1}} \\ &\geq 2\|X_{b,t}\|_\infty \frac{6s_0\sigma x_{\max}}{1/(32\zeta K)} \sqrt{\frac{2\log(t-1) + 2\log d}{t-1}}. \end{aligned}$$

By solving the above inequality, we have that for $t \geq \mathcal{O}(\frac{\sigma^2 x_{\max}^4 s_0^2 \zeta^2 K^2 \log d}{\Delta_*^2})$, we will always select the optimal arm a_t^* when both Γ_t and $\mathcal{G}_{3,t}$ hold. □

Proof. (Proposition 2) Now we consider time such that $t \geq T'' = \mathcal{O}(\log T + s_0 K \log K \log d + \frac{\sigma^2 x_{\max}^4 s_0^2 \zeta^2 K^2 \log d}{\Delta_*^2})$. From the above proof, we know that we will select the optimal arm with high probability.

Suppose $t \geq 2T''$ and $v \in \mathbb{R}^d$, then we have

$$\begin{aligned} \frac{1}{t} \sum_{s=1}^{t} \langle X_{\hat{a}_s,s}, v \rangle^2 &\geq \frac{1}{t} \sum_{s=T''+1}^{t} \langle X_{\hat{a}_s,s}, v \rangle^2 \\ &\geq \frac{1}{t} \sum_{s=T''+1}^{t} \langle X_{\hat{a}_s,s}, v \rangle^2 \mathbb{1}_{\{\Gamma_s \cap \mathcal{G}_{3,t}\}} \end{aligned}$$

$$= \frac{1}{t}\sum_{s=T''+1}^{t} \langle X_{a_s^*,s}, v\rangle^2 \mathbb{1}_{\{\Gamma_s\}}\mathbb{1}_{\{\mathcal{G}_{3,t}\}} \tag{39}$$

Since event $\mathcal{G}_{3,t}$ only depends on the history up to time $t-1$, the random variables $\{X_{a_t^*,t}|\mathcal{G}_{3,t}\}$ are i.i.d. Moreover, since $\{\mathbb{1}_{\{\Gamma_s\}} : s \geq T''\}$ are i.i.d Bernoulli random variables, then according the Hoeffding's Lemma, we have

$$\sum_{s=T''+1}^{t} \mathbb{1}_{\{\Gamma_s\}} \leq \frac{1}{2}\cdot\frac{1}{2}(t-T'') \quad \text{w.p.} \leq \mathcal{O}(e^{-c_4 t}), \tag{40}$$

where c_4 is some constant.

Now we consider the sum of $\langle X_{a_s^*,s}, v\rangle^2$ over times when event Γ_t holds

$$\sum_{s:\Gamma_s \text{ holds}} \langle X_{a_s^*,s}, v\rangle^2 - \mathbb{E}[\langle X_{a_s^*,s}, v\rangle^2 \big| \Gamma_t].$$

Define $Z_s = \langle X_{a_s^*,s}, v\rangle^2 - \mathbb{E}[\langle X_{a_s^*,s}, v\rangle^2 \big| \Gamma_t]$, then Z_s are i.i.d. sub-exponential random variables. Moreover, $\mathbb{E}[\langle X_{a_s^*,s}, v\rangle^2 \big| \Gamma_t] \geq \phi_0^2\|v\|_2^2$ according to Assumption (c), and from Lemma 8, we have condition on event Γ_t

$$\|Z_s\|_{\psi_1} \leq 2\|\langle X_{a_s^*,s}, v\rangle^2\|_{\psi_1} \leq 4\|\langle X_{a_s^*,s}, v\rangle\|_{\psi_2}^2 = 4v^T\mathbb{E}[X_{a_t^*}X_{a_t^*}^T|\Gamma_t]v \leq 4\Lambda_1^2.$$

Then applying Bernstein inequality for sub-exponential random variables, we can have a lower bound for a fixed vector v and $\epsilon = \frac{\phi_0^2}{2}$.

$$\begin{aligned}
&P\left(|\sum_{s:\Gamma_s \text{ holds}} Z_s| \geq \epsilon m \Bigg| \sum_{s=T''+1}^{t} \mathbb{1}_{\{\Gamma_s\}} = m\right)\\
\leq& 2\exp\left(-\tilde{c}_5 \min(\frac{\epsilon^2}{\|Z_s\|_{\psi_1}^2}, \frac{\epsilon}{\|Z_s\|_{\psi_1}})m\right)\\
=& 2\exp\left(-\tilde{c}_5 \min(\frac{\phi_0^4}{4\|Z_s\|_{\psi_1}^2}, \frac{\phi_0^2}{2\|Z_s\|_{\psi_1}})m\right) = \mathcal{O}(e^{-c_5 m}),
\end{aligned} \tag{41}$$

where $\tilde{c}_5, c_5$ are constants.

Now combining inequalities (39)–(41), we have for any v in ϵ'-net $\mathcal{N}_1$ for $S^{d-1} \cap (\cup_{|J|=C's}E_J)$, where ϵ and C' need to be specified, and $t \geq 2T''$ that

$$\frac{1}{t}\sum_{s=1}^{t}\langle X_{\hat{a}_s,s}, v\rangle^2 \geq \frac{1}{t}\sum_{s=T''+1}^{t}\langle X_{a_s^*,s}, v\rangle^2\mathbb{1}_{\{\Gamma_s\}} \geq \frac{1}{t}\frac{\phi_0^2}{2}\frac{\frac{1}{2}\cdot(t-T'')}{2}\|v\|_2^2 \geq \frac{\phi_0^2}{16},$$

with probability greater than $1 - \mathcal{O}(e^{-c_4(t-T'')} + e^{-c_5\cdot\frac{1}{4}(t-T'')})$.

Similar as before, we can extend the above inequality to all vectors in $\mathbb{C}_3(S)$ by using the approximation of $\mathcal{N}_1$, where $C = \mathcal{O}(\log K)$ and $\epsilon = \mathcal{O}(\frac{1}{\log K})$. Then for any $v \in \mathbb{C}_3(S)$ and $t \geq 2T''$, we have

$$\hat{\Sigma}_t \in \mathcal{C}(supp(\beta_*), \frac{\phi_0}{8})$$

with probability greater than $1-\mathcal{O}(e^{s_0 \log K \log d}[e^{-c_4(t-T'')}+e^{-c_5\cdot\frac{1}{4}(t-T'')}]) = 1-\mathcal{O}(e^{-t+s_0 \log K \log d})$.

Then combine Proposition 1 on $\mathcal{G}_{3,t}$ and the above inequality, we can prove the Proposition 2. □

B.3. Proof of Theorem 2

Define T'' as the threshold in Proposition 2, then we divide the cumulative regret into three groups:

(a) Initialization when $t \le T'' = \mathcal{O}(\log T + s_0 K \log K \log d + \frac{\sigma^2 x_{\max}^4 s_0^2 \zeta^2 K^2 \log d}{\Delta_*^2})$.
(b) Times $t > T''$ when $\hat{\Sigma}_t \notin \mathcal{C}(supp(\beta_*), \phi_*)$, where $\phi_* = \frac{\phi_0}{8}$.
(c) Times $t > T''$ when $\hat{\Sigma}_t \in \mathcal{C}(supp(\beta_*), \phi_*)$, where $\phi_* = \frac{\phi_0}{8}$.

The cumulative regret J_1 from time periods in group (a) at time T is bounded by at most $2bx_{\max}T'' = \mathcal{O}(2bx_{\max}[\log T + s_0 K \log K \log d + \frac{\sigma^2 x_{\max}^4 s_0^2 \zeta^2 K^2 \log d}{\Delta_*^2}])$.

According to Proposition 2, we can bound the cumulative regret in group (b) at time T.

$$
\begin{aligned}
J_2 = \sum_{t=T''}^{T} \mathbb{E}[r_t \mathbb{1}_{\{\hat{\Sigma}_t \notin \mathcal{C}(supp(\beta_*),\phi_*)\}}] &\le \sum_{t=T''}^{T} 2bx_{\max} P\left(\hat{\Sigma}_t \notin \mathcal{C}(supp(\beta_*),\phi_*)\right) \\
&\lesssim \sum_{t=T''}^{T} 2bx_{\max}\left[\frac{1}{t} + e^{-t+s_0 \log k \log d}\right] = \mathcal{O}(x_{\max} b \log T).
\end{aligned}
$$

To prove the bound of the cumulative regret in group (c), we first define events:

$$
\Gamma_{\xi_t} = \{\omega \in \Omega : \langle X_{a_t^*,t}, \theta_* \rangle \ge \langle X_{b,t}, \theta_* \rangle + \xi_t,\ \forall b \ne a_t^*\}.
$$

In addition, we define $A_t := \{\hat{\Sigma}_t \in \mathcal{C}(supp(\beta_*), \phi_*)\}$ for simplicity. Then the cumulative regret at time T in group (c) can be writen as:

$$
J_3 = \sum_{t=T''}^{T} \mathbb{E}[r_t \mathbb{1}_{\{A_t \cap \Gamma_{\xi_t}\}}] + \sum_{t=T''}^{T} \mathbb{E}[r_t \mathbb{1}_{\{A_t \cap \Gamma_{\xi_t}^c\}}].
$$

Now we set $\xi_t = 2x_{\max}\tau_{t-1} = 2x_{\max}\tau_0\sqrt{\frac{\log d + \log(t-1)}{t-1}}$ with $\tau_0 = \frac{384\sqrt{2}s_0\sigma x_{\max}}{\phi_0^2}$, then when both Γ_{ξ_t} and A_t hold, we have for any $k \ne a_t^*$ that

$$
\begin{aligned}
&\max_{\beta \in \mathcal{C}_t}\langle X_{a_t^*,t}, \beta \rangle - \max_{\beta \in \mathcal{C}_t}\langle X_{k,t}, \beta \rangle \\
\ge &\langle X_{a_t^*,t}, \beta_* \rangle - 2x_{\max}\tau_{t-1} - \langle X_{k,t}, \beta_* \rangle \ge 0.
\end{aligned}
$$

So we will always select the optimal arm a_t^* under the event $A_t \cap \Gamma_{\xi_t}$, and the first part of the cumulative regret in J_3 will be zero.

Now we consider the time when both A_t and $\Gamma^c_{\xi_t}$ hold. Then if a sub-optimal arm $k \neq a^*_t$ is selected, the regret incurred will be

$$\langle X_{a^*_t,t} - X_{k,t}, \beta_* \rangle \leq \max_{\beta \in \mathcal{C}_t} \langle X_{a^*_t,t}, \beta \rangle - \max_{\beta \in \mathcal{C}_t} \langle X_{k,t}, \beta \rangle + 2x_{\max}\tau_{t-1} \leq 2x_{\max}\tau_{t-1} = \xi_t.$$

Moreover, from Assumption 1(b), we have

$$P\left(\langle X_{a^*_t,t}, \beta_* \rangle - \max_{b \neq a^*_t} \langle X_{b,t}, \beta_* \rangle \leq \xi_t \right) \leq \frac{1}{2}(\frac{\xi_t}{\Delta_*})^{\alpha}.$$

(i) For the $\alpha \in [0,1]$ case, we can bound the second term in J_3 as

$$\begin{aligned} &\sum_{t=T''}^{T} \mathbb{E}[r_t \mathbb{1}_{\{A_t \cap \Gamma^c_{\xi_t}\}}] \leq \sum_{t=T''}^{T} \mathbb{E}[r_t \mathbb{1}_{\{\text{select arm } i \neq s^*_t\}}] \\ &\leq \sum_{t=T''}^{T} \xi_t P\left(\langle X_{a^*_t,t}, \beta_* \rangle - \max_{b \neq a^*_t} \langle X_{b,t}, \beta_* \rangle \leq \xi_t \right) \\ &\leq \sum_{t=T''}^{T} \frac{1}{2\Delta_*^{\alpha}} \xi_t^{\alpha+1} \\ &\leq \sum_{t=1}^{T} \frac{2^{\alpha}(x_{\max}\tau_0)^{\alpha+1}}{\Delta_*^{\alpha}} \left(\frac{\log d + \log t}{t} \right)^{\frac{\alpha+1}{2}} \\ &= \begin{cases} C_2 \frac{s_0^{\alpha+1}\sigma^{\alpha+1}x_{\max}^{2(\alpha+1)}}{\Delta_*^{\alpha}\phi_0^{2(\alpha+1)}} (\log d)^{\frac{\alpha+1}{2}} T^{\frac{1-\alpha}{2}}, & \text{when} \quad \alpha \in [0,1), \\ C_2 \frac{s_0^2\sigma^2 x_{\max}^4}{\Delta_* \phi_0^4} [\log d + \log T] \log T, & \text{when} \quad \alpha = 1, \end{cases} \end{aligned}$$

(ii) For the $\alpha \in (1, +\infty)$ case, the second term in J_3 can be bounded as

$$\begin{aligned} &\sum_{t=T''}^{T} \mathbb{E}[r_t \mathbb{1}_{\{A_t \cap \Gamma^c_{\xi_t}\}}] \\ &\leq \sum_{t=T''}^{T} \xi_t P\left(\langle X_{a^*_t,t}, \beta_* \rangle - \max_{b \neq a^*_t} \langle X_{b,t}, \beta_* \rangle \leq \xi_t \right) \\ &\leq \sum_{t=T''}^{T} \xi_t \cdot \min\left(1, \frac{\xi_t^{\alpha}}{2\Delta_*^{\alpha}}\right). \end{aligned}$$

Moreover, according to the definition of ξ_t, we have that

$$\frac{\xi_t^{\alpha}}{2\Delta_*^{\alpha}} \geq 1 \iff t \lesssim T''' := \frac{4^{1-1/\alpha} x_{\max}^2 \tau_0^2 \log d}{\Delta_*^2}.$$

Then, the above summation can be decomposed into two parts:

$$\sum_{t=T''}^{T} \xi_t \cdot \min\left(1, \frac{\xi_t^{\alpha}}{2\Delta_*^{\alpha}}\right)$$

$$\begin{aligned}
&\leq \sum_{t=T''}^{T'''} \xi_t + \sum_{t=T'''}^{T} \frac{1}{2\Delta_*^\alpha} \xi_t^{\alpha+1} \\
&\leq x_{\max} b T''' + \frac{2^\alpha (x_{\max}\tau_0)^{\alpha+1}}{\Delta_*^\alpha} \sum_{t=T'''}^{T} \left(\frac{\log d + \log t}{t} \right)^{\frac{\alpha+1}{2}}.
\end{aligned} \tag{42}$$

Here for $t \geq T'''$, we have that

$$\frac{\log d + \log t}{t} \leq \frac{\log d + \log T'''}{T'''} = \mathcal{O}\left(\frac{\Delta_*^2}{4^{1-1/\alpha} x_{\max}^2 \tau_0^2} \right).$$

We define the constant $M := \frac{\Delta_*^2}{4^{1-1/\alpha} x_{\max}^2 \tau_0^2}$. Then, the second term in the right-hand side of inequality (42) can be bounded by

$$\begin{aligned}
&\sum_{t=T'''}^{T} \left(\frac{\log d + \log t}{t} \right)^{\frac{\alpha+1}{2}} \\
&\leq \sum_{t=1}^{T} \left(\frac{\log d + \log t}{t} \right)^{\frac{\alpha+1}{2}} \mathbb{1}_{\left\{ \frac{\log d + \log t}{t} \leq M \right\}} \\
&\lesssim \sum_{t=1}^{T} \left(\frac{\log d}{t} \right)^{\frac{\alpha+1}{2}} \mathbb{1}_{\left\{ \frac{\log d}{t} \leq M \right\}} \\
&\lesssim \int_{1}^{T} \left(\frac{\log d}{t} \right)^{\frac{\alpha+1}{2}} \mathbb{1}_{\left\{ \frac{\log d}{t} \leq M \right\}} \, \mathrm{d}t \\
&= \int_{\log d/T}^{\min(\log d, M)} \log d \cdot u^{\frac{\alpha+1}{2}-2} \, \mathrm{d}u \\
&\leq \frac{2}{\alpha-1} \log d \cdot M^{\frac{\alpha-1}{2}} \\
&= \frac{2\Delta_*^{\alpha-1} \log d}{(\alpha-1) 4^{(\alpha-1)^2/2\alpha} x_{\max}^{\alpha-1} \tau_0^{\alpha-1}}.
\end{aligned}$$

The equality in the above is due to the change of variable by taking $u = \frac{\log d}{t}$. Therefore, the second term in J_3 can be bounded for $\alpha \in (1, +\infty)$ as

$$\begin{aligned}
&\sum_{t=T''}^{T} \mathbb{E}[r_t \mathbb{1}_{\{A_t \cap \Gamma_{\xi_t}^c\}}] \\
&\leq x_{\max} b T''' + \frac{2^\alpha (x_{\max}\tau_0)^{\alpha+1}}{\Delta_*^\alpha} \cdot \frac{2\Delta_*^{\alpha-1} \log d}{(\alpha-1) 4^{(\alpha-1)^2/2\alpha} x_{\max}^{\alpha-1} \tau_0^{\alpha-1}} \\
&\leq x_{\max} b T''' + \frac{2^{3-1/\alpha} x_{\max}^2 \tau_0^2 \log d}{(\alpha-1)\Delta_*} \\
&\leq C_2 \left[\frac{b s_0^2 \sigma^2 x_{\max}^5 \log d}{\Delta_*^2 \phi_0^4} + \frac{s_0^2 \sigma^2 x_{\max}^4 \log d}{(\alpha-1)\Delta_* \phi_0^4} \right].
\end{aligned}$$

(iii) For the $\alpha = +\infty$ case, Assumption 1(b) implies that for $\xi_t < \Delta_*$, the following inequality holds.

$$P\left(\langle X_{a_t^*,t}, \beta_* \rangle - \max_{b \neq a_t^*} \langle X_{b,t}, \beta_* \rangle \leq \xi_t \right) = 0.$$

Therefore, the second term in J_3 can be bounded as

$$\begin{aligned}\sum_{t=T''}^{T} \mathbb{E}[r_t \mathbb{1}_{\{A_t \cap \Gamma_{\xi_t}^c\}}] &\leq \sum_{t=T''}^{T} \mathbb{E}[r_t \mathbb{1}_{\{\text{select arm } i \neq s_t^*\}}] \\ &\leq \sum_{t=T''}^{T} \xi_t P\left(\langle X_{a_t^*,t}, \beta_* \rangle - \max_{b \neq a_t^*} \langle X_{b,t}, \beta_* \rangle \leq \xi_t \right) \\ &\leq \sum_{t=T''}^{T} \xi_t \mathbb{1}_{\{\xi_t \geq \Delta_*\}} \\ &\overset{(a)}{\leq} \sum_{t=T''}^{T} \frac{\xi_t^{\gamma+1}}{\Delta_*^{\gamma}} \\ &= \sum_{t=T''}^{T} \frac{2^{\gamma+1}(x_{\max}\tau_0)^{\gamma+1}}{\Delta_*^{\gamma}} \left(\frac{\log d + \log t}{t}\right)^{\frac{\gamma+1}{2}}\end{aligned}$$

where constant $\gamma \geq 0$. The inequality (a) above is due to $\xi_t/\Delta_* \geq 1$. By letting $\gamma \to 1_+$, we can obtain an upper bound for the $\alpha = \infty$ case, i.e.,

$$\sum_{t=T''}^{T} \mathbb{E}[r_t \mathbb{1}_{\{A_t \cap \Gamma_{\xi_t}^c\}}] \leq C_2 \frac{s_0^2 \sigma^2 x_{\max}^4 \log d}{\Delta_* \phi_0^4}.$$

By summing the cumulative regret in three groups, we can obtain the upper bound for the total expected cumulative regret of our ℓ_1-confidence ball based method up to time T.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Casaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Yasin Abbasi-Yadkori, Dávid Pál, and Casaba Szepesvári. Online-to-confidence-set conversions and application to sparse stochastic bandits. *AISTATS*, pages 1–9, 2012.

Naoki Abe, Alan W. Biermann, and Philip M. Long. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):236–293, 2003. MR2010092

Jean-Yves Audibert, Alexandre B Tsybakov, et al. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35:608–633, 2007. MR2336861

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, pages 397–422, 2003. MR1984023

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002a.

Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2019. MR4059503

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011. MR2807761

Emmanuel Candes and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005. MR2243152

Wei Chu, Lihong Li, Lev Reyzin, and Robert E Schapire. Contextual bandits with linear payoff functions. *AISTATS*, pages 208–214, 2011.

International Warfarin Pharmacogenetics Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8):753, 2009.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. *Conference On Learning Theory*, pages 355–366, 2008.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96:1348–1360, 2001. MR1946581

Alexander Goldenshluger and Assaf Zeevi. Woodroofe's one-armed bandit problem revisited. *The Annals of Applied Probability*, 19(4):1603 – 1633, 2009. MR2538082

Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1):230 – 261, 2013. MR3353472

Gi-Soo Kim and Myunghee Cho Paik. Doubly-robust lasso bandit. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Roberto Imbuzeiro Oliveira. The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties. *Probability Theory and Related Fields*, 166(3):1175–1194, 2016. MR3568047

Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59(6):3434–3447, 2013. MR3061256

Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010. MR2674726

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. MR1379242

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and GittaEditors Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge University

Press, 2012. MR2963170

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019. MR3967104

Xue Wang, Mingcheng Wei, and Tao Yao. Minimax concave penalized multi-armed bandit model with high-dimensional covariates. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5200–5208. PMLR, 10–15 Jul 2018.

Cun-Hui Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38:894–942, 2010. MR2604701