Abstract—We study the linear contextual bandit problem with finite action sets. When the problem dimension is $d$, the time horizon is $T$, and there are $n \leq 2^{d/2}$ candidate actions per time period, we 1) show that the minimax expected regret is $\Omega(\sqrt{dT(\log T)(\log n)})$ for every algorithm, and 2) introduce a Variable-Confidence-Level (VCL) SupLinUCB algorithm whose regret matches the lower bound up to iterated logarithmic factors. Our algorithmic result saves two $\sqrt{\log T}$ factors from previous analysis, and our information-theoretical lower bound also improves previous results by one $\sqrt{\log T}$ factor, revealing a regret scaling quite different from classical multi-armed bandits in which no logarithmic $T$ term is present in minimax regret. Our proof techniques include variable confidence levels and a careful analysis of layer sizes of SupLinUCB on the upper bound side, and delicately constructed adversarial sequences showing the tightness of elliptical potential lemmas on the lower bound side.

Index Terms—Linearly parameterized bandits, minimax-optimal regret.

I. INTRODUCTION

The stochastic multi-armed bandit (MAB) problem is a sequential experiment in which sequential decisions are made over $T$ time periods in order to maximize the expected cumulative reward of the made decisions. First studied by [2], [3] and many more works thereafter [4], [5], [6], and [7], the MAB problems are one of the simplest yet most popular frameworks to study exploration-exploitation tradeoffs in sequential experiments.

In real-world applications such as advertisement selection [8], recommendation systems [9] and information retrieval [10], side information is most of the time available for each possible actions. Contextual bandit models are thus proposed to incorporate such contextual information into sequential decision making. While the study of general contextual bandit models is certainly of great interest [11], [12], [13], many research efforts have also been devoted into an important special case of the contextual bandit model, in which the mean rewards of actions are parameterized by linear functions [7], [8], [14], [15], [16], [17], [18]. We refer the readers to Sec. I-B for a more detailed accounts of existing results along this direction.

In this paper, we consider the linear contextual bandit problem with finite action sets, known time horizon and oblivious action context. We derive upper and lower bounds on the best worst-case cumulative regret any policy can achieve, that match each other except for iterated logarithmic terms (see Table I for details and comparison with existing works). Many new proof techniques and insights are generated, as we discuss in Sec. I-C.

A. Problem Formulation and Minimax Regret

There are $T \geq 1$ time periods, conveniently denoted as $\{1, 2, \ldots, T\}$, and a fixed but unknown $d$-dimensional regression model $\theta$. Throughout this paper we will assume the model is normalized, meaning that $\|\theta\|_2 \leq 1$. At each time period $t$, a policy $\pi$ is presented with an action set $A_t = \{x_{it}\} \subseteq \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$, where $i$ is the index of for the candidate action in $A_t$. An adversary will choose the action sets $A_1, \ldots, A_T$ before the policy is executed, in an arbitrary way. The policy then chooses, based on the feedback from previous time periods $\{1, 2, \ldots, t-1\}$, either deterministically or randomly an action $x_{it} \in A_t$ and receives a reward $r_t = x_{it}^\top \theta + \epsilon_t$, where $\{\epsilon_t\}$ are independent centered sub-Gaussian random variables with variance proxy 1, representing noise during the reward collection procedure. The objective is to design a good policy $\pi$ that tries to maximize its expected cumulative regret $\mathbb{E} \sum_{t=1}^{T} r_t$.

More specifically, a policy $\pi$ designed for $d$-dimensional vectors, $T$ time periods and maximum action set size $n = \max_{t \leq T} |A_t|$ can be parameterized as $\pi = (\pi_1, \pi_2, \ldots, \pi_T)$ such that

$$i_t = \begin{cases} \pi_1(\nu, A_1), & t = 1; \\ \pi_t(\nu, A_1, r_1, \ldots, A_{t-1}, r_{t-1}, A_t), & t = 2, \ldots, T, \end{cases}$$

where $\nu$ is a random quantity defined over a probability space that generates randomness in policy $\pi$. We use $\Pi_{T,n,d}$ to denote the class of all policies defined above.

To evaluate the performance of a policy $\pi$, we consider its expected regret $\mathbb{E}[R_T]$, defined as the sum of the differences of the rewards between the policy’s choosing actions and the optimal action in hindsight. More specifically, for a policy $\pi$ and a pre-specified action sets sequence $A_1, \ldots, A_T$, the
expected regret is defined as
\[ E[R^T] = E \left[ \sum_{t=1}^{T} \max_{x_t \in A_t} x_t^\top \theta - x_t^\top \bar{\theta} \right]. \tag{1} \]

Clearly, the expected regret defined in Eq. (1) depends both on the policy \( \pi \) and the environment \( \theta \), \( \{A_t\} \). Hence, a policy that has small regret for one set of environment parameters might incur large regret for other sets of environment parameters. To provide a unified evaluation criterion, we adopt the concept of worst-case regret and aim to find a policy that has the smallest possible worst-case regret. More specifically, we are interested in the following defined minimax regret
\[ R(T; n, d) := \inf_{\pi \in \Pi_{T, n, d}} \sup_{\theta \in \mathbb{R}^d} E[R^T]. \tag{2} \]

Note that for \( n = \infty \), the supremum is taken over all closed \( A_t = \{ x \in \mathbb{R}^d : ||x||_2 \leq 1 \} \) for all \( t \).

The minimax framework has been increasingly popular in identifying information-theoretical limits of learning and statistics problems [19], [20], [21] and was applied to bandit problems as well [22].

Note also that, as described in Eq. (2), the problem instances we are considering in this paper are oblivious [23] with finite horizons, meaning that the regression model \( \theta \) and action sets sequences \( \{A_t\}_{t=1}^{T} \) are chosen adversarially before the execution of the policy \( \pi \), and the policy knows the time horizon \( T \) before the first time period \( t = 1 \).

a) Asymptotic notations: For two sequences \( \{a_n\} \) and \( \{b_n\} \), we write \( a_n = O(b_n) \) or \( a_n \lesssim b_n \) if there exists a universal constant \( C \) \( n \to \infty \) such that \( \limsup_{n \to \infty} |a_n|/|b_n| \leq C \). Similarly, we write \( a_n = \Omega(b_n) \) or \( a_n \gtrsim b_n \) if there exists a universal constant \( c > 0 \) such that \( \liminf_{n \to \infty} |a_n|/|b_n| \geq c \). We write \( a_n = \Theta(b_n) \) or \( a_n \approx b_n \) if both \( a_n \lesssim b_n \) and \( a_n \gtrsim b_n \) hold. In asymptotic notations, we will drop base notations of logarithms and use instead \( \log x \) for both \( \ln x \), \( \log_2 x \) as well as logarithms with other constant base numbers. In non-asymptotic scenarios, however, base notations will not be dropped and \( \ln x \) refers specifically to \( \log_e x \).

B. Related Works

The linear contextual bandit setting was introduced by Abe et al. [8]. Auer [7] and Chu et al. [14] proposed the SupLinRel and SupLinUCB algorithms respectively, both of which achieve \( O(\sqrt{dT} \log^{3/2}(nT)) \) regret. When there are \( n = \Theta(d) \) arms per round, Chu et al. [14] showed an \( \Omega(\sqrt{dT}) \) minimax regret lower bound. A detailed account of these results are given in Table I.

Note that our problem requires that there are only finitely many candidate actions per round. When the number of candidate actions is not bounded, Dani et al. [17] and Rusmevichientong and Tsitsiklis [18] showed algorithms that achieve \( O(\sqrt{dT} \log^{3/2}(Td)) \) regret. This bound was later improved to \( O(\sqrt{dT} \log T) \) by Abbasi-Yadkori et al. [15].

Dani et al. [17] also showed an \( \Omega(d \sqrt{T}) \) regret lower bound when there are 2\( \Theta(d) \) candidate actions. Our lower bound, on the other hand, implies an \( \Omega(d \sqrt{T \log T}) \) lower bound for the infinite-action case when the action space changes over time. A detailed discussion is given in Corollary 23 in Sec. III-E.

While this paper focuses on the regret minimization task for linear contextual bandits, the pure exploration scenario also attracts much research attention in both the ordinary bandit setting (e.g. [24], [25], [26], [27]) and the linear contextual setting (e.g. [28], [29], [30]). It is also worth noting that for the ordinary multi-armed bandit problem (where the \( n \) arms are independent and not associated with contextual information), the MOSS algorithm [22] achieves \( O(\sqrt{nT}) \) expected regret; and the matching lower bound was proved by Auer et al. [31]. The idea of using adaptive confidence levels in upper confidence bands has also been extensively studied in the context of finite-armed bandit settings to remove additional logarithmic factors; see, for example, [22], [32], [33], [34]. To our knowledge, the adaptive confidence intervals idea has not yet been applied to contextual bandit problems with the objective of refined regret analysis.

C. Our Results

The main results of this paper are the following two theorems that upper and lower bound the minimax regret \( R(T; n, d) \) for various problem parameter values.

**Theorem 1 (Upper Bound):** For any \( n < \infty \), the minimax regret \( R(T; n, d) \) can be asymptotically upper bounded by \( \text{poly}(\log \log(nT)) \cdot O(\sqrt{dT} \log(T) \log(n)) \).

**Theorem 2 (Lower Bound):** For any small constant \( \epsilon > 0 \), and any \( n, d \), such that \( n \leq 2^{d/2} \) and \( T \geq d (\log_2 n)^{1+\epsilon} \), the minimax regret \( R(T; n, d) \) can be asymptotically lower bounded by \( \Omega(1) \cdot \sqrt{dT \log(T) \log(T/d)} \).

**Remark 3:** In Theorem 1, \( \text{poly}(\log \log(nT)) = (\log \log(nT))^{\gamma} \) for some constant \( \gamma > 0 \); in Theorem 2, the \( \Omega(1) \) notation hides constants that depend on \( \epsilon > 0 \).

Comparing Theorems 1 and 2, we see that the upper and lower bounds nearly match each other up to iterated logarithmic terms when \( n \) (the number of actions per time period) is not too large. While Theorems 1 and 2 technically only apply to finite \( n \) cases, we will also extend the lower bound (Theorem 2) to the \( n = \infty \) case directly (see Corollary 23.
for the detailed statement), and improve the previous result by [17].

So far as we are aware, our Theorem 2 provides the first $\sqrt{T \log T}$-style lower bound under gap-free settings in multi-armed bandit literature. Even when the degrees of freedom for unknown parameters are constants for both problems (i.e., $n = d = O(1)$), our theorem shows that linear bandits is harder than ordinary multi-armed bandits, because of the variation of arms over the time periods, which marks a separation between the two problems.

D. Techniques and Insights

On the upper bound side, we use two main techniques to remove additional logarithmic factors from previous analysis. Our first technique is to use variable confidence levels, by allowing the failure probability to increase as the policy progresses, because late fails usually lead to smaller additionally incurred regret. Our second idea to remove unnecessary logarithmic factors is to use a more careful analysis of each “layer” in the SupLinUCB algorithm [14]. Previous analysis like [7] and [14] uses the total number of time periods $T$ to upper bound the sizes of each layer, resulting in an addition $O(\sqrt{T \log T})$ term as there are $\Theta(\log T)$ layers. In our analysis, we develop a more refined theoretical control over the sizes of each layer, and show that the layer sizes have an exponentially increasing property. With such a property we are able to remove an additional $O(\sqrt{T \log T})$ term from the regret upper bounds.

On the lower bound side, we consider a carefully designed sequence $\{z_t\}$ (see the proof of Lemma 11 for details) which shows the tightness of the elliptical potential lemma, a key technical step in the proof of all previous analysis of linearly parameterized bandits and their variants [7], [14], [15], [17], [18], [35], [36]. The constructed sequence $\{z_t\}$ not only shows the tightness of existing analysis, but also motivated our construction of adversarial problem instances that lower bound regret of general bandit algorithms.

II. Upper Bounds

We propose Variable-Confidence-Level (VCL) SupLinUCB, a variant of the SupLinUCB algorithm [7], [14] that uses variable confidence levels in the construction of confidence intervals at different stages of the algorithm. We then derive an upper regret bound that is almost tight in terms of dependency on the problem parameters, especially the time horizon parameter $T$.

A. The VCL-SupLinUCB Algorithm

Algorithm 1 describes our proposed VCL-SupLinUCB algorithm. The algorithm is a variant of the SupLinUCB algorithm proposed in [7] and [14], with variable confidence levels at different time periods.

1) High-Level Intuitions and Structures of VCL-SupLinUCB: Our proposed VCL-SupLinUCB algorithm is based on the classical SupLinUCB algorithm, which uses the idea of “layered data partitioning” to resolve delicate data dependency problems in sequential decision making. In this subsection we give a high-level description of the core ideas behind SupLinUCB [14].

Recall that in linear contextual bandit, at each time $t \in \{1, 2, \ldots, T\}$ a set of $n$ contextual/feature vectors $\{x_{i,t}\}_{i=1}^n$ are given, and the algorithm is tasked with selecting one vector $i_t \in [n]$ in the hope of maximizing the expected linear payoff $\langle x_{i_t, t}, \theta \rangle$. The most natural way is to obtain an ordinary-least squares (OLS) estimate $\hat{\theta}_t$ on all payoff data collected prior to time $t$, and then select $i_t \in [n]$ that maximizes $\langle x_{i_t, t}, \hat{\theta}_t \rangle$ plus upper confidence bands. One way to understand this choice is because that the OLS estimate is the maximum-likelihood estimate of $\theta$, when the noises $\{\epsilon_t\}$ are i.i.d. Gaussian variables, which, under certain non-degeneration conditions, is consistent and asymptotically efficient, meaning that it reaches the Cramér-Rao lower bound in an asymptotic sense [37]. To deal with the degenerated case when the information matrix $\sum_t x_{i, t} x_{i, t}^\top$ contains too small eigenvalues, the LinUCB-type algorithms ([9], [14]) add an identity matrix to the information matrix as commonly done in ridge regression. We also follow this method in in our VCL-SupLinUCB algorithm. In the original LinRel-type algorithms by Auer [7], a slightly different variant of OLS is adopted to treat this degeneration.

One major disadvantage of this approach, however, is the implicit statistical correlation hidden in the OLS estimate $\hat{\theta}_t$, preventing rigorous analysis attaining tight regret bounds. Indeed, to our knowledge the best regret upper bounds established for such procedures are $O(d \sqrt{T \log^2 T})$ [16], which is not tight in either $O(\log T)$ or $O(d)$ terms.

The work of Auer [7], with follow-ups in [14] and [35], uses the idea of “layered data partitioning” to develop a significantly more complex algorithm (also known as SupLinRel and SupLinUCB) to resolve the data dependency problem. Instead of using all payoff data prior to time $t$ for an OLS estimate, the time periods prior to $t$ are divided into disjoint “layers” $\zeta \in \{0, 1, \ldots, \zeta_0\}$, or more specifically $\{X_{\zeta,t}\}_{\zeta=0}^{\zeta_0}$ such that $X_{\zeta\cap\zeta',t} = \emptyset$ for $\zeta \neq \zeta'$ and $\bigcup_{\zeta=0}^{\zeta_0} X_{\zeta,t} = \{1, 2, \ldots, t-1\}$. For larger values of $\zeta$, the partitioned data subset $X_{\zeta,t}$ is likely to contain more data points prior to $t$, hence the resulting OLS estimate on data from $X_{\zeta,t}$ is likely to be more accurate.

At time $t$, the layers $\zeta = 0, 1, \ldots, \zeta_0$ are visited sequentially and within each layer an OLS estimate $\hat{\theta}_{\zeta,t}$ is calculated on data collected solely during periods in $X_{\zeta,t}$. More specifically, the algorithm starts from $\zeta = 0$ (corresponding to the widest confidence bands) and increases $\zeta$, while in the mean time knocking out all sub-optimal actions using OLS estimates from $X_{\zeta,t}$. The procedure is carefully designed so that the “stopping” layer $\zeta_t$ produced at time $t$ only depends on $\{X_{\zeta,t}\}_{\zeta=\zeta_t}$, which decouples the statistical correlation in OLS estimates. A more careful and rigorous statement of this statistical decoupling property is given in Proposition 1.

2) Notations: In our pseudo-code description of VCL-SupLinUCB (Algorithm 1) there are many notations to help define and clarify this subtle procedure. To help with readability and checking with the technical analysis, we provide a
The analysis of [7] and [14] we are able to pinpoint the sources of
proved again in this paper for completeness. It is essentially the same as that of Lemma 14 in [7], which is
\[ \{X_{\xi,t} \} \]
the UCB algorithm:
from [14] on regret upper bounds of the traditional SupLinUCB
algorithm description and analysis, in Table II.

\section{B. Tight Regret Analysis}

In this section we sketch our regret analysis of Algorithm 1 that gives rise to almost tight \( T \) dependency. To shed lights on the novelty of our analysis, we first review existing results from [14] on regret upper bounds of the traditional SupLinUCB algorithm:

\textbf{Theorem 4 [14]}: The expected cumulative regret of the classical SupLinUCB algorithm can be upper bounded by
\[ O(\sqrt{dT \log^3(nT)}) \]
It is immediately noted that the regret upper bound in Theorem 4 has three \( O(\sqrt{\log T}) \) terms. Digging into the analysis of [7] and [14] we are able to pinpoint the sources of each of the \( O(\sqrt{\log T}) \) terms:

1. One \( O(\sqrt{d}) \) term arises from a union bound over all \( T \) time periods;
2. One \( O(\sqrt{d \log T}) \) term arises from the elliptical potential lemma bounding the summation of squared confidence interval lengths;
3. One \( O(\sqrt{\log T}) \) term arises from the \( O(\log T) \) levels of \( \zeta \in \{0, 1, \ldots, \zeta_0\} \).

In this section, we will focus primarily on our techniques to remove the first and the third \( O(\sqrt{\log T}) \) term, while in the next section we prove a lower bound showing that the \( O(\sqrt{\log T}) \) term arising from the second source cannot be eliminated for any algorithm.

As a first step of our analysis, we revisit the crucial property of statistical independence across resolution levels \( \zeta \), a core property used in prior analysis of SupLinUCB type algorithms [7], [14], [35]. The proof of Proposition 1 is essentially the same as that of Lemma 14 in [7], which is proved again in this paper for completeness.

\textbf{Proposition 1}: For any \( \zeta \) and \( t \), conditioned on \( \{X_{\zeta,t}\}_{t} \leq \zeta \), \( \{\lambda_{\zeta,t}\}_{t} \leq \zeta \) and \( \{\lambda_{\zeta,t}\}_{t} < \zeta \), the random variables \( \{\varepsilon_{\varepsilon'}\}_{\varepsilon' \in \mathcal{C}_{\zeta}} \) are independent, centered sub-Gaussian random variables with variance proxy 1.

\textbf{Proof}: By definition, we have that \( \lambda_{\zeta,t} = \sum_{v \in \mathcal{X}_{\zeta,t}} x_{v}^{T} \theta + \varepsilon_{v} \), where \( \varepsilon_{v} \) are selected (potentially adversarially) before the execution of the bandit algorithm, to prove Proposition 1 it suffices to prove that, conditioned on \( \{X_{\zeta,t}\}_{t} \leq \zeta \), \( \{\varepsilon_{\varepsilon'}\}_{\varepsilon' \in \mathcal{C}_{\zeta}} \) are independent, centered sub-Gaussian random variables.

In the rest of this proof we define \( \mathcal{Y}_{\xi}^{<} := \mathcal{X}_{\xi,t} \cup \cdots \cup \mathcal{X}_{\xi-1,t} \) and \( \mathcal{Y}_{\xi}^{>} := \mathcal{X}_{\xi,t} \cup \cdots \cup \mathcal{X}_{\xi-1,t} \). Define also, for \( t'' \in \mathcal{Y}_{\xi}^{<} \), \( z_{v''} \in \{0, 1, \ldots, \zeta\} \) as the unique integer such that \( v'' \in \mathcal{X}_{z_{v''},t''} \). Let also \( \mathcal{C}(\mathcal{Y}_{\xi}^{<}) \) be any function that is a constant function with respect to \( \{\varepsilon_{\varepsilon'}\}_{\varepsilon' \in \mathcal{C}_{\zeta}} \). We then have, by definition of conditional probability,
\[
P[\{\varepsilon_{\varepsilon'}\}_{\varepsilon' \in \mathcal{C}_{\zeta}} | \{X_{\zeta,t}\}_{t} \leq \zeta, \{\varepsilon_{\varepsilon'}\}_{\varepsilon' \in \mathcal{C}_{\zeta}}] = C(\varepsilon_{\varepsilon'}) \cdot P[\{X_{\zeta,t}\}_{t} \leq \zeta, \{\varepsilon_{\varepsilon'}\}_{\varepsilon' \in \mathcal{C}_{\zeta}}].
\]
For any \( t' \leq t \), define the history \( \mathcal{H}_{t'} := \{X_{\zeta,t'} \}_{t'} = \{\varepsilon_{\varepsilon'}\}_{\varepsilon' \in \mathcal{C}_{\zeta}} \}. Note that \( \mathcal{H}_{t'} \) is
only a “partial” history of the algorithm execution up to time period \( t' \), because only those time periods \( t'' \leq t' \) whose resolution levels \( \zeta_{t''} \) are smaller than or equal to \( \zeta \) are included. The joint distribution of \( \{ X_{t'}, t \leq t \} \), \( \{ i_{t'}, t \leq t \} \), \( \{ \epsilon_{t'}, t \leq t \} \), \( \{ \varphi_{t'}, t \leq t \} \), can then be decomposed as

\[
P[\{ X_{t'}, t \leq t \}, \{ i_{t'}, t \leq t \}, \{ \epsilon_{t'}, t \leq t \}]
= \prod_{t'=1}^{t} \varphi_{t'}(\{ \epsilon_{t'}, t \leq t \}) \times \psi_{t'}(1\{ \varphi_{t'} \leq \zeta \}),
\]

where

\[
\varphi_{t'} := P[\zeta_{t'} = \zeta, \epsilon_{t'}, i_{t'} | \mathcal{H}_{t-1}],
\psi_{t'} := P[\zeta_{t'} > \zeta | \mathcal{H}_{t-1}].
\]

It is now time to review how \( \zeta_{t'} \) and \( i_{t'} \) are selected at time \( t' \) by Algorithm 1. First, note that to decide whether \( \zeta_{t'} \leq \zeta \) or \( \zeta_{t'} > \zeta \), only \( \{ i_{t'}, \epsilon_{t'} \} \in \mathcal{Y}_{t'} \) is used in such a decision rule, because at resolution level \( \zeta_{t'} = \zeta \) the condition in clause 2 of the if-else-if-else loop in Algorithm 1 does not depend on \( \{ \epsilon_{t'}, i_{t'} \} \), whose success would determine whether \( \zeta_{t'} > \zeta \) (when the condition in clause 2 holds) or \( \zeta_{t'} \leq \zeta \) (when the condition in clause 2 fails to hold). Similarly, the decision \( i_{t'} \) made when \( \zeta_{t'} \leq \zeta \) does not depend on \( \{ \epsilon_{t'}, i_{t'} \} \), because in neither the first nor the third clause of the if-else-if-else loop is the estimate \( \hat{\theta}_{t', t} \) used, if \( \zeta_{t'} \leq \zeta \). Both cases imply that

\[
\varphi_{t'} := P[\zeta_{t'} = \zeta, \epsilon_{t'}, i_{t'} | \mathcal{H}_{t-1}],
\psi_{t'} := P[\zeta_{t'} > \zeta | \mathcal{H}_{t-1}].
\]

Now combine Eqs. (3,4) and note that, for every \( t' \leq t \) with \( \zeta_{t'} = \zeta \), the variable \( \epsilon_{t'} \) is not included in the history to be conditioned on in later product probabilities. Hence, we have

\[
P[\{ X_{t'}, t \leq t \}, \{ i_{t'}, t \leq t \}, \{ \epsilon_{t'}, t \leq t \}]
= C(-\{ \epsilon_{t'}, t \leq t \}) \times P[\{ i_{t'}, t \leq t \}],
= C(-\{ \epsilon_{t'}, t \leq t \}) \times \prod_{t' \in X_{t'}} P[\epsilon_{t'}],
\]

which is to be demonstrated.

We next detail how we would remove two of the three \( O(\sqrt{T}) \) terms mentioned in the previous paragraph.

1) Removing the First \( O(\sqrt{T}) \) Term: To remove the first \( O(\sqrt{T}) \) term arising from a union bound over all \( T \) time periods, our main idea is to use variable confidence levels depending on the (square root of the) quadratic form \( \omega_{t', t} = \sqrt{x_{t', t}} x_{t, t} \) instead of constant confidence levels \( 1/poly(T) \) used in [7] and [14]. The following lemma gives an upper bound on the regret of VCL-SupLinUCB:

**Lemma 5:** The sequence of actions \( \{ i_{t'} \}_{t=1}^{T} \) produced by Algorithm 1 satisfies

\[
E[R^T] \leq \sqrt{d T} + E \left[ \sum_{t=1}^{T} \alpha_{t', t} \cdot \omega_{t', t} \right]
\leq \sqrt{d T} + \sqrt{\log(n \log T)}
\cdot \mathbb{E} \left[ \sum_{t=1}^{T} \sqrt{\max(1, \log(T \omega_{t', t}^2/d))} \cdot \omega_{t', t} \right] + \mathbb{E} \left[ \sqrt{\log(T \omega_{t', t}^2/d)} \cdot \omega_{t', t} \right],
\]

(5)

where \( \zeta_i \in \{0, 1, \ldots, \zeta_0 \} \) is the resolution level at time period \( t \) and \( \alpha_{t', t} \cdot \omega_{t', t} \) are defined in Algorithm 1.

Compared to similar lemmas in existing analytical framework [7], [14], the major improvement is the reduction from \( \log(T) \) to \( \log(T \omega_{t', t}^2/d) \) in the multiplier before the main confidence interval length term \( \omega_{t, t} \), meaning that when the \( \{ \theta_{t, t} \} \) shrink as more observations are collected, the overall confidence interval length also decreases. This helps reduce the \( \log(T) \) term, which eventually disappears when \( \omega_{t, t} \) is sufficiently small.

To state our proof of Lemma 5 we define some notations and also present an intermediate lemma. For any \( \zeta \) such that \( \zeta \leq \zeta_i \), define \( m_{t, t} := \max_{i \in \mathcal{N}_{t, t}} x_i \theta \) and \( m_{t, t} := \min_{i \in \mathcal{N}_{t, t}} x_i \theta \) as the largest and smallest mean reward for actions within action subset \( \mathcal{N}_{t, t} \). For convenience, we also define \( \bar{m}_{t, t} := \bar{m}_{t, t} \) and \( \underline{m}_{t, t} := \underline{m}_{t, t} \) for all \( \zeta > \zeta_i \). The following lemma is central to our proof of Lemma 5:

**Lemma 6:** For all \( t \) and \( \zeta = 0, 1, \ldots, \zeta_0 \), it holds that

\[
E[|m_{t, t} - m_{t, t}|] \leq 2 \sqrt{\log \left( \frac{\zeta_0 T}{d} \right)};
\]

\[
E[\max\{m_{t, t} - m_{t, t} - 2\zeta - \epsilon_i \} \cdot b \{ \zeta \} \leq \sqrt{\log \left( \frac{\zeta_0 T}{d} \right)}.
\]

(7)

At a higher level, Eq. (6) states that by reducing the candidate set from \( \mathcal{N}_{t, t} \) to \( \mathcal{N}_{t, t} \), the action corresponding to large rewards is preserved (up to an error term of \( 2 \sqrt{\log \left( \frac{\zeta_0 T}{d} \right)} \)). Eq. (7) further gives an exponentially decreasing upper bound on the differences between the best and the worst actions within \( \mathcal{N}_{t, t} \), corroborating the intuition that as \( \zeta \) increases and we go to more refined levels, the action set \( \mathcal{N}_{t, t} \) should “zoom in” onto the actions with the best potential rewards.

**Proof of Lemma 6:** We prove Eqs. (6) and (7) separately.

Proof of Eq. (6): Note that \( \bar{m}_{t, t} = \bar{m}_{t, t} \) and \( \underline{m}_{t, t} = \underline{m}_{t, t} \) as \( \zeta \geq \zeta_i \) and we therefore only need to prove the inequality conditioned on \( \zeta < \zeta_i \). That is, we only need to upper bound \( E[|m_{t, t} - m_{t, t}|] \mid \zeta \leq \zeta_i \), because

\[
E[|m_{t, t} - m_{t, t}|] \leq E[|m_{t, t} - m_{t, t}|] + E[|m_{t, t} - m_{t, t}|] \mid \zeta \leq \zeta_i] \cdot p \{ \zeta \leq \zeta_i \}
\]

\[
= E[|m_{t, t} - m_{t, t}|] \mid \zeta \leq \zeta_i] \cdot p \{ \zeta \leq \zeta_i \}.
\]

Define \( I^* := \arg \max_{i \in \mathcal{N}_{t, t}} \{ x_i \theta \} \) and \( J^* := \arg \max_{i \in \mathcal{N}_{t, t}} \{ x_i \theta \} \), representing the “optimal” action in \( \mathcal{N}_{t, t} \) under the true model \( \theta \) and the estimated model \( \hat{\theta}_{t, t} \), respectively. If \( I^* \in \mathcal{N}_{t, t} \), then \( \bar{m}_{t, t} = \bar{m}_{t, t} \) because \( \mathcal{N}_{t, t} \subseteq \mathcal{N}_{t, t} \). Therefore, we only need to consider the case of \( I^* \notin \mathcal{N}_{t, t} \) for the sake of proving Eq. (6).

Note that \( J^* \in \mathcal{N}_{t, t} \) always holds, because \( J^* \) maximizes \( x_i \theta \) in \( \mathcal{N}_{t, t} \), and by our algorithm the maximum of \( x_i \theta \), \( i \in \mathcal{N}_{t, t} \) will never be removed unless a decision on \( i \) can already be made. We then have with probability one that

\[
E[|m_{t, t} - m_{t, t}|] \leq \mathbb{E} \left[ (x_{I^*, t} - x_{J^*, t})^2 \right] \theta,
\]

(8)
where the second inequality holds because \( J^* \in \mathcal{N}_{\zeta+1,t} \) and therefore \( x_{j,t}^* \leq \max_{i \in \mathcal{N}_{\zeta+1,t}} x_{i,t}^\theta \).

For any \( \zeta, t \) and \( i \in \mathcal{N}_{\zeta,t} \), define \( E_{\zeta,t}^i := \{ x_{i,t}(\hat{\theta}_{i,t} - \theta) \leq \omega_{\zeta,t}^i \} \) as the success event in which the estimation error of \( x_{i,t}^\theta \) is within the confidence interval \( \omega_{\zeta,t}^i \). By definition, with probability one it holds that

\[
x_{j,t}^\top \theta \leq x_{j,t}^\top \hat{\theta}_{j,t} + \omega_{\zeta,t}^j + 1 \{-E_{\zeta,t}^j \} \cdot |x_{j,t}^\top (\hat{\theta}_{j,t} - \theta)|.
\tag{9}
\]

(10)

Also, conditioned on the event \( I^* \notin \mathcal{N}_{\zeta+1,t} \), the procedure of Algorithm 1 implies

\[
x_{j,t}^\top \hat{\theta}_{j,t} < x_{j,t}^\top \hat{\theta}_{j,t} - 2^{1-\zeta}.
\tag{11}
\]

Subtracting Eq. (10) from Eq. (9) and considering Eq. (11), we have

\[
\begin{align*}
(x_{t,t}^* - x_{j,t}^*)^\top \theta & \leq (x_{j,t}^\top \hat{\theta}_{j,t} - x_{j,t}^\top \hat{\theta}_{j,t}) + \omega_{\zeta,t}^j + \omega_{\zeta,t}^j + \sum_{i \in \{1, J^*\}} 1 \{-E_{\zeta,t}^i \} \cdot |x_{i,t}^\top (\hat{\theta}_{i,t} - \theta)| \\
& \leq \omega_{\zeta,t}^j + \omega_{\zeta,t}^j - 2^{1-\zeta} + \sum_{i \in \{1, J^*\}} 1 \{-E_{\zeta,t}^i \} \cdot |x_{i,t}^\top (\hat{\theta}_{i,t} - \theta)|,
\end{align*}
\tag{12}
\]

where the last inequality holds because \( \omega_{\zeta,t}^j \leq 2^{-\zeta} \) for all \( i \in \mathcal{N}_{\zeta,t} \), if the algorithm is executed to resolution level \( \zeta + 1 \). Combining Eqs. (8,12) and taking expectations, we obtain

\[
\begin{align*}
\mathbb{E}[\bar{m}_{\zeta,t} - \bar{m}_{\zeta+1,t} | \zeta < \zeta_t] & \leq \mathbb{E}\left[ 1 \{ I^* \notin \mathcal{N}_{\zeta+1,t} \} \cdot \left( 1 \{-E_{\zeta,t}^i \} |x_{i,t}^\top (\hat{\theta}_{i,t} - \theta)| + 1 \{-E_{\zeta,t}^i \} |x_{j,t}^\top (\hat{\theta}_{j,t} - \theta)| \right) \right] \leq 2^{2-\zeta} + \omega_{\zeta,t}^j + \omega_{\zeta,t}^j + \sum_{i \in \mathcal{N}_{\zeta,t}} 1 \{-E_{\zeta,t}^i \} \cdot |x_{i,t}^\top (\hat{\theta}_{i,t} - \theta)|,
\end{align*}
\tag{13}
\]

The following lemma gives an upper bound on the two terms in Eq. (13):

**Lemma 7**: For any \( \zeta, t \) and \( i \), we have that

\[
\mathbb{E}[1 \{-E_{\zeta,t}^i \} \cdot |x_{i,t}^\top (\hat{\theta}_{i,t} - \theta)| | \zeta < \zeta_t, i \in \mathcal{N}_{\zeta,t}] \leq 2\sqrt{\pi d/(n\zeta_0 \sqrt{T})}.
\]

Lemma 7 can be proved by using the statistical independence between \( \{E_{\zeta,t}^i \} \forall i \in \mathcal{N}_{\zeta,t} \) and \( \{x_{i,t}^\top \} \forall i \in \mathcal{N}_{\zeta,t} \), (Proposition 1), and integration of least-squares estimation errors. As its proof is technical but rather routine, we defer it to the appendix. With Lemma 7 and Eq. (13), we have that

\[
\begin{align*}
\mathbb{E}[\bar{m}_{\zeta,t} - \bar{m}_{\zeta+1,t} | \zeta < \zeta_t] & \leq 2\sqrt{\pi d/(n\zeta_0 \sqrt{T})} \\
& \leq 2\sqrt{\pi d/(n\zeta_0 \sqrt{T})} + \sum_{i \in \mathcal{N}_{\zeta,t}} 1 \{-E_{\zeta,t}^i \} \cdot |x_{i,t}^\top (\hat{\theta}_{i,t} - \theta)| \leq 2^{2-\zeta} + \omega_{\zeta,t}^j + \omega_{\zeta,t}^j + \sum_{i \in \mathcal{N}_{\zeta,t}} 1 \{-E_{\zeta,t}^i \} \cdot |x_{i,t}^\top (\hat{\theta}_{i,t} - \theta)|.
\end{align*}
\tag{14}
\]

In addition, because both \( I^* \) and \( J^* \) belong to \( \mathcal{N}_{\zeta,t} \), Line 12 of Algorithm 1 implies that

\[
x_{j,t}^\top \hat{\theta}_{j,t} \geq x_{j,t}^\top \hat{\theta}_{j,t} - 2^{1-\zeta} \geq x_{j,t}^\top \hat{\theta}_{j,t} - 2^{2-\zeta}.
\tag{15}
\]

(16)

Subtracting Eq. (14) from Eq. (15) and applying Eq. (16), we have

\[
\begin{align*}
x_{j,t}^\top \theta - x_{j,t}^\top \theta & \leq x_{j,t}^\top \hat{\theta}_{j,t} - x_{j,t}^\top \hat{\theta}_{j,t} + \omega_{\zeta,t}^j + \omega_{\zeta,t}^j \leq 2^{2-\zeta} + \omega_{\zeta,t}^j + \omega_{\zeta,t}^j + \sum_{i \in \mathcal{N}_{\zeta,t}} 1 \{-E_{\zeta,t}^i \} \cdot |x_{i,t}^\top (\hat{\theta}_{i,t} - \theta)|, \tag{17}
\end{align*}
\]

where the last inequality is because of Line 11 of Algorithm 1, such that \( \omega_{\zeta,t}^j \leq 2^{2-\zeta} \).

Let \( \bar{E}_{\zeta,t} \) be the event that \( \zeta - 1 < \zeta_t \) and \( \bar{E}_{\zeta,t} \) be the event that \( \zeta - 1 < \zeta_t \) and \( i \in \mathcal{N}_{\zeta+1,t} \). In total, we have

\[
\begin{align*}
\mathbb{E}\left[ \max\left\{ x_{j,t}^\top \theta - x_{j,t}^\top \theta - 2^{2-\zeta}, 0 \right\} \left| \zeta < \zeta_t \right. \right] & \leq \mathbb{E}\left[ \max\left( 0, \sum_{i \in \mathcal{N}_{\zeta+1,t}} 1 \{-E_{\zeta,t}^i \} \cdot |x_{i,t}^\top (\hat{\theta}_{i,t} - \theta)| \right) \left| \bar{E}_{\zeta,t} \right. \right] \\
& = \mathbb{E}\left[ 1 \{-E_{\zeta,t}^i \} \cdot |x_{i,t}^\top (\hat{\theta}_{i,t} - \theta)| \left| \bar{E}_{\zeta,t} \right. \right].
\end{align*}
\tag{18}
\]
If the third clause is active, we have that Eq. (7) is then proved.

In the above derivation, all steps are straightforward except for the second inequality in Eq. (19), which we explain in more details here. This inequality is derived by a case analysis on how \( i_t \) is selected. According to Algorithm 1, there are only two cases in which \( i_t \) is decided: the first clause or the third clause in the “if-elseif-else” loop in Algorithm 1. If the first clause is active, we have that

\[
2^{3^{-\zeta_t}} \leq 8\sqrt{d/T}.
\]

If the third clause is active, we have that

\[
\alpha_{\zeta_{t+1}, t}^{i_t} \geq 2^{-\zeta_t}.
\]

Combining Eqs. (17, 18, 20), we have

\[
E[R^T] \leq E \sum_{t=1}^{T \wedge \tau} \left( \sqrt{d/T} + \alpha_{\zeta_{t}, t}^{i_t} \omega_{\zeta, t}^{i_t} \right)
\]

which is to be demonstrated.

2) Removing the Third \( O(\sqrt{\log T}) \) Term: In order to remove the third source of \( O(\sqrt{\log T}) \) term, our analysis goes one step beyond the classical elliptical potential analysis (see Lemma 9 in later sections) to have more refined controls of the cumulative regret within each resolution level \( \zeta \). More specifically, we establish the following main lemma upper bounding the sums of confidence band lengths:

**Lemma 8:** For any \( \zeta \), let \( \mathcal{X}_t = \mathcal{X}_{\zeta, t} \) be all time periods \( t \) such that \( \zeta_t = \zeta \), and define \( T_\zeta = |\mathcal{X}_\zeta| \). Then the following hold with probability one:

\[
\sum_{t \in \mathcal{X}_\zeta} \alpha_{\zeta_t, t}^{i_t} \omega_{\zeta, t}^{i_t} \leq 2^{1^{-\zeta}} T_\zeta
\]

\[
\sum_{t \in \mathcal{X}_\zeta} \alpha_{\zeta_t, t}^{i_t} \omega_{\zeta, t}^{i_t} \leq \sqrt{dT_\zeta \log(T_\zeta) \log(eT/T_\zeta) \log n}
\]

\[
\times \text{poly}(\log(\log(nT)))
\]

\( \forall 0 \leq \zeta \leq \zeta_0; \)

\( T_\zeta \leq 4^\zeta d \log^4(nT) \leq \zeta \leq \zeta_0; \)

We remark on some interesting aspects of the results in Lemma 8. First, we improve the \( \log(nT) \) term that is common to previous elliptical potential lemma (Lemma 9) analysis to \( \sqrt{\log(T_\zeta) \log(T/T_\zeta) \log(n)} \) by exploiting the power of Lemma 5 and an application of Jensen’s inequality on \( f(x) = \sqrt{x \ln(Tx/d)} \) instead of the more commonly used \( f(x) = \sqrt{x} \). We also impose an additional upper bound of \( 2^{1^{-\zeta}} T_\zeta \) and an exponentially-increasing upper bound on \( T_\zeta \) by carefully analyzing the procedures of Algorithm 1.

**Proof of Lemma 8:** We prove the three inequalities in Lemma 8 separately. We first prove, for all \( \zeta > 0 \), that\n
\[
\sum_{t \in \mathcal{X}_\zeta} \alpha_{\zeta_t, t}^{i_t} \omega_{\zeta, t}^{i_t} \leq 2^{1^{-\zeta}} T_\zeta.
\]

Because \( \zeta > 0 \), we have that \( \alpha_{\zeta_t, t}^{i_t} \omega_{\zeta, t}^{i_t} = \omega_{\zeta, t}^{i_t} \leq 2^{-1^{-\zeta}} \) for all \( t \in \mathcal{X}_\zeta \) by the second clause of the if-elseif-else loop of Algorithm 1. The inequality immediately follows.

We next prove the second inequality in Lemma 8. Below we state a version of the celebrated elliptical potential lemma, key to many existing analysis of linearly parameterized bandit problems [7], [14], [15], [35], [36]. For completeness, we also include its proof here.

**Lemma 9** [15]: For any vectors \( y_1, y_2, \ldots, y_T \) with \( \ell_2 \)-norms upper bounded by 1, define \( U_0 = I \) and \( U_t = U_{t-1} + y_t y_t^\top \) for \( t \geq 1 \). It then holds that\n
\[
\sum_{t=1}^{T \wedge \tau} y_t^\top U_{t-1}^{-1} y_t \leq 2 \ln(\det(U_{t-1})).
\]

**Proof:** For each \( t \geq 1 \), by the definition of \( U_t \) and noting that \( U_{t-1} \) is positive definite (and therefore invertible), we have that

\[
\det U_t = \det(U_{t-1} + y_t y_t^\top)
\]

\[
= \det \left[ U_{t-1}^{1/2} (I + U_{t-1}^{-1/2} y_t y_t^\top U_{t-1}^{-1/2}) U_{t-1}^{-1/2} \right]
\]

\[
= (\det(U_{t-1})) \cdot \det(I + U_{t-1}^{-1/2} y_t y_t^\top U_{t-1}^{-1/2}).
\]
Since $\det \left( I + \left( U_{t-1}^{-1/2} y_t \right) \left( U_{t-1}^{-1/2} y_t \right)^T \right) = 1 + \| U_{t-1}^{-1/2} y_t \|_2^2$, we further have that
\[
\det U_t = (\det U_{t-1})(1 + \| U_{t-1}^{-1/2} y_t \|_2^2) \\
\geq (\det U_{t-1}) \exp \left( \frac{1}{2} \| U_{t-1}^{-1/2} y_t \|_2^2 \right) \tag{21}
\]
holds for every $t \geq 1$, where the inequality is due to that $1+x \geq \exp(x/2)$ for all $x \in [0,1]$ and $\| U_{t-1}^{-1/2} y_t \|_2^2 \leq 1$. Unrolling Eq. (21) for $t = T, T-1, T_2, \ldots, 1$, we have that
\[
\det U_T \geq (\det U_0) \exp \left( \frac{1}{2} \sum_{t=1}^{T} \| U_{t-1}^{-1/2} y_t \|_2^2 \right) \\
= \exp \left( \frac{1}{2} \sum_{t=1}^{T} \| U_{t-1}^{-1/2} y_t \|_2^2 \right).
\]
Taking the natural logarithm and multiply both sides by a factor of 2, we prove the desired inequality.

With Lemma 9, we can prove the second inequality by applying Jensen’s inequality and the concavity of $f(x) = \sqrt{x}$ and $f(x) = x \log(Tx/d)$. More specifically, let $X^+ = \{ t \in X^\alpha | \omega^\alpha_{t,t} \geq \sqrt{T/d} \}$, and $T^+ = |X^+|$. Note that, for all $t \in X^\alpha \setminus X^+$, because $\omega^\alpha_{t,t} < \sqrt{T/d}$, it holds that $\max\{ 1, \log[ T(\omega^\alpha_{t,t})^2 / d ] \} \leq 1$. Subsequently, by definition of $\alpha^\alpha_{t,t}$ and $\omega^\alpha_{t,t}$, we have
\[
\sum_{t \in X^\alpha} \alpha^\alpha_{t,t} \omega^\alpha_{t,t} \leq \sqrt{\log(n \log T)} \\
\cdot \sum_{t \in X^\alpha} \sqrt{\max\{ 1, \log[ T(\omega^\alpha_{t,t})^2 / d ] \} \cdot \omega^\alpha_{t,t}} \\
\leq \sqrt{\log(n \log T)} \left( T^+ \cdot \frac{1}{T^+} \sum_{t \in X^\alpha \setminus X^+} \omega^\alpha_{t,t} \right) + T^+ \left( \frac{1}{T^+} \sum_{t \in X^\alpha} \sqrt{\log[ T(\omega^\alpha_{t,t})^2 / d ]} \cdot \omega^\alpha_{t,t} \right) \\
\leq \sqrt{\log(n \log T)} \left( T^+ \cdot \frac{1}{T^+} \sum_{t \in X^\alpha \setminus X^+} \omega^\alpha_{t,t} \right) + T^+ \left( \frac{1}{T^+} \sum_{t \in X^\alpha} (\omega^\alpha_{t,t})^2 \right) \\
\leq \sqrt{\log(n \log T)} \left( T^+ \cdot \frac{1}{T^+} \sum_{t \in X^\alpha \setminus X^+} (\omega^\alpha_{t,t})^2 \right) + T^+ \left( \frac{1}{d} \frac{1}{T^+} \sum_{t \in X^\alpha \setminus X^+} (\omega^\alpha_{t,t})^2 \right) \\
\leq \sqrt{\log(n \log T)} \left( T^+ \cdot \frac{1}{T^+} \sum_{t \in X^\alpha \setminus X^+} (\omega^\alpha_{t,t})^2 \right) + T^+ \left( \frac{1}{d} \frac{1}{T^+} \sum_{t \in X^\alpha \setminus X^+} (\omega^\alpha_{t,t})^2 \right),
\]
and noting that $\ln \det(\Lambda_{\alpha,T}) \leq d \ln(T^+ + 1)$ because $\| x_t \|_2^2 \leq 1$ for all $t$, we have
\[
\sum_{t \in T^+} \alpha^\alpha_{t,t} \omega^\alpha_{t,t} \leq \sqrt{\log(n \log T)} \\
\cdot \left( \sqrt{dT_T \log(T^+)} + \sqrt{\log[ (T \log T^+)^2 ]} \cdot \sqrt{dT_T \log(T^+)} \right).
\]
Since $T^+ \log[ (T \log T^+)^2 ] \leq T^+ + T_T \log( (T \log T^+)^2 )$ holds for $T^+ \leq T_T$, we further have
\[
\sum_{t \in T^+} \alpha^\alpha_{t,t} \omega^\alpha_{t,t} \leq \sqrt{\log(n \log T)} \\
\cdot \left( \sqrt{dT_T \log(T^+)} + \sqrt{\log[ (T \log T^+)^2 ]} \cdot \sqrt{dT_T \log(T^+)} \right),
\]
which proves the second inequality in Lemma 8. Note that, unlike the other two inequalities, this inequality holds for the first resolution level $\alpha = 0$ as well.

We next prove the last inequality in Lemma 8 which upper bounds $T_T$ for $\alpha > 0$. By the second clause of the if-elseif-else line of Algorithm 1, we know that $\omega^\alpha_{t,t} = \alpha^\alpha_{t,t} \omega^\alpha_{t,t} \geq 2^{1-\alpha}$ for all $t \in X^\alpha$. Subsequently, $$(2^{-\alpha-1})^2 \cdot T_T \leq \sum_{t \in X^\alpha} \omega^\alpha_{t,t}^2 \leq \max_{t \in X^\alpha} \alpha^\alpha_{t,t} \omega^\alpha_{t,t} \leq 2^{\alpha} \cdot \sum_{t \in X^\alpha} (\omega^\alpha_{t,t})^2 \leq \log(T^+ \cdot d) \cdot \sqrt{\log(T^+ \cdot d)} \cdot \sqrt{dT_T \log(T^+)} \cdot \sqrt{dT_T \log(T^+)},$$
where the last inequality holds by applying Lemma 9. Re-arranging the terms we obtain $T_T \leq 4^{d} \log^{4}(nT)$, which is to be demonstrated.

3) Putting it Together: We are now ready to combine Lemmas 5, 8 to prove our main result in Theorem 1. We first divide the resolution levels $\alpha \in \{ 0, 1, \ldots, \zeta_0 \}$ into three different sets: $Z_0 := \{ 0 \}$, $Z_1 := \{ 1, \ldots, \zeta_0 \}$ and $Z_2 := \{ \zeta : \zeta^* < \zeta \leq \zeta_0 \}$, where $\zeta^*$ is an integer to be defined later. Clearly $Z_0$, $Z_1$ and $Z_2$ partition $\{ 0, \ldots, \zeta_0 \}$. The summation $E(\sum_{t=1}^{T} \alpha^\alpha_{t,t} \cdot \omega^\alpha_{t,t})$ on the right-hand side of Eq. (5) in Lemma 5 can then be carried out separately (to simplify notations we denote $\gamma_{n,T} := \text{poly}(\log(nT))$; all inequalities below hold with probability one, following Lemma 8):
\[
\sum_{\zeta \in Z_0} \sum_{t \in X^\alpha} \alpha^\alpha_{t,t} \omega^\alpha_{t,t} \leq \gamma_{n,T} \sqrt{dT_T \log(T^+ \cdot d \log( T^+ / T_0 ))} \log n \leq \gamma_{n,T} \sqrt{dT_T \log T \log n}; \tag{22}
\]
\[
\sum_{\zeta \in Z_1} \sum_{t \in X^\alpha} \alpha^\alpha_{t,t} \omega^\alpha_{t,t} \leq \sum_{\zeta = 1}^{\zeta_0} \left( 2^{-\zeta} \cdot 4^{d} \log^{4}(nT) \right) \cdot \gamma_{n,T} \leq 2^{\zeta+1} \cdot d \log^{4}(nT) \cdot \gamma_{n,T}; \tag{23}
\]
\[
\sum_{\zeta \in Z_2} \sum_{t \in X^\alpha} \alpha^\alpha_{t,t} \omega^\alpha_{t,t} \leq \sum_{\zeta \in Z_2} \sqrt{dT_T \log T \log(eT) \log n} \cdot \gamma_{n,T} \leq \sqrt{|Z_2| \cdot d \sum_{\zeta \in Z_2} T_T} \log T \log( eT \cdot |Z_2| / T_T ) \log n \cdot \gamma_{n,T} \tag{24}
\]
Eq. (22) holds by a case analysis on $T_0$: if $T_0 \leq T/\log T$ then $d T_0 \log(T_0) \log(eT/T_0) \log n \leq \sqrt{\log T \log n}$; if, on the other hand, $T/\log T < T_0 \leq T$, then $\sqrt{dT \log T \times \log(eT/T_0) \log n} \leq \sqrt{dT \log T \times \log(eT/T) \log n}$ 

Eq. (24) holds by applying Jensen’s inequality on the convex function $x \mapsto \sqrt{x \log(eT/x)}$. Eq. (25) holds by applying the monotonicity of the function $x \mapsto \sqrt{x \log(eT/x)}$ and the fact that $x := \sum_{z \in \mathbb{Z}} T z \leq T$.

Recall that $c_0 = [\sqrt{\log_4(T/d)}]$ and therefore $T/d \leq 2c_0 \leq 2\sqrt{T/d}$. Select $c^* = c_0 - 4 \log_2 e \cdot \log(\log(T))$; we have that $|\mathbb{Z}| = O(\log(\log(T)))$ and $c^* \leq 2\sqrt{T/(\log^4(T))}$.

Adding Eqs. (23,25) and using Lemma 5, we obtain the main upper bound result of this paper (Theorem 1).

### III. LOWER BOUNDS

In this section we establish our main lower bound result (Theorem 2). To simplify our analysis, we shall prove instead the following lower bound result, which places more restrictions on the problem parameters $n,d,$ and $T$:

**Theorem 10**: Suppose $T \geq d^5$ and $n = 2d^2/2$. Then $\Omega(n \cdot d \sqrt{T \log T}.$

Theorem 10 can be easily extended to the case of $n < 2d^2/2$ and $T < d^3$ as well, by a zero-filling trick and reducing the effective dimensionality of the constructed adversarial instances. We place the proof of this extension (which eventually leads to a proof of Theorem 2) in Sec. III-D and shall focus on proving Theorem 10 first.

In Sec. III-A, we provide a short argument on the tightness of the elliptical potential lemma which is critically used in most existing analysis for linear bandit algorithms. This is done via a novel construction of the sequence $\{z_t\}$ and intuitively explains the necessity of an $O(\log n)$ factor in all known regret bounds whose analysis is based on the potential lemma. However, it requires several new ideas to show the desired lower bound for all algorithms.

In Sec. III-B, we first prove Theorem 10 for the special case when $d = 2$, as a warmup. We will demonstrate how to use the sequence $\{z_t\}$ to construct a collection of instances. Thanks to the properties of $\{z_t\}$, a suboptimal pull at each round will contribute $\Omega(\sqrt{T \log T})$ regret, and we will show that if we randomly choose a constructed adversarial instance, any policy will have $\Omega(1)$ probability to make a suboptimal pull at each round, and hence the $\Omega(\sqrt{T \log T})$ total regret is proved.

In Sec. III-C, we extend the construction and analysis to general $d$. The construction for general $d$ is obtained via a direct-product fashion operation to $d/2$ copies of the adversarial instances constructed for $d = 2$ so that there are $n = 2^d/2$ arms at each round. We will show that if we randomly choose an adversarial instance from the constructed class, any policy will suffer $\Omega(d \sqrt{T \log T})$ regret.

#### A. Tightness of the Elliptical Potential Lemma

To motivate our construction of adversarial bandit instances, in this section we give a warm-up exercise showing the critical elliptical potential lemma (Lemma 9) used heavily in our analysis and existing analysis of linear contextual bandit problems is in fact tight [7], [14], [15], [35], [36], even for the univariate case.

**Lemma 11**: For any $T \geq 1$, there exists a sequence $z_1, z_2, \ldots, z_T \in [0,1]$, such that if we let $V_0 = 1$ and $V_t = V_{t-1} + z_t z_t^T$ for $t \geq 1$, then

$$\sum_{t=1}^{T} \sqrt{z_t^2/V_{t-1}} \geq \sqrt{\frac{T}{2} \cdot \ln T}.$$  

As a remark, using Lemma 9 and Cauchy-Schwarz inequality we easily have

$$\sqrt{\sum_{t=1}^{T} z_t^2/V_{t-1}} \leq \sqrt{T \log V_T} \leq \sqrt{T \log n}$$

for all sequences $z_1, \ldots, z_T \in [0,1]$. Lemma 11 essentially shows that this argument cannot be improved, and therefore current analytical frameworks of SupLinUCB [14] or SupLinRel [7] cannot hope to get rid of all $O(\log T)$ terms. While such an argument is not a rigorous lower bound proof as it only applies to specific analysis of certain policies, we find still the results of Lemma 11 very insightful, which also inspires the construction of adversarial problem instances in our formal lower bound proof later.

**Proof of Lemma 11**: Let $S_t = (1 + \ln T/2T)^t$ for all $t \geq 0$, and let $z_t = \sqrt{S_{t-1} \ln T/2T}$ for all $t \geq 1$. Note that $z_t$ is a monotonically increasing function of $t$; and for $T \geq 1$ it holds that $S_{t-1} = (1 + \ln T/2T)^{t-1} \leq T$. Therefore, for any $t \leq T$, we may verify that $z_t \in [0,1]$ since

$$z_t \leq z_T = \sqrt{S_{T-1} \ln T/2T} = \sqrt{T \ln T/2T} < 1.$$  

Now we verify Eq. (26). Note that for any $t \leq T$, we have

$$V_t = 1 + \sum_{j=1}^{t} z_j z_j^T = 1 + \ln T/2T \sum_{j=1}^{t} \left(1 + \ln T/2T\right)^{j-1} = \left(1 + \ln T/2T\right)^t = S_t.$$  

Therefore, we have

$$\sum_{t=1}^{T} \sqrt{z_t^2/V_{t-1}} = \sum_{t=1}^{T} |V_{t-1}^{-1/2} z_t| = \sum_{t=1}^{T} \sqrt{\ln T/2T} = \sqrt{T \ln T/2T}.$$  

\[\square\]

#### B. Warmup: the Lower Bound Theorem for $d = 2$

In this subsection, we first prove the lower bound theorem (Theorem 10) for $d = 2$ as a warmup. In this special case, there are only $n = 2d^2/2 = 2$ arms during each time period.

1. **Construction of Adversarial Problem Instances:** We will construct a finite set of bandit problem instances that will serve as the adversarial construction of our lower bound proof for general policies. We start with the definitions of stages and intervals.
that unique in stage 1, which is the leaf nodes/intervals then correspond to a regression model evenly partitioned into 3 parts and the middle part is left out. Each of the $d$-Cantor set $S$ intervals partitioned into 3

The following lemma lower bounds the regret incurred by each suboptimal policy makes a suboptimal pull at time $t$, we set $\|\tilde{\theta}(u)\|_2 \leq 1$ for every $u$. We then construct the set of context vectors $\{x^{(u)}_{t,i}\}_{i \in \{0,1\}, t \in [T]}$. For any round $t$ that belongs to stage $j$, we set $x^{(u)}_{t,i} = (z_t, 0)$ and $x^{(u)}_{1,t} = (0, ((\alpha_u^{-1} + \beta_u^{-1})/2) \cdot 2z_t)$. On may verify (using Eq. (27)) that for sufficiently large $T$, we have that $\|x^{(u)}_{t,i}\|_2 \leq 1$ for all $i$ and $t$.

2) The Analysis for Suboptimal Pulls: Since there are only two candidate actions at each time $t$, we call the action with the smaller expected reward to be a suboptimal pull. The following lemma lower bounds the regret incurred by each suboptimal pull.

**Lemma 12:** For any instance $B^{(u)}$ and any time $t$, if a policy makes a suboptimal pull at time $t$, then the incurred regret is at least $\sqrt{nT}/(36\sqrt{T})$.

**Proof:** Assuming time $t$ is in stage $j$, we have that the incurred regret is

$$\left(\|z_{t}\|^{-1} - 0, ((\alpha_u^{-1} + \beta_u^{-1})/2) \cdot 2z_t\right)^{\top} \theta(u) = z_t \cdot \left|\gamma_u - \frac{\alpha_u^{-1} + \beta_u^{-1}}{2} \right| \geq z_t \cdot \left|\frac{\alpha_u^{-1} - \beta_u^{-1}}{6}\right|.$$
Since \( z_t = \sqrt{\frac{S_t-1}{2T} \ln T} \) and \( |\alpha_t^{-1} - \beta_t^{-1}| = 3^{-j} \), we have
\[
z_t \cdot \frac{|\alpha_t^{-1} - \beta_t^{-1}|}{\sqrt{S_t-1}} \geq \frac{1}{6 \cdot 3^j} \sqrt{\frac{S_t-1}{2T} \ln T} \geq \sqrt{\ln T} \frac{1}{36T},
\]
where the last inequality is because of Eq. (30).

For any policy \( \pi \), underlying model \( \theta(u) \), let \( p_t^{u,\pi} \) be the probability that \( \pi \) makes a suboptimal pull at time \( t \). We have the following corollary.

**Corollary 13:** \( E[R^T] \geq \sum_{t=1}^T p_t^{u,\pi} \cdot \sqrt{\ln T}(36/T) \).

The following lemma lower bounds \( p_t^{u,\pi} \).

**Lemma 14:** For any stage \( j \), let \( u, u' \in \{0, 1, 2, \ldots, 2k-1\} \) be two parameters such that \( \tau_j^{-1} = \tau_j'^{-1} \) but \( \tau_j^{\prime} \neq \tau_j^u \). (The definition of \( \tau \) can be found in Sec. III-B1.) Then for any policy \( \pi \) and time period \( t \) in stage \( j \), it holds that \( p_t^{u,\pi} + p_t^{u',\pi} \geq 1/2 \).

Intuitively, Lemma 14 holds because at any time and in stage \( j \), there is always a suboptimal action. If the model parameter is \( \theta^{u} \), and \( \theta^{u'} \) is the same, then the differences of the mean rewards at any time \( t' > t \) are the same.

**Proof of Lemma 14:** By our construction, we have that
\[
(1, 0)^T (\theta^{u} - \theta^{u'}) \leq \left( \frac{1}{3} \right)^j.
\]
Therefore, Claim 15 (proved below), for any event \( E \) at time \( t \) before the end of stage \( j \), we have that
\[
|\Pr[E|u] - \Pr[E|u']| \leq \frac{1}{2} \left( 1, 0 \right)^T (\theta^{u} - \theta^{u'}) \sqrt{S_t}
\]
\[
\leq \frac{1}{2} \left( \frac{1}{3} \right)^j \sqrt{S_t} \leq \frac{1}{2}.
\]
(32)

The last inequality holds because at any time \( t \) in stage \( j \), it holds that \( S_t \leq 9^j \).

Note that, at any time \( t \), if the model parameter is \( u \), the difference between the rewards of the two possible actions is
\[
(x_{0,t}^{(u)} - x_{1,t}^{(u)})^T \left( \theta_{2x-1,t}^{(u)} - \theta_{2x,t}^{(u)} \right) = \frac{z_t}{2} (2\gamma_u - \alpha_u^{-1} - \beta_u^{-1}).
\]
This value is greater than 0 if and only if \( 2\gamma_u > \alpha_u^{-1} + \beta_u^{-1} \). Since \( \tau_j^{u} = \tau_j^{u'} \) and \( \tau_j^{u'} = \tau_j^{u'} \), by our construction Eq. (31), we have that exactly one of \( \gamma_u \) and \( \gamma_{u'} \) is greater than \( \frac{1}{2} (\alpha_u^{-1} + \beta_u^{-1}) \). In other words, at time \( t \), any arm that is suboptimal for the model parameter is \( u \) is not suboptimal for parameter \( u' \), and vice versa. In light of this, let \( E \) be the event that at time \( t \) policy \( \pi \) pulls an arm that is \( s \)-suboptimal for parameter \( u' \), and we have that the complement event \( \bar{E} \) is that at time \( t \) policy \( \pi \) pulls an arm that is \( s \)-suboptimal for parameter \( u ' \). By Eq. (32), we have
\[
p_t^{u,\pi} + p_t^{u',\pi} = \Pr[E|u] + \Pr[\bar{E}|u'] = 1 + \Pr[E|u] - \Pr[E|u'] \geq \frac{1}{2}.
\]

**Claim 15:** For any \( u, u' \), let \( j \) be the largest number such that \( \tau_j^{-1} = \tau_j'^{-1} \). For any time \( t \leq t_j \) and any event \( E \) that happens at time \( t \), we have
\[
|\Pr[E|u] - \Pr[E|u']| \leq \frac{1}{2} (1, 0)^T (\theta^{(u)} - \theta^{(u')}) \sqrt{S_t}.
\]

**Proof:** Note that by our construction, at any time \( t \leq t_j \), the contextual vectors of both \( B^{(u)} \) and \( B^{(u')} \) are the same. Moreover, for any hidden vector and any arm, the reward distribution is a shifted standard Gaussian with variance \( 1 \).

For any time \( t \leq t_j \), let \( D_1 \) be the product of the arm reward distributions at and before round \( t \) when the hidden vector is \( \theta^{(u)} \), and let \( D_2 \) be the same product distribution when the hidden vector is \( \theta^{(u')} \). Since the second dimensions of \( \theta^{(u)} \) and \( \theta^{(u')} \) are the same, the differences of the mean rewards at any time \( t' > t \) is either \( \theta^{(u)}_{z_{t'}} - \theta^{(u')}_{z_{t'}} \) (if the first arm is pulled) or 0 (if the second arm is pulled), where \( \theta^{(u)}_{z_{t'}} \) and \( \theta^{(u')}_{z_{t'}} \) denote the first dimensions of the corresponding vectors. Note that the KL divergence between two variance-1 Gaussians with means \( \mu_1 \) and \( \mu_2 \) is \( |\mu_1 - \mu_2|^2 / 2 \). Therefore, we have
\[
\text{KL}(D_1||D_2) \leq \frac{1}{2} \sum_{t'=1}^t \left( \theta^{(u)}_{z_{t'}} - \theta^{(u')}_{z_{t'}} \right)^2 = \frac{1}{2} (1, 0)^T (\theta^{(u)} - \theta^{(u')})^2 \sum_{t'=1}^t z_{t'}^2.
\]
\[
\leq \frac{1}{2} (1, 0)^T (\theta^{(u)} - \theta^{(u')})^2 \left( 1 + \sum_{t'=1}^t z_{t'}^2 \right) = \frac{1}{2} (1, 0)^T (\theta^{(u)} - \theta^{(u')})^2 S_t.
\]

Therefore, at time \( t \), and for any event \( E \), we have
\[
|\Pr[E|u] - \Pr[E|u']| \leq \frac{1}{2} \text{KL}(D_1||D_2) \leq \frac{1}{2} (1, 0)^T (\theta^{(u)} - \theta^{(u')})^2 \sqrt{S_t}
\]
where the first inequality holds because of Pinsker’s inequality (Lemma 25).

3) The Average-Case Analysis: We are now ready to prove Theorem 10 assuming \( d = 2 \). Recall that \( \{0, 1, 2, \ldots, 2k-1\} \) is the finite collection of the parameters \( u \) for the adversarial bandit instances we constructed in Sec. III-B1, and \( p_t^{u,\pi} \) is the probability of an suboptimal pull at time \( t \). The minimax regret \( \mathfrak{R}(T; n = 2, d = 2) \) can then be lower bounded by
\[
\mathfrak{R}(T; 2, 2) \geq \inf_{\pi} \max_{U \in U_t} E[R^T] \geq \inf_{\pi} \max_{u \in \{0, 1, 2, \ldots, 2k-1\}} \frac{\sqrt{\ln T}}{30\sqrt{T}} \sum_{t=1}^T p_t^{U,\pi},
\]
\[
\geq \inf_{\pi} \frac{1}{30} \sum_{u=0}^{2k-1} \frac{\sqrt{\ln T}}{30\sqrt{T}} \sum_{t=1}^T p_t^{U,\pi}.
\]

Here, Eq. (33) holds by applying Corollary 13, and Eq. (34) holds because the average regret always lower bounds the worst-case regret.
Recall that $\oplus$ denotes the binary bit-wise exclusive or (XOR) operator. For any time $t$, suppose it is in stage $j$, for any parameter $u$, if we let $u = u \oplus 2^{d/j+1}$ by Lemma 14 and Observation 1, we have that $p^U,\pi + p^V,\pi \geq 1/2$ for all policies $\pi$. Let $q^U,\pi$ be the expected number of suboptimal pulls made by policy $\pi$ in all time periods of stage $j$. Then

\[
q^U,\pi + q^V,\pi \geq \frac{1}{2}(t_j - t_{j-1}), \quad \forall \text{ policy } \pi. \tag{35}
\]

We next compute the average expected number of suboptimal pulls made by any policy $\pi$ over all time periods $t$.

\[
\frac{1}{2d} \sum_{i=0}^{2d-1} \sum_{i=1}^{d/2} p^u,\pi = \frac{1}{2d} \sum_{i=0}^{2d-1} \sum_{i=1}^{d/2} q^u,\pi + q^u,\pi = \frac{1}{2d} \sum_{i=1}^{d/2} (t_j - t_{j-1}) = T/4, \tag{36}
\]

where the inequality holds because of Eq. (35). Combining Eqs. (36,34) we complete the proof of Theorem 10 for $n = 2$ and $d = 2$.

**C. The Lower Bound Proof for General $d$**

1) Construction of Adversarial Problem Instances: We will use the definitions of stages and intervals introduced in Sec. III-B1, and will need to define dimension groups as follows.

a) Dimension groups: Without loss of generality, we assume that $d$ is an even number. We also divide the $d$ dimensions into $d/2$ groups, where the $s$-th group ($s \in \{1, 2, 3, \ldots, d/2\}$) corresponds to the $(2s - 1)$-th and $2s$-th dimension.

Now we describe the set of adversarial instances for general $d$. The set of the instances can be viewed as a $(d/2)$-time direct product of the $d = 2$ special case.

b) Bandit instances $B^U$: Now we are ready to construct our lower bound instances. We will consider many bandit instances that are parameterized by $U = (u_1, u_2, \ldots, u_{d/2})$ where each $u_s$ ($s \in [d/2]$) is indexed by $\{0, 1, 2, \ldots, 2d - 1\}$. Let $U$ denote the set of all possible $U$’s, and we have $|U| = 2kd/2$. For each $U \in U$, the bandit instance $B^U$ consists of a hidden vector $\theta^U$ and a set of context vectors $\{x_i^U\}$. In all bandit instances, we set the noise $e_t$ to be independent Gaussian with variance $1$.

We first construct the hidden vectors $\theta^U$. For each $U = (u_1, u_2, \ldots, u_{d/2})$, we set $\theta_{2s-1}^U = \frac{2u_s}{\sqrt{d}}$ and $\theta_{2s}^U = \frac{1}{2\sqrt{d}}$ for all $s \in [d/2]$. By our construction, we have $\|\theta^U\|_2 \leq 1$ for every $U$.

We then construct the set of context vectors $\{x_i^U\}$. We label the $n$ arms by $0, 1, 2, \ldots, 2d/2 - 1$. For each arm $i \in \{0, 1, 2, \ldots, 2d/2 - 1\}$ and each dimension group $s \in [d/2]$, let $b_s(i)$ be the $s$-th least significant bit in the binary representation of $i$. At any time $t$, the context vectors of two arms $i_1$ and $i_2$ may differ at the $s$-th dimension group only when $b_s(i_1) \neq b_s(i_2)$. For any round $t$ that belongs to stage $j$, and for each dimension group $s \in [d/2]$, we set the corresponding entries of the context vector of Arm $i$ to be $(x_i^U)_{2s-1} = (x_i^U)_{2s} = (z_t \cdot \sqrt{d}, 0)$ if $b_s(i) = 0$, and set the corresponding entries to be $((x_i^U)_{2s-1} = (x_i^U)_{2s} = (\alpha_{u_s}^{-1} + \beta_{u_s}^{-1})/2 \cdot z_t \cdot \sqrt{d})$ if $b_s(i) = 1$. For $T \geq d^3$, one may easily verify (using Eq. (27)) that $\|x_i^U\|_2 \leq 1$ for $i$ and $t$.

2) s-Suboptimal Pulls and Their Implications: In our construction of adversarial bandit instances $B^U$, for each dimension group $s \in [d/2]$ the policy has to choose between two potential actions (corresponding to this group $s$) of $((x_i^U)_{2s-1}, (x_i^U)_{2s}) = (0, (\alpha_{u_s}^{-1} + \beta_{u_s}^{-1})z_t \cdot \sqrt{d})$ and $((x_i^U)_{2s-1}, (x_i^U)_{2s}) = (z_t \cdot \sqrt{d}, 0)$. One of the actions would lead to a larger expected reward depending on the unknown model $\theta$, and a policy should try to identify and execute such action for as many times as possible. This motivates us to define the concept of $s$-suboptimal pulls, which counts the number of times a policy plays a suboptimal action.

Definition 16 (s-Suboptimal Pull): For any $s \in [d/2]$, we say a policy makes one $s$-suboptimal pull at time period $t$ if the policy picks an action corresponding to the lesser expected reward.

We also break up the regret incurred by a policy at time period $t$ into $d/2$ terms, each corresponding to a dimension group $s \in [d/2]$.

Definition 17 (s-Segment Regret): For any $s \in [d/2]$ and time period $t$, define

\[
x_s(t) = \left((x_i^U)_{2s-1, t}, (x_i^U)_{2s, t}\right) - \left((x_i^U)_{2s-1, t'}, (x_i^U)_{2s, t'}\right) \left(\theta_{2s-1, t'}, \theta_{2s, t'}\right),
\]

where $i_t$ is the action the policy plays and $i_t^*$ is the optimal action at time $t$.

By definition, the regret incurred at time period $t$ can be expressed as $\sum_{s=1}^{d/2} x_s(t)$. Also, intuitions behind the definition of $s$-optimal pulls suggest that the more $s$-optimal pulls a policy makes, the larger $s$-segment regret it should incur. The following lemma quantifies this intuition by giving a lower bound of $s$-segment regret using the number of $s$-optimal pulls. The lemma is an analogue of Lemma 12 and its proof is deferred to Sec. III-D.

Lemma 18: For any instance $B^U$, any coordinate group $s$, and any time $t$, if a policy makes an $s$-suboptimal pull at time $t$, then $x_s(t) \geq \frac{1}{2}(t_j - t_{j-1})$. As a corollary, the expected regret of a policy $\pi$ can be explicitly lower bounded by the expected number of $s$-optimal pulls the policy makes.

Corollary 19: For policy $\pi$, underlying model $\theta^U$ and dimension group $s$, let $p_{s,\pi}$ be the probability of an $s$-suboptimal pull at time $t$. Then $E[R^T] \geq \sum_{t=1}^{T} \sum_{s=1}^{d/2} p_{s,\pi} \cdot \frac{1}{2}(t_j - t_{j-1})$. As a corollary, the expected regret of a policy $\pi$ can be explicitly lower bounded by the expected number of $s$-optimal pulls the policy makes.

3) Average-Case Analysis: Recall that $U$ is the finite collection of the parameters $U$ for the adversarial bandit instances we constructed in Sec. III-C1, and $p_{s,\pi}$ be the probability of an $s$-suboptimal pull at time $t$ defined in Sec. III-C2. The
minimax regret $\mathcal{R}(T; n, d)$ can then be lower bounded by
\begin{align}
\mathcal{R}(T; n, d) \geq & \inf_{\pi \in \Pi_{T, n, d}} \max_{U \subseteq \{1, \ldots, d\}} \mathbb{E}[R^\pi_T] \\
\geq & \inf_{\pi \in \Pi_{T, n, d}} \max_{U \subseteq \{1, \ldots, d\}} \sqrt{\frac{\ln T}{36T}} \cdot \sum_{t=1}^{T} \sum_{s,j=t}^{d/2} p_{s,j}^{U, \pi} \\
\geq & \inf_{\pi \in \Pi_{T, n, d}} \frac{1}{|U|} \sum_{U \subseteq \{1, \ldots, d\}} \sqrt{\frac{\ln T}{36T}} \cdot \sum_{t=1}^{T} \sum_{s,j=t}^{d/2} p_{s,j}^{U, \pi}. \tag{37}
\end{align}

Here, Eq. (37) holds by applying Corollary 19, and Eq. (38) holds because the average regret always lower bounds the worst-case regret.

The following lemma lower bounds $\{p_{s,j}^{U, \pi}\}$ for particular pairs of parameterizations.

**Lemma 20:** For any stage $j$ and any group $s$, let $U = (u_1, u_2, \ldots, u_d/2)$ and $U' = (u'_1, u'_2, \ldots, u'_d/2)$ be two parameters such that $\tau_{s,j}^{\pi} = \tau_{s,j}^{\pi'}$ but $\tau_{s,j}^{\pi} \neq \tau_{s,j}^{\pi'}$, and $u_a = u'_a$ for every $a \neq s$. (The definition of $\tau$ can be found in Sec. III-B1.) Then for any policy $\pi$ and time period $t$ in stage $j$, it holds that $p_{s,j}^{U, \pi} + p_{s,j}^{U', \pi} \geq 1/2$.

Lemma 20 is an analogue of Lemma 14 and its proof is deferred to Sec. III-D.

We are now ready to prove Theorem 10. For any parameter $U = (u_1, u_2, \ldots, u_d/2)$ and any dimension group $s$, we may write $U = (u_s, u_{-s})$ where $u_{-s} = (u_1, \ldots, u_{s-1}, u_{s+1}, \ldots, u_d/2)$. Recall that $\oplus$ denotes the binary bit-wise exclusive or (XOR) operator. For any time $t$, suppose it is in stage $j$. If we let $U' = (u_s \oplus 2^{k-j+1}, u_{-s})$, by Lemma 20 and Observation 1, we have that $p_{s,j}^{U, \pi} + p_{s,j}^{U', \pi} \geq 1/2$ for all policies $\pi$. Let $q_{s,j}^{U, \pi}$ be the expected number of $s$-suboptimal pulls made by policy $\pi$ in all time periods of stage $j$. Then
\[ q_{s,j}^{U, \pi} + q_{s,j}^{U', \pi} \geq \frac{1}{2} (t_j - t_{j-1}), \quad \forall \text{ policy } \pi. \tag{39} \]

We next compute the average expected number of $s$-suboptimal pulls made by any policy $\pi$ over all time periods $t$.
\begin{align}
\frac{1}{|U|} \sum_{U \subseteq \{1, \ldots, d\}} \sum_{t=1}^{T} \sum_{s,j=t}^{d/2} p_{s,j}^{U, \pi} &= \frac{1}{|U|} \sum_{U \subseteq \{1, \ldots, d\}} \sum_{t=1}^{T} \sum_{s,j=t}^{d/2} q_{s,j}^{U, \pi} \\
= & \frac{1}{|U|} \sum_{j=1}^{k} \sum_{s,j=t}^{d/2} \sum_{u_s=0}^{2^{k-j}-1} \sum_{u_{-s}=0}^{2^{k-j+1} - 1} \left( q_{s,j}^{u_s, u_{-s}, \pi} + q_{s,j}^{u_s \oplus 2^{k-j+1}, u_{-s}, \pi} \right) \\
= & \frac{1}{2|U|} \sum_{j=1}^{k} \sum_{s,j=t}^{d/2} \sum_{u_s=0}^{2^{k-j}-1} \sum_{u_{-s}=0}^{2^{k-j+1} - 1} \left( q_{s,j}^{u_s, u_{-s}, \pi} + q_{s,j}^{u_s \oplus 2^{k-j+1}, u_{-s}, \pi} \right) \\
\geq & \frac{1}{4|U|} \sum_{j=1}^{k} \sum_{s,j=t}^{d/2} \sum_{u_s=0}^{2^{k-j}-1} \sum_{u_{-s}=0}^{2^{k-j+1} - 1} (t_j - t_{j-1}) = \frac{T}{4} \cdot \frac{d}{2}, \tag{40}
\end{align}
where the inequality holds because of Eq. (39). Combining Eqs. (40),(38) we complete the proof of Theorem 10.

**D. Additional Proofs**

We first remark that the $T \geq d^{\epsilon}$ condition in Theorem 10 can be relaxed to $T \geq d^{2+\epsilon}$ for any small constant $\epsilon > 0$, which leads to the following theorem.

**Theorem 21:** For any small constant $\epsilon > 0$ and sufficiently large $d$, suppose $T \geq d^{2+\epsilon}$ and $n = 2d/2$. Then $\mathcal{R}(T; n, d) = \Omega(1) \cdot d \sqrt{T \log T}$.

Theorem 21 can be proved using the identical argument of the proof of Theorem 10, where the difference is that first we redefine
\[ S_t = \left( 1 + \frac{(\epsilon/2) \cdot \ln T}{2T} \right)^t \quad \text{and} \quad z_t = \sqrt{\frac{(\epsilon/2) \cdot S_{t-1} \ln T}{2T}} \]

for all $t \in [T]$. Then we list the following changes to the calculations in the proof of Theorem 10.

- In (27), we have
  \[ z_t \leq z_T \leq \sqrt{\frac{(\epsilon/2) S_{T-1} \ln T}{2T}} \leq \sqrt{\frac{\epsilon}{4} \cdot T^{\epsilon/4-1} \ln T}. \]

- At the end of Sec. III-C1., we verify that $\|x_{i,t}^{(U)}\|_2 \leq 1$ since
  \[ \|x_{i,t}^{(U)}\|_2 = \frac{d}{4} \cdot 4d \cdot \frac{\epsilon}{4} \cdot T^{\epsilon/4-1} \ln T, \]
  and for $T \geq d^{2+\epsilon}$, this value is at most $d^{2/4-\epsilon/2} \ln d \cdot T^{\epsilon/2} \leq 1$ (for large enough $d$).

- At the end of the proof of Lemma 18, we have
  \[ (41) = \frac{1}{6} \cdot 3^{t+1} \sqrt{\frac{(\epsilon/2) S_{T-1} \ln T}{2T}} \geq \sqrt{\ln T \cdot \frac{36}{36T}}. \]

Therefore, the corresponding lower bounds in Lemma 18, Corollary 19, and the final regret lower bound in the theorem will be multiplied by a factor of $\sqrt{\epsilon/2}$.

We also remark that the requirement that $T \geq d^{2+\epsilon}$ is essentially necessary for the $\Omega(d/\sqrt{T \log T})$ regret lower bound. Indeed, if $T \leq d^2$, we have $\Omega(d/\sqrt{T \log T}) \geq \Omega(T \sqrt{T}) = \omega(T)$, while the regret of any algorithm is at most $T$.

We now use in Theorem 21 to establish the regret lower bound for $n \leq 2d$, proving Theorem 2.

**Proof of Theorem 2:** To simplify the presentation, we assume without loss of generality that $n$ is an integer power of 2 and $d$ is a multiple of $\log_2 n$. We divide the time horizon into $d/\log_2 n$ phases, where phase $j \in [d/\log_2 n]$ consists of rounds $t \in \left\{ (T-j) \cdot \log_2 n, T \cdot \log_2 n \right\}$. During each phase $j$, the hidden vector and the context vectors are constructed in the same way as Theorem 21 for dimensions $s \in ((j-1) \cdot \log_2 n, j \cdot \log_2 n]$. The entries of the context vectors for the rest of the dimensions (i.e., $s \notin ((j-1) \cdot \log_2 n, j \cdot \log_2 n]$) are set to 0.

By our phase construction, $\pi$ can be viewed as a sub-policy in a $\log_2 n$-dimensional space with $n$ arms during phase $j$. One may verify that the length of the $j$-th phase is
\[ T \cdot \log_2 n / d \geq \frac{d (\log_2 n)^{1+\epsilon} \cdot \log_2 d}{d} = (\log_2 n)^{2+\epsilon}, \]
satisfying the condition in Theorem 21. Therefore, by Theorem 21, the regret of $\pi$ incurred during phase $j$ is $\Omega(\log n \sqrt{\epsilon T \log n / d})$. Therefore, the total regret of the $d/\log_2 n$ phases is at least $\Omega(\sqrt{\epsilon d T \log n / d})$. \(\square\)
Lemma 19 (Restated): For any instance $B(U)$, any coordinate group $s$, and any time $t$, if a policy makes an $s$-suboptimal pull at time $t$, then $\tau_{a_i}^{(t)} \geq \sqrt{\ln T/(36v(T))}$.

Proof: Let $U = (u_1, u_2, \ldots, u_{d/2})$. Assuming that time $t$ is in stage $j$, we verify this claim by calculating the difference of two possible expected reward contributions made by the $s$-th coordinate group, as follows.

$$
\left( z_t \cdot d, 0 \right) ^\top \begin{pmatrix} \theta_{2s-1}^{(U)} \\
\theta_{2s}^{(U)} 
\end{pmatrix}
- \begin{pmatrix} 0, (\alpha_{a_i}^{(s)} - \beta_{a_i}^{(s)}) z_t \cdot d \n= z_t \cdot \left( \gamma_{a_i}^{(s)} - \frac{\alpha_{a_i}^{(s)} - \beta_{a_i}^{(s)}}{2} \right) \geq z_t \cdot \frac{\gamma_{a_i}^{(s)} - \beta_{a_i}^{(s)}}{6}.
$$

Since $z_t = \sqrt{\frac{S_{t-1} \ln T}{2T}}$ and $|\gamma_{a_i}^{(s)} - \beta_{a_i}^{(s)}| = 3^{-j}$, we have

$$
(41) = \frac{1}{6 \cdot 3^{j}} \frac{S_{t-1} \ln T}{2T} \geq \frac{1}{6 \cdot 3^{j}} \frac{S_{t-1} \ln T}{2T} \geq \frac{\ln T}{36 \sqrt{T}},
$$

where the last inequality is because of Eq. (30).

Lemma 21 (Restated): For any stage $j$ and any group $s$, let $U = (u_1, u_2, \ldots, u_{d/2})$ and $U' = (u_1', u_2', \ldots, u_{d/2}')$ be two parameters such that $\tau^{U}_{a_i} = \tau^{U'}_{a_i}$ but $\tau_{a_i}' \neq \tau_{a_i}'$, and $u_0 = u_0'$ for every $a \neq s$. (The definition of $\tau$ can be found in Sec. III-B1.) Then for any policy $\pi$ and period time $t$ in stage $j$, it holds that $p_{s,t}^{(U, \pi)} + p_{s,t}^{(U', \pi)} \geq 1/2$.

Proof: By our construction, we have that

$$
|\theta_{2s-1}^{(U)} - \theta_{2s}^{(U')}| \leq \frac{1}{\sqrt{d}} \cdot \left( \frac{1}{3} \right)^j.
$$

Therefore, by Claim 22 (proved below, which is an analogue of Claim 15), for any event $E$ at time $t$ before the end of stage $j$, we have

$$
|\Pr[E[U]] - \Pr[E[U']]| \leq \frac{\sqrt{d}}{2} \left( \sum_{s'=1}^{d/2} |\theta_{2s'-1}^{(U)} - \theta_{2s'}^{(U')}| \right) \sqrt{S_t} \leq \frac{1}{2} \left( \frac{1}{3} \right)^j \sqrt{S_t} \leq \frac{1}{2}.
$$

(42)

The last inequality holds because at any time $t$ in stage $j$, it holds that $S_t \leq \frac{d}{2}$. Let $v_1 := (z_t \cdot d, 0)$ and $v_2 := (0, (\alpha_{a_i}^{(s)} - \beta_{a_i}^{(s)}) z_t \cdot d)$. Note that, at any time $t$, the difference between two possible reward contributed by the $s$-th dimension group is

$$
(v_1 - v_2) ^\top \begin{pmatrix} \theta_{2s-1}^{(U)} \n\theta_{2s}^{(U')} 
\end{pmatrix} = \frac{z_t}{2} (2\gamma_{a_i}^{(s)} - \alpha_{a_i}^{(s)} - \beta_{a_i}^{(s)}) \cdot 2.
$$

This value is greater than 0 if and only if $2\gamma_{a_i}^{(s)} > \alpha_{a_i}^{(s)} + \beta_{a_i}^{(s)}$. Since $\tau^{U}_{a_i} = \tau^{U'}_{a_i}$ and $\tau_{a_i}' \neq \tau_{a_i}'$, by our construction Eq. (31), we have that exactly one of $\gamma_{a_i}^{(s)}$ and $\gamma_{a_i}^{(s)}$ is greater than $\frac{1}{2} (\alpha_{a_i}^{(s)} + \beta_{a_i}^{(s)})$. In other words, at time $t$, any arm that is $s$-suboptimal for parameter is $U$ is not $s$-suboptimal for parameter $U'$, and vice versa. In light of this, let $E$ be the event that at time $t$ policy $\pi$ pulls an arm that is $s$-suboptimal for parameter $U$, and we have that the complement event $E$ is that at time $t$ policy $\pi$ pulls an arm that is $s$-suboptimal for parameter $U'$. By Eq. (42), we have

$$
p_{s,t}^{U, \pi} + p_{s,t}^{U', \pi} = \Pr[E[U]] + \Pr[E[U']] = 1 + \Pr[E[U]] - \Pr[E[U']] \geq \frac{1}{2}.
$$

Claim 22: For any $U, U'$, let $j$ be the largest number such that $\tau^{U}_{a_i} = \tau^{U'}_{a_i}$ holds for every $s$. For any time $t \leq t_j$ and any event $E$ that happens at time $t$, we have

$$
|\Pr[E[U]] - \Pr[E[U']]| \leq \frac{\sqrt{d}}{2} \left( \sum_{s'=1}^{d/2} |\theta_{2s'-1}^{(U)} - \theta_{2s'}^{(U')}| \right) \sqrt{S_t}.
$$

Proof: Note that by our construction, at any time $t \leq t_j$, the contextual vectors of both $B(U)$ and $B(U')$ are the same. Moreover, for any hidden vector and any arm, the reward distribution is a shifted standard Gaussian with variance 1.

For any time $t \leq t_j$, let $D_1$ be the product of the arm reward distributions at and before round $t$ when the hidden vector is $\theta^{(U)}$, and let $D_2$ be the same product distribution when the hidden vector is $\theta^{(U')}$.

Since $\theta^{(U)} = \theta^{(U')}$, for all $s' \in [d/2]$, the difference of the mean rewards at any time $t' : 1 \leq t' \leq t$ for $\theta^{(U)}$ and $\theta^{(U')}$ is at most $\sum_{s'=1}^{d/2} \tau^{U}_{2s'-1} \sqrt{d} - \tau^{U'}_{2s'-1} \sqrt{d}$. Note that the KL divergence between two variance-1 Gaussians with means $\mu_1$ and $\mu_2$ is $|\mu_1 - \mu_2|^{2}/2$. Therefore, we have

$$
\text{KL} (D_1 || D_2) \leq \frac{1}{2} \sum_{i=1}^{t} \left( \sum_{s'=1}^{d/2} \tau^{U}_{2s'-1} \sqrt{d} - \sum_{s'=1}^{d/2} \tau^{U'}_{2s'-1} \sqrt{d} \right)^2
$$

= $\frac{d}{2} \left( \sum_{s'=1}^{d/2} \tau^{U}_{2s'-1} \sqrt{d} \right)^2 \left( \sum_{s'=1}^{d/2} \tau^{U'}_{2s'-1} \sqrt{d} \right)^2
$$

= $\frac{d}{2} \left( \sum_{s'=1}^{d/2} \tau^{U}_{2s'-1} \sqrt{d} \right) \left( 1 + \sum_{s'=1}^{d/2} \tau^{U'}_{2s'-1} \sqrt{d} \right)
$$

= $\frac{d}{2} \left( \sum_{s'=1}^{d/2} \tau^{U}_{2s'-1} \sqrt{d} \right) \left( 1 + \sum_{s'=1}^{d/2} \tau^{U'}_{2s'-1} \sqrt{d} \right)
$$

Therefore, at time $t$, and for any event $E$, we have

$$
|\Pr[E[U]] - \Pr[E[U']]| \leq \frac{1}{2} \text{KL}(D_1 || D_2)
$$

= $\frac{1}{2} \sqrt{d} \left( \sum_{s'=1}^{d/2} \tau^{U}_{2s'-1} \sqrt{d} \right) \sqrt{S_t}
$$

where the first inequality holds because of Pinsker’s inequality (Lemma 25).

E. Lower Bound for Infinite Action Spaces

The following corollary establishes an $\Omega(d\sqrt{T \log T})$ lower bound for the case in which the action spaces for each time period are infinite and changing over time.
Corollary 23: For any policy $\pi$, there exists a bandit instance with regression model $\theta \in \mathbb{R}^d$, $\|\theta\|_2 \leq 1$ and closed action spaces $A_1, \ldots, A_T \subseteq \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$, $|A_i| = \infty$, such that for sufficiently large $T$,

$$\mathbb{E}[R^T] = \Omega(d\sqrt{T\log T}).$$

Proof: We prove the corollary by contradiction. Suppose the opposite, that there exists a policy $\pi'$ such that for all $\theta \in \mathbb{R}^d$, $\|\theta\|_2 \leq 1$ and measurable action spaces $A_1, \ldots, A_T \subseteq \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ such that $|A_i| = \infty$, it holds that $\mathbb{E}[R^T] = o(d\sqrt{T\log T})$. We will show that there exists a policy $\pi''$ to achieve $o(d\sqrt{T\log T})$ on all bandit instances with $\theta \in \mathbb{R}^d$, $\|\theta\|_2 \leq 1$, and action sets $A_1', \ldots, A_T' \subseteq \{y \in \mathbb{R}^d : \|y\|_2 \leq 1, y^T \theta \geq 0\}$ such that $|A_i'| = 2^{d/2}$, contradicting Theorem 10 and Theorem 21.

For action sets $A_1', \ldots, A_T' \subseteq \{y \in \mathbb{R}^d : \|y\|_2 \leq 1, y^T \theta \geq 0\}$ such that $|A_i'| = 2^{d/2}$, the policy $\pi''$ construct $A_i = \{x(\lambda, y) = \lambda y : y \in A_i', \lambda \in [0, 1]\}$ and simulate policy $\pi$ when the candidate action space is $A_i$. Clearly, $A_i$ is closed and $|A_i| = \infty$. If policy $\pi$ decides to play $x_t = x(\lambda, y_t) \in A_i$, the policy $\pi'$ plays $y_t \in A_i'$, observes the reward $r_t$, and feeds $\lambda r_t$ as reward to the policy $\pi$. Since $r_t - y_t^T \theta$ is a centered sub-Gaussian with variance proxy 1, we have that $\lambda r_t - \lambda x_t^T \theta$ is also a centered sub-Gaussian with variance proxy (at most) 1.

Since $y^T \theta \geq 0$ for all $y \in A_i'$ and all $t$, we upper bound the expected regret incurred by $\pi''$ as follows.

$$\mathbb{E}[R^T] = \sum_{t=1}^T \mathbb{E}\left[\max_{y \in A_i'} y^T \theta - y_t^T \theta\right]$$

$$= \sum_{t=1}^T \mathbb{E}\left[\sup_{x \in A_i} x^T \theta - y_t^T \theta\right]$$

$$\leq \sum_{t=1}^T \mathbb{E}\left[\sup_{x \in A_i} x^T \theta - x_t^T \theta\right] = \mathbb{E}[R^T] \leq o(d\sqrt{T\log T}),$$

where we use the subscript in $R^T$ to denote whether the regret is incurred by the policy $\pi$ or $\pi''$, and the first inequality is because, if at time $t$, $x_t = x(\lambda, y_t)$ is chosen by policy $\pi$, then $y_t^T \theta = x_t^T \theta / \lambda \geq x_t^T \theta$.

IV. RECENT DEVELOPMENTS AND FUTURE DIRECTIONS

There have been a few new related works after the original version of this paper [1]. We summarize them as follows. Li et al. [39] extended the algorithmic results of the paper to the infinite-armed linear contextual bandits and showed that $n\Omega(T; n, d)$ can be asymptotically upper bounded by $\text{poly}(\log \log T)\cdot O(d\sqrt{T\log T})$ for even $n = \infty$. For moderate time horizon $T$ (e.g., $T \leq \text{poly}(d)$), this new result improves our Theorem 1 by a factor at least $\log d$ when $n \geq \exp(\Omega(d))$.

In another recent work, Simchi-Levi and Xu [40] showed an $O(\sqrt{d + \log T})$ regret upper bound for a special case of our problem, namely the stochastic context setting, where the action set (a.k.a., the context set) $A_i$ is i.i.d. drawn from an unknown distribution $D$. In comparison, the action sets in our problem are determined by an oblivious adversary and can be very different at different time periods. The regret bound by Simchi-Levi and Xu becomes only $O(\sqrt{dT})$ when $n = O(1)$ and $T \leq \exp(O(d))$, which seems to “contradict” our lower bounds. This is because our lower bound instances crucially rely on the ability to switch action sets across time periods. This contrast also reveals an interesting separation between the stochastic context setting and the general linear contextual bandit problem.

We also discuss several related open questions for future research, which we strongly believe are fundamental to the field and would call for substantially new techniques. One natural problem is to eliminate the $\text{poly}(\log \log(nT))$-factor gap left in this paper and prove the optimal minimax regret bounds (only up to constant factors). To this end, we observe that the $\zeta_0 = \lceil \log_2(\sqrt{T}/d) \rceil$ layers in our VCL-SupLinUCB algorithm are the main source of the $\text{poly}(\log \log T)$ factors and one might need to come up with a “layer-free” algorithm to reduce the iterative-log factors in the regret. In general, it is both practically useful and technically interesting to develop simpler algorithms to the “layered data partitioning” technique while the new algorithm should also achieve the provably optimal minimax regret. Finally, we note that the algorithm in this paper only works for an oblivious adversary (and so do the SupLinUCB/SupLinRel algorithms). It is intriguing to study whether we can prove a similar regret bound for an adaptive adversary, where the finite action set at each time period $t$ is decided by the adversary based on the entire history of the first $(t-1)$ time periods.

Finally, we note that in the current VCL-SupLinUCB framework, the ordinary least squares estimator is adopted on adaptively chosen covariates. It is an interesting question whether more advanced statistical estimation methods could be employed, due to the fact that covariates are adaptively chosen and only ellipsoidal estimation errors are required, demanding smaller estimation errors on directions in which the covariates are observed more frequently. For example, non-linear shrinkage methods such as thresholding [41], Stein’s shrinkage [42], and Pinsker’s estimator [43], [44] might be able to further bring down the estimation errors and eventually close the remaining gaps between upper and lower bounds in the current paper.

APPENDIX

Probability tools.

The following lemma is the Hoeffding’s concentration inequality for sub-Gaussian random variables, which can be found in for example [45].

Lemma 24: Let $X_1, \ldots, X_n$ be independent centered sub-Gaussian random variables with sub-Gaussian parameter $\sigma^2$. Then for any $\xi > 0$,

$$\Pr\left[\sum_{i=1}^n X_i \geq \xi\right] \leq 2 \exp\left(-\frac{\xi^2}{2n\sigma^2}\right).$$

The following lemma states Pinsker’s inequality [38].

Lemma 25: If $P$ and $Q$ are two probability distributions on a measurable space $(X, \Sigma)$, then for any measurable event
A ∈ Σ, it holds that
\[ |P(A) − Q(A)| ≤ \sqrt{\frac{1}{2} \text{KL}(P∥Q)}, \]
where
\[ \text{KL}(P∥Q) = \int_X \left( \frac{dP}{dQ} \right) dP \]
is the Kullback–Leibler divergence.

Additional proofs in Sec. II.

Below we provide the proof of Lemma 7.

Lemma 7 (Restated): For any ζ, t and i, we have that
\[ \mathbb{E}[1\{−E_i^t\} · |x_i^T(\hat{\theta}_{i,t} − θ)| |ζ < ζ_i, i ∈ N_{ζ,t}] \leq \sqrt{2πd/(nζ_0 \sqrt{T})}. \]

Proof: The entire proof is carried out conditioned on
\[ \{X_{i,t−1}\}_{i≤ζ_i}, \{it\}_{t ∈ Ξ_{i−1,ζ−1,1∪· · ·∪X_{i,t−1,ζ−1}}, Λ_{ζ−1,1}, Λ_{ζ−1,2−1,1}, · · · , Λ_{t,1−1}, Λ_{ζ−1,2−1,1}, Λ_{ζ−1,3−1,1}, · · · , Λ_{t,1−1}. \]
This renders the quantities \( α^i_{ζ,t}, ω^i_{ζ,t} \) deterministic. Note that
the event \( ζ < ζ_i \) and the set \( N_{ζ,t} \) also become deterministic.
Therefore, we only need to prove that \( \mathbb{E}[1\{−E_i^t\} · |x_i^T(\hat{\theta}_{i,t} − θ)|] \leq \sqrt{2πd/(nζ_0 \sqrt{T})} \) assuming that \( ζ < ζ_i \) and \( i ∈ N_{ζ,t} \).

Note also that, by Proposition 1, the quantities \( \{ε_{i′}\}_{i′ ∈ Ξ_{i−1,ζ−1}} \) remain independent, centered sub-Gaussian random variables.

We first derive an upper bound on the tail of \( |x_i^T(\hat{\theta}_{i,t} − θ)| \).
Note that \( X_{ζ−1,t−1} \) is the set of all time periods \( τ < t \) such that \( ζ_τ = ζ \). By elementary algebra, we have
\[
\begin{align*}
&x_i^T(\theta − \hat{\theta}_{i,t}) = x_i^T(\theta − Λ_{ζ−1,t−1} \hat{\theta}_{i,t−1}) \\
&= x_i^T(\theta − Λ_{ζ−1,t−1} \sum_{t′ \in Ξ_{i−1,t−1}} x_{i,t′}(x_{i,t′}^T \theta + ε_{i_t′})) \\
&= x_i^T(\theta − Λ_{ζ−1,t−1} (Λ_{ζ−1,t−1} − I) \theta − Λ_{ζ−1,t−1} \sum_{t′ \in Ξ_{i−1,t−1}} x_{i,t′} ε_{i_t′}) \\
&= x_i^T Λ_{ζ−1,t−1} \theta − Λ_{ζ−1,t−1} \sum_{t′ ∈ Ξ_{i−1,t−1}} x_{i,t′} ε_{i_t′}. \\
\end{align*}
\]
Subsequently,
\[
(\theta − \hat{\theta}_i)^T x_i \leq |x_i^T Λ_{ζ−1,t−1} θ| + \sum_{t′ ∈ Ξ_{i−1,t−1}} x_i^T Λ_{ζ−1,t−1} x_{i,t′} ε_{i_t′}. \tag{43}
\]

For the first term in the RHS (right-hand side) of Eq. (43), applying Cauchy-Schwarz inequality and the facts that \( Λ_{ζ−1,t−1} \succeq I, \|θ\|_2 \leq 1 \) we have \( |x_i^T Λ_{ζ−1,t−1} θ| \leq \|Λ_{ζ−1,t−1} x_i\|_2 \|Λ_{ζ−1,t−1} θ\|_2 \leq \sqrt{\text{tr}(Λ_{ζ−1,t−1}^2 x_i^2) \text{tr}(Λ_{ζ−1,t−1}^2 \|θ\|_2^2)} \). For the second term in the RHS of Eq. (43), because \( \{ε_{i_t′}\}_{i′ ∈ Ξ_{i−1,1−1}} \) are centered sub-Gaussian variables with sub-Gaussian parameter 1 and \( \{ε_{i_t′}\}_{i′ ∈ Ξ_{i−1,t−1}} \) are statistically independent even after the conditioning (By Proposition 1), we conclude that
\[
\sum_{t′ ∈ Ξ_{i−1,t−1}} x_i^T Λ_{ζ−1,t−1} x_{i,t′} ε_{i_t′} \text{ is also a centered sub-Gaussian random variable with sub-Gaussian parameter upper bounded by}
\]
\[
\begin{align*}
&\sum_{t′ ∈ Ξ_{i−1,t−1}} (x_i^T Λ_{ζ−1,t−1} x_{i,t′} ε_{i_t′})^2 \\
&= x_i^T Λ_{ζ−1,t−1} \left( \sum_{t′ ∈ Ξ_{i−1,t−1}} x_{i,t′} x_i^T ε_{i_t′} ε_{i_t′} \right) Λ_{ζ−1,t−1} x_i \\
&= x_i^T Λ_{ζ−1,t−1} x_i. \tag{44}
\end{align*}
\]
Combining Eqs. (43), (44), and using standard concentration inequalities of sub-Gaussian random variables (see for example Lemma 24), we have for every \( δ ∈ (0, 1) \)
\[
\text{Pr}\left[(\theta − \hat{\theta}_{i,t−1})^T x \geq \left(\sqrt{2\ln(δ−1)} + 1\right) \sqrt{x_i^T Λ_{ζ−1,t−1} x_i}\right] \leq δ, \tag{45}
\]
which is equivalent to
\[
\text{Pr}\left[(\theta − \hat{\theta}_{i,t−1})^T x ≥ βω^i_{ζ,t} \right] ≤ e^{−(β−1)2/2}; \quad \forall β ≥ 1. \tag{46}
\]

Integrating both sides of Eq. (46) from \( α^i_{ζ,t}ω^i_{ζ,t} \) to \( +∞ \) we obtain
\[
\begin{align*}
\mathbb{E}[1\{−E_i^t\} · |x_i^T(\hat{\theta}_{i,t} − θ)|] & \leq \int_{α^i_{ζ,t}ω^i_{ζ,t}}^{+∞} \text{Pr}\left[|x_i^T(\hat{\theta}_{i,t} − θ)| ≥ u \right] du \\
& \leq \int_{α^i_{ζ,t}ω^i_{ζ,t}}^{+∞} e^{−(u/ω^i_{ζ,t})^2/2} du \\
& = \sqrt{2πω^i_{ζ,t}} \int_{α^i_{ζ,t}ω^i_{ζ,t}}^{+∞} \frac{1}{2π} e^{−v^2/2} dv \\
& ≤ \sqrt{2πω^i_{ζ,t}} e^{−(α^i_{ζ,t}ω^i_{ζ,t})^2/2}, \tag{47}
\end{align*}
\]
where the last inequality again holds by tail bounds of Gaussian random variables (Lemma 24). Plugging in the expression of \( α^i_{ζ,t} \) in Algorithm 1, the right-hand side of Eq. (47) can be upper bounded by
\[
\sqrt{2πω^i_{ζ,t}} \cdot \exp\left\{-\frac{\max\{1, \ln(T(ω^i_{ζ,t})^2/d)\} \ln(n^2ζ^2_{i,t})}{2}\right\} \\
\leq \sqrt{2πω^i_{ζ,t}} \sqrt{\frac{d}{T(ω^i_{ζ,t})^2 n^2ζ^2_{i,t}}} \leq \sqrt{2πd/nζ_0 \sqrt{T}},
\]
which is to be demonstrated. \( \square \)

ACKNOWLEDGMENT

Author names are listed in alphabetical order.

REFERENCES

[1] Y. Li, Y. Wang, and Y. Zhou, “Nearly minimax-optimal regret for linearly parameterized bandits,” in Proc. Conf. Learn. Theory (COLT), 2019, pp. 2173–2174.

[2] W. R. Thompson, “On the likelihood that one unknown probability exceeds another in view of the evidence of two samples,” Biometrika, vol. 25, nos. 3–4, pp. 283–294, Dec. 1933.

[3] H. Robbins, “Some aspects of the sequential design of experiments,” Bull. Amer. Math. Soc., vol. 58, no. 5, pp. 527–535, 1952.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
