Tight Regret Bounds for Infinite-armed Linear Contextual Bandits

Yingkai Li
Northwestern University
yingkai.li@u.northwestern.edu

Yining Wang
University of Florida
yining.wang@warrington.ufl.edu

Xi Chen
New York University
xc13@stern.nyu.edu

Yuan Zhou
University of Illinois at Urbana-Champaign
yuanz@illinois.edu

Abstract

Linear contextual bandit is an important class of sequential decision making problems with a wide range of applications to recommender systems, online advertising, healthcare, and many other machine learning related tasks. While there is a lot of prior research, tight regret bounds of linear contextual bandit with infinite action sets remain open. In this paper, we address this open problem by considering the linear contextual bandit with (changing) infinite action sets. We prove a regret upper bound on the order of $O(\sqrt{d^2T\log T}) \times \mathrm{poly}(\log\log T)$ where d is the domain dimension and T is the time horizon. Our upper bound matches the previous lower bound of $\Omega(\sqrt{d^2T\log T})$ in [Li et al., 2019] up to *iterated* logarithmic terms.

1 Introduction

Linear contextual bandit is an important class of sequential decision making problems with an extensive history of research in both machine learning and operations research [Abbasi-Yadkori et al., 2011, Chu et al., 2011, Auer, 2002, Rusmevichientong and Tsitsiklis, 2010, Dani et al., 2008, Li et al., 2019]. In the linear contextual bandit problem, a player makes sequential decisions over T time periods. At each time period t, an *action set* $D_t \subseteq \mathbb{R}^d$ is provided; the player selects an *action* $x_t \in D_t$, and subsequently receives a *reward* r_t parameterized as

$$r_t = \langle x_t, \theta\rangle + \xi_t,$$

where $\theta \in \mathbb{R}^d$, $\|\theta\|_2 \le 1$ is a fixed but unknown parameter vector, and $\{\xi_t\}$ are independent centered sub-Gaussian noise variables with the variance proxy 1. The performance is evaluated by the cumulative *regret*, defined as

$$R_T := \sum_{t=1}^{T} \sup_{x\in D_t} \langle x, \theta\rangle - \langle x_t, \theta\rangle.$$

The objective of this paper is to design an algorithm that achieves the *optimal* expected regret under the worst case, when the action sets $\{D_t\}$ are *infinite* (i.e., $|D_t| = \infty$). In the next sections, we give a rigorous definitions of policy and action domains studied in this paper. We also discuss (informally) our main results, and compare them with existing results in the literature.

1.1 Definition of policy and action domains

Suppose that there are T time periods and the problem dimension is d. A policy π can be represented as $\pi = (\phi_1, \phi_2, \cdots, \phi_T)$ where $\phi_t : (x_1, y_1, \cdots, x_{t-1}, y_{t-1}, D_t) \mapsto x_t \in D_t$ is a randomized function that maps the data collected from prior episodes $\{1, 2, \cdots, t-1\}$ to an action $x_t \in D_t$ to be selected at time period t. Note that future feasible sets D_{t+1}, D_{t+2} are *not* revealed to the policy π when it is making an action decision at time t.

Let $\mathcal{S}_d := \big\{S : S \text{ is closed}, S \subseteq \{x \in \mathbb{R}^d : \|x\|_2 \le 1\}\big\}$ be the set of all closed subsets of the unit d-dimensional ℓ_2 ball. The domains $D_1, \cdots, D_T \in \mathcal{S}_d$ are chosen arbitrarily, *before* any policy π is executed. We remark that this setting is known in the literature as the "oblivious" setting.

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

1.2 Existing work and our results

A summary of our results as well as existing results are given in Table 1. The regularity conditions that $\|\theta\|_2 \leq 1$ and $D_t \subseteq \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ are imposed, so that $|\mathbb{E}[r_t]| = |\langle x_t, \theta\rangle| \leq 1$ holds for all $x_t \in D_t$. Additionally, as suggested by the title, we consider the *infinite-armed* case in which $|D_t| = \infty$ for all t. We also impose the regularity condition that the action sets D_t are *closed*, so that the supremum over the sets can always be achieved by an action.

[Dani et al., 2008] derived an algorithm based on confidence balls of prediction errors of θ, achieving a worst-case expected regret of $O(\sqrt{d^2 T \log^3 T})$. [Abbasi-Yadkori et al., 2011] further improved the analysis and obtained $O(\sqrt{d^2 T \log^2 T})$ regret. On the lower bound side, [Dani et al., 2008] proved a regret lower bound of $\Omega(\sqrt{d^2 T})$ for all policies, which was later improved to $\Omega(\sqrt{d^2 T \log T})$ by [Li et al., 2019] as a direct corollary of regret lower bounds for finite-armed linear contextual bandits. While [Li et al., 2019] derived matching upper bounds for the finite-armed case, their results and techniques cannot be directly applied to the infinite-armed case even if computational issues are disregarded, as covering nets of $\{D_t\}$ up to $1/\mathrm{poly}(T)$ accuracy would incur additional logarithmic terms in T.

In this paper, we prove the following main result:

Theorem 1 (Informal)**.** *There is a policy whose worst-case expected regret is asymptotically upper bounded by* $O(\sqrt{d^2 T \log T}) \times \mathrm{poly}(\log \log T)$.

Comparing with the lower bound $\Omega(\sqrt{d^2 T \log T})$, the upper bound in Theorem 1 is tight up to iterated logarithmic terms. Our results thus close the $O(\sqrt{\log T})$ gap between the upper bound (in [Abbasi-Yadkori et al., 2011]) and the lower bound in infinite-armed linear contextual bandit. In addition, the idea behind our varying confidence level (VCL) UCB algorithm and a number of technical tools developed in the proof might also be useful for other contextual bandit problems.

1.3 Proof techniques

Sharp tail bounds of self-normalized empirical processes. Due to the inherent statistical dependency between the chosen actions $\{a_t\}$ and noise variables $\{\xi_t\}$, the estimation error of θ at each time step cannot be analyzed using standard closed-forms of linear regression estimators. The work of [Abbasi-Yadkori et al., 2011] pioneered the use of *self-normalized empirical processes* to understand the estimation and prediction errors at each time step.

In this paper, we make use of sharp tail bounds on the supremum of self-normalized empirical processes in high-dimensional probability (Lemma 3). By exploiting such tail bounds we have a much more refined control of failure probabilities at each time step, which lays the foundation of our improved regret analysis.

Varying confidence levels in UCB-type algorithms. Most existing methods on linear contextual bandit can be categorized as *Upper-Confidence-Bound (UCB)* or *Optimism-in-Face-of-Uncertainty (OFU)* type algorithms, which build confidence bands/balls around unknown parameters at each time step and then pick actions in the most optimistic way.

While most existing algorithms set constant confidence levels (corresponding to failure probabilities at each time), in this paper we consider *varying confidence levels* (VCL), with higher failure probabilities towards the end of the time horizon T. The intuition is that later fails would incur much less regret. Similar ideas were also employed in previous works [Audibert and Bubeck, 2009, Li et al., 2019, Wang et al., 2018] to improve regret guarantees in bandit problems.

2 Algorithm design and main results

Algorithm 1, named VCL-SupLinUCB, is the main algorithm of this paper which combines the varying confidence levels (VCL) design with the existing SupLinUCB algorithm [Auer, 2002, Chu et al., 2011]. The basic idea of Algorithm 1 is to classify the time periods into different layers such that the chosen context is statistically independent with the noise in the reward distribution. Then the algorithm estimates $\widehat{\theta}_{\zeta,t}$ for each layer ζ and eventually selects the arm with largest upper confidence bound according to the rules specified in Lines 7-12 in Algorithm 1. Those ideas are similar to the previous SupLinUCB algorithm, and the major difference between Algorithm 1 and previous approaches is the *varying* confidence levels (reflected by the inclusion of $\omega_{x,t}$ in $\alpha_{x,t}$), which allows for sharper regret bounds.

The following theorem is the main result of this paper:

Theorem 2. *Suppose the universal constant $C > 0$ in the input of Algorithm 1 is sufficiently large. Then there exists constants $C_1, C_2 > 0$ that only depend on C such that for all $\|\theta\|_2 \leq 1$ and $\{D_t \subseteq \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}\}$, the regret R_T satisfies the following inequality for any $\delta \in (0,1)$,*

$$\mathbb{E}\left[\max\left\{R_T - C_1 d\sqrt{T \log T \log(1/\delta)} \cdot \log\log(T/\delta), 0\right\}\right] \leq C_2 \delta d \sqrt{T}.$$

Table 1: Summary of results. Both θ and $\{D_t\}$ belong to the unit ball $\{x \in \mathbb{R}^d : \|x\|_2 \le 1\}$, and $|D_t| = \infty$ for all t. Upper and lower bounds are for $\mathbb{E}[R_T]$ under the worst case. $O(\cdot)$ and $\Omega(\cdot)$ notations hide universal constants only, and $\mathrm{poly}(\log\log T)$ means $(\log\log T)^{O(1)}$.

	[Dani et al., 2008]	[Abbasi-Yadkori et al., 2011]	[Li et al., 2019]	**this paper**
Upper bound	$O(\sqrt{d^2 T \log^3 T})$	$O(\sqrt{d^2 T \log^2 T})$	N/A	$O(\sqrt{d^2 T \log T}) \times \mathrm{poly}(\log\log T)$
Lower bound	$\Omega(\sqrt{d^2 T})$	N/A	$\Omega(\sqrt{d^2 T \log T})$	N/A

Input: $\zeta_0 = \lceil \log_2(\sqrt{T/d}/\delta) \rceil$, Time horizon T, confidence parameter δ, domain dimension d, universal constant $C \ge 1$;
Initialization: $\mathcal{X}_{\zeta,0} = \emptyset$, $\Lambda_{\zeta,0} = I_{d\times d}$, $\lambda_{\zeta,0} = \vec{0}_d$, $\widehat{\theta}_{\zeta,0} = \vec{0}_d$ for $\zeta \le \zeta_0$;
for $t = 1, 2, \cdots, T$ **do**
  Observe D_t, and set $\zeta = 0$ and $\mathcal{N}_{\zeta,t} = D_t$;
  while a choice x_t has yet to be made **do**
    Compute $\widehat{\theta}_{\zeta,t} = \Lambda_{\zeta,t-1}^{-1}\lambda_{\zeta,t-1}$, and for every $x \in \mathcal{N}_{\zeta,t}$, compute $\omega_{\zeta,t}^x = \sqrt{x^\top \Lambda_{\zeta,t-1}^{-1} x}$, $\alpha_{\zeta,t}^x = \sqrt{\max\{1, \ln[(T\ln^4 T \ln^2(1/\delta))(\omega_{\zeta,t}^x)^2/(d\delta^2)]\}}$ and $\varpi_{\zeta,t}^x = C \cdot \sqrt{d} \cdot \alpha_{\zeta,t}^x \omega_{\zeta,t}^x$;
    if $\zeta = \zeta_0$ **then**
      Find $x_t \in \mathcal{N}_{\zeta,t}$ that maximizes $\min\{1, x_{it}^\top \widehat{\theta}_{\zeta,t} + \varpi_{\zeta,t}^x\}$ and set $\zeta_t = \zeta$;
    else if $\varpi_{\zeta,t}^x \le 2^{-\zeta}$ for all $x \in \mathcal{N}_{\zeta,t}$ **then**
      Update $\mathcal{N}_{\zeta+1,t} = \mathcal{N}_{\zeta,t} \cap \{x : x^\top \widehat{\theta}_{\zeta,t} \ge \max_{y\in\mathcal{N}_{\zeta,t}} y^\top \widehat{\theta}_{\zeta,t} - 2^{1-\zeta}\}$, $\zeta \leftarrow \zeta + 1$;
    else
      Select any $x_t \in \mathcal{N}_{\zeta+1,t}$ such that $\varpi_{\zeta,t}^x \ge 2^{-\zeta}$, and set $\zeta_t = \zeta$;
    end if
  end while
  Select action x_t and observe feedback $r_t = x_t^\top \theta + \xi_t$;
  Update: $\mathcal{X}_{\zeta_t,t} = \mathcal{X}_{\zeta_t,t-1} \cup \{x_t\}$, $\Lambda_{\zeta_t,t} = \Lambda_{\zeta_t,t-1} + x_t x_t^\top$, $\lambda_{\zeta_t,t} = \lambda_{\zeta_t,t-1} + r_t x_t$, and $\mathcal{X}_{\zeta',t} = \mathcal{X}_{\zeta',t-1}$, $\Lambda_{\zeta',t} = \Lambda_{\zeta',t-1}$, $\lambda_{\zeta',t} = \lambda_{\zeta',t-1}$ for any $\zeta' \ne \zeta_t$;
end for

Algorithm 1: The VCL-SupLinUCB algorithm

Theorem 2 implies the following two statements. In the following, $\lesssim$ means that the constants in the inequality are omitted.

i. The expected regret $\mathbb{E}[R_T] \lesssim d\sqrt{T\log T \log(1/\delta)} \cdot \log\log(T/\delta)$. In particular, if we take $\delta = \Omega(1)$, we have that $\mathbb{E}[R_T] \lesssim d\sqrt{T\log T} \cdot \log\log T$.

ii. With probability at least $1-\delta$, it holds that $R_T \lesssim d\sqrt{T\log T\log(1/\delta)} \cdot \log\log(T/\delta)$. This is because, by Markov's inequality, $\Pr[R_T - C_1 d\sqrt{T\log T\log(1/\delta)} \cdot \log\log(T/\delta) > C_2 d\sqrt{T}] \le \mathbb{E}[\max\{R_T - C_1 d\sqrt{T\log T\log(1/\delta)} \cdot \log\log(T/\delta), 0\}]/(C_2 d\sqrt{T}) \le \delta$.

While neither of the two statements implies each other, we note that, if iterated logarithmic factor is left out, statement ii) is stronger than the high probability bound proved by [Abbasi-Yadkori et al., 2011], where the regret is at most $O(d\sqrt{T\log T\log(T/\delta)})$ with probability at least $1-\delta$.

The proof of Theorem 2 is stated in the next section.

3 Proof of Theorem 2

3.1 Uniform confidence region for $\widehat{\theta}_{\zeta,t}$

We first present a lemma that upper bounds the errors $|\langle x, \widehat{\theta}_{\zeta,t} - \theta\rangle|$ with high probability.

Lemma 3. *For any $t \in [T]$, any layer $\zeta \in \{0, 1, 2, \ldots, \zeta_0\}$, and any $\gamma \in (0, 1/2]$, with probability $1-\gamma$ it holds that*

$$\sup_{x\in\mathbb{R}^d} (\omega_{\zeta,t}^x)^{-1} \big|x^\top(\widehat{\theta}_{\zeta,t} - \theta)\big| \lesssim \sqrt{d} + \sqrt{\ln(1/\gamma)}.$$

The proof of Lemma 3 can be roughly divided into three steps. First, the closed-form expression of Ridge regression to express $\widehat{\theta}_{\zeta,t}$ in terms of θ and ξ. At the second step, a self-normalized empirical process is derived by manipulating and normalizing the expression

derived in the first step. Finally, sharp tail bounds of sub-Gaussian processes are invoked to prove Lemma 3.

Proof of Lemma 3. Let

$$\mathcal{T}_{\zeta,t-1} := \{\tau : \tau \le t-1 \text{ and } \zeta_\tau = \zeta\}$$

and let

$$n_{\zeta,t-1} := |\mathcal{T}_{\zeta,t-1}|.$$

Note that we also have $n_{\zeta,t-1} = |\mathcal{X}_{\zeta,t-1}|$.

Let $X_{\zeta,t-1}$ be a $n_{\zeta,t-1} \times d$ matrix constructed by stacking all $x \in \mathcal{X}_{\zeta,t-1}$ together, i.e.,

$$\Lambda_{\zeta,t-1} = X_{\zeta,t-1}^\top X_{\zeta,t-1} + I.$$

Let $\Xi_{\zeta,t-1}$ be the $n_{\zeta,t-1}$-dimensional vector that contains all noises ξ_τ such that $\tau \in \mathcal{T}_{\zeta,t-1}$. We also let

$$r_{\zeta,t-1} = X_{\zeta,t-1}\theta + \Xi_{\zeta,t-1}$$

be the $n_{\zeta,t-1}$-dimensional vector by concatenating all rewards for time periods $\tau \in \mathcal{T}_{\zeta,t-1}$.

Define also $\|x\|_A := \sqrt{x^\top A x}$ for d-dimensional vectors x and $d \times d$ positive-semidefinite matrices A. Then

$$\begin{aligned}\widehat{\theta}_{\zeta,t} &= (X_{\zeta,t-1}^\top X_{\zeta,t-1} + I)^{-1} X_{\zeta,t-1}^\top (X_{\zeta,t-1}\theta + \Xi_{\zeta,t-1}) \\ &= (I - \Lambda_{\zeta,t-1}^{-1})\theta + \Lambda_{\zeta,t-1}^{-1} X_{\zeta,t-1}^\top \Xi_{\zeta,t-1}.\end{aligned}$$

Subtracting one θ and left multiplying with $(\widehat{\theta}_{\zeta,t} - \theta)^\top \Lambda_{\zeta,t-1}$ on both sides of the above identity, we obtain

$$\|\widehat{\theta}_{\zeta,t} - \theta\|_{\Lambda_{\zeta,t-1}}^2 = -(\widehat{\theta}_{\zeta,t} - \theta)^\top \theta + (\widehat{\theta}_{\zeta,t} - \theta)^\top X_{\zeta,t-1}^\top \Xi_{\zeta,t-1}. \tag{1}$$

Note that

$$|(\widehat{\theta}_{\zeta,t} - \theta)^\top \theta| \le \|\theta\|_2 \|\widehat{\theta}_{\zeta,t} - \theta\|_2 \le \|\widehat{\theta}_{\zeta,t} - \theta\|_{\Lambda_{\zeta,t-1}}$$

because $\|\theta\|_2 \le 1$ and $\Lambda_{\zeta,t-1} \succeq I$. Dividing both sides of Eq. (1) by $\|\widehat{\theta}_{\zeta,t} - \theta\|_{\Lambda_{\zeta,t-1}}$, we have

$$\begin{aligned}&\|\widehat{\theta}_{\zeta,t} - \theta\|_{\Lambda_{\zeta,t-1}} \le 1 + \phi^\top X_{\zeta,t-1}^\top \Xi_{\zeta,t-1}, \\ &\text{where } \phi = (\widehat{\theta}_{\zeta,t} - \theta)/\|\widehat{\theta}_{\zeta,t} - \theta\|_{\Lambda_{\zeta,t-1}}.\end{aligned} \tag{2}$$

It is easy to verify that ϕ satisfies $\|\phi\|_{\Lambda_{\zeta,t-1}} \le 1$. Consider linear transforms $\widetilde{x}_\tau = \Lambda_{\zeta,t-1}^{-1/2} x_\tau$ for all $\tau \in \mathcal{T}_{\zeta,t-1}$ and $\widetilde{\phi} = \Lambda_{\zeta,t-1}^{1/2}\phi$. Then $\widetilde{\phi}$ satisfies $\|\widetilde{\phi}\|_2 \le 1$. Subsequently, Eq. (2) can be re-formulated as

$$\|\widehat{\theta}_{\zeta,t} - \theta\|_{\Lambda_{\zeta,t-1}} \le 1 + \sup_{\|\widetilde{\phi}\|_2 \le 1} G_{\widetilde{\phi}}, \tag{3}$$

where $G_{\widetilde{\phi}} = \sum_{\tau \in \mathcal{T}_{\zeta,t-1}} \xi_\tau \langle \widetilde{x}_\tau, \widetilde{\phi}\rangle$.

We next show that $G_\cdot$ is a sub-Gaussian process with respect to $\|\cdot\|_2$. Since $\{\xi_\tau\}_{\tau \in \mathcal{T}_{\zeta,t-1}}$ and $\{x_\tau\}_{\tau \in \mathcal{T}_{\zeta,t-1}}$ are statistically independent [Chu et al., 2011, Auer, 2002], we have that $\{\xi_\tau\}_{\tau \in \mathcal{T}_{\zeta,t-1}}$ and $\{\widetilde{x}_\tau\}_{\tau \in \mathcal{T}_{\zeta,t-1}}$ are statistically independent. Therefore, for any ϕ, ϕ', $G_\phi - G_{\phi'} = \sum_{\tau \in \mathcal{T}_{\zeta,t-1}} \xi_\tau \langle \widetilde{x}_\tau, \phi - \phi'\rangle$ is a centered sub-Gaussian random variable with variance proxy

$$\begin{aligned}&\sum_{\tau \in \mathcal{T}_{\zeta,t-1}} |\langle \widetilde{x}_\tau, \phi - \phi'\rangle|^2 \\ &= (\phi - \phi')^\top \Big(\sum_{\tau \in \mathcal{T}_{\zeta,t-1}} \widetilde{x}_\tau \widetilde{x}_\tau^\top \Big)(\phi - \phi') \\ &= (\phi - \phi')^\top \Lambda_{\zeta,t-1}^{-1/2} \Big(\sum_{\tau \in \mathcal{T}_{\zeta,t-1}} x_\tau x_\tau^\top \Big) \Lambda_{\zeta,t-1}^{-1/2} (\phi - \phi') \\ &\le \|\phi - \phi'\|_2^2.\end{aligned}$$

Subsequently, invoking Lemma 10, we have with probability $1-\gamma$ that

$$\begin{aligned}&\|\widehat{\theta}_{\zeta,t} - \theta\|_{\Lambda_{\zeta,t-1}} \\ \lesssim& 1 + \int_0^\infty \sqrt{\ln N(\{x \in \mathbb{R}^d : \|x\|_2 \le 1\}; \|\cdot\|_2, \epsilon)}\mathrm{d}\epsilon \\ &+ \sqrt{\ln(1/\gamma)} \\ \lesssim& 1 + \int_0^2 \sqrt{d \ln(1/\epsilon)}\mathrm{d}\epsilon + \sqrt{\ln(1/\gamma)} \lesssim \sqrt{d} + \sqrt{\ln(1/\gamma)}.\end{aligned}$$

Finally, Lemma 3 is proved by the Cauchy-Schwarz inequality:

$$\begin{aligned}\big|x^\top(\widehat{\theta}_{\zeta,t} - \theta)\big| &\le \|x\|_{\Lambda_{\zeta,t-1}^{-1}} \|\widehat{\theta}_{\zeta,t} - \theta\|_{\Lambda_{\zeta,t-1}} \\ &\le \omega_{\zeta,t}^x \|\widehat{\theta}_{\zeta,t} - \theta\|_{\Lambda_{\zeta,t-1}}, \quad \forall x \in \mathbb{R}^d,\end{aligned}$$

which is to be demonstrated. □

In this paper, we only use the following weaker version of Lemma 3.

Corollary 4. *For any $t \in [T]$, any layer $\zeta \in \{0, 1, 2, \ldots, \zeta_0\}$, and any $\gamma \in (0, 1/2]$, with probability $1-\gamma$ it holds that*

$$\sup_{x \in \mathbb{R}^d} (\omega_{\zeta,t}^x)^{-1} \big|x^\top(\widehat{\theta}_{\zeta,t} - \theta)\big| \lesssim \sqrt{d \cdot \ln(1/\gamma)}.$$

3.2 Regret upper bound at a single time step

For each time t, we first bound the expected error of the estimation of any arm that lies out of its confidence band using the sharp bound we obtained in Corollary 4.

Lemma 5. *There exists a sufficiently large universal constant $C > 0$ such that for each layer $\zeta \in \{0, 1, 2, \ldots, \zeta_0\}$, for each time $t \in [T]$, and for any*

$\delta \in (0, 1/2]$, *it holds that*

$$\mathbb{E}\left[\max_{x\in D_t}\left\{\mathbf{1}\left[\left|x^\top(\widehat{\theta}_{\zeta,t}-\theta)\right| \geq C\sqrt{d}\cdot\alpha^x_{\zeta,t}\cdot\omega^x_{\zeta,t}\right] \cdot\left|x^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|\right\}\right] \lesssim \delta d/\sqrt{T\ln^2 T}. \quad (4)$$

To prove Lemma 5, we adopt a novel argument that partitions the action set according to the geometric scale of the confidence levels of the actions. Using Corollary 4, for each partition, we derive a uniform error bound for the expected reward of the actions in the partition with empirical estimate $\widehat{\theta}_{\zeta,t}$. Since we have no control on the index of the partition that the maximizer x^* belongs to, we finally employ a union bound argument to combine the error bounds for every partition and complete the proof.

Proof of Lemma 5. For each layer ζ and for each time t, consider a partition of D_t, namely $\{\mathcal{A}^\kappa_{\zeta,t}\}_{\kappa\in\{1,2,3,\ldots,K\}}$, where $K = \lceil \log_2(T^2/\delta^2)\rceil + 1$, and we define

$$\mathcal{A}^\kappa_{\zeta,t} = \begin{cases} \{x\in D_t : \omega^x_{\zeta,t}\in(2^{-\kappa}, 2^{-\kappa+1}]\} & \text{when } \kappa < K \\ \{x\in D_t : \omega^x_{\zeta,t}\in(0, 2^{-\kappa+1}]\} & \text{when } \kappa = K \end{cases}$$

For each κ, we let

$$m^\kappa_{\zeta,t} = \sup_{i\in\mathcal{A}^\kappa_{\zeta,t}}\left\{\left|x^\top_{it}(\widehat{\theta}_{\zeta,t}-\theta)\right|\right\}$$

be the maximum estimation error for the context vectors in $\mathcal{A}^\kappa_{\zeta,t}$ and next we first provide bounds on $m^\kappa_{\zeta,t}$. By Corollary 4, there exists a universal constant C, such that for all $\beta \geq \sqrt{\ln 2}$, we have that

$$\begin{aligned}
&\Pr\left[m^\kappa_{\zeta,t} \geq C\cdot 2^{-\kappa}\sqrt{d}\cdot\beta\right] \\
&= \Pr\left[m^\kappa_{\zeta,t} \geq (C/2)\cdot 2^{-\kappa+1}\sqrt{d}\cdot\beta\right] \\
&\leq \Pr\left[\exists i\in\mathcal{A}^\kappa_{\zeta,t} : \left|x^\top_{it}(\widehat{\theta}_{\zeta,t}-\theta)\right| \geq (C/2)\cdot\omega^i_{\zeta,t}\sqrt{d}\cdot\beta\right] \\
&\leq e^{-\beta^2}.
\end{aligned}$$

Now we let

$$\alpha^\kappa_t = \sqrt{\max\{1, \ln[T\ln^4 T\ln^2(1/\delta)\cdot 2^{-2\kappa}/(d\delta^2)]\}},$$

and use $\mathbf{1}[\cdot]$ to denote the indicator function. For each κ, it holds that

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}\left[m^\kappa_{\zeta,t} \geq C\cdot 2^{-\kappa}\sqrt{d}\cdot\alpha^\kappa_t\right]\cdot m^\kappa_{\zeta,t}\right] \\
\leq\ & \Pr\left[m^\kappa_{\zeta,t} \geq C\cdot 2^{-\kappa}\sqrt{d}\cdot\alpha^\kappa_t\right]\cdot\left(C\cdot 2^{-\kappa}\sqrt{d}\cdot\alpha^\kappa_t\right) \\
&\quad + \int_{C\cdot 2^{-\kappa}\sqrt{d}\cdot\alpha^\kappa_t}^{+\infty}\Pr[m^\kappa_{\zeta,t}\geq z]dz \\
\leq\ & \exp\left(-(\alpha^\kappa_t)^2\right)\cdot C\cdot 2^{-\kappa}\sqrt{d}\cdot\alpha^\kappa_t \\
&\quad + C\cdot 2^{-\kappa}\sqrt{d}\cdot\int_{\alpha^\kappa_t}^{\infty} e^{-\beta^2}d\beta \\
\lesssim\ & \exp\left(-(\alpha^\kappa_t)^2\right)\cdot 2^{-\kappa}\sqrt{d}\cdot\alpha^\kappa_t. \quad (5)
\end{aligned}$$

We now upper bound Eq. (5) by considering the following two cases.

In the first case, when $\alpha^\kappa_t = 1$, we have that

$$T\ln^4 T\ln^2(1/\delta)\cdot 2^{-2\kappa}/(d\delta^2) \leq e,$$

which means that,

$$2^{-\kappa} \lesssim \sqrt{d\delta^2/(T\ln^4 T\ln^2(1/\delta))}.$$

Therefore, Eq. (5) is upper bounded by

$$e^{-1}\cdot 2^{-\kappa}\sqrt{d} \lesssim \delta d/\sqrt{T\ln^4 T\ln^2(1/\delta)}.$$

In the second case, when $\alpha > 1$, we have that

$$T\ln^4 T\ln^2(1/\delta)\cdot 2^{-2\kappa}/(d\delta^2) = \exp((\alpha^\kappa_t)^2)$$

and therefore

$$2^{-\kappa} = \delta\sqrt{d/(T\ln^4 T\ln^2(1/\delta))}\cdot\exp((\alpha^\kappa_t)^2/2).$$

We can upper bound Eq. (5) by

$$\exp\left(-(\alpha^\kappa_t)^2(1-1/2)\right)\cdot\delta\sqrt{d/(T\ln^4 T\ln^2(1/\delta))}\cdot\sqrt{d}\cdot\alpha^\kappa_t \lesssim \delta d/\sqrt{T\ln^4 T\ln^2(1/\delta)}.$$

Summarizing the two cases, we have

$$\mathbb{E}\left[\mathbf{1}\left[m^\kappa_{\zeta,t} \geq C\cdot 2^{-\kappa}\sqrt{d}\cdot\alpha^\kappa_t\right]m^\kappa_{\zeta,t}\right] \lesssim \delta d/\sqrt{T\ln^4 T\ln^2(1/\delta)}. \quad (6)$$

We now work with the Left-Hand Side of Eq. (4). Let x^* be the maximizer in the LHS of Eq. (4), and let κ^* be the index of the partition such that $x^*\in\mathcal{A}^{\kappa^*}_{\zeta,t}$. We

have

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}\left[\left|(x^*)^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|\geq C\sqrt{d}\cdot\alpha_{\zeta,t}^{x^*}\omega_{\zeta,t}^{x^*}\right]\right.\\
&\qquad\left.\times\left|(x^*)^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|\right]\\
&\leq\mathbb{E}\left[\mathbf{1}\left[\kappa^*=K\right]\cdot\left|(x^*)^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|\right]\\
&\quad+\mathbb{E}\left[\mathbf{1}\left[\kappa^*<K\right]\times\mathbf{1}\left[\left|(x^*)^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|\geq\right.\right.\\
&\qquad\left.\left.C\sqrt{d}\cdot\alpha_{\zeta,t}^{x^*}\omega_{\zeta,t}^{x^*}\right]\times\left|(x^*)^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|\right]. \quad (7)
\end{aligned}$$

We first focus on the first term in the Right-Hand Side of Eq. (7). When $\kappa^*=K$, we have $\omega_{\zeta,t}^{x^*}\leq 2/(T/\delta)^2$. Therefore,

$$\begin{aligned}
&\Pr\left[\kappa^*=K\text{ and }\left|(x^*)^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|\geq\delta/\sqrt{T\ln^4 T\ln^2(1/\delta)}\right]\\
&\leq\Pr\left[(\omega_{\zeta,t}^{x^*})^{-1}\left|(x^*)^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|>T^{1.5}/(2\delta\ln^2 T\ln(1/\delta))\right]\\
&\lesssim\delta^3\cdot\exp(-T),
\end{aligned}$$

where the last inequality is due to Corollary 4 and for $T\gtrsim\sqrt{d}$. Therefore, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}\left[\kappa^*=K\right]\left|(x^*)^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|\right]\lesssim\delta/\sqrt{T\ln^4 T\ln^2(1/\delta)}\\
&\quad+\mathbb{E}\left[\mathbf{1}\left[\kappa^*=K\text{ and }\left|(x^*)^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|\right.\right.\\
&\quad\left.\left.>\delta/\sqrt{T\ln^4 T\ln^2(1/\delta)}\right]\cdot\left|(x^*)^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|\right]\\
&\leq\delta/\sqrt{T\ln^4 T\ln^2(1/\delta)}\\
&\quad+\left\{\Pr\left[\kappa^*=K\text{ and }\left|(x^*)^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|\right.\right.\\
&\quad\left.\left.>\delta/\sqrt{T\ln^4 T\ln^2(1/\delta)}\right]\cdot\mathbb{E}\left[\left|(x^*)^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|^2\right]\right\}^{1/2}\\
&\lesssim\delta/\sqrt{T\ln^4 T\ln^2(1/\delta)}, \quad (8)
\end{aligned}$$

where the second inequality due to Cauchy-Schwartz, and the last inequality is because of

$$\begin{aligned}
\mathbb{E}\left[\left|(x^*)^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|^2\right]&\leq\mathbb{E}\left[\left\|\widehat{\theta}_{\zeta,t}-\theta\right\|_2^2\right]\\
&\leq\mathbb{E}\left[\left(\left\|\widehat{\theta}_{\zeta,t}\right\|_2+1\right)^2\right]\lesssim T^2.
\end{aligned}$$

Now we work with the second term in the Right-Hand Side of Eq. (7). When $\kappa^*<K$, we have $2^{-\kappa^*}<\omega_{\zeta,t}^{x^*}$ and $\alpha_t^{\kappa^*}\leq\alpha_{\zeta,t}^{x^*}$, and therefore

$$\begin{aligned}
&\mathbf{1}\left[\kappa^*<K,\left|(x^*)^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|\geq C\sqrt{d}\cdot\alpha_{\zeta,t}^{x^*}\cdot\omega_{\zeta,t}^{x^*}\right]\\
&\qquad\times\left|(x^*)^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|\\
&\leq\ \mathbf{1}\left[\left|(x^*)^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|\geq C\sqrt{d}\cdot\alpha_{\zeta,t}^{x^*}\cdot 2^{-\kappa^*}\right]\\
&\qquad\cdot\left|(x^*)^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|\\
&\leq\mathbf{1}\left[m_t^{\kappa^*}\geq C\sqrt{d}\cdot\alpha_{\zeta,t}^{x^*}\cdot 2^{-\kappa^*}\right]\cdot m_t^{\kappa^*}\\
&\leq\sum_{\kappa=1}^{K-1}\mathbf{1}\left[m_t^{\kappa}\geq C\sqrt{d}\cdot\alpha_{\zeta,t}^{x^*}\cdot 2^{-\kappa^*}\right]\cdot m_{\zeta,t}^{\kappa}. \quad (9)
\end{aligned}$$

Taking expectation and invoking Eq. (6), we have

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}\left[\kappa^*<K,\left|(x^*)^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|\geq C\sqrt{d}\cdot\alpha_{\zeta,t}^{x^*}\cdot\omega_{\zeta,t}^{x^*}\right]\right.\\
&\qquad\left.\cdot\left|(x^*)^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|\right]\\
&\leq\sum_{\kappa=1}^{K-1}\mathbb{E}\left[\mathbf{1}\left[m_t^{\kappa}\geq C\sqrt{d}\cdot\alpha_{\zeta,t}^{x^*}\cdot 2^{-\kappa^*}\right]\cdot m_{\zeta,t}^{\kappa}\right]\\
&\lesssim\delta d/\sqrt{T\ln^2 T}. \quad (10)
\end{aligned}$$

Combining Eq. (7), Eq. (8), and Eq. (10), we have

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}\left[\left|(x^*)^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|\geq C\sqrt{d}\cdot\alpha_{\zeta,t}^{x^*}\cdot\omega_{\zeta,t}^{x^*}\right]\right.\\
&\qquad\left.\times\left|(x^*)^\top(\widehat{\theta}_{\zeta,t}-\theta)\right|\right]\lesssim\delta d/\sqrt{T\ln^2 T},
\end{aligned}$$

which is to be demonstrated. □

For any layer $\zeta\in\{0,1,2,\ldots,\zeta_0\}$ and any time period $t\in[T]$, we define

$$\overline{m}_{\zeta,t}:=\max_{x\in\mathcal{N}_{\zeta,t}}\{x^\top\theta\},\quad\text{and}\quad\underline{m}_{\zeta,t}:=\min_{x\in\mathcal{N}_{\zeta,t}}\{x^\top\theta\}$$

as the largest and smallest mean reward for actions in the action subset $\mathcal{N}_{\zeta,t}$. For convenience, we also define

$$\overline{m}_{\zeta_0+1,t}:=\overline{m}_{\zeta_0,t},\quad\text{and}\quad\underline{m}_{\zeta_0+1,t}:=\underline{m}_{\zeta_0,t}.$$

Note that $\max_{x\in D_t}\{x^\top\theta\}=\overline{m}_{0,t}$ and $x_t^\top\theta\geq\underline{m}_{\zeta_t,t}$ (due to $x_t\in\mathcal{N}_{\zeta_t,t}$), we have that the regret incurred at time t is

$$\begin{aligned}
\max_{x\in D_t}\{x^\top\theta\}-x_t^\top\theta&\leq\left((\overline{m}_{0,t}-\overline{m}_{\zeta_t,t})+(\overline{m}_{\zeta_t,t}-\underline{m}_{\zeta_t,t})\right)\\
&\leq(\overline{m}_{\zeta_t,t}-\underline{m}_{\zeta_t,t})+\sum_{\zeta=1}^{\zeta_t}(\overline{m}_{\zeta-1,t}-\overline{m}_{\zeta,t}). \quad (11)
\end{aligned}$$

In the following lemma, we provide upper bounds for the expressions in the Right-Hand Side of Eq. (11). Intuitively, the following lemma states that the expected

maximum rewards between adjacent layers are close to each other, and the expected differences between any pair of actions inside any layer are small and exponentially decreasing as the layer increases.

Lemma 6. *For all t and $\zeta = 0, 1, \cdots, \zeta_0$, it holds that*

$$\mathbb{E}\left[\max\{\overline{m}_{\zeta,t} - \overline{m}_{\zeta+1,t}, 0\}\right] \lesssim \delta d/\sqrt{T\ln^2 T}; \quad (12)$$

$$\mathbb{E}\left[\max\{\overline{m}_{\zeta,t} - \underline{m}_{\zeta,t} - 2^{3-\zeta}, 0\} \cdot \mathbf{1}\{\zeta \le \zeta_t\}\right] \lesssim \delta d/\sqrt{T\ln^2 T}. \quad (13)$$

Proof. We first prove Eq. (12). Let

$$y_t^* := \arg\max_{y\in\mathcal{N}_{\zeta,t}} \{y^\top\theta\}$$

and

$$z_t^* := \arg\max_{z\in\mathcal{N}_{\zeta,t}} \{z^\top\widehat{\theta}_{\zeta,t}\}.$$

If $y_t^* \in \mathcal{N}_{\zeta+1,t}$, then $\overline{m}_{\zeta,t} = \overline{m}_{\zeta+1,t}$ because $\mathcal{N}_{\zeta+1,t} \subseteq \mathcal{N}_{\zeta,t}$. On the other hand, if $y_t^* \notin \mathcal{N}_{\zeta+1,t}$, note that $z_t^* \in \mathcal{N}_{\zeta+1,t}$ because z_t^* maximizes $z^\top\widehat{\theta}_{\zeta,t}$ in $\mathcal{N}_{\zeta,t}$. Summarizing both cases of $y_t^* \in \mathcal{N}_{\zeta+1,t}$ (in which $\overline{m}_{\zeta+1,t} = \overline{m}_{\zeta,t}$) and $y_t^* \notin \mathcal{N}_{\zeta+1,t}$ (in which $\overline{m}_{\zeta+1,t} \ge (z_t^*)^\top\theta$ as $z_t^* \in \mathcal{N}_{\zeta+1,t}$), we have

$$\overline{m}_{\zeta,t} - \overline{m}_{\zeta+1,t} \le \mathbf{1}\{y_t^* \notin \mathcal{N}_{\zeta+1,t}\} \cdot (y_t^* - z_t^*)^\top\theta. \quad (14)$$

For any ζ, t and $y \in \mathcal{N}_{\zeta,t}$, define

$$\mathcal{E}_{\zeta,t}^y := \{|y^\top(\widehat{\theta}_{\zeta,t} - \theta)| \le \varpi_{\zeta,t}^y\}$$

as the success event in which the estimation error of $y^\top\widehat{\theta}_{\zeta,t}$ for $y^\top\theta$ is within the confidence interval $\varpi_{\zeta,t}^y$. By definition,

$$(y_t^*)^\top\theta \le (y_t^*)^\top\widehat{\theta}_{\zeta,t} + \varpi_{\zeta,t}^{y_t^*} + \mathbf{1}\{\neg\mathcal{E}_{\zeta,t}^{y_t^*}\} \cdot \left|(y_t^*)^\top(\widehat{\theta}_{\zeta,t} - \theta)\right|; \quad (15)$$

$$(z_t^*)^\top\theta \ge (z_t^*)^\top\widehat{\theta}_{\zeta,t} - \varpi_{\zeta,t}^{z_t^*} - \mathbf{1}\{\neg\mathcal{E}_{\zeta,t}^{z_t^*}\} \cdot \left|(z_t^*)^\top(\widehat{\theta}_{\zeta,t} - \theta)\right|. \quad (16)$$

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

$$(y_t^*)^\top\widehat{\theta}_{\zeta,t} < (z_t^*)^\top\widehat{\theta}_{\zeta,t} - 2^{1-\zeta}. \quad (17)$$

Subtracting Eq. (16) from Eq. (15) and considering Eq. (17), we have

$$\begin{aligned}
&(y_t^* - z_t^*)^\top\theta \\
\le&\varpi_{\zeta,t}^{y_t^*} + \varpi_{\zeta,t}^{z_t^*} - 2^{1-\zeta} + \sum_{x\in\{y_t^*, z_t^*\}} \mathbf{1}\{\neg\mathcal{E}_{\zeta,t}^x\} \cdot \left|x^\top(\widehat{\theta}_{\zeta,t} - \theta)\right| \\
\le& \sum_{x\in\{y_t^*, z_t^*\}} \mathbf{1}\{\neg\mathcal{E}_{\zeta,t}^x\} \cdot \left|x^\top(\widehat{\theta}_{\zeta,t} - \theta)\right|, \quad (18)
\end{aligned}$$

where the last inequality holds because $\varpi_{\zeta,t}^x \le 2^{-\zeta}$ for all $x \in \mathcal{N}_{\zeta,t}$, if the algorithm is executed to level $\zeta + 1$. Combining Eqs. (14,18) and Lemma 5, taking expectations, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\max\{\overline{m}_{\zeta,t} - \overline{m}_{\zeta+1,t}, 0\}\right] \\
\le& \mathbb{E}\left[\mathbf{1}\{y_t^* \notin \mathcal{N}_{\zeta+1,t}\} \cdot \Big(\mathbf{1}\{\neg\mathcal{E}_{\zeta,t}^{y_t^*}\}\left|(y_t^*)^\top(\widehat{\theta}_{\zeta,t} - \theta)\right| \right.\right. \\
&\qquad \left.\left. + \mathbf{1}\{\neg\mathcal{E}_{\zeta,t}^{z_t^*}\}\left|(z_t^*)^\top(\widehat{\theta}_{\zeta,t} - \theta)\right|\Big)\right] \\
\le& \mathbb{E}\left[\mathbf{1}\{\neg\mathcal{E}_{\zeta,t}^{y_t^*}\}\left|(y_t^*)^\top(\widehat{\theta}_{\zeta,t} - \theta)\right|\right] \\
&\quad + \mathbb{E}\left[\mathbf{1}\{\neg\mathcal{E}_{\zeta,t}^{z_t^*}\}\left|(z_t^*)^\top(\widehat{\theta}_{\zeta,t} - \theta)\right|\right] \\
\lesssim& \delta d/\sqrt{T\ln^2 T}.
\end{aligned}$$

Now we focus on Eq. (13). We only need to prove the equation for $\zeta > 0$ since it is trivially true for $\zeta = 0$. Let

$$w_t^* := \arg\min_{w\in\mathcal{N}_{\zeta,t}} \{w^\top\theta\}.$$

Clearly, we have that

$$\overline{m}_{\zeta,t} - \underline{m}_{\zeta,t} = (y_t^* - w_t^*)^\top\theta.$$

Similar to Eqs. (15,16), we can establish that

$$\begin{aligned}
(y_t^*)^\top\theta \le& (y_t^*)^\top\widehat{\theta}_{\zeta-1,t} + \varpi_{\zeta-1,t}^{y_t^*} \\
&+ \mathbf{1}\{\neg\mathcal{E}_{\zeta-1,t}^{y_t^*}\} \cdot \left|(y_t^*)^\top(\widehat{\theta}_{\zeta-1,t} - \theta)\right|; \quad (19)\\
(w_t^*)^\top\theta \ge& (w_t^*)^\top\widehat{\theta}_{\zeta-1,t} - \varpi_{\zeta-1,t}^{w_t^*} \\
&- \mathbf{1}\{\neg\mathcal{E}_{\zeta-1,t}^{w_t^*}\} \cdot \left|(w_t^*)^\top(\widehat{\theta}_{\zeta-1,t} - \theta)\right|. \quad (20)
\end{aligned}$$

In addition, because both y_t^* and w_t^* belong to $\mathcal{N}_{\zeta,t} \subseteq \mathcal{N}_{\zeta-1,t}$, the second step of Algorithm 1 implies that conditional on $\zeta \le \zeta_t$,

$$\begin{aligned}
(y_t^*)^\top\widehat{\theta}_{\zeta-1,t} &\le (w_t^*)^\top\widehat{\theta}_{\zeta-1,t} - 2^{1-(\zeta-1)} \\
&\le (w_t^*)^\top\widehat{\theta}_{\zeta-1,t} - 2^{2-\zeta}, \quad (21)
\end{aligned}$$

and

$$\varpi_{\zeta-1,t}^{y_t^*} \le 2^{-(\zeta-1)}, \qquad \varpi_{\zeta-1,t}^{w_t^*} \le 2^{-(\zeta-1)}. \quad (22)$$

Subtracting Eq. (19) from Eq. (20) and applying Eqs. (21,22), we get for any $\zeta \le \zeta_t$,

$$\begin{aligned}
&\overline{m}_{\zeta,t} - \underline{m}_{\zeta,t} = (y_t^* - w_t^*) \\
&\le 2^{2-\zeta} + 2\cdot 2^{-(\zeta-1)} + \mathbf{1}\{\neg\mathcal{E}_{\zeta-1,t}^{y_t^*}\} \cdot \left|(y_t^*)^\top(\widehat{\theta}_{\zeta-1,t} - \theta)\right| \\
&\qquad + 1\{\neg\mathcal{E}_{\zeta-1,t}^{w_t^*}\} \cdot \left|(w_t^*)^\top(\widehat{\theta}_{\zeta-1,t} - \theta)\right| \\
&= 2^{3-\zeta} + \mathbf{1}\{\neg\mathcal{E}_{\zeta-1,t}^{y_t^*}\} \cdot \left|(y_t^*)^\top(\widehat{\theta}_{\zeta-1,t} - \theta)\right| \\
&\qquad + 1\{\neg\mathcal{E}_{\zeta-1,t}^{w_t^*}\} \cdot \left|(w_t^*)^\top(\widehat{\theta}_{\zeta-1,t} - \theta)\right|.
\end{aligned}$$

Therefore, since the right hand-side of the above inequality is non-negative, we have

$$\begin{aligned}
&\mathbb{E}\left[\max\{\overline{m}_{\zeta,t} - \underline{m}_{\zeta,t} - 2^{3-\zeta}, 0\} \cdot \mathbf{1}\{\zeta \le \zeta_t\}\right] \\
&\le \mathbb{E}\left[\mathbf{1}\{\neg\mathcal{E}_{\zeta-1,t}^{y_t^*}\} \times \left|(y_t^*)^\top(\widehat{\theta}_{\zeta-1,t} - \theta)\right| \right. \\
&\qquad \left. + \mathbf{1}\{\neg\mathcal{E}_{\zeta-1,t}^{w_t^*}\} \cdot \left|(w_t^*)^\top(\widehat{\theta}_{\zeta-1,t} - \theta)\right|\right].
\end{aligned}$$

We finally apply Lemma 5 and prove Eq. (13). □

3.3 The elliptical potential lemma, and putting everything together

Lemma 7. *If the parameter C in Algorithm 1 is a large enough universal constant, then we have*

$$\mathbb{E}\left[\max\left\{R_T - 8\cdot\sum_{t=1}^{T}\varpi_{\zeta_t,t}^{x_t}, 0\right\}\right] \lesssim \delta d\sqrt{T}. \quad (23)$$

Note that instead of a high probability bound, which is usual in the previous analysis (e.g., [Dani et al., 2008, Abbasi-Yadkori et al., 2011]), our upper bound is in an expectation form. This crucially helps us to avoid the extra $\log T$ factor due to the union bound argument.

Proof of Lemma 7. Since

$$R_T = \sum_{t=1}^{T}(\max_{x\in D_t}\{x^\top\theta\} - x_t^\top\theta),$$

by Eq. (11) we have

$$R_T \le \sum_{t=1}^{T}\left((\overline{m}_{0,t} - \overline{m}_{\zeta_t,t}) + (\overline{m}_{\zeta_t,t} - \underline{m}_{\zeta_t,t})\right). \quad (24)$$

By Eq. (13) in Lemma 6, we have that for any time t,

$$\begin{aligned}&\mathbb{E}\left[\max\{\overline{m}_{\zeta_t,t} - \underline{m}_{\zeta_t,t}, -2^{3-\zeta_t}, 0\}\right]\\ &\le \sum_{\zeta=0}^{\zeta_0}\mathbb{E}\left[\max\{\overline{m}_{\zeta,t} - \underline{m}_{\zeta,t} - 2^{3-\zeta}, 0\}\cdot\mathbf{1}\{\zeta\le\zeta_t\}\right]\\ &\lesssim \delta d/\sqrt{T},\end{aligned}$$

Together with Eq. (12) in Lemma 6, we have that

$$\begin{aligned}&\mathbb{E}\left[\max\{(\overline{m}_{0,t} - \overline{m}_{\zeta_t,t}) + (\overline{m}_{\zeta_t,t} - \underline{m}_{\zeta_t,t}) - 2^{3-\zeta_t}, 0\}\right]\\ &\le \sum_{\zeta=0}^{\zeta_0} E\left[\max\{\overline{m}_{\zeta,t} - \overline{m}_{\zeta+1,t}, 0\}\right]\\ &\quad + \mathbb{E}\left[\max\{\overline{m}_{\zeta_t,t} - \underline{m}_{\zeta_t,t}, -2^{3-\zeta_t}, 0\}\right] \lesssim \delta d/\sqrt{T}.\end{aligned} \quad (25)$$

Summing up (25) for all $t\in[T]$ and together with (24), we have that

$$\mathbb{E}\left[\max\left\{R_T - \sum_{t=1}^{T}2^{3-\zeta_t}, 0\right\}\right] \lesssim \delta d\sqrt{T}. \quad (26)$$

Note that $\varpi_{\zeta,t}^{x_t} \ge 2^{-\zeta_t} - \delta\sqrt{d/T}$ by the first and the third cases of the if-elseif-else loop of Algorithm 1. Therefore, Eq. (26) implies the lemma statement. □

Below we state a version of the celebrated *elliptical potential lemma*, key to many existing analysis of linearly parameterized bandit problems [Auer, 2002, Filippi et al., 2010, Abbasi-Yadkori et al., 2011, Chu et al., 2011, Li et al., 2017].

Lemma 8 ([Abbasi-Yadkori et al., 2011])**.** *Let $U_0 = I$ and $U_t = U_{t-1} + y_t y_t^\top$ for $t\ge 1$. For any vectors $y_1, y_2, \ldots, y_T$, it holds that*

$$\sum_{t=1}^{T} y_t^\top U_{t-1}^{-1} y_t \le 2\ln(\det(U_T)).$$

Using Lemma 8, we prove the following Lemma 9. The proof Lemma 9 follows the similar lines of Lemma 6 in [Li et al., 2019] and we defer it to Appendix B. At a high level, the proof exploits the power of variated confidence levels (i.e., the specially designed $\alpha_{\zeta,t}^i$ quantity in Algorithm 1) and relies on an application of Jensen's inequality to the concave function $f(\tau) = \sqrt{\tau\ln((T\ln^4 T\ln^2(1/\delta))\tau/(d\delta^2))}$, as well as the commonly used $f(\tau) = \sqrt{\tau}$.

Lemma 9. *It holds that*

$$\sum_t \varpi_{\zeta_t,t}^{x_t} \lesssim d\sqrt{T\log T\log(1/\delta)}\cdot\log\log(T/\delta).$$

Combining Lemma 7 and Lemma 9, we prove Theorem 2.

4 Conclusions

In this paper we study the linearly parameterized contextual bandit problem and develop algorithms that achieve minimax-optimal regret up to iterated logarithmic terms. Future directions include generalizing the proposed approach to contextual bandits with generalized linear models, as well as other variants of contextual bandit problems.

Acknowledgment

Xi Chen would like to thank the support from NSF IIS-1845444. Yuan Zhou would like to thank the support from NSF CCF-2006526.

References

[Abbasi-Yadkori et al., 2011] Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*.

[Audibert and Bubeck, 2009] Audibert, J.-Y. and Bubeck, S. (2009). Minimax policies for adversarial and stochastic bandits. In *Proceedings of the Conference on Learning Theory (COLT)*.

[Auer, 2002] Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

[Chu et al., 2011] Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*.

[Dani et al., 2008] Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback. In *Proceedings of the Conference on Learning Theory (COLT)*.

[Filippi et al., 2010] Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. (2010). Parametric bandits: The generalized linear case. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*.

[Li et al., 2017] Li, L., Lu, Y., and Zhou, D. (2017). Provable optimal algorithms for generalized linear contextual bandits. In *Proceedings of the International Conference of Machine Learning (ICML)*.

[Li et al., 2019] Li, Y., Wang, Y., and Zhou, Y. (2019). Nearly minimax-optimal regret for linearly parameterized bandits. In *Proceedings of the annual Conference on Learning Theory (COLT)*.

[Rusmevichientong and Tsitsiklis, 2010] Rusmevichientong, P. and Tsitsiklis, J. N. (2010). Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411.

[van Handel, 2014] van Handel, R. (2014). Probability in high dimension. Technical report, Princeton University.

[Wang et al., 2018] Wang, Y., Chen, X., and Zhou, Y. (2018). Near-optimal policies for dynamic multinomial logit assortment selection models. In *Proceedings of the advances in Neural Information Processing Systems (NeurIPS)*.