On the Interplay Between Misspecification and Sub-optimality Gap in Linear Contextual Bandits

Weitong Zhang [1] **Jiafan He** [1] **Zhiyuan Fan** [2] **Quanquan Gu** [1]

Abstract

We study linear contextual bandits in the misspecified setting, where the expected reward function can be approximated by a linear function class up to a bounded misspecification level $\zeta > 0$. We propose an algorithm based on a novel data selection scheme, which only selects the contextual vectors with large uncertainty for online regression. We show that, when the misspecification level ζ is dominated by $\widetilde{\mathcal{O}}(\Delta/\sqrt{d})$ with Δ being the minimal sub-optimality gap and d being the dimension of the contextual vectors, our algorithm enjoys the same gap-dependent regret bound $\widetilde{\mathcal{O}}(d^2/\Delta)$ as in the well-specified setting up to logarithmic factors. In addition, we show that an existing algorithm SupLinUCB (Chu et al., 2011) can also achieve a gap-dependent constant regret bound without the knowledge of sub-optimality gap Δ. Together with a lower bound adapted from Lattimore et al. (2020), our result suggests an interplay between misspecification level and the sub-optimality gap: (1) the linear contextual bandit model is efficiently learnable when $\zeta \leq \widetilde{\mathcal{O}}(\Delta/\sqrt{d})$; and (2) it is not efficiently learnable when $\zeta \geq \widetilde{\Omega}(\Delta/\sqrt{d})$. Experiments on both synthetic and real-world datasets corroborate our theoretical results.

1. Introduction

Linear contextual bandits (Li et al., 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011; Agrawal & Goyal, 2013) have been extensively studied when the reward function can be represented as a linear function of the contextual vectors. However, such a well-specified linear model assumption sometimes does not hold in practice. This motivates the study of misspecified linear models. In particular, we only assume that the reward function can be approximated by a linear function up to some worst-case error ζ called *misspecification level*. Existing algorithms for misspecified linear contextual bandits (Lattimore et al., 2020; Foster et al., 2020) can only achieve an $\widetilde{\mathcal{O}}(d\sqrt{K} + \zeta K\sqrt{d}\log K)$ regret bound, where K is the total number of rounds and d is the dimension of the contextual vector. Such a regret, however, suggests that the performance of these algorithms will degenerate to be linear in K when K is sufficiently large. The reason for this performance degeneration is because existing algorithms, such as OFUL (Abbasi-Yadkori et al., 2011) and linear Thompson sampling (Agrawal & Goyal, 2013), utilize all the collected data without selection. This makes these algorithms vulnerable to "outliers" caused by the misspecified model. Meanwhile, the aforementioned results do not consider the sub-optimality gap in the expected reward between the best arm and the second best arm. Intuitively speaking, if the sub-optimality gap is smaller than the misspecification level, there is no hope to obtain a sublinear regret. Therefore, it is sensible to take into account the sub-optimality gap in the misspecified setting, and pursue a gap-dependent regret bound.

The same misspecification issue also appears in reinforcement learning with linear function approximation, when a linear function cannot exactly represent the transition kernel or value function of the underlying MDP. In this case, Du et al. (2019) provided a negative result showing that if the misspecification level is larger than a certain threshold, any RL algorithm will suffer from an exponentially large sample complexity. This result was later revisited in the stochastic linear bandit setting by Lattimore et al. (2020), which shows that a large misspecification error will make the bandit model not efficiently learnable. However, these results cannot well explain the tremendous success of deep reinforcement learning on various tasks (Mnih et al., 2013; Schulman et al., 2015; 2017), where the deep neural networks are used as function approximators with misspecification error.

In this paper, we aim to understand the role of model misspecification in linear contextual bandits through the lens of sub-optimality gap. By proposing a new algorithm with data selection, we can achieve a constant regret bound for such a

[1]Department of Computer Science, University of California, Los Angeles, California, USA [2]IIIS, Tsinghua University, Beijing, China. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

problem. We also shows that the existing algorithm, SupLinUCB (Chu et al., 2011) can be also viewed as a boostrapped version of our proposed algorithm. Our contributions are highlighted as follows:

- We propose a new algorithm called DS-OFUL (Data Selection OFUL). DS-OFUL only learns from the data with large uncertainty. We prove an $\widetilde{\mathcal{O}}(d^2\Delta^{-1})$ constant gap-dependent regret[1] bound independent from K when the misspecification level is small (i.e., $\zeta = \widetilde{\mathcal{O}}(\Delta/\sqrt{d})$) and the minimal sub-optimality gap Δ is known. Our regret bound even improves upon the gap-dependent regret in the well-specified setting (Abbasi-Yadkori et al., 2011) from $\log(K)$ to constant regret bound. To the best of our knowledge, this is the first constant gap-dependant regret bound for misspecified linear contextual bandits as well as the well-specified linear bandit without any prior assumptions.
- We show that an existing algorithm, SupLinUCB (Chu et al., 2011), can be viewed as a multi-level version of our proposed algorithm. With a fine-grained analysis, we are able to show that SupLinUCB can achieve $\widetilde{\mathcal{O}}(d^2\Delta^{-1})$ constant regret under the same condition of misspecification level without knowing the sub-optimality gap.
- We also prove a gap-dependent lower bound following the lower bound proof techniques in Du et al. (2019); Lattimore et al. (2020). This, together with the upper bound, suggests an interplay between the misspecification level and the sub-optimality gap: the linear contextual bandit is efficiently learnable if $\zeta \leq \widetilde{\mathcal{O}}(\Delta/\sqrt{d})$ while it is not efficiently learnable if $\zeta \geq \widetilde{\Omega}(\Delta/\sqrt{d})$.
- Finally, we conduct experiments on the linear contextual bandit with both synthetic and real datasets, and demonstrate the superior performance of DS-OFUL algorithm and the effectiveness of SupLinUCB. This corroborates our theoretical results.

Notation. Scalars and constants are denoted by lower and upper case letters, respectively. Vectors are denoted by lower case boldface letters $\mathbf{x}$, and matrices by upper case boldface letters $\mathbf{A}$. We denote by $[k]$ the set $\{1, 2, \cdots, k\}$ for positive integers k. For two non-negative sequence $\{a_n\}, \{b_n\}$, $a_n = \mathcal{O}(b_n)$ means that there exists a positive constant C such that $a_n \leq Cb_n$, and we use $\widetilde{\mathcal{O}}(\cdot)$ to hide the log factor in $\mathcal{O}(\cdot)$ other than number of rounds T or episode K; $a_n = \Omega(b_n)$ means that there exists a positive constant C such that $a_n \geq Cb_n$, and we use $\widetilde{\Omega}(\cdot)$ to hide the log factor. For a vector $\mathbf{x} \in \mathbb{R}^d$ and a positive semi-definite matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, we define $\|\mathbf{x}\|_{\mathbf{A}}^2 = \mathbf{x}^\top \mathbf{A}\mathbf{x}$. For any set $\mathcal{C}$, we use $|\mathcal{C}|$ to denote its cardinality.

[1] we use notation $\widetilde{\mathcal{O}}(\cdot)$ to hide the log factor other than number of rounds K

2. Related Work

In this section, we review the related work for misspecified linear bandits and misspecified reinforcement learning.

Linear Contextual Bandits. There is a large body of literature on linear contextual bandits. For example, Auer (2002); Chu et al. (2011); Agrawal & Goyal (2013) studied linear contextual bandits when the number of arms is finite. Abbasi-Yadkori et al. (2011) proposed an algorithm called OFUL to deal with the infinite arm set. All these works come with an $\widetilde{\mathcal{O}}(\sqrt{K})$ problem-independent regret bound, and an $\mathcal{O}(d^2\Delta^{-1}\log(K))$ gap-dependent regret bound is also given by Abbasi-Yadkori et al. (2011).

Misspecified Linear Bandits. Ghosh et al. (2017) is probably the first work considering the misspecified linear bandits, which shows that the OFUL (Abbasi-Yadkori et al., 2011) algorithm cannot achieve a sublinear regret in the presence of misspecification. They, therefore, proposed a new algorithm with a hypothesis testing module for linearity to determine whether to use OFUL (Abbasi-Yadkori et al., 2011) or the multi-armed UCB algorithm. Their algorithm enjoys the same performance guarantee as OFUL in the well-specified setting and can avoid the linear regret under certain misspecification setting. Lattimore et al. (2020) proposed a phase-elimination algorithm for misspecified stochastic linear bandits, which achieves an $\widetilde{\mathcal{O}}(\sqrt{dK} + \zeta K\sqrt{d})$ regret bound. For contextual linear bandits, both Lattimore et al. (2020) and Foster et al. (2020) proved an $\widetilde{\mathcal{O}}(d\sqrt{K} + \zeta K\sqrt{d})$ regret bound under misspecification. Takemura et al. (2021) showed that SupLinUCB can achieve a similar regret bound without the knowledge of the misspecification level. Van Roy & Dong (2019) proved a lower bound of sample complexity, which suggests when $\zeta\sqrt{d} \geq \sqrt{8\log|\mathcal{D}|}$, any best arm identification algorithm will suffer a $\Omega(2^d)$ sample complexity, where $\mathcal{D}$ is the decision set. When the reward is deterministic and does not contain noise, they provided an algorithm using $\widetilde{\mathcal{O}}(d)$ sample complexity to identify a Δ-optimal arm when $\zeta \leq \Delta/\sqrt{d}$. Lattimore et al. (2020) also mentioned that if $\zeta\sqrt{d} \leq \Delta$, there exists a best arm identification algorithm that only needs to pull $\widetilde{\mathcal{O}}(d)$ arms to find a Δ-optimal arm with the knowledge of ζ. Note that although the exponential sample complexity lower bound for best-arm identification can be translated into a regret lower bound in linear contextual bandits, the algorithms for best-arm identification and the corresponding upper bounds cannot be easily extended to linear contextual bandits. Besides these works on misspecification, He et al. (2022) studied the linear contextual bandits with adversarial corruptions, where the reward for each round can be corrupted arbitrarily. They assumed that the summation of the corruption up to K rounds is bounded by $C > 0$ and proposed an algorithm achieving $\widetilde{\mathcal{O}}(d\sqrt{K} + dC)$ regret bound with the known C. Since the

corruption level $C = K\zeta$ in the misspecification setting, their result directly implied an $\mathcal{O}(d\sqrt{K} + dK\zeta)$ linear regret, which differs from the optimal guarantee with a extra $O(\sqrt{d})$ factor. Besides these series of work, Camilleri et al. (2021) also studied the robustness of kernel bandits with misspecification.

3. Preliminaries of Linear Contextual Bandits

We consider a linear contextual bandit problem. In round $k \in [K]$, the agent receives a decision set $\mathcal{D}_k \subset \mathbb{R}^d$ and selects an arm $\mathbf{x}_k \in \mathcal{D}_k$ then observes the reward $r_k = r(\mathbf{x}_k) + \varepsilon_k$, where $r(\cdot) : \mathbb{R}^d \mapsto [0, 1]$ is a deterministic expected reward function and ε_k is a zero-mean R-sub-Gaussian random noise. i.e., $\mathbb{E}[\mathrm{e}^{\lambda\varepsilon_k}|\mathbf{x}_{1:k}, \varepsilon_{1:k-1}] \leq \exp(\lambda^2 R^2/2), \forall k \in [K], \lambda \in \mathbb{R}$.

In this work, we assume that all contextual vector $\mathbf{x} \in \mathcal{D}_k$ satisfies $\|\mathbf{x}\|_2 \leq L$ and the reward function $r(\cdot) : \mathbb{R}^d \to [0, 1]$ can be approximated by a linear function $r(\mathbf{x}) = \mathbf{x}^\top\boldsymbol{\theta}^* + \eta(\mathbf{x})$, where $\eta(\cdot) : \mathbb{R}^d \mapsto [-\zeta, \zeta]$ is an unknown misspecification error function. We further assume $\|\boldsymbol{\theta}^*\|_2 \leq B$ and for simplicity, we assume $B, L \geq 1$. We denote the optimal reward at round k as $r_k^* = \max_{\mathbf{x}\in\mathcal{D}_k} r(\mathbf{x})$ and the optimal arm $\mathbf{x}_k^* = \operatorname{argmax}_{\mathbf{x}\in\mathcal{D}_k} r(\mathbf{x})$. Our goal is to minimize the regret defined by $\mathrm{Regret}(K) := \sum_{k=1}^K r_k^* - r(\mathbf{x}_k)$.

In this paper, we focus on the minimal sub-optimality gap condition.

Definition 3.1 (Minimal sub-optimality gap)**.** For each $\mathbf{x} \in \mathcal{D}_k$, the sub-optimality gap $\Delta_k(\mathbf{x})$ is defined by $\Delta_k(\mathbf{x}) := r_k^* - r(\mathbf{x})$ and the minimal sub-optimality gap Δ is defined by $\Delta := \min_{k\in[K],\mathbf{x}\in\mathcal{D}_k}\{\Delta_k(\mathbf{x}) : \Delta_k(\mathbf{x}) > 0\}$.

Then we further assume this minimal sub-optimality gap is strictly positive, i.e., $\Delta > 0$.

4. Constant Regret Bound with Known Sub-Optimality Gap Δ

4.1. Algorithm

In this subsection, we propose our algorithm, DS-OFUL, in Algorithm 1. The algorithm runs for K rounds. At each round, the algorithm first estimates the underlying parameter $\boldsymbol{\theta}^*$ by solving the following ridge regression problem in Line 4

$$\boldsymbol{\theta}_k = \operatorname{argmin}_{\boldsymbol{\theta}} \sum_{i\in\mathcal{C}_{k-1}} \left(r_i - \mathbf{x}_i^\top\boldsymbol{\theta}\right)^2 + \lambda\|\boldsymbol{\theta}\|_2^2,$$

where $\mathcal{C}_{k-1}$ is the index set of the selected contextual vectors for regression and is initialized as an empty set at the beginning. After receiving the contextual vectors set $\mathcal{D}_k$, the algorithm selects an arm from the optimistic estimation powered by the Upper Confidence Bound (UCB) bonus in Line 6. In line 8, the algorithm adds the index of current round into $\mathcal{C}_k$ if the UCB bonus of the chosen arm $\mathbf{x}_k$, denoted by $\|\mathbf{x}_k\|_{\mathbf{U}_k^{-1}}$, is greater than the threshold Γ. Intuitively speaking, since the UCB bonus reflects the uncertainty of the model about the given arm $\mathbf{x}$, Line 8 discards the data that brings little uncertainty ($\|\mathbf{x}\|_{\mathbf{U}_k^{-1}}$) to the model. Finally, we denote the total number of selected data in Line 8 by $|\mathcal{C}_K|$. We will declare the choices of the parameter Γ, β and λ in the next section.

Algorithm 1 Data Selection OFUL (DS-OFUL)

Input: Threshold Γ, radius β and regularizer λ

1: Initialize $\mathcal{C}_0 = \emptyset, \mathbf{U}_0 = \lambda\mathbf{I}, \boldsymbol{\theta}_0 = \mathbf{0}$
2: **for** $k = 1, \ldots, K$ **do**
3: Set $\mathbf{U}_k = \lambda\mathbf{I} + \sum_{i\in\mathcal{C}_{k-1}} \mathbf{x}_i\mathbf{x}_i^\top$.
4: Set $\boldsymbol{\theta}_k = \mathbf{U}_k^{-1}\sum_{i\in\mathcal{C}_{k-1}} r_i\mathbf{x}_i$.
5: Receive the decision set $\mathcal{D}_k$.
6: Select $\mathbf{x}_k = \operatorname{argmax}_{\mathbf{x}\in\mathcal{D}_k}\left\{\mathbf{x}^\top\boldsymbol{\theta}_k + \beta\|\mathbf{x}\|_{\mathbf{U}_k^{-1}}\right\}$.
7: Receive reward r_k
8: **if** $\|\mathbf{x}_k\|_{\mathbf{U}_k^{-1}} \geq \Gamma$ **then** $\mathcal{C}_k = \mathcal{C}_{k-1} \cup \{k\}$ **else** $\mathcal{C}_k = \mathcal{C}_{k-1}$
9: **end for**

4.2. Regret Bound

In this subsection, we provide the regret upper bound of Algorithm 1 and the regret lower bound for learning the misspecified linear contextual bandit.

Theorem 4.1 (Upper Bound)**.** For any $0 < \delta < 1$, let $\lambda = B^{-2}$ and $\Gamma = \Delta/(2\sqrt{d}\iota_1)$ where $\iota_1 = (24 + 18R)\log((72 + 54R)LB\sqrt{d}\Delta^{-1}) + \sqrt{8R^2\log(1/\delta)}$. Set $\beta = 1 + 4\sqrt{d\iota_2} + R\sqrt{2d\iota_3}$ where $\iota_2 = \log(3LB\Gamma^{-1})$, $\iota_3 = \log((1 + 16L^2B^2\Gamma^{-2}\iota_2)/\delta)$. If the misspecification level is bounded by $2\sqrt{d}\zeta\iota_1 \leq \Delta$, then with probability at least $1 - \delta$, the cumulative regret of Algorithm 1 is bounded by

$$\mathrm{Regret}(K) \leq \frac{32\beta\sqrt{2d^3\iota_2\log(1 + 16d\Gamma^{-2}\iota_2)}\iota_1}{\Delta}.$$

Remark 4.2. Since $\beta = \widetilde{\mathcal{O}}(\sqrt{d})$, Theorem 4.1 suggests an $\widetilde{\mathcal{O}}(d^2\Delta^{-1})$ constant regret bound independent of the total number of rounds K when $\zeta \leq \widetilde{\mathcal{O}}(\Delta/\sqrt{d})$, which improves the logarithmic regret $\widetilde{\mathcal{O}}(d^2\Delta^{-1}\log K)$ in Abbasi-Yadkori et al. (2011) to a constant regret[2]. Note that our constant regret bound relies on the knowledge of the minimal sub-optimality gap Δ, while the OFUL algorithm in Abbasi-Yadkori et al. (2011) does not need prior knowledge about the minimal sub-optimality gap Δ.

[2]When we say constant regret, we ignore the $\log(1/\delta)$ factor in the regret as we choose δ to be a constant.

Remark 4.3. Our *high probability* constant regret bound does not violate the lower bound proved in Hao et al. (2020), which says that certain diversity condition on the contexts is necessary to achieve an *expected* constant regret bound (Papini et al., 2021). Here we only provide a high-probability constant regret bound. When extending this high probability constant regret bound to expected regret bound, we have

$$\mathbb{E}[\text{Regret}(K)] \leq \widetilde{\mathcal{O}}(d^2\Delta^{-1}\log(1/\delta))(1-\delta) + \delta K,$$

which depends on K. To obtain a sub-linear expected regret, we can choose $\delta = 1/K$, which yields a logarithmic regret $\widetilde{\mathcal{O}}(d^2\Delta^{-1}\log(K))$ and does not violate the lower bound in Hao et al. (2020).

Remark 4.4. Notably, Papini et al. (2021) can achieve a constant expected regret bound under certain diversity condition, which requires the contexts of arms span the whole $\mathbb{R}^d$ space. In contrast, our constant regret bound does not need such an assumption and is a high-probability constant regret bound.

4.3. Key Proof Techniques

Here we present the key proof techniques for achieving the constant regret with the knowledge of sub-optimality gap Δ. The detailed proof is deferred to Appendix B.

Regret decomposition The total regret over all K rounds can be decomposed as follows

$$\text{Regret}(K) = \sum_{k\in\mathcal{C}_K}\big(r_k^* - r(\mathbf{x}_k)\big) + \sum_{k\notin\mathcal{C}_K}\big(r_k^* - r(\mathbf{x}_k)\big). \tag{4.1}$$

Finite samples collected in $\mathcal{C}_k$ Since we only adding the contextual arm with large uncertainty (i.e., $\|\mathbf{x}\|_{\mathbf{U}_k^{-1}} \geq \Gamma$) into the set $\mathcal{C}_k$, we can bound the number of samples in $\mathcal{C}_k$ as $\mathcal{C}_k = \widetilde{\mathcal{O}}(d\Gamma^{-2})$ which is claimed in the following lemma.

Lemma 4.5. Given $0 < \Gamma \leq 1$, set $\lambda = B^{-2}$. For any $k \in [K]$, $|\mathcal{C}_k| \leq 16d\Gamma^{-2}\log(3LB\Gamma^{-1})$.

Then the following lemma suggests that a finite regression set $\mathcal{C}_k$ can lead to a small confidence set with misspecification.

Lemma 4.6. Let $\lambda = B^{-2}$. For all $\delta > 0$, with probability at least $1-\delta$, for all $\mathbf{x} \in \mathbb{R}^d, k \in [K]$, the prediction error is bounded by:

$$|\mathbf{x}^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}^*)| \leq \left(1 + R\sqrt{2d\iota} + \zeta\sqrt{|\mathcal{C}_k|}\right)\|\mathbf{x}\|_{\mathbf{U}_k^{-1}},$$

where $\iota = \log((d + |\mathcal{C}_k|L^2B^2)/(d\delta))$ and $|\mathcal{C}_k|$ is the total number of data used in regression at the k-th round.

Comparing the confidence radius $\widetilde{\mathcal{O}}(R\sqrt{d} + \zeta\sqrt{|\mathcal{C}_k|})$ here with the conventional radius $\widetilde{\mathcal{O}}(R\sqrt{d})$ in OFUL, one can find that the misspecification error will affect the radius by an $\sqrt{|\mathcal{C}_K|}$ factor. If we use all the data to do regression, the confidence radius will be in the order of $\widetilde{\mathcal{O}}(\sqrt{K})$ and therefore will lead to a $\mathcal{O}(K\sqrt{\log K})$ regret bound (see Lemma 11 in Abbasi-Yadkori et al. (2011)). This makes the regret bound vacuous. In contrast, in our algorithm, the confidence radius is only $\sqrt{|\mathcal{C}_K|}$ where $|\mathcal{C}_K|$ is finite given Lemma 4.5. As a result, our regret bound will not grow with K as in OFUL and will be smaller.

Skipped rounds are optimal Given the fact that the selected arm set $\mathcal{C}_k$ is finite, the rest of the proof is simply showing that the skipped rounds $k \notin \mathcal{C}_k$ are optimal and will not incur regret. Since we have $\|\mathbf{x}\|_{\mathbf{U}_k^{-1}} \leq \Gamma$ for those skipped rounds, the sub-optimality is bounded by the following (informal) lemma.

Lemma 4.7. The instantaneous regret for round $k \notin \mathcal{C}_k$ is bounded by

$$\Delta_k(\mathbf{x}_k) \leq 2\zeta + 2\beta\|\mathbf{x}_k\|_{\mathbf{U}_k^{-1}} \leq \widetilde{\Theta}(\zeta + \Delta + \sqrt{d}\Gamma),$$

Setting $\Gamma = \widetilde{\Theta}(\Delta/\sqrt{d})$ suggests that the instantaneous regret $\Delta_k(\mathbf{x}_k) \leq \Delta$, which means no instantaneous regret occurs on round k.

Achieving the constant regret To wrap up, as (4.1) suggests, for rounds $k \in \mathcal{C}_K$, we can follow the gap-dependent regret analysis in Abbasi-Yadkori et al. (2011) and obtain an $\widetilde{\mathcal{O}}(d^2\log(|\mathcal{C}_K|)/\Delta)$ gap-dependent regret bound, which is independent of K according to Lemma 4.5. For rounds $k \notin \mathcal{C}_K$, Lemma 4.7 guarantees a zero instantaneous regret. Putting them together yields the claimed constant regret bound.

5. Constant Regret Bound with Unknown Sub-Optimality Gap Δ

5.1. Algorithm

Although Algorithm 1 can achieve a constant regret, it requires the knowledge of sub-optimality gap Δ. To tackle this problem, we propose a new algorithm that does not require the knowledge of sub-optimality gap Δ.

The algorithm is described in Algorithm 2. It inherits the arm elimination method from SupLinUCB (Chu et al., 2011). A similar algorithm is also presented for misspecified linear bandits in Takemura et al. (2021).

Algorithm 2 works as follows. At each round $k \in [K]$, the algorithm maintains l levels of ridge regression with different set $\mathcal{C}_{k-1}^l$, where the estimation error for the l-th level

is about $\beta(l)2^{-l}$ (we will prove this in the latter analysis). Then starting from the first level $l = 1$ and the received decision set $\mathcal{D}_k$, if there exists an arm in the decision set with a large uncertainty (i.e., $\|\mathbf{x}\|_{(\mathbf{U}_k^l)^{-1}} \geq 2^{-l}$), the algorithm directly selects that arm (Line 10). According to Lemma 4.5 in the analysis of DS-OFUL, the number of selected contexts at each level should be bounded. If the uncertainty for all arms is smaller than the threshold 2^{-l}, the algorithm follows the arm elimination rule, which reduces the decision set into

$$\mathcal{D}_k^{l+1} = \left\{\mathbf{x} : \mathbf{x} \in \mathcal{D}_k^l, r_k^l(\mathbf{x}_k^l) - r_k^l(\mathbf{x}) \leq 3\beta(l)2^{-l}\right\}. \tag{5.1}$$

Then the algorithm enters the next level $l+1$ until it reaches $\log(k)$-th level as Line 13 suggests. For the level $l \geq \log(k)$, the algorithm directly selects the arm with highest optimistic reward on Line 14 and does not add the index k to the regression set $\mathcal{C}_k^l$ as on Line 15 since the uncertainty is small enough.

Algorithm 2 can be viewed as the multi-level version of Algorithm 1 boosted by the peeling technique. Algorithm 2 does not require the knowledge of the sub-optimality gap Δ: if Δ is known, one can directly jump to a specific level $l_\Delta = \widetilde{\mathcal{O}}(\log(d/\Delta))$, where the prediction error is bounded by $2\beta(l_\Delta)2^{-l_\Delta} = \widetilde{\mathcal{O}}(\Delta)$ and is sufficient to achieve zero-instantaneous regret. However, when the Δ is unknown, Algorithm 2 has to do a grid search over $2^{-1}, 2^{-2}, \cdots 2^{-l_\Delta}, \cdots$ and waste some of the samples to learn the first $l_\Delta - 1$ levels. We will revisit and compare the difference between these two algorithms in the later regret analysis.

5.2. Regret Bound

This subsection provides the regret upper bound for Algorithm 2.

Theorem 5.1 (Upper Bound)**.** For any $0 < \delta < 1$, let $\lambda = B^{-2}$. For every integer $l > 0$, set $\beta(l) = 1 + R\sqrt{2d\iota_2(l)}$ where $\iota_2(l) = \log((d2^l + 16L^2B^2 8^l \iota_1(l))/(d\delta))$ and $\iota_1(l) = \log\left(3LB2^l\right)$. If the misspecification level is bounded by $4l_\Delta\zeta\left(1 + 4\sqrt{d\iota_1(l_\Delta)}\right) < \Delta$ where l_Δ is the minimal solution to $l_\Delta > \log(8\beta(l_\Delta)/\Delta)$, then with probability at least $1-\delta$, the cumulative regret of Algorithm 1 is bounded by

$$\text{Regret}(K) \leq \frac{2^{14}d\beta^2(l_\Delta)\iota_1(l_\Delta)}{\Delta}.$$

Remark 5.2. Since $\beta(l) = \widetilde{\mathcal{O}}(\sqrt{dl})$ and $l_\Delta = \widetilde{\mathcal{O}}(\log(d/\Delta))$, Theorem 5.1 suggests that SupLinUCB enjoys a constant regret bound $\widetilde{\mathcal{O}}(d^2\Delta^{-1})$ when $\zeta \leq \widetilde{\mathcal{O}}(\Delta/\sqrt{d})$, which is independent of the total number of rounds K. Note that in Algorithm 2, the choices of λ and β_l do not depend on the sub-optimality gaps Δ and misspecification level ζ.

Algorithm 2 SupLinUCB

Input: Regularization λ, confidence radius $\beta(\cdot)$
1: Initialize $\mathcal{C}_0^l = \emptyset$ for all $l \in [\lceil \log(K) \rceil]$
2: **for** $k = 1, 2, \cdots K$ **do**
3: Set $\mathcal{D}_k^1 = \mathcal{D}_k$ and $l = 1$
4: **repeat**
5: Set $\mathbf{U}_k^l = \lambda\mathbf{I} + \sum_{i \in \mathcal{C}_{k-1}^l} \mathbf{x}_i\mathbf{x}_i^\top$
6: Set $\boldsymbol{\theta}_k^l = (\mathbf{U}_k^l)^{-1} \sum_{i \in \mathcal{C}_{k-1}^l} r_i\mathbf{x}_i$
7: Set $r_k^l(\mathbf{x}) = \mathbf{x}^\top\boldsymbol{\theta}_k^l + \beta(l)\|\mathbf{x}\|_{(\mathbf{U}_k^l)^{-1}}$
8: Select action $\mathbf{x}_k^l = \text{argmax}_{\mathbf{x} \in \mathcal{D}_k^l} r_k^l(\mathbf{x})$
9: **if** $\max_{\mathbf{x} \in \mathcal{D}_k^l} \|\mathbf{x}\|_{(\mathbf{U}_k^l)^{-1}} \geq 2^{-l}$ **then**
10: Choose $\mathbf{x}_k = \text{argmax}_{\mathbf{x} \in \mathcal{D}_k^l} \|\mathbf{x}\|_{(\mathbf{U}_k^l)^{-1}}$
11: Update $\mathcal{C}_k^l = \mathcal{C}_{k-1}^l \cup \{k\}$
12: Keep $\mathcal{C}_k^{l'} = \mathcal{C}_{k-1}^{l'}$ for all $l' \neq l$
13: **else if** $k \leq 4^l d$ **then**
14: Choose $\mathbf{x}_k = \mathbf{x}_k^l$
15: Keep $\mathcal{C}_k^{l'} = \mathcal{C}_{k-1}^{l'}$ for all $l' \geq 1$
16: **else**
17: Set $\mathcal{D}_k^{l+1}$ according to (5.1)
18: Increase $l = l + 1$
19: **end if**
20: **until** $\mathbf{x}_k$ is chosen
21: Take action $\mathbf{x}_k$ and receive reward r_k
22: **end for**

Remark 5.3. When $\zeta \geq \Delta/\sqrt{d}$, it is hard to provide a gap-dependent regret bound due to the large misspecification level ζ. However, a gap-independent regret bound of $\widetilde{\mathcal{O}}(\sqrt{dK} + \sqrt{d}\zeta K\log(K))$ is proved in Takemura et al. (2021), which suggests the performance of SupLinUCB algorithm will not significantly decrease when the condition on misspecification does not hold.

Remark 5.4. Comparing the constant factors of DS-OFUL (Algorithm 1) and SupLinUCB (Algorithm 2) on the dominating terms $\widetilde{\mathcal{O}}(\beta^2 d/\Delta)$, one can find that the constant factors of SupLinUCB is significantly larger than DS-OFUL. This is because it takes more samples to learn the first $l_\Delta - 1$ levels in SupLinUCB while DS-OFUL directly learns the l_Δ-th level. Therefore, despite having the same order of constant regret bound (in big-O notation), one can expect that SupLinUCB has a worse performance than DS-OFUL (when Δ is known or can be estimated by grid search).

5.3. Key Proof Techniques

Here we provide additional proof techniques besides the techniques discussed in Section 4.3. First of all, Lemmas 4.5 and 4.6, which are built on a single level selected

by $\|\mathbf{x}\|_{\mathbf{U}_k^{-1}} \geq \Gamma$, can be generalized to the following lemmas for all levels l. The detailed proof are deferred to Appendix D.

Lemma 5.5. Set $\lambda = B^{-2}$, for any $k \in [K]$ and $l > 0$, $|\mathcal{C}_k^l| \leq 16d4^l\iota_1(l)$, where $\iota_1(l) = \log\left(3LB2^l\right)$.

Lemma 5.6. Set $\lambda = B^{-2}$. For any level $l > 0$, for any $\delta > 0$, with probability at least $1-\delta$, for all $k \in [K]$, the prediction error is bounded by

$$\left|\mathbf{x}^\top(\boldsymbol{\theta}_k^l - \boldsymbol{\theta}^*)\right| \leq \left(1 + R\sqrt{2d\iota_2(l)} + \zeta\sqrt{|\mathcal{C}_k^l|}\right)\|\mathbf{x}\|_{(\mathbf{U}_k^l)^{-1}},$$

for all $\mathbf{x}$ such that $\|\mathbf{x}\|_2 \leq L$, where $\iota_2(l) = \log((d + |\mathcal{C}_k^l|L^2B^2)/(d\delta))$.

The following two proof techniques are crucial to prove constant regret bound of Algorithm 2.

Optimal arm is never eliminated Considering the optimal arm in the eliminated set, which is defined by $\mathbf{x}_k^{l,*} = \operatorname{argmax}_{\mathbf{x}\in\mathcal{D}_l} r(\mathbf{x})$. Obviously $\mathbf{x}_k^{1,*} = \mathbf{x}_k^*$. The following (informal) lemma says that the decision set always contains a nearly optimal action $\mathbf{x}_k^{l,*}$:

Lemma 5.7 (informal)**.** For any level $l > 0$, assume some good events hold, then there exists $\mathbf{x}_k^{l,*} \in \mathcal{D}_k^l$, such that $r(\mathbf{x}_k^*) - r(\mathbf{x}_k^{l,*}) \leq 2(l-1)\zeta\left(1 + 4\sqrt{d\iota_1(l)}\right)$ where $\iota_1(l) = \log\left(3LB2^l\right)$.

Given the result of Lemma 5.7 and the existence of the sub-optimality gap Δ, we have $\mathbf{x}_k^{l,*} = \mathbf{x}_k^*$ when l is not too large. This means that the optimal arm is never eliminated from the decision set $\mathcal{D}^l$.

Sub-optimal arms are all eliminated Intuitively speaking, at level l, the prediction error is bounded by $\widetilde{\mathcal{O}}(\beta(l) \cdot 2^{-l})$ with some additional misspecification term ζ. Therefore, when we eliminate the arms at level l, the sub-optimality of the arms in $\mathcal{D}^l$ is bounded by the following (informal) lemma:

Lemma 5.8 (informal)**.** For any level $l > 0$, for any arm $\mathbf{x} \in \mathcal{D}_k^l$, $r(\mathbf{x}_k^*) - r(\mathbf{x}) \leq 6\beta(l)2^{-l} + 2l\zeta\left(1 + 4\sqrt{d\iota_1(l)}\right)$ where $\iota_1(l) = \log\left(3LB2^l\right)$.

Given Lemma 5.8, we know that when l is sufficiently large (e.g., larger than l_Δ), all $\mathbf{x} \in \mathcal{D}_k^l$ enjoys a sub-optimality less than Δ. Combining with the existence of sub-optimality gap Δ, we know that all of the sub-optimal arms are eliminated after level l_Δ.

Regret decomposition Given Lemma 5.5 and Lemma 5.8, the regret over all K rounds can be decomposed into

$$\begin{aligned}\text{Regret}(K) &= \sum_{k=1}^{K}\left(r(\mathbf{x}_k^*) - r(\mathbf{x}_k)\right)\\ &= \sum_{l\geq 1}\sum_{k\in\mathcal{C}_K^l}\left(r(\mathbf{x}_k^*) - r(\mathbf{x}_k)\right)\\ &= \sum_{l=1}^{l_\Delta}\sum_{k\in\mathcal{C}_K^l}\left(r(\mathbf{x}_k^*) - r(\mathbf{x}_k)\right),\end{aligned}$$

where the last equality is due to the fact that no regret occurs after $l > l_\Delta$. For each level $l \leq l_\Delta$, the summation of the instantaneous regret within $k \in \mathcal{C}_K^l$ can be bounded following the gap-dependent regret bound of Abbasi-Yadkori et al. (2011) to obtain a $\widetilde{\mathcal{O}}(d^2\log|\mathcal{C}_K^l|/\Delta)$ regret bound which is independent from K. Then taking the summation over $l \leq l_\Delta$ yields the claimed constant regret bound.

6. Lower Bound

Following a similar idea in Lattimore et al. (2020), we prove a gap-dependent lower bound for misspecified stochastic linear bandits. Note that stochastic linear bandit can be seen as a special case of linear contextual bandits with a fixed decision set $\mathcal{D}_k = \mathcal{D}$ across all round $k \in [K]$. Similar results and proof can be found in Du et al. (2019) for episodic reinforcement learning.

Theorem 6.1 (Lower Bound)**.** Given the dimension d and the number of arms $|\mathcal{D}|$, for any $\Delta \leq 1$ and $\zeta \geq 3\Delta\sqrt{8\log(|\mathcal{D}|)/(d-1)}$, there exists a set of stochastic linear bandit problems $\boldsymbol{\Theta}$ with minimal sub-optimality gap Δ and misspecification error level ζ, such that for any algorithm that has a sublinear expected regret bound for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, i.e., $\mathbb{E}[\text{Regret}_{\boldsymbol{\theta}}(K)] \leq CK^\alpha$ with $C > 0$ and $0 \leq \alpha < 1$, we have

- When $K \leq \mathcal{O}(|\mathcal{D}|)$, the expected regret is lower bounded by $\mathbb{E}_{\boldsymbol{\theta}\sim\text{Unif.}(\boldsymbol{\Theta})}[\text{Regret}_{\boldsymbol{\theta}}(K)] \geq K\Delta$.
- When $K \geq \Omega(|\mathcal{D}|)$, the expected regret is lower bounded by $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\mathbb{E}[\text{Regret}_{\boldsymbol{\theta}}(K)] \geq \widetilde{\Omega}(|\mathcal{D}|\log(K)\Delta^{-1})$.

Remark 6.2. Theorem 6.1 shows two regimes under the case $\zeta \geq \widetilde{\Omega}(\Delta/\sqrt{d})$. In the first regime $K \leq \mathcal{O}(|\mathcal{D}|)$ where the decision set is large (e.g., $|\mathcal{D}| = d^{100}$), any algorithm will suffer from a linear regret $\widetilde{\mathcal{O}}(\Delta K)$, which suggests that the regime cannot be efficiently learnable. In the second regime $K \geq \Omega(|\mathcal{D}|)$, Theorem 6.1 suggests an $\widetilde{\Omega}(|\mathcal{D}|\Delta^{-1}\log(K))$ regret lower bound, which is matched by the multi-armed bandit algorithm with an upper bound $\widetilde{\mathcal{O}}(|\mathcal{D}|\Delta^{-1}\log(K))$ (Lattimore & Szepesvári, 2020).

Therefore, in this easier regime, linear function approximation cannot provide any performance improvement and one can simply adopt the multi-armed bandit algorithm to learn the bandit model.

Remark 6.3. Theorems 4.1 and 6.1 provide a holistic picture about the role of misspecification in linear contextual bandits. Here we focus on the more difficult regime $K \leq |\mathcal{D}|$. In the regime $K \leq |\mathcal{D}|$, when $\zeta \leq \widetilde{\mathcal{O}}(\Delta/\sqrt{d})$, Theorem 4.1 suggests that the bandit problem is efficiently learnable, and our algorithm DS-OFUL can achieve a constant regret, which improves upon the logarithmic regret bound in the well-specified setting (Abbasi-Yadkori et al., 2011). On the other hand, when $\zeta \geq \widetilde{\Omega}(\Delta/\sqrt{d})$, Theorem 6.1 provides a linear regret lower bound suggesting that the bandit model can not be efficiently learned.

7. Experiments

To verify the performance improvement by data selection using the UCB bonus in Algorithm 1 and the effectiveness of the parameter-free algorithm Algorithm 2, we conduct experiments for bandit tasks on both synthetic and real-world datasets, which we will describe in detail below.

7.1. Synthetic Dataset

The synthetic dataset is composed as follows: we set $d = 16$ and generate parameter $\boldsymbol{\theta}^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and contextual vectors $\{\mathbf{x}_i\}_{i=1}^N \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ where $N = 100$. The generated parameter and vectors are later normalized to be $\|\boldsymbol{\theta}^*\|_2 = \|\mathbf{x}_i\|_2 = 1$. The reward function is calculated by $r_i = \langle \boldsymbol{\theta}^*, \mathbf{x}_i \rangle + \eta_i$ where $\eta_i \sim \text{Unif}\{-\zeta, \zeta\}$. The contextual vectors and reward function is fixed after generated. The random noise on the receiving rewards ε_t are sampled from the standard normal distribution.

We set the misspecification level $\zeta = 0.02$ and verified that the sub-optimality gap over the N contextual vectors $\Delta \approx 0.18$. We do a grid search for $\beta = \{1, 3, 10\}$, $\lambda = \{1, 3, 10\}$ [3] and report the cumulative regret of Algorithm 1 with different parameter $\Gamma = \{0, 0.02, 0.05, 0.08, 0.18\}$ over 8 independent trials with total rounds $K = 10000$. It is obvious that when $\Gamma = 0$, our algorithm degrades to the standard OFUL algorithm (Abbasi-Yadkori et al., 2011) which uses data from all rounds into regression.

Besides the OFUL algorithm, we also compare with the algorithm (LSW) in Equation (6) of Lattimore et al. (2020) and the RLB in Ghosh et al. (2017) in Figure 1 and Table 1. For Lattimore et al. (2020), the estimated reward is updated by $r(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\theta}_k + \beta \|\mathbf{x}\|_{\mathbf{U}_k^{-1}} + \varepsilon \sum_{s=1}^k |\mathbf{x}^\top \mathbf{U}_k^{-1} \mathbf{x}_s^{-1}|$. However, since the time complexity of the LSW algorithm is $\widetilde{\mathcal{O}}(K^2)$ due to the hardness of calculating $\varepsilon \sum_{s=1}^k |\mathbf{x}^\top \mathbf{U}_k^{-1} \mathbf{x}_s^{-1}|$ incrementally w.r.t. k. In our setting it takes more than 7 hours for 10000 rounds.

[3] By "grid search", we tune the parameter $(\beta, \lambda) = (1, 1), (1, 3), \cdots, (10, 3), (10, 10)$ and see their results.

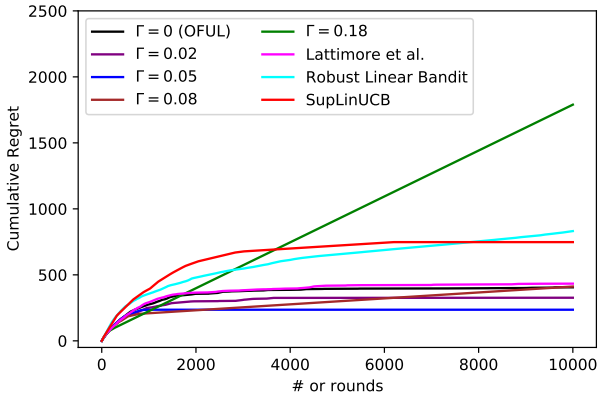


Figure 1. Cumulative regret comparison of DS-OFUL (with difference choices of Γ), SupLinUCB, Lattimore et al. (2020) and Robust Linear Bandit Ghosh et al. (2017) over 10000 rounds. Results are averaged over 8 replicates.

For the RLB algorithm in Ghosh et al. (2017), we did the hypothesis test for $k = 10$ rounds and then decided whether to use OFUL or multi-armed UCB. The results show that both LSW and RLB achieve a worse regret than OFUL since in our setting ζ is relatively small.

The result is shown in Figure 1 and the average cumulative regret on the last round is reported in Table 1 with its variance over 8 trials. We can see that by setting $\Gamma \approx \Delta/\sqrt{d} \approx 0.18/\sqrt{16} \approx 0.05$, Algorithm 1 can achieve less cumulative regret compared with OFUL ($\Gamma = 0$). The algorithm with a proper choice of Γ also convergences to zero instantaneous regret faster than OFUL. It is also evident that a too large $\Gamma = 0.18 \approx \Delta$ will cause the algorithm to fail to learn the contextual vectors and induce a linear regret. Also, our algorithm shows that using a larger Γ can significantly boost the speed of the algorithm by reducing the number of regressions needed in the algorithm.

Besides the performance improvement achieved by Algorithm 1, the experiments also demonstrates the effectiveness of Algorithm 2. As Table 1 suggests, SupLinUCB achieves a zero cumulative regret over the last 1000 steps. However, as discussed in Remark 5.4, the total regret of SupLinUCB is much higher than the DS-OFUL and OFUL since it takes more samples to learn the first $l_\Delta - 1$ levels which is not used by DS-OFUL. This constant larger sample complexity could also be verified by a longer elapsed time for executing the SubLinUCB comparing to DS-OFUL.

7.2. Real-world Dataset

To demonstrate that the proposed algorithm can be easily applied to modern machine learning tasks, we carried

Table 1. Averaged cumulative regret and elapsed time of DS-OFUL over 8 runs. The **bold face** value indicates the best (low regret or low elapsed time) for all the algorithm configurations

Algorithm Configuration, (Γ)	Regret (mean$\pm$std.)	Regret in last 1k steps	Elapsed Time(sec)
OFUL (Abbasi-Yadkori et al., 2011), $\Gamma = 0$	405.4 ± 76.5	4.94	15.06
DS-OFUL (Algorithm 1), $\Gamma = 0.02$	326.5 ± 68.0	**0.0**	8.59
DS-OFUL (Algorithm 1), $\Gamma = 0.05$	$\mathbf{235.75 \pm 40.3}$	**0.0**	6.30
DS-OFUL (Algorithm 1), $\Gamma = 0.08$	411.6 ± 566.7	22.44	5.97
DS-OFUL (Algorithm 1), $\Gamma = 0.13$	1789.5 ± 1918.8	173.67	**5.56**
Eq. (6) in Lattimore et al. (2020)	433.36 ± 64	1.79	≥ 7 hrs.
Robust Linear Bandit (Ghosh et al., 2017)	831.5 ± 880.4	42.58	12.85
SupLinUCB (Algorithm 2)	747.9 ± 329.5	**0.0**	31.86

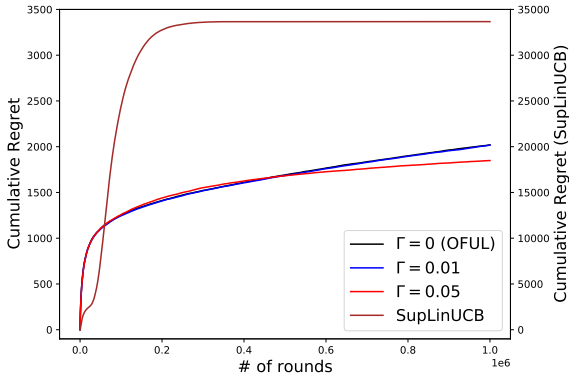


Figure 2. Cumulative regret of DS-OFUL on the Asirra dataset over 1M rounds with different Γ under misspecification level $\zeta = 0.01$. Results are averaged over 8 runs. The cumulative regret of DS-OFUL (as well as OFUL) can be read from the y-axis on the left. The cumulative regret of SupLinUCB algorithm can be read from the y-axis on the right.

out experiments on the Asirra dataset (Elson et al., 2007). The task of agent is to distinguish the image of cats from the image of dogs. At each round k, the agent receives the feature vector $\boldsymbol{\phi}_{1,k} \in \mathbb{R}^{512}$ of a cat image and another feature vector $\boldsymbol{\phi}_{2,k} \in \mathbb{R}^{512}$ of a dog image. Both feature vectors are generated using ResNet-18 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009). We normalize $\|\boldsymbol{\phi}_{1,k}\|_2 = \|\boldsymbol{\phi}_{2,k}\|_2 = 1$. The agent is required to select the cat from these two vectors. It receives reward $r_t = 1$ if it selects the correct feature vector, and receives $r_t = 0$ otherwise. It is trivial that the sub-optimality gap of this task is $\Delta = 1$. To better demonstrate the influence of misspecification on the performance of the algorithm, we only select the data with $|\boldsymbol{\phi}_i^\top \boldsymbol{\theta}^* - r_i| \leq \zeta$ with $r_i = 1$ if it is a cat and $r_i = 0$ otherwise. $\boldsymbol{\theta}^*$ is a pretrained parameter on the whole dataset using linear regression $\boldsymbol{\theta}^* = \operatorname{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^N (\boldsymbol{\phi}_i^\top \boldsymbol{\theta} - r_i)^2$, which the agent does not know. For hyper-parameter tuning, we select $\beta = \{0.1, 0.3, 1\}$ and $\lambda = \{1, 3, 10\}$ by doing a grid search [4] and repeat the experiments for 8 times over 1M rounds for each parameter configuration. As shown in Figure 2, when $\zeta = 0.01$, setting $\Gamma = 0.05 \approx \Delta/\sqrt{d}$ will eventually have a better performance comapred with OFUL algorithm (setting $\Gamma = 0$). On the other hand, the SupLinUCB algorithm (Algorithm 2) will suffer from a much higher, but constant regret bound, which is well aligned with our theoretical result especially Remark 5.4. We skip the Robust Linear Bandit (Ghosh et al., 2017) algorithm since it is for stochastic linear bandit with fixed contextual features for each arm while here the contextual features are sampled and not fixed. The LSW (Equation (6) in Lattimore et al. (2020) is skipped due to the infeasible executing time.

As a sensitivity analysis, we also set $\zeta = \{0.5, 0.1, 0.05\}$ to test the impact of misspecification on the performance of algorithm choices of Γ. More experiment configurations and results are deferred to Appendix A.

8. Conclusion and Future Work

We study the misspecified linear contextual bandit from a gap-dependent perceptive. We propose an algorithm and show that if the misspecification level $\zeta \leq \widetilde{\mathcal{O}}(\Delta/\sqrt{d})$, the proposed algorithm, DS-ODUL, can achieve the same gap-dependent regret bound as in the well-specified case. Along with Lattimore et al. (2020); Du et al. (2019), we provide a complete picture on the interplay between misspecification and sub-optimality gap, in which $\Delta/\sqrt{d}$ plays an important role on the phase transition of ζ to decide if the bandit model can be efficiently learned.

Besides the aforementioned constant regret result, DS-OFUL algorithm requires the knowledge of sub-optimality ap Δ. We prove that the SupLinUCB algorithm (Chu et al., 2011) can be viewed as a multi-level version of our algorithm and can also achieve a constant regret with our fine-grained analysis without the knowledge of Δ. Experiments

[4] By "grid search", we tune the parameter $(\beta, \lambda) = (0.1, 1), (0.1, 3), \cdots, (1, 3), (1, 10)$ and see their results.

are conducted to demonstrate the performance of the DS-OFUL algorithm and verify the effectiveness of SupLinUCB algorithm.

The promising result suggests a few interesting directions for future research. For example, it would be interesting to incorporate the Lipschitz continuity or smoothness properties of the reward function to derive fine-grained results.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. WZ, JH and QG are supported in part by the National Science Foundation CAREER Award 1906169 and research fund from UCLA-Amazon Science Hub. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pp. 127–135. PMLR, 2013.

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

Camilleri, R., Jamieson, K., and Katz-Samuels, J. High-dimensional experimental design and kernel bandits. In *International Conference on Machine Learning*, pp. 1227–1237. PMLR, 2021.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Du, S. S., Kakade, S. M., Wang, R., and Yang, L. F. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2019.

Elson, J., Douceur, J. J., Howell, J., and Saul, J. Asirra: A captcha that exploits interest-aligned manual image categorization. In *Proceedings of 14th ACM Conference on Computer and Communications Security (CCS)*. Association for Computing Machinery, Inc., October 2007.

Foster, D. J., Gentile, C., Mohri, M., and Zimmert, J. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33, 2020.

Ghosh, A., Chowdhury, S. R., and Gopalan, A. Misspecified linear bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Hao, B., Lattimore, T., and Szepesvari, C. Adaptive exploration in linear contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, pp. 3536–3545. PMLR, 2020.

He, J., Zhou, D., and Gu, Q. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pp. 4171–4180. PMLR, 2021a.

He, J., Zhou, D., and Gu, Q. Uniform-PAC bounds for reinforcement learning with linear function approximation. In *Advances in Neural Information Processing Systems*, 2021b.

He, J., Zhou, D., Zhang, T., and Gu, Q. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. In *Advances in Neural Information Processing Systems*, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Lattimore, T., Szepesvari, C., and Weisz, G. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pp. 5662–5670. PMLR, 2020.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Papini, M., Tirinzoni, A., Restelli, M., Lazaric, A., and Pirotta, M. Leveraging good representations in linear contextual bandits. In *International Conference on Machine Learning*, pp. 8371–8380. PMLR, 2021.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Takemura, K., Ito, S., Hatano, D., Sumita, H., Fukunaga, T., Kakimura, N., and Kawarabayashi, K.-i. A parameter-free algorithm for misspecified linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 3367–3375. PMLR, 2021.

Van Roy, B. and Dong, S. Comments on the du-kakade-wang-yang lower bounds. *arXiv preprint arXiv:1911.07910*, 2019.

Zanette, A., Lazaric, A., Kochenderfer, M. J., and Brunskill, E. Learning near optimal policies with low inherent bellman error. In *ICML*, 2020.

A. Experiment Details and Additional Results

A.1. Experiment Configuration

The experiment on synthetic dataset is conducted on Google Colab with a 2-core Intel® Xeon® CPU @ 2.20GHz. The experiment on the real-world Asirra dataset (Elson et al., 2007) is conducted on an AWS p2-xlarge instance.

A.2. Data Preprocessing for the Asirra Dataset

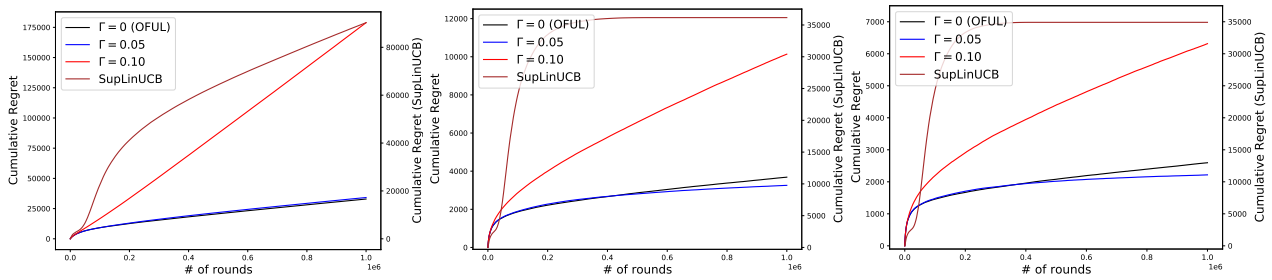
To demonstrate how our algorithm can deal with different levels of misspecification, we do data preprocessing before feeding the data into the agent. As described in Section 7.2, the remaining data with expected misspecification level ζ are shown in Table 2. It can be verified that even with the smallest misspecification level, there are still more than 10% of the data is selected.

Table 2. The number of remaining data samples after data processing with expected misspecification level

ζ	# of cats	# of dogs
∞ (without preprocessing)	12500	12500
0.5 (linear separable)	10316	10511
0.1	3182	3248
0.05	2408	2442
0.01	1886	1905

A.3. Additional Result on the Asirra Dataset

As a sensitivity analysis, we change the misspecification level in the preprocessing part in the Asirra dataset. The result is shown in Figure 3. This result suggests that when the misspecification is small enough, setting $\Gamma = \Delta/\sqrt{d}$ can deliver a reasonable result and SupLinUCB Chu et al. (2011) can achieve a constant regret bound when $\zeta \leq 0.1$. It is aligned with the parameter setting in our Theorem 4.1 and the result in our Theorem 5.1. Meanwhile, we found that when $\zeta = 0.5$, which means it is strictly larger than the threshold $\Delta/\sqrt{d}$, the algorithm cannot achieve a similar performance with of $\zeta < 0.1$, regardless of the setting of parameter Γ. This also verifies the theoretical understanding of how a large misspecification level will harm the performance of the algorithm.



(a) $\zeta = 0.5$ (b) $\zeta = 0.1$ (c) $\zeta = 0.05$

Figure 3. The performance of DS-OFUL under different misspecification levels ζ. Results are averaged over 8 runs, with standard errors shown as shaded areas.

B. Detailed Proof of Theorem 4.1

In this section, we provide detailed proof for Theorem 4.1. First, we present a technical lemma to bound the total number of data used in the online linear regression in Algorithm 1.

Lemma B.1 (Restatement of Lemma 4.5)**.** Given $0 < \Gamma \leq 1$, set $\lambda = B^{-2}$. For any $k \in [K]$, $|\mathcal{C}_k| \leq 16d\Gamma^{-2}\log(3LB\Gamma^{-1})$.

Lemma B.1 suggests that up to $\widetilde{O}(d\Gamma^{-2})$ contextual vectors have a UCB bonus greater than Γ. A similar result is also provided in He et al. (2021b), suggesting an $\widetilde{\mathcal{O}}(\Gamma^{-2})$ Uniform-PAC sample complexity. Lemma B.1 also suggests that the numbers of data points added into the regression set $\mathcal{C}$ is finite. Thus, the impact of the noise and the misspecification on the linear regression estimator can be well-controlled.

For a linear regression with up to $|\mathcal{C}_k|$ data points, the next lemma controls the prediction error under misspecification.

Lemma B.2 (Formal statement of Lemma 4.6)**.** Let $\lambda = B^{-2}$. For all $\delta > 0$, with probability at least $1-\delta$, for all $\mathbf{x} \in \mathbb{R}^d, k \in [K]$, the prediction error is bounded by:

$$|\mathbf{x}^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}^*)| \leq \left(1 + R\sqrt{2d\iota} + \zeta\sqrt{|\mathcal{C}_k|}\right)\|\mathbf{x}\|_{\mathbf{U}_k^{-1}},$$

where $\iota = \log((d + |\mathcal{C}_k|L^2B^2)/(d\delta))$ and $|\mathcal{C}_k|$ is the total number of data used in regression at the k-th round.

Lemma B.2 provides a similar confidence bound as the well-specified linear contextual bandits algorithms like OFUL (Abbasi-Yadkori et al., 2011). Comparing the confidence radius here $\widetilde{\mathcal{O}}(R\sqrt{d} + \zeta\sqrt{|\mathcal{C}_{k-1}|})$ with the conventional radius in OFUL $\widetilde{\mathcal{O}}(R\sqrt{d})$, one can find that there is an additional term $\zeta\sqrt{|\mathcal{C}_k|}$ that is caused by the misspecification. If we directly use all data to do the regression, the resulting confidence radius will be in the order of $\widetilde{\mathcal{O}}(\sqrt{K})$ and therefore will lead to a $\mathcal{O}(K\sqrt{\log K})$ regret bound (see Lemma 11 in Abbasi-Yadkori et al. (2011)). This makes the regret bound vacuous. In our algorithm, however, the confidence radius is only $\sqrt{|\mathcal{C}_k|}$ where $|\mathcal{C}_k|$ is bounded by Lemma B.1. As a result, our regret bound will not be vacuous (i.e., superlinear in K).

When the misspecification level is well bounded by $\zeta = \widetilde{\mathcal{O}}(\Delta/\sqrt{d})$, the following corollary is a direct result of Lemmas B.2 by replacing the term $|\mathcal{C}_k|$ with its upper bound provided in Lemma B.1.

Corollary B.3. Suppose $2\sqrt{d}\zeta\iota_1 \leq \Delta$, let $\lambda = B^{-2}$ and $0 < \Gamma \leq 1$. Let $\beta = 1 + 2\Delta\Gamma^{-1}\sqrt{\iota_2}/\iota_1 + R\sqrt{2d\iota_3}$ where $\iota_2 = \log(3LB\Gamma^{-1})$, $\iota_3 = \log((1 + 16L^2B^2\Gamma^{-2}\iota_2)/\delta)$, then with probability at least $1-\delta$, for all $\mathbf{x} \in \mathbb{R}^d, k \in [K]$, the estimation error for all $k \in [K]$ is bounded by: $|\mathbf{x}^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}^*)| \leq \beta\|\mathbf{x}\|_{\mathbf{U}_k^{-1}}$.

Proof. By Lemma B.1, replacing $|\mathcal{C}_k|$ with its upper bound yields

$$|\mathbf{x}^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}^*)| \leq (1 + 4\sqrt{d}\zeta\Gamma^{-1}\sqrt{\iota_2} + R\sqrt{2d\iota_3})\|\mathbf{x}\|_{\mathbf{U}_k^{-1}} \leq \beta\|\mathbf{x}\|_{\mathbf{U}_k^{-1}},$$

where the second inequality is due to the condition $2\sqrt{d}\zeta \leq \Delta/\iota_1$. □

Next we introduce an auxiliary lemma controlling the instantaneous regret bound using the UCB bonus and the misspecification level.

Lemma B.4 (Formal statement of Lemma 4.7)**.** Suppose Corollary B.3 holds, for all $k \in [K]$, the instantaneous regret at round k is bounded by

$$\Delta_k(\mathbf{x}_k) = r_k^* - r(\mathbf{x}_k) \leq 2\zeta + 2\beta\|\mathbf{x}_k\|_{\mathbf{U}_k^{-1}}.$$

The next technical lemma from He et al. (2021a) bounds the summation of a subset of the bonuses.

Lemma B.5 (Lemma 6.6, He et al. 2021a)**.** For any subset $\mathcal{G} = \{c_1, \cdots, c_i\} \subseteq \mathcal{C}_K$, we have

$$\sum_{k \in \mathcal{G}} \|\mathbf{x}_k\|^2_{\mathbf{U}_k^{-1}} \leq 2d\log(1 + |\mathcal{G}|L^2/\lambda).$$

The next auxiliary lemma is used to control the dominating terms.

Lemma B.6. Let $\iota_1 = (24 + 18R)\log((72 + 54R)LB\sqrt{d}\Delta^{-1}) + \sqrt{8R^2\log(1/\delta)}$, $\Gamma = \Delta/(2\sqrt{d}\iota_1)$, $\iota_2 = \log(3LB\Gamma^{-1}), \iota_3 = \log((1 + 16L^2B^2\Gamma^{-2}\iota_2)/\delta)$, we have $\iota_1 > 2 + 4\sqrt{\iota_2} + R\sqrt{2\iota_3}$.

Equipped with these lemmas, we can start the proof of Theorem 4.1.

Proof of Theorem 4.1. First, note that by setting $\Gamma = \Delta/(2\sqrt{d}\iota_1)$, the confidence radius β becomes $1 + 4\sqrt{d\iota_2} + R\sqrt{2d\iota_3}$. Then our proof starts by assuming that Corollary B.3 holds with probability at least $1-\delta$. We decompose the index set $[K]$ into two subsets. The first set is the set of not selected data $[K] \setminus \mathcal{C}_K$, and the second set is the set of selected data $\mathcal{C}_K$. We will bound the cumulative regret within these two sets separately.

First, for those non-selected data $k \notin \mathcal{C}_k$, i.e. $\|\mathbf{x}_k\|_{\mathbf{U}_k^{-1}} < \Gamma$, combining Lemma B.4 with Corollary B.3 yields

$$r_k^* - r(\mathbf{x}_k) < 2\zeta + 2\beta\Gamma = 2\zeta + \frac{\Delta}{\sqrt{d}\iota_1} + \frac{\sqrt{2\iota_3}R\Delta}{\iota_1} + \frac{4\Delta\sqrt{\iota_2}}{\iota_1}, \tag{B.1}$$

where $\iota_1, \iota_2, \iota_3$ are the same as Theorem 4.1, and the equality is due to $\Gamma = \Delta/(2\sqrt{d}\iota_1)$. When misspecification condition $2\sqrt{d}\zeta \leq \Delta/\iota_1$ holds, (B.1) suggests that

$$r_k^* - r(\mathbf{x}_k) < \frac{2\Delta}{\sqrt{d}\iota_1} + \frac{4\Delta\sqrt{\iota_2}}{\iota_1} + \frac{\sqrt{2\iota_3}R\Delta}{\iota_1}. \tag{B.2}$$

Lemma B.6 suggests that when $\iota_1 = (24+18R)\log((72+54R)LB\sqrt{d}\Delta^{-1}) + \sqrt{8R^2\log(1/\delta)}\,\iota_1 > 2 + 4\sqrt{\iota_2} + R\sqrt{2\iota_3}$, (B.2) yields that the instantaneous regret $r_k^* - r(\mathbf{x}_k) < \Delta$ at round k. By Definition 3.1, the instantaneous regret is zero for all $k \notin \mathcal{C}_k$, indicating the non-selected data incur zero instantaneous regret.

In addition, Lemma B.4 suggests that the instantaneous regret for those $k \in \mathcal{C}_K$ is bounded by

$$\begin{aligned}
\sum_{k\in\mathcal{C}_K} r_k^* - r(\mathbf{x}_k) &\leq \sum_{k\in\mathcal{C}_K} \left(2\beta\|\boldsymbol{\phi}_k\|_{\mathbf{U}_k^{-1}} + 2\zeta\right) \\
&\leq 2\beta\sqrt{|\mathcal{C}_K|}\sqrt{\sum_{k\in\mathcal{C}_K}\|\boldsymbol{\phi}_k\|^2_{\mathbf{U}_k^{-1}}} + 2|\mathcal{C}_K|\zeta \\
&\leq 8\beta\Gamma^{-1}\sqrt{d\iota_2}\sqrt{2d\log(1+16d\Gamma^{-2}\iota_2)} + 32\zeta d\Gamma^{-2}\iota_2 \\
&\leq 16\beta\sqrt{2d^3\iota_2\log(1+16d\Gamma^{-2}\iota_2)}\iota_1/\Delta + 64\sqrt{d^3}\iota_1\iota_2/\Delta \\
&\leq 32\beta\sqrt{2d^3\iota_2\log(1+16d\Gamma^{-2}\iota_2)}\iota_1/\Delta,
\end{aligned} \tag{B.3}$$

where the second inequality follows the Cauchy-Schwarz inequality, the third one yields from Lemma B.5 while the fourth utilizes the fact that $\Gamma = \Delta/(2\sqrt{d}\iota_1)$ and $\zeta \leq \Delta/(2\sqrt{d}\iota_1)$. The last one is due to the fact that the second term in the fourth inequality is dominated by the first one.

To warp up, the cumulative regret can be decomposed by

$$\text{Regret}(K) = \sum_{k\notin\mathcal{C}_K}(r_k^* - r(\mathbf{x}_k)) + \sum_{k\in\mathcal{C}_K}(r_k^* - r(\mathbf{x}_k)) \leq 0 + \frac{32\beta\sqrt{2d^3\iota_2\log(1+16d\Gamma^{-2}\iota_2)}\iota_1}{\Delta},$$

where the first two zeros are given by the fact that for $k \notin \mathcal{C}_K$, we have $r_k^* - r(\mathbf{x}_k) = 0$. the regret bound for $k \in \mathcal{G}$ is given by (B.3). □

C. Proof of Technical Lemmas in Appendix B

C.1. Proof of Lemma B.1

To prove this lemma, we introduce the well-known elliptical potential lemma (Abbasi-Yadkori et al., 2011)

Lemma C.1 (Lemma 11, Abbasi-Yadkori et al. 2011)**.** Let $\{\boldsymbol{\phi}_i\}_{i=1}^I$ be a sequence in $\mathbb{R}^d$, define $\mathbf{U}_i = \lambda\mathbf{I} + \sum_{j=1}^i \boldsymbol{\phi}_j\boldsymbol{\phi}_j^\top$, then

$$\sum_{i=1}^I \min\left\{1, \|\boldsymbol{\phi}_i\|^2_{\mathbf{U}_{i-1}^{-1}}\right\} \leq 2d\log\left(\frac{\lambda d + IL^2}{\lambda d}\right).$$

The following auxiliary lemma and its corollary are useful

Lemma C.2 (Lemma A.2, Shalev-Shwartz & Ben-David 2014)**.** Let $a \geq 1$ and $b > 0$. Then $x \geq 4a\log(2a) + 2b$ yields $x \geq a\log(x) + b$.

Lemma C.2 can easily indicate the following lemma.

Lemma C.3. Let $a \geq 1$. Then $x \geq 4\log(2a) + a^{-1}$ yields $x \geq \log(1 + ax)$.

Proof. Let $y = 1 + ax, x = (y-1)/a$. Then $x \geq 4\log(2a) + a^{-1}$ is equivalent with $y \geq 4a\log(2a) + 2$. By Lemma C.2, this implies $y \geq a\log(y) + 1$ which is exactly $x \geq \log(1 + ax)$. □

Equipped with these technical lemmas, we can start our proof.

Proof of Lemma B.1. Since the cardinality of set $\mathcal{C}_k$ is monotonically increasing w.r.t. k, we fix k to be K in the proof and only provide the bound of $\mathcal{C}_K$. For all selected data $k \in \mathcal{C}_K$, we have $\|\boldsymbol{\phi}_k\|_{\mathbf{U}_k^{-1}} \geq \Gamma$. Therefore, when $\Gamma \leq 1$, the summation of the bonuses over data $k \in \mathcal{C}_K$ is lower bounded by

$$\sum_{k \in \mathcal{C}_K} \min\left\{1, \|\boldsymbol{\phi}_k\|^2_{\mathbf{U}_k^{-1}}\right\} \geq |\mathcal{C}_K| \min\{1, \Gamma^2\} = |\mathcal{C}_K|\Gamma^2. \tag{C.1}$$

On the other hand, Lemma C.1 implies

$$\sum_{k \in \mathcal{C}_K} \min\left\{1, \|\boldsymbol{\phi}_k\|^2_{\mathbf{U}_k^{-1}}\right\} \leq 2d\log\left(\frac{\lambda d + |\mathcal{C}_K|L^2}{\lambda d}\right). \tag{C.2}$$

Combining (C.2) and (C.1), the total number of the selected data points $|\mathcal{C}_K|$ is bounded by

$$\Gamma^2|\mathcal{C}_K| \leq 2d\log\left(\frac{\lambda d + |\mathcal{C}_K|L^2}{\lambda d}\right).$$

This result can be re-organized as

$$\frac{\Gamma^2|\mathcal{C}_K|}{2d} \leq \log\left(1 + \frac{2L^2}{\Gamma^2\lambda}\frac{\Gamma^2|\mathcal{C}_K|}{2d}\right). \tag{C.3}$$

Let $\lambda = B^{-2}$ and since $2L^2B^2 \geq 2 \geq \Gamma^2$, by Lemma C.3, if

$$\frac{\Gamma^2|\mathcal{C}_K|}{2d} > 4\log\left(\frac{4L^2B^2}{\Gamma^2}\right) + 1 \geq 4\log\left(\frac{4L^2B^2}{\Gamma^2}\right) + \frac{\Gamma^2}{2L^2B^2},$$

then (C.3) will not hold. Thus the necessary condition for (C.3) to hold is

$$\frac{\Gamma^2|\mathcal{C}_K|}{2d} \leq 4\log\left(\frac{4L^2B^2}{\Gamma^2}\right) + 1 = 8\log\left(\frac{2LB}{\Gamma}\right) + \log(e) = 8\log\left(\frac{2LBe^{\frac{1}{8}}}{\Gamma}\right) < 8\log\left(\frac{3LB}{\Gamma}\right).$$

By basic calculus we get the claimed bound for $|\mathcal{C}_K|$ and complete the proof. □

C.2. Proof of Lemma B.2

The proof follows the standard technique for linear bandits, we first introduce the self-normalized bound for vector-valued martingales from Abbasi-Yadkori et al. (2011).

Lemma C.4 (Theorem 1, Abbasi-Yadkori et al. 2011)**.** Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\varepsilon_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that ε_t is $\mathcal{F}_t$-measurable and ε_t is conditionally R-sub-Gaussian for some $R \geq 0$. Let $\{\boldsymbol{\phi}_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that $\boldsymbol{\phi}_t$ is $\mathcal{F}_{t-1}$ measurable and $\|\boldsymbol{\phi}\|_2 \leq L$ for all t. For any $t \geq 0$, define $\mathbf{U}_t = \lambda\mathbf{I} + \sum_{k=1}^{t} \boldsymbol{\phi}_k\boldsymbol{\phi}_k$. Then for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 0$

$$\left\|\sum_{k=1}^{t} \boldsymbol{\phi}_k\varepsilon_k\right\|^2_{\mathbf{U}_t^{-1}} \leq 2R^2\log\left(\frac{\sqrt{\det(\mathbf{U}_t)}}{\sqrt{\det(\mathbf{U}_0)}\delta}\right).$$

Lemma C.5 (Lemma 8, Zanette et al. 2020)**.** Let $\{\mathbf{a}_i\}_{i=1}^d$ be any sequence of vectors in $\mathbb{R}^d$ and $\{b_i\}_{i=1}^d$ be any sequence of scalars such that $|b_i| \leq \zeta$. For any $\lambda > 0$:

$$\left\|\sum_{i=1}^n \mathbf{a}_i b_i\right\|^2_{\left[\sum_{i=1}^n \mathbf{a}_i\mathbf{a}_i^\top + \lambda\mathbf{I}\right]^{-1}} \leq n\zeta^2.$$

The next lemma is to bound the perturbation of the misspecification

Lemma C.6. Let $\{\eta_k\}_k$ be any sequence of scalars such that $|\eta_k| \leq \zeta$ for any $k \in [K]$. For any index subset $\mathcal{C} \subseteq [K]$, define $\mathbf{U} = \lambda\mathbf{I} + \sum_{k\in\mathcal{C}} \mathbf{x}_k\mathbf{x}_k^\top$, then for any $\mathbf{x} \in \mathbb{R}^d$, we have

$$\left|\mathbf{x}^\top\mathbf{U}^{-1}\sum_{k\in\mathcal{C}}\mathbf{x}_k\eta_k\right| \leq \zeta\sqrt{|\mathcal{C}|}\|\mathbf{x}\|_{\mathbf{U}^{-1}}.$$

Proof. By Cauchy-Schwartz inequality we have

$$\left|\mathbf{x}^\top\mathbf{U}^{-1}\sum_{k\in\mathcal{C}}\mathbf{x}_k\eta_k\right| \leq \|\mathbf{x}\|_{\mathbf{U}^{-1}}\left\|\sum_{k\in\mathcal{C}}\mathbf{x}_k\eta_k\right\|_{\mathbf{U}^{-1}} \leq \zeta\sqrt{|\mathcal{C}|}\|\mathbf{x}\|_{\mathbf{U}^{-1}},$$

where the second inequality dues to lemma C.5. □

The next lemma is the Determinant-Trace inequality.

Lemma C.7. Suppose sequence $\{\mathbf{x}_k\}_{k=1}^K \subset \mathbb{R}^d$ and for any $k \in [K]$, $\|\mathbf{x}_k\|_2 \leq L$. For any index subset $\mathcal{C} \subseteq [K]$, define $\mathbf{U} = \lambda\mathbf{I} + \sum_{k\in\mathcal{C}} \mathbf{x}_k\mathbf{x}_k^\top$ for some $\lambda > 0$, then $\det(\mathbf{U}) \leq (\lambda + |\mathcal{C}|L^2/d)^d$.

Proof. The proof of this lemma is almost the same as Lemma 10 in Abbasi-Yadkori et al. (2011) by replacing the index set $[K]$ with any subset $\mathcal{C}$. We refer the readers to Abbasi-Yadkori et al. (2011) for details. □

Equipped with these lemmas, we can start our proof.

Proof of Lemma B.2. For any $k \in [K]$, considering the data samples $k' \in \mathcal{C}_{k-1}$ used for regression at round k. Following the update rule of $\mathbf{U}_k$ and $\boldsymbol{\theta}_k$ yields

$$\begin{aligned}
\mathbf{U}_k(\boldsymbol{\theta}_k - \boldsymbol{\theta}^*) &= \mathbf{U}_k\mathbf{U}_k^{-1}\left(\sum_{k'\in\mathcal{C}_{k-1}}\mathbf{x}_{k'}r_{k'}\right) - \left(\lambda\mathbf{I} + \sum_{k'\in\mathcal{C}_{k-1}}\mathbf{x}_{k'}\mathbf{x}_{k'}^\top\right)\boldsymbol{\theta}^* \\
&= \sum_{k'\in\mathcal{C}_{k-1}}\mathbf{x}_{k'}r_{k'} - \lambda\boldsymbol{\theta}^* - \sum_{k'\in\mathcal{C}_{k-1}}\mathbf{x}_{k'}\mathbf{x}_{k'}^\top\boldsymbol{\theta}^* \\
&= -\lambda\boldsymbol{\theta}^* + \sum_{k'\in\mathcal{C}_{k-1}}\mathbf{x}_{k'}(r_{k'} - \mathbf{x}_{k'}^\top\boldsymbol{\theta}^*) \\
&= -\lambda\boldsymbol{\theta}^* + \sum_{k'\in\mathcal{C}_{k-1}}\mathbf{x}_{k'}\varepsilon_{k'} + \sum_{k'\in\mathcal{C}_{k-1}}\mathbf{x}_{k'}\eta_{k'},
\end{aligned}$$

where the first equation is due to the fact that $\mathbf{U}_k = \lambda\mathbf{I} + \sum_{k'\in\mathcal{C}_{k-1}}\mathbf{x}_k\mathbf{x}_k^\top$ and $\boldsymbol{\theta}_k = \mathbf{U}_k^{-1}\sum_{k'\in\mathcal{C}_{k-1}}\mathbf{x}_{k'}r_{k'}$. The last equation follows the fact that $r_{k'}$ is generated from $r_{k'} = r(\mathbf{x}_{k'}) + \varepsilon_{k'} = \mathbf{x}_{k'}^\top\boldsymbol{\theta}^* + \eta(\mathbf{x}_{k'}) + \varepsilon_{k'}$, where we denote $\eta(\mathbf{x}_{k'})$ as $\eta_{k'}$ for the model misspecification error and $\varepsilon_{k'}$ is the random noise. Therefore, consider any contextual vector $\mathbf{x} \in \mathbb{R}^d$, we have

$$\begin{aligned}
\left|\mathbf{x}^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}^*)\right| &= \left|\mathbf{x}^\top\mathbf{U}_k^{-1}\mathbf{U}_k(\boldsymbol{\theta}_k - \boldsymbol{\theta}^*)\right| \\
&\leq \lambda\underbrace{\left|\mathbf{x}^\top\mathbf{U}_k^{-1}\boldsymbol{\theta}^*\right|}_{q_1} + \underbrace{\left|\mathbf{x}^\top\mathbf{U}_k^{-1}\sum_{k'\in\mathcal{C}_{k-1}}\boldsymbol{\phi}_{k'}\varepsilon_{k'}\right|}_{q_2} + \underbrace{\left|\mathbf{x}^\top\mathbf{U}_k^{-1}\sum_{k'\in\mathcal{C}_{k-1}}\boldsymbol{\phi}_{k'}\eta_{k'}\right|}_{q_3},
\end{aligned}$$

where the inequality is due to the triangle inequality. Lemma C.6 yields $q_3 \leq \zeta\sqrt{|\mathcal{C}_{k-1}|}\|\mathbf{x}\|_{\mathbf{U}_k^{-1}}$. From the fact that $|\mathbf{x}^\top \mathbf{A}\mathbf{y}| \leq \|\mathbf{x}\|_{\mathbf{A}}\|\mathbf{y}\|_{\mathbf{A}}$, we can bound term q_1 by

$$q_1 \leq \|\mathbf{x}\|_{\mathbf{U}_k^{-1}}\|\boldsymbol{\theta}^*\|_{\mathbf{U}_k^{-1}} \leq \lambda^{-1/2}B\|\mathbf{x}\|_{\mathbf{U}_k^{-1}}. \tag{C.4}$$

where the last inequality is due to the fact that $\mathbf{U}_k^{-1} \preceq \lambda^{-1}\mathbf{I}$. Term q_2 is also bounded as

$$q_2 \leq \|\mathbf{x}\|_{\mathbf{U}_k^{-1}}\left\|\sum_{k'\in\mathcal{C}_{k-1}}\mathbf{x}_{k'}\varepsilon_{k'}\right\|_{\mathbf{U}_k^{-1}} = \|\mathbf{x}\|_{\mathbf{U}_k^{-1}}\underbrace{\left\|\sum_{k'=1}^{K}\mathbb{1}\left[k'\in\mathcal{C}_{k-1}\right]\mathbf{x}_{k'}\varepsilon_{k'}\right\|_{\mathbf{U}_k^{-1}}}_{I_1}, \tag{C.5}$$

where the second equation uses the indicator function to rewrite the summation over subset $\mathcal{C}_{k-1}$. Denoting $\mathbf{y}_{k'} = \mathbb{1}\left[k'\in\mathcal{C}_{k-1}\right]\mathbf{x}_{k'}$, noticing that $\|\mathbf{y}_{k'}\|_2 \leq \|\mathbf{x}_{k'}\|_2 \leq L$ and

$$\mathbf{U}_k = \sum_{k'\in\mathcal{C}_{k-1}}\mathbf{x}_{k'}\mathbf{x}_{k'}^\top = \sum_{k'=1}^{K}\mathbb{1}\left[k'\in\mathcal{C}_{k-1}\right]\mathbf{x}_{k'}\mathbf{x}_{k'}^\top = \sum_{k'=1}^{K}\mathbf{y}_{k'}\mathbf{y}_{k'}^\top,$$

by Lemma C.4, I_1 can be further bounded by

$$I_1 \leq \sqrt{2R^2\log\left(\frac{\sqrt{\det(\mathbf{U}_k)}}{\sqrt{\det(\mathbf{U}_0)}\delta}\right)} \leq R\sqrt{2\log\left(\frac{\det(\mathbf{U}_k)}{\det(\mathbf{U}_0)\delta}\right)} = R\sqrt{2\log\left(\frac{\det(\mathbf{U}_k)}{\lambda^d\delta}\right)}, \tag{C.6}$$

where the second inequality follows the fact that $\det(\mathbf{U}_k) \geq \det(\mathbf{U}_0) = \lambda^d$. Notice that $\mathbf{U}_k = \lambda\mathbf{I} + \sum_{k'\in\mathcal{C}_{k-1}}\mathbf{x}_{k'}\mathbf{x}_{k'}^\top$. Lemma C.7 suggests that $\det(\mathbf{U}_k) \leq (\lambda + |\mathcal{C}_{k-1}|L^2/d)^d$, plugging this into (C.6), we obtain

$$I_1 \leq R\sqrt{2\log\left(\frac{(\lambda+|\mathcal{C}_{k-1}|L^2/d)^d}{\lambda^d\delta}\right)} \leq R\sqrt{2d\log\left(\frac{d\lambda+|\mathcal{C}_{k-1}|L^2}{d\lambda\delta}\right)}.$$

Plugging the bound of I_1 into (C.5) and combining with (C.4) and Lemma C.6 together, replacing $|\mathcal{C}_{k-1}|$ with its upper bound $|\mathcal{C}_K|$ we have with probability at least $1-\delta$, for all $k\in[K], \mathbf{x}\in\mathbb{R}^d$,

$$|\mathbf{x}^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}^*)| \leq \left(R\sqrt{2d\log\left(\frac{d\lambda+|\mathcal{C}_K|L^2}{d\lambda\delta}\right)} + B\lambda^{-1/2} + \zeta\sqrt{|\mathcal{C}_K|}\right)\|\boldsymbol{\phi}\|_{\mathbf{U}_k^{-1}}.$$

Letting $\lambda = B^{-2}$ we get the claimed results. □

C.3. Proof of Lemma B.4

Proof. According to the definition of expected reward function $r(\mathbf{x})$, we have for all $k\in[K]$, suppose the condition in Lemma B.2 holds, then

$$\begin{aligned}
r_k^* - r_k &= \eta(\mathbf{x}_k^*) - \eta(\mathbf{x}_k) + (\mathbf{x}_k^*)^\top\boldsymbol{\theta}^* - \mathbf{x}_k^\top\boldsymbol{\theta}^* \\
&\leq 2\zeta + (\mathbf{x}_k^*)^\top\boldsymbol{\theta}^* - \mathbf{x}_k^\top\boldsymbol{\theta}^* \\
&= 2\zeta + (\mathbf{x}_k^*)^\top\boldsymbol{\theta}_k + (\mathbf{x}_k^*)^\top(\boldsymbol{\theta}^* - \boldsymbol{\theta}_k) - \mathbf{x}_k^\top\boldsymbol{\theta}_k + \mathbf{x}_k^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}^*) \\
&\leq 2\zeta + (\mathbf{x}_k^*)^\top\boldsymbol{\theta}_k + \beta\|\mathbf{x}_k^*\|_{\mathbf{U}_k^{-1}} - \mathbf{x}_k^\top\boldsymbol{\theta}_k + \beta\|\mathbf{x}_k\|_{\mathbf{U}_k^{-1}} \\
&\leq 2\zeta + \mathbf{x}_k^\top\boldsymbol{\theta}_k + \beta\|\mathbf{x}_k\|_{\mathbf{U}_k^{-1}} - \mathbf{x}_k^\top\boldsymbol{\theta}_k + \beta\|\mathbf{x}_k\|_{\mathbf{U}_k^{-1}} \\
&\leq 2\zeta + 2\beta\|\mathbf{x}_k\|_{\mathbf{U}_k^{-1}},
\end{aligned}$$

where the first inequality utilize the fact that $|\eta(\mathbf{x})| \leq \zeta$ for all $\mathbf{x}\in\mathcal{D}_k$, the second inequality follows from Corollary B.3, the third inequality is due to the fact that $\mathbf{x}_k = \operatorname{argmax}_{\mathbf{x}\in\mathcal{D}_k}\mathbf{x}^\top\boldsymbol{\theta}_k + \beta\|\mathbf{x}\|_{\mathbf{U}_k^{-1}}$, which is executed in Line 6 of Algorithm 1. □

C.4. Proof of Lemma B.6

Proof. First it is clear to see that $\sqrt{2\iota_3} = \sqrt{2\log(1+16L^2B^2\Gamma^{-2}\iota_2) + 2\log(1/\delta)}$. Using the fact that $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$, it can be further bounded by

$$\sqrt{2\iota_3} \le \sqrt{2\log(1+16L^2B^2\Gamma^{-2}\iota_2)} + \sqrt{2\log(1/\delta)}.$$

Assuming $L \ge 1, B \ge 1, \Gamma = \Delta/(2\sqrt{d}\iota_1) \le 1$ yields $LB\Gamma^{-1} \ge 1$, then by basic calculus one can verify that

$$2 + 4\sqrt{\iota_2} \le 6\log(3LB\Gamma^{-1}), \quad \sqrt{2\log(1+16L^2B^2\Gamma^{-2}\iota_2)} \le 3\log(3LB\Gamma^{-1}),$$

therefore we have that

$$\begin{aligned} 2 + 4\sqrt{\iota_2} + R\sqrt{2\iota_3} &\le (6+3R)\log(3LB\Gamma^{-1}) + \sqrt{2\log(1/\delta)}R \\ &= (6+3R)\log(6LB\sqrt{d}\Delta^{-1}\iota_1) + \sqrt{2\log(1/\delta)}R, \end{aligned}$$

where the last equality is from the fact that $\Gamma = \Delta/(2\sqrt{d}\iota_1)$. Lemma C.2 suggests that the necessary condition for

$$\underbrace{(6LB\sqrt{d}\Delta^{-1})\iota_1}_{x} \ge \underbrace{(6LB\sqrt{d}\Delta^{-1})(6+3R)}_{a}\log(6LB\sqrt{d}\Delta^{-1}\iota_1) + \underbrace{(6LB\sqrt{d}\Delta^{-1})\sqrt{2\log(1/\delta)}R}_{b} \tag{C.7}$$

is that

$$\begin{aligned} (6LB\sqrt{d}\Delta^{-1})\iota_1 \ge{}& 4(6LB\sqrt{d}\Delta^{-1})(6+3R)\log(2(6LB\sqrt{d}\Delta^{-1})(6+3R)) \\ &+ 2(6LB\sqrt{d}\Delta^{-1})\sqrt{2\log(1/\delta)}R, \end{aligned}$$

which suggests that setting

$$\iota_1 = (24+18R)\log((72+54R)LB\sqrt{d}\Delta^{-1}) + \sqrt{8R^2\log(1/\delta)}$$

implies the fact that $\iota_1 \ge 2 + 4\sqrt{\iota_2} + R\sqrt{2\iota_3}$ □

D. Detailed Proof of Theorem 5.1

The first lemma shows that the contexts selected to l-th level are bounded independent from K

Lemma D.1 (Restatement of Lemma 5.5)**.** Set $\lambda = B^{-2}$. For any $k \in [K]$ and $l > 0$, $|\mathcal{C}_k^l| \le 16d4^l\iota_1(l)$ where $\iota_1(l) = \log\left(3LB2^l\right)$.

Proof. The proof is similar to the proof of Lemma B.1 by repalcing $\Gamma = 2^{-l}$. □

The next lemma provides a fluctuation control as well as the concentration in the ridge regression

Lemma D.2 (Restatement of Lemma 5.6)**.** Set $\lambda = B^{-2}$. For any level $l > 0$, for any $\delta > 0$, with probability at least $1-\delta$, for all $k \in [K]$, the estimation error is bounded by

$$\left|\mathbf{x}^\top(\boldsymbol{\theta}_k^l - \boldsymbol{\theta}^*)\right| \le \left(1 + R\sqrt{2d\iota_2(l)} + \zeta\sqrt{\left|\mathcal{C}_k^l\right|}\right)\|\mathbf{x}\|_{(\mathbf{U}_k^l)^{-1}},$$

for all $\mathbf{x}$ such that $\|\mathbf{x}\|_2 \le L$, where $\iota_2(l) = \log((d + |\mathcal{C}_k^l|L^2B^2)/(d\delta))$.

Proof. The proof is similar to the proof of Lemma B.2 □

Combining Lemma D.1 and Lemma D.2, we have the following corollary.

Corollary D.3. Set $\lambda = B^{-2}$. For any $\delta > 0$, with probability at least $1-\delta$, for all round $k \in [K]$ and any level $l > 0$, the prediction error is bounded by

$$\left|\mathbf{x}^\top(\boldsymbol{\theta}_k^l - \boldsymbol{\theta}^*)\right| \le \left(\beta(l) + 4\zeta 2^l\sqrt{d\iota_1(l)}\right)\|\mathbf{x}\|_{(\mathbf{U}_k^l)^{-1}},$$

for all $\mathbf{x}$ such that $\|\mathbf{x}\|_2 \le L$, where $\beta(l) = 1 + R\sqrt{2d\iota_2(l)}$, $\iota_2(l) = \log((d2^l + 16L^2B^2 8^l\iota_1(l))/(d\delta))$, and $\iota_1(l) = \log\left(3LB2^l\right)$.

Proof. The proof is simply by plugging the result in Lemma D.1 into Lemma D.2 and replacing the δ with $\delta/2^l$. By the union bound over $l \in \mathbb{N}^+$ and the fact that $\sum_{l=1}^\infty \delta/2^l = \delta$ yields the claimed result. □

Now, we are about to control $\mathcal{D}_k^l$, which means here we only consider the case where $\|\mathbf{x}\|_{(\mathbf{U}_k^l)^{-1}} \le 2^{-l}$ for all $\mathbf{x} \in \mathcal{D}_k^l$ and assuming the high-probability event in previous subsection always holds. The following lemma suggests that the decision set always keeps a nearly optimal action $\mathbf{x}_k^{l,*}$. Let $\mathcal{G}_K$ be the event that the high probability statement in Corollary D.3 holds.

Lemma D.4 (Formal statement of Lemma 5.7)**.** For any level $l > 0$, assume event $\mathcal{G}_K$ holds, then there exists $\mathbf{x}_k^{l,*} \in \mathcal{D}_k^l$, $r(\mathbf{x}_k^*) - r(\mathbf{x}_k^{l,*}) \le 2(l-1)\zeta\left(1 + 4\sqrt{d\iota_1(l)}\right)$ where $\iota_1(l) = \log\left(3LB2^l\right)$.

Proof. We would prove the statement by induction. Since $\mathcal{D}_k^1 = \mathcal{D}_k$, we have $\mathbf{x}_k^* \in \mathcal{D}_k^1$ and thus the induction basis holds according to $r(\mathbf{x}_k^*) - r(\mathbf{x}_k^{l,*}) = 0$. Now we assume the statement holds for level l, that is, there exists $\mathbf{x}_k^{l,*} \in \mathcal{D}_k^l$ such that $\mathbf{x}_k^{l,*} \in \mathcal{D}_k^l$, $r(\mathbf{x}_k^*) - r(\mathbf{x}_k^{l,*}) \le 2(l-1)\zeta\left(1 + 4\sqrt{d\iota_1(l)}\right)$.

If $\mathbf{x}_k^{l,*} \in \mathcal{D}_k^{l+1}$, then the desired statement directly holds by choosing $\mathbf{x}_k^{l,*} = \mathbf{x}_k^{l-1,*}$. Otherwise $\mathbf{x}_k^{l,*}$ is eliminated by some action $\mathbf{x}_k^{l+1,*} \in \mathcal{D}_k^l$ that $r_k^l(\mathbf{x}_k^{l+1,*}) \ge r_k^l(\mathbf{x}_k^{l,*}) + 2\beta(l)2^{-l}$. Moreover, from the definition of estimator $r_k^l(\cdot)$, we have

$$r_k^l(\mathbf{x}_k^{l+1,*}) - r(\mathbf{x}_k^{l+1,*}) \le \zeta + \left\langle \mathbf{x}_k^{l+1,*}, \theta_k^l - \theta^* \right\rangle + \beta(l)\left\|\mathbf{x}_k^{l+1,*}\right\|_{(\mathbf{U}_k^l)^{-1}} \tag{D.1}$$

and

$$r(\mathbf{x}_k^{l,*}) - r_k^l(\mathbf{x}_k^{l,*}) \le \zeta - \left\langle \mathbf{x}_k^{l,*}, \theta_k^l - \theta^* \right\rangle - \beta(l)\left\|\mathbf{x}_k^{l,*}\right\|_{(\mathbf{U}_k^l)^{-1}}. \tag{D.2}$$

Combining (D.1) and (D.2) and the fact that $r_k^l(\mathbf{x}_k^{l+1,*}) \ge r_k^l(\mathbf{x}_k^{l,*}) + 3\beta(l)2^{-l}$ gives that

$$\begin{aligned}
r(\mathbf{x}_k^{l,*}) - r(\mathbf{x}_k^{l+1,*}) &\le -3\beta(l)2^{-l} + 2\zeta + \left\langle \mathbf{x}_k^{l+1,*} - \mathbf{x}_k^{l,*}, \theta_k^l - \theta^* \right\rangle - \beta(l)\left\|\mathbf{x}_k^{l+1,*}\right\|_{(\mathbf{U}_k^l)^{-1}} + \beta(l)\left\|\mathbf{x}_k^{l,*}\right\|_{(\mathbf{U}_k^l)^{-1}} \\
&\le -3\beta(l)2^{-l} + 2\zeta + 2\cdot 2^{-l}\left(\beta(l) + 4\zeta 2^l\sqrt{d\iota_1(l)}\right) + \beta(l)2^{-l} \\
&\le 2\zeta\left(1 + 4\sqrt{d\iota_1(l)}\right),
\end{aligned}$$

where the second inequality is suggested by Corollary D.3 and $\|\mathbf{x}\|_{(\mathbf{U}_k^l)^{-1}} \le 2^{-l}$ for all $\mathbf{x} \in \mathcal{D}_k^l$. The desired statement can then be reached using the induction hypothesis. □

Then, the following lemma suggests that the performance of the actions in the decision set is guaranteed.

Lemma D.5 (Formal statement of Lemma 5.8)**.** For any level $l > 0$, assume event $\mathcal{G}_K$ holds, then for any action $\mathbf{x} \in \mathcal{D}_k^l$, $r(\mathbf{x}_k^*) - r(\mathbf{x}) \le 6\beta(l)2^{-l} + 2l\zeta\left(1 + 4\sqrt{d\iota_1(l)}\right)$ where $\iota_1(l) = \log\left(3LB2^l\right)$.

Proof. Let $\mathbf{x}_k^{l,*} \in \mathcal{D}_k^l$ be the optimal action given in Lemma D.4. According to the elimination process, for any action $\mathbf{x} \in \mathcal{D}_k^l$, it holds that $r_k^l(\mathbf{x}) \ge r_k^l(\mathbf{x}_k^{l,*}) - 3\beta(l)2^{-l}$. Moreover, from the definition of estimator $r_k^l(\cdot)$, we have

$$r_k^l(\mathbf{x}) - r(\mathbf{x}) \le \zeta + \left\langle \mathbf{x}, \theta_k^l - \theta^* \right\rangle + \beta(l)\|\mathbf{x}\|_{(\mathbf{U}_k^l)^{-1}}$$

and

$$r(\mathbf{x}_k^{l,*}) - r_k^l(\mathbf{x}_k^{l,*}) \leq \zeta - \left\langle \mathbf{x}_k^{l,*}, \theta_k^l - \theta^* \right\rangle - \beta(l) \left\| \mathbf{x}_k^{l,*} \right\|_{(\mathbf{U}_k^l)^{-1}}.$$

Combining the above three inequalities give

$$\begin{aligned} r(\mathbf{x}_k^{l,*}) - r(\mathbf{x}) &\leq 3\beta(l)2^{-l} + 2\zeta + 2^{-l} + \left\langle \mathbf{x} - \mathbf{x}_k^{l,*}, \theta_k^l - \theta^* \right\rangle - \beta(l) \left\| \mathbf{x}_k^{l,*} \right\|_{(\mathbf{U}_k^l)^{-1}} + \beta(l) \left\| \mathbf{x}_k^{l-1,*} \right\|_{(\mathbf{U}_k^l)^{-1}} \\ &\leq 3\beta(l)2^{-l} + 2\zeta + 2 \cdot 2^{-l} \left(\beta(l) + 4\zeta 2^l \sqrt{d\iota_1(l)} \right) + \beta(l)2^{-l} \\ &\leq 6\beta(l)2^{-l} + 2\zeta \left(1 + 4\sqrt{d\iota_1(l)} \right), \end{aligned}$$

where the second inequality is suggested by Corollary D.3 and $\|\mathbf{x}\|_{(\mathbf{U}_k^l)^{-1}} \leq 2^{-l}$ for all $\mathbf{x} \in \mathcal{D}_k^l$. The desired statement can then be reached by combining Lemma D.4. □

Proof of Theorem 5.1. Consider the case that event $\mathcal{G}_K$ holds. Let l_Δ be the smallest integer solution to $l_\Delta > \log(8\beta(l_\Delta)\Delta^{-1})$. Note this relation ensures $4\beta(l_\Delta)2^{-l_\Delta} < \Delta/2$. In case that the misspecification level is bounded by $2l_\Delta\zeta \left(1 + 4\sqrt{d\iota_1(l_\Delta)}\right) < \Delta/2$, it holds that $6\beta(l_\Delta)2^{-l_\Delta} + 2l_\Delta\zeta \left(1 + 4\sqrt{d\iota_1(l_\Delta)}\right) < \Delta$. According to Lemma D.5, it satisfies that

$$r(\mathbf{x}_k^*) - r(\mathbf{x}) \leq 6\beta(l_\Delta)2^{-l_\Delta} + 2l_\Delta\zeta \left(1 + 4\sqrt{d\iota_1(l_\Delta)}\right)$$

for any $\mathbf{x} \in \mathcal{D}_k^{l_\Delta}$. According to the process of arm elimination, we have $\mathcal{D}_k^l \subseteq \mathcal{D}_k^{l_\Delta}$ for any $l \geq l_\Delta$. Thus, it holds that $r(\mathbf{x}_k^*) - r(\mathbf{x}) < \Delta$ for any $\mathbf{x} \in \mathcal{D}_k^l, l \geq l_\Delta$. Note that according to the definition of Δ, we have $r(\mathbf{x}_k^*) - r(\mathbf{x}) > \Delta$ for all $\mathbf{x} \in \mathcal{D}_k^l$ that $r(\mathbf{x}_k^*) \neq r(\mathbf{x})$. These two statements together restrict $r(\mathbf{x}_k^*) = r(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{D}_k^l$ on every $l > l_\Delta$, that is, any action that remains in the decision sets on higher levels are optimal. Let $\mathcal{U}_K^l$ be the set of index k that action $\mathbf{x}_k$ is chosen from layer l. We have $|\mathcal{U}_K^l| \leq |\mathcal{C}_K^l| + 4^l d$. Thus, we could decompose the total regret by

$$\begin{aligned} \mathrm{Regret}(K) &= \sum_{l \geq 1} \sum_{k \in \mathcal{U}_K^l} (r(\mathbf{x}_k^*) - r(\mathbf{x})) = \sum_{l=1}^{l_\Delta - 1} \sum_{k \in \mathcal{U}_K^l} (r(\mathbf{x}_k^*) - r(\mathbf{x})) \\ &\leq \sum_{l=1}^{l_\Delta - 1} (|\mathcal{C}_K^l| + 4^l d) \cdot \left(6\beta(l)2^{-l} + 2l\zeta \left(1 + 4\sqrt{d\iota_1(l)}\right) \right) \\ &\leq \sum_{l=1}^{l_\Delta - 1} 16d4^l\iota_1(l) \cdot \left(6\beta(l)2^{-l} + 2l\zeta \left(1 + 4\sqrt{d\iota_1(l)}\right) \right) \\ &\leq 96d \sum_{l=1}^{l_\Delta - 1} \beta(l)2^l\iota_1(l) + 32d\zeta \sum_{l=1}^{l_\Delta - 1} l4^l\iota_1(l) \left(1 + 4\sqrt{d\iota_1(l)}\right) \\ &\leq 96d\beta(l_\Delta)2^{l_\Delta}\iota_1(l_\Delta) + 32dl_\Delta 4^{l_\Delta}\iota_1(l_\Delta)\zeta \left(1 + 4\sqrt{d\iota_1(l_\Delta)}\right) \\ &\leq 1536d\beta^2(l_\Delta)\iota_1(l_\Delta)/\Delta + 8192d\beta^2(l_\Delta)\iota_1(l_\Delta)/\Delta \\ &\leq 2^{14}d\beta^2(l_\Delta)\iota_1(l_\Delta)/\Delta \end{aligned}$$

where the second equality is given by Lemma D.5, the second inequality is given by Lemma D.1, the third last inequality holds since $\beta(\cdot)$ and $\iota_1(\cdot)$ are monotone increase and the second inequality since $2^{l_\Delta - 1} \leq 8\beta(l_\Delta - 1)\Delta^{-1} \leq 8\beta(l_\Delta)\Delta^{-1}$ and $2l_\Delta\zeta \left(1 + 4\sqrt{d\iota_1(l_\Delta)}\right) < \Delta/2$.

□

E. Proof of Theorem 6.1

To begin with, we introduce the lemma providing a sparse vector set in $\mathbb{R}^d$.

Lemma E.1 (Lemma 3.1, Lattimore et al. 2020)**.** For any $\varepsilon > 0$ and $d < [|\mathcal{D}|]$ such that $d \geq \lceil 8\log(|\mathcal{D}|)\varepsilon^{-2}\rceil$, there exists a vector set $\mathcal{D} \subset \mathbb{R}^d$ such that $\|\mathbf{x}\|_2 = 1$ for all $\mathbf{x} \in \mathcal{D}$ and $|\langle \mathbf{x}, \mathbf{y}\rangle| \leq \varepsilon$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{D}$ and $\mathbf{x} \neq \mathbf{y}$.

Next, we present the Bretagnolle–Huber inequality providing the lower bound to distinguish a system.

Lemma E.2 (Bretagnolle–Huber inequality)**.** Let P and Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$, let $\mathcal{A} \in \mathcal{F}$ be an arbitary event. Then

$$P(\mathcal{A}) + Q(\mathcal{A}^c) \geq \frac{1}{2}\exp(-\mathrm{KL}(P, Q)).$$

For stochastic linear bandit problem with finite arm, we can denote $T_i(k)$ as the number of rounds the algorithm visit the i-th arm over total k rounds. Then We have the KL-divergence decomposition lemma.

Lemma E.3 (Lemma 15.1, Lattimore & Szepesvári (2020))**.** Let $\nu = (P_1, \cdots, P_n)$ be the reward distributions associated with one n-armed bandit and let $\nu' = (P_1', \cdots, P_n')$ be another n-armed bandit. Fix some algorithm π and let $\mathbb{P}_\nu = \mathbb{P}_{\nu\pi}, \mathbb{P}_{\nu'} = \mathbb{P}_{\nu',\pi}$ be the probability measures on the canonical bandit model induced by the k-round interconnection of π and ν (respectively, π and ν'). Then $\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{i=1}^n \mathbb{E}_\nu[T_i(n)]\mathrm{KL}(P_i, P_i')$

Proof of Theorem 6.1. The proof starts from inheriting the idea from Lattimore et al. (2020). Given dimension d and the number of arms $|\mathcal{D}|$, setting $\varepsilon = \sqrt{8\log(|\mathcal{D}|)/(d-1)}$, we can provide the contextual vector set $\mathcal{D}$ such that

$$\|\mathbf{x}\|_2 = 1, \forall \mathbf{x} \in \mathcal{D}, |\langle \mathbf{x}, \mathbf{y}\rangle| \leq \sqrt{\frac{8\log(|\mathcal{D}|)}{d-1}}, \forall \mathbf{x}, \mathbf{y} \in \mathcal{D}, \mathbf{x} \neq \mathbf{y},$$

For simplicity, we index the decision set as $\mathbf{x}_1, \cdots, \mathbf{x}_{|\mathcal{D}|}$. Given the minimal sub-optimality gap Δ, we provide the parameter set $\boldsymbol{\Theta}$ as follows:

$$\boldsymbol{\Theta} = \left\{\boldsymbol{\theta}_{(i,j)} = \Delta\mathbf{x}_i + 2\Delta\mathbf{x}_j, \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}, i \neq j\right\} \bigcup \{\boldsymbol{\theta}_i = \Delta\mathbf{x}_i, \mathbf{x}_i \in \mathcal{D}\}.$$

It can be verified that $\boldsymbol{\Theta}$ contains two kinds of $\boldsymbol{\theta}$. The first one $\boldsymbol{\theta}_{(i,j)}$ is a mixture of two different contexts $\mathbf{x}_i, \mathbf{x}_j$ with different strength Δ and 2Δ. The second one is $\boldsymbol{\theta}_i$ which only contains features from one context $\mathbf{x}_i$. We can further verify that the size of $|\boldsymbol{\Theta}| = |\mathcal{D}|^2$ and $\|\boldsymbol{\theta}\|_2 \leq \sqrt{5}\Delta$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. For different parameter $\boldsymbol{\theta}$, the reward function is sampled from a Gaussian distribution $\mathcal{N}(r_{\boldsymbol{\theta}}(\mathbf{x}), 1)$, where the expected reward function is defined as

$$r_{\boldsymbol{\theta}_{(i,j)}}(\mathbf{x}) = \begin{cases} 2\Delta \text{ if } \mathbf{x} = \mathbf{x}_j \\ \Delta \text{ if } \mathbf{x} = \mathbf{x}_i \\ 0 \text{ otherwise} \end{cases}, r_{\boldsymbol{\theta}_i}(\mathbf{x}) = \begin{cases} \Delta \text{ if } \mathbf{x} = \mathbf{x}_i \\ 0 \text{ otherwise} \end{cases}.$$

We can verify that the minimal sub-optimality of all these bandit problem is Δ. For different parameter $\boldsymbol{\theta}$ and input $\mathbf{x}$, by utilizing the sparsity of the set $\mathcal{D}$ (i.e. $|\mathbf{x}^\top y| \leq \varepsilon$ if $\mathbf{x} \neq \mathbf{y}$), we can verify the misspecification level as

$$|r_{\boldsymbol{\theta}_{(i,j)}}(\mathbf{x}) - \boldsymbol{\theta}_{(i,j)}^\top \mathbf{x}| = \begin{cases} |2\Delta - 2\Delta\mathbf{x}_j^\top\mathbf{x} - \Delta\mathbf{x}_i^\top\mathbf{x}| \leq \Delta\varepsilon \text{ if } \mathbf{x} = \mathbf{x}_j \\ |\Delta - 2\Delta\mathbf{x}_j^\top\mathbf{x} - \Delta\mathbf{x}_i^\top\mathbf{x}| \leq 2\Delta\epsilon \text{ if } \mathbf{x} = \mathbf{x}_i \\ |0 - 2\Delta\mathbf{x}_j^\top\mathbf{x} - \Delta\mathbf{x}_i^\top\mathbf{x}| \leq 3\Delta\varepsilon \text{ otherwise} \end{cases}$$

$$|r_{\boldsymbol{\theta}_i}(\mathbf{x}) - \boldsymbol{\theta}_i^\top(\mathbf{x})| = \begin{cases} |\Delta - \Delta\mathbf{x}_i^\top\mathbf{x}| = 0 \text{ if } \mathbf{x} = \mathbf{x}_i \\ |0 - \Delta\mathbf{x}_i^\top\mathbf{x}| \leq \Delta\varepsilon \text{ otherwise.} \end{cases}$$

Therefore we have verified that the misspecification level is bounded by $\zeta = 3\Delta\varepsilon$.

The provided bandit structure is hard for any linear algorithm to learn since any algorithm cannot get any information before it encounters non-zero expected rewards, even regardless of the noise of the rewards. We following the same method in Lattimore & Szepesvári (2020). If the algorithm choose arm i at the first round, there would be $|\mathcal{D}|$ parameters (i.e. $\boldsymbol{\theta}_i, \boldsymbol{\theta}_{(i,\cdot)}$ receiving a non-zero expected reward. On the second round if the algorithm choose a different arm j, there would

be $|\mathcal{D}|$ parameters (i.e. $\boldsymbol{\theta}_j, \boldsymbol{\theta}_{(j,k:k\neq i)}$ receiving a non-zero expected reward. Therefore the average time of receiving zero expected reward should be

$$\begin{aligned}
|\mathcal{D}|^{-2}\sum_{i=1}^{|\mathcal{D}|}(i-1)(|\mathcal{D}|-i+1) &= |\mathcal{D}|^{-2}\sum_{i=0}^{|\mathcal{D}|-1} i(|\mathcal{D}|-i) \\
&= |\mathcal{D}|^{-2}\left(|\mathcal{D}|\sum_{i=0}^{|\mathcal{D}|-1} i - \sum_{i=0}^{|\mathcal{D}|-1} i^2\right) \\
&= |\mathcal{D}|^{-2}\left(\frac{|\mathcal{D}|^2(|\mathcal{D}|-1)}{2} - \frac{|\mathcal{D}|(|\mathcal{D}|-1)(2|\mathcal{D}|-1)}{6}\right) \\
&= \frac{|\mathcal{D}|-1}{2}\left(1-\frac{2|\mathcal{D}|-1}{3|\mathcal{D}|}\right) \\
&\geq \frac{|\mathcal{D}|-1}{6},
\end{aligned}$$

where the third equation is from the fact that $\sum_{i=1}^n i = n(n+1)/2$ and $\sum_{i=1}^n i^2 = n(n+1)(2n+1)/6$. The last inequality is from the fact that $2|\mathcal{D}|-1)/(3|\mathcal{D}|) \leq 2/3$. Therefore, even without of the random noise, any algorithm is expected to receive $\min\{K, (|\mathcal{D}|-1)/6\}$ uninformative data with expected reward to be zero. Therefore any algorithm will receive a $\Delta\min\{K, (|\mathcal{D}|-1)/6\}$ regret considers the suboptimality as Δ.

Next, we consider the effect of random noise. For any algorithm running on this parameter set $\boldsymbol{\Theta}$, we find two parameter $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_{i,j}$ where $j \neq i$. Define the event as $\mathcal{A} = \{T_j(k) \geq k/2\}$ and $\mathcal{A}^c = \{T_j(k) < k/2\}$. By Lemma E.2 and Lemma E.3,

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\theta}_i}\left(T_j(k) \geq \frac{k}{2}\right) + \mathbb{P}_{\boldsymbol{\theta}_{(i,j)}}\left(T_j(k) < \frac{k}{2}\right) &\geq \frac{1}{2}\exp(-\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}_i}, \mathbb{P}_{\boldsymbol{\theta}_{(i,j)}})) \\
&\geq \frac{1}{2}\exp\left(-\sum_{n\in\mathcal{D}}\mathbb{E}_{\boldsymbol{\theta}_i}[T_n(k)]\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\theta}_{(i,j)},n}, \mathbb{P}_{\boldsymbol{\theta}_j,n}\right)\right). \qquad \text{(E.1)}
\end{aligned}$$

Noticing the minimal sub-optimality gap is Δ. Also the j-th arm is the sub-optimal arm for parameter $\boldsymbol{\theta}_i$. Therefore, once $T_j(k) \geq k/2$, the algorithm will at least suffer from $\Delta k/2$ regret for parameter $\boldsymbol{\theta}_i$. Also, since the j-th arm is the optimal arm for bandit $\boldsymbol{\theta}_{(i,j)}$. If $T_j(k) < k/2$, the algorithm will also at least suffer from $\Delta k/2$ regret for $\boldsymbol{\theta}_{(i,j)}$. Denoting $\mathcal{R}_{\boldsymbol{\theta}}(k)$ as the expected cumulative regret over k rounds, that is to say

$$\mathcal{R}_{\boldsymbol{\theta}_i}(k) \geq \frac{\Delta k}{2}\mathbb{P}_{\boldsymbol{\theta}_i}(T_j(k) \geq k/2) \quad \mathcal{R}_{\boldsymbol{\theta}_j}(k) \geq \frac{\Delta k}{2}\mathbb{P}_{\boldsymbol{\theta}_i}(T_j(k) < k/2). \qquad \text{(E.2)}$$

On the other hand since the bandit using $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ only differ in the j-th arm. Since standard Gaussian noise is adapted, $\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}_i,n}, \mathbb{P}_{\boldsymbol{\theta}_{(i,j)},n}) = \Delta^2\,\mathbb{1}[n=j]/2$. Combining this with (E.2), (E.1) suggests that

$$\mathcal{R}_{\boldsymbol{\theta}_i}(k) + \mathcal{R}_{\boldsymbol{\theta}_j}(k) \geq \frac{\Delta k}{2}\exp\left(-\frac{\Delta^2}{2}\mathbb{E}_{\boldsymbol{\theta}_i}\left[T_j(k)\right]\right),$$

which suggests that

$$\mathbb{E}_{\boldsymbol{\theta}_i}\left[T_j(k)\right] \geq \frac{\log(\Delta k) - \log 2 - \log(\mathcal{R}_{\boldsymbol{\theta}_i}(k) + \mathcal{R}_{\boldsymbol{\theta}_j}(k))}{\Delta^2/2}, \qquad \text{(E.3)}$$

For any algorithm seeking to get a sublinear expected regret bound of $\mathcal{R}_{\boldsymbol{\theta}}(k) \leq Ck^\alpha$ with $C > 0, 0 \leq \alpha < 1$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, (E.3) becomes

$$\mathbb{E}_{\boldsymbol{\theta}_i}\left[T_j(k)\right] \geq \frac{\log(\Delta k) - \log 2 - \log(2Ck^\alpha)}{\Delta^2/2} = \frac{\log(\Delta k) - \log(4C) - \alpha\log k}{\Delta^2/2}. \qquad \text{(E.4)}$$

Since that the regret on $\boldsymbol{\theta}_i$ can be decomposed by

$$\mathcal{R}_{\boldsymbol{\theta}_i}(k) = \Delta \sum_{n=1, n\neq i}^{|\mathcal{D}|} T_n(k), \tag{E.5}$$

combining (E.5) with (E.4) yields

$$\mathcal{R}_{\boldsymbol{\theta}_i}(k) \geq \frac{2(|\mathcal{D}|-1)}{\Delta} \max\left\{\log(\Delta k) - \log(4C) - \alpha \log k, 0\right\},$$

where the $\max$ operator is trivially taken for $\mathcal{R}_{\boldsymbol{\theta}}(k) \geq 0$.

□