Breaking the $\sqrt{T}$ Barrier: Instance-Independent Logarithmic Regret in Stochastic Contextual Linear Bandits

Avishek Ghosh [1] **Abishek Sankararaman** [2]

Abstract

We prove an instance independent (poly) logarithmic regret for stochastic contextual bandits with linear payoff. Previously, in (Chu et al., 2011), a lower bound of $\mathcal{O}(\sqrt{T})$ is shown for the contextual linear bandit problem with arbitrary (adversarily chosen) contexts. In this paper, we show that stochastic contexts indeed help to reduce the regret from $\sqrt{T}$ to $\mathsf{polylog}(T)$. We propose Low Regret Stochastic Contextual Bandits (LR-SCB), which takes advantage of the stochastic contexts and performs parameter estimation (in ℓ_2 norm) and regret minimization simultaneously. LR-SCB works in epochs, where the parameter estimation of the previous epoch is used to reduce the regret of the current epoch. The (poly) logarithmic regret of LR-SCB stems from two crucial facts: (a) the application of a norm adaptive algorithm to exploit the parameter estimation and (b) an analysis of the shifted linear contextual bandit algorithm, showing that shifting results in increasing regret. We have also shown experimentally that stochastic contexts indeed incurs a regret that scales with $\mathsf{polylog}(T)$.

1. INTRODUCTION

Contextual bandits are sequential decision making systems, where a learner is typically equipped with K actions (also called "arms"). At each round $t \in [T]$ the learner picks an action in the presence of contextual side information. Algorithms for these class of problems typically employ a decision rule that maps the context information to the action chosen. The goal of the learner is to maximize the reward (or in other words, minimize the regret with respect to the best mapping in the hindsight). Contextual bandit paradigm is typically used in advertisement placement (Li et al., 2010), clinical trials (Tewari & Murphy, 2017) and recommendation systems (Agarwal et al., 2016).

The problem of contextual bandits with linear payoffs has a rich body of existing literature. This framework was introduced by (Abe et al., 2003; Auer, 2002) and further developed in (Li et al., 2010; Chu et al., 2011). The framework of linear payoff—although simple, is expressive enough to capture several practical real world problems, as explained in (Abe et al., 2003; Li et al., 2010). In particular, (Chu et al., 2011) proposes a learning algorithm based on the UCB based optimistic idea. The resulting algorithm, namely SupLinUCB considers arbitrary contexts (i.e., contexts are generated by an adversary) and obtains a high probability regret of $\mathcal{O}(\sqrt{dT\log^3(KT)})$, where d is the dimension of the contexts. In the same paper, it is shown that if the contexts are adversarially generated, any contextual bandit algorithm with linear payoff will incur $\Omega(\sqrt{dT})$ regret. Moreover, several variants of contextual bandits are also studied, for example, in supervised learning (Beygelzimer et al., 2011), balanced exploration (Dimakopoulou et al., 2019) and in delayed systems (Zhou et al., 2019).

The contextual bandit paradigm has also been investigated beyond linear rewards. As an instance, (Agarwal et al., 2012) and (Agarwal et al., 2014) consider the K-armed generic contextual bandit system and analyzes a regressor elimination type and projection smoothing based learning algorithms respectively, which attains a regret guarantee of $\tilde{\mathcal{O}}(\sqrt{KT})$. These algorithms are computationally inefficient and depend on an oracle. Furthermore, (Foster & Rakhlin, 2020) converts the generic contextual bandit problem to an online regression problem, and obtains similar regret. Recently, (Simchi-Levi & Xu, 2021) proposes a learning algorithm, namely FALCON, that obtains $\tilde{\mathcal{O}}(\sqrt{KT})$ regret in the presence of an offline regression oracle. Moreover, (Zhou et al., 2020) proposes a neural net based learning for contextual bandits.

In this paper, we stick to the framework of stochastic contextual bandits with linear payoff, and ask the following

[1]Halıcıoğlu Data Science Institute (HDSI), UC San Diego, USA [2]Dept. of Electrical Engg. and Computer Sciences, UC Berkeley, USA (Abishek is currently with Amazon AWS AI, Palo Alto, USA but work done outside the scope of Amazon). Correspondence to: Avishek Ghosh <a2ghosh@ucsd.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

"Can (structured) stochastic contexts help in reducing the regret of linear contextual bandits?"

It turns out, the answer to this question is an astounding *yes*. In fact, if the stochastic contexts satisfy a few regularity conditions, it is possible to break the $\Omega(\sqrt{T})$ regret barrier of (Chu et al., 2011), and obtain an instance-independent regret of $\mathcal{O}(\text{polylog}\, T)$. We crucially exploit the stochasticity of the contexts. The regularity conditions we impose (formally written in equation 1) enable us to do statistical estimation (inference) and regret minimization simultaneously.

We emphasize that bandits with stochastic contexts are also studied quite extensively for contextual linear bandits; for example (Gentile et al., 2014) uses it for clustering in multi-agent systems, (Chatterji et al., 2020) uses it for binary model selection between linear and standard multi-armed bandits, (Ghosh et al., 2021b) uses it for model selection and (Ghosh et al., 2021c) uses it for collaboration and personalization in multi-agent systems. Furthermore, for generic contextual bandit problems beyond linear payoffs, the assumption of stochastic contexts is quite common (see (Agarwal et al., 2014; 2012; Simchi-Levi & Xu, 2021)).

In this work, we propose an epoch based learning algorithm, namely Low Regret Stochastic Contextual Bandits (LR-SCB). In Theorem 5.1, we show that the (instance independent) regret of our proposed algorithm scales as[1] $\mathcal{O}(\text{polylog}(T))$. We leverage the concurrent inference and regret minimization aspect to obtain poly-logarithmic regret. Note that previously, in (Gentile et al., 2014; Chatterji et al., 2020; Ghosh et al., 2021c), this simultaneous estimation and regret minimization condition is used to perform additional tasks (on top of regret minimization) such as *clustering, model selection and personalization*.

In LR-SCB, we break the learning horizon into epochs of increasing length. At each epoch, we simultaneously minimize regret and form an estimate of the underlying parameter. Let us assume the underlying parameter for the linear contextual bandit is θ^*. In the first epoch, we play the standard contextual bandit algorithm, OFUL of (Chatterji et al., 2020)[2] with stochastic contexts and learn an estimate $\widehat{\theta}$ of θ^*. Subsequently, in the next epoch, we modify the reward of the learning algorithm in a specific way, such that underlying parameter we need to learn is $\theta^* - \widehat{\theta}$. Hence, the sifted parameter will learn will have a small norm, i,e., $\|\theta^* - \widehat{\theta}\|$ is small, since $\widehat{\theta}$ is an estimate of θ^*. In order to exploit this, we use the norm adaptive algorithm, ALB-norm of (Ghosh et al., 2021b), which gives regret proportional to the parameter norm. Note that, owing to the proper shift, the norm of the shifted parameter is small, which in turn results in a small regret. We keep on doing this over multiple epochs, and shift the underlying parameter accordingly. With an appropriate choice of epoch lengths, it turns out that this phase based algorithm attains a regret of $\mathcal{O}(\text{polylog}(T))$.

[1] We have a worse dependence on the context dimension d.

[2] In fact, we play a variation of the OFUL algorithm, see Section 4. For completeness, we reproduce this in Algorithm 2.

1.1. Our Contributions:

1.1.1. Algorithmic

We propose an epoch based learning algorithm for stochastic contextual bandits. Our algorithm, LR-SCB introduces proper shifts to the underlying unknown parameters, and uses a norm adaptive algorithm, ALB-norm repeatedly over epochs. We obtain an instance independent $\text{polylog}(T)$ regret for the stochastic contextual linear bandit, thus breaking the $\sqrt{T}$ barrier shown in (Chu et al., 2011). We show that stochastic contexts indeed help in reducing the regret. To the best of our knowledge, this is the first work to show a (poly) logarithmic instance independent regret for stochastic contextual bandits.

1.1.2. Technical novelty

A key technical challenge we encounter is the characterization of ALB-norm under shifts. We argue in Appendix A that it is sufficient to understand the behavior of the shifted OFUL system, and in Section 6 as well as in Appendix D, we rigorously analyze the shifted OFUL (which might also be of independent interest). For this, we derive an anti-concentration property for the contexts, and in conjunction with independence, we show that OFUL is indeed robust to shifts, and shifting can only increase the regret.

Furthermore, we also require ALB-norm to yield parameter estimation guarantee, similar to OFUL, and in Appendix C, we show that indeed, ALB-norm outputs the required guarantees.

1.1.3. Experiments

We validate our theoretical findings via experiments. In particular for different context dimension, we characterize the regret of LR-SCB with respect to $\log T$, and compare it with OFUL as a baseline. We observe that LR-SCB outperforms OFUL in terms of regret. Furthermore, to understand the regret scaling of LR-SCB better, we plot log regret with respect to $\log \log T$, and obtain a straight line with slope around 2. This implies that the regret of LR-SCB is indeed $\text{polylog}(T)$, which confirms our theoretical result.

2. RELATED WORK

Contextual Bandits: The literature on contextual bandits is quite rich, starting from (Auer, 2002; Abe et al., 2003). Around 2010, with the motivation of recommendation, the

study of contextual bandits got some momentum with seminal papers like (Li et al., 2010; Chu et al., 2011). Most of these papers assume arbitrary, adversarially generated contexts and obtain regret rates of $\mathcal{O}(\sqrt{T})$. Furthermore, several variants of contextual bandits is studied in the literature, for example, in delayed systems (Zhou et al., 2019) and in supervised learning.

Apart from this linear contextual bandits, there has been a significant effort to understand the generic contextual bandits (Agarwal et al., 2012; 2016). Most of these algorithms are non-implementable and very recently (Foster & Rakhlin, 2020; Simchi-Levi & Xu, 2021) proposes a reduction of the generic contextual bandit problem to an online and offline regression respectively. Very recently, stochastic contexts are used in linear contextual bandits, for example (Chatterji et al., 2020; Gentile et al., 2014). The regret guarantee for these algorithms also scale with $\mathcal{O}(\sqrt{T})$. On the other hand, in this work we exploit the stochastic contexts to simultaneously estimate and minimize regret and as a result, we obtain a regret of $\mathsf{polylog}(T)$, thus breaking the $\sqrt{T}$ barrier.

Adaptive Bandit Algorithms: As explained in Section 1, the use of an adaptive algorithm that exploits the small norm enables our learning algorithm to obtain logarithmic regret. Adaptive algorithms in bandits have gained a lot of interest in the recent years, for example in (Ghosh et al., 2021b), the authors define parameter norm and sparsity as complexity parameters for stochastic linear bandit and adapt to those without any apriori knowledge. (Foster et al., 2019) also adapts to the sparsity in a linear bandit problem, whereas (Pacchiano et al., 2020) uses the corrall framework of (Agarwal et al., 2017) to obtain adaptive algorithms for bandits and reinforcement learning. In the corralling framework, the base algorithms are treated as bandit arms, and a learning algorithm is played to choose the correct model. Very recently, the adaptation question is also addressed for generic contextual bandits (Krishnamurthy & Athey, 2021; Ghosh et al., 2021d). Apart from this, in reinforcement learning, a few recent works have started inquiring the question of adaptation, for example (Lee et al., 2021) in the framework of function approximation and (Ghosh et al., 2021a) for generic (but separable) reinforcement learning.

2.1. Notation

Throughout the paper, for positive integer r, the notation $[r]$ refers to the set $\{1, 2, \ldots, r\}$. We use the notation $\tilde{\mathcal{O}}(.)$ to ignore the logarithmic factors. Moreover, $\|.\|$ refers to the ℓ_2 norm unless otherwise specified. Further, $\mathsf{polylog}(.)$ captures only the poly-logarithmic dependencies. $C, C_1, \ldots, c, c_1, \ldots$ represents universal constants, the value of which may differ from instance to instance.

2.2. Organization

We state the problem of contextual linear bandit problem along with the structural assumptions on the stochastic contexts in Section 3. In Section 4, we propose our algorithm, namely `LR-SCB` and discuss its implications. Furthermore, in Section 5, we obtain regret guarantees for `LR-SCB` and discuss special cases, followed by a concise proof sketch. In Section 6, we analyze a the performance of a shifted linear bandit—an ingredient which was crucial for obtaining the main results. This analysis of shifted linear bandits may be of independent interest as well. Finally, in Section 7, we verify our theoretical findings via experiments. We conclude with a few open problems in Section 8.

3. PROBLEM SETUP

We consider the setup of stochastic contextual bandit with linear payoffs (Chu et al., 2011; Chatterji et al., 2020). At the beginning of each round $t \in [T]$, the learner chooses one of the K available arms, and gets a reward. To help the learner make the choice of the arm, at each round, the learner is handed K context vectors, d dimensional each, denoted by $\beta_t = [\beta_{1,t}, \ldots, \beta_{K,t}] \in^{d \times K}$. When the learner chooses arm i, the reward obtained is given by $\langle \beta_{i,t}, \theta^* \rangle + \xi_t$, where θ^* is the d-dimensional unknown parameter, with $\|\theta^*\| \leq 1$, and $\{\xi_t\}_{t=1}^T$ denote the noise.

Stochastic Assumptions: We assume that the contexts are stochastic, following the framework of (Chatterji et al., 2020; Ghosh et al., 2021b). We denote the sigma algebra generated by all noise random variables upto and including time $t-1$ by $\mathcal{F}_{t-1}$. Moreover, by $\mathbb{E}_{t-1}(.)$ and $\mathbb{V}_{t-1}(.)$, we denote the as the conditional expectation and conditional variance operators respectively with respect to $\mathcal{F}_{t-1}$. We further assume that the noise parameter, $(\xi_t)_{t \geq 1}$ are conditionally sub-Gaussian noise with known parameter σ, conditioned on all the arm choices and realized rewards in the system upto and including time $t-1$, and without loss of generality, let $\sigma = 1$ throughout.

The contexts $\{\beta_t\}_{t=1}^T$ are assumed to be bounded—in particular, we let the contexts be drawn from $[-c/\sqrt{d}, c/\sqrt{d}]^{\otimes d}$, where c is a universal constant and the $1/\sqrt{d}$ scaling is without loss of generality, so that the norm of the contexts are $\mathcal{O}(1)$. Moreover, the contexts $\beta_{i,t}$ are assumed to be drawn independent of the past and $\{\beta_{j,t}\}_{j \neq i}$, from a distribution satisfying

$$\mathbb{E}_{t-1}[\beta_{i,t}] = 0 \qquad \mathbb{E}_{t-1}[\beta_{i,t}\, \beta_{i,t}^\top] \succeq \rho_{\min} I. \tag{1}$$

Furthermore, for any fixed $z \in \mathbb{R}^d$, with unity norm, the random variable $(z^\top \beta_{i,t})^2$ is conditionally sub-Gaussian, for all i, with $\mathbb{V}_{t-1}[(z^\top \beta_{i,t})^2)] \leq 4\rho_{\min}$. This means that the conditional mean of the covariance matrix is zero and the conditional covariance matrix is positive definite with

minimum eigenvalue at least $\rho_{\min}$. Furthermore, the conditional variance bound assumption is for technical reasons and is crucially required to apply (1) for contexts of (random) bandit arms selected by our learning algorithm (see Lemma 1 of (Gentile et al., 2014)).

Note this this above set of assumptions on context vectors is not new and the exact set of assumptions were used in (Gentile et al., 2017; Chatterji et al., 2020; Ghosh et al., 2021c;b)[3]. In (Gentile et al., 2017), the authors introduced the above-mentioned set of assumptions and use them for parametric inference on top of regret minimization for online clustering problem with bandit information. (Chatterji et al., 2020) uses the same context assumptions for binary model selection between simple multi-armed and contextual linear bandits. Furthermore, (Ghosh et al., 2021b) uses the identical assumptions to obtain an adaptive problem complexity adaptive regret guarantees for linear bandits and (Ghosh et al., 2021c) uses these assumptions to ensure personalization for multi-agent linear bandits. Apart from the above mentioned papers, (Foster et al., 2019) uses similar assumptions for stochastic linear bandits and (Ghosh et al., 2021a) uses it for model selection in Reinforcement learning problems with function approximation. In all of the above papers, the authors need parametric inference in conjunction with regret minimization, which is a harder task. If the stochastic contexts are structured, these two tasks can be performed simultaneously. It turns out that the above-mentioned set of assumptions are sufficient to ensure this.

Example: Although we present here the technical conditions needed on contexts, this include simple examples as well. As an instance, it includes the simple setting where the contexts evolve according to a random process independent of the actions and rewards from the learning algorithm. Hence, any zero mean (full rank) iid random variables drawn from a (coordinate-wise) bounded space, generated exogenous to the actions of the agents can be taken as stochastic contexts. As an example, random vectors drawn in an i.i.d manner across rounds from $\mathsf{Unif}[-c_0/\sqrt{d}, c_0/\sqrt{d}]^{\otimes d}$ for a constant c_0. For this we have $\rho_{\min} = c_1/d$, where c_1 is a constant. In Section 5, we take this uniform distribution as a special case and completely characterize its perfromance.

Note that the above-mentioned framework of generating contexts are quite standard in the generic contextual bandit literature(Agarwal et al., 2012; 2014) as well, where at each round nature picks a context sampled i.i.d in each round from a fixed and known distribution.

Performance Metric: At time t, we denote $B_t \in [K]$ as the arm played by the agent. We want to compete with the optimal arm. Since we do not know θ^*, we are bound to incur some error characterized by an equivalent regret term. The regret, over a time horizon of T is given by

$$R_i(T) = \sum_{t=1}^{T} \max_{j\in[K]} \langle \beta_{j,t}, \theta^* \rangle - \langle \beta_{B_t,t}, \theta^* \rangle \qquad (2)$$

4. Low Regret Stochastic Contextual Bandits (`LR-SCB`)

Throughout this paper, we refer OFUL as the optimistic learning algorithm of (Abbasi-yadkori et al., 2011) for linear bandits. In fact (Chatterji et al., 2020) uses this in the finite armed contextual framework, and we use a variation of their OFUL algorithm, without arm biases. For completeness, we reproduce this in Algorithm 2. We use OFUL as a black box in Algorithm 2.

We now present the algorithm for the stochastic contextual bandit. We divide the learning horizon into epochs of length $T_1, T_2, \ldots, T_N$, where N is the number of epochs. In the first phase T_1, we aim to minimize regret and estimate the parameter θ^* simultaneously for T_1 rounds. At the end of this phase, we obtain an estimate $\widehat{\theta}_{T_1}$, of θ^*.

Subsequently, in the second phase, which lasts for T_2 rounds, our goal is to utilize the estimate $\widehat{\theta}_{T_1}$. Here, we aim to learn the parameter $\theta^* - \widehat{\theta}_{T_1}$. Note that, the norm of $\theta^* - \widehat{\theta}_{T_1}$ is small since we spend the previous epoch to learn θ^*. Hence, in this epoch, instead of using the OFUL algorithm, we use an adaptive algorithm that exploits the small norm. In particular, we use a modified version (reproduced in Algorithm 3) of the Adaptive Linear Bandits-norm (`ALB-norm`) of (Ghosh et al., 2021b), that exploits the small norm of $\theta^* - \widehat{\theta}_{T_1}$ to obtain a reduced regret, which depends linearly on $\|\theta^* - \widehat{\theta}_{T_1}\|$. As seen in Algorithm 1, the learning of $\theta^* - \widehat{\theta}_{T_1}$ is achieved by shifting the reward by the inner product of the estimate $\widehat{\theta}_{T_1}$. By exploiting the anti-concentration of measure along with some standard results from optimization, we show, in Section 6 as well as in Appendix D that the regret of the shifted system is worse than the regret of the original system (in high probability)[4].

We now continue the above-mentioned estimation procedure in the third epoch as well, which lasts for T_3 rounds. Here, we exploit the fact that at the end of the second epoch, we obtain $\widehat{\theta}_{T_2}$, which is an estimate of $\theta^* - \widehat{\theta}_{T_1}$. In Appendix C, we show that similar to the OFUL algorithm, `ALB-norm` also constructs an estimate of the parameter under consideration. Basically, `ALB-norm` is equivalent to playing the OFUL algorithm in successive epochs with norm refine-

[3]The conditional variance assumption is implicitly used in (Chatterji et al., 2020) without explicit statement.

[4]This is intuitive since, otherwise one can find *appropriate shifts* to reduce the regret of OFUL, which contradicts the optimality of OFUL.

Algorithm 1 Low Regret Stochastic Contextual Bandits (LR−SCB)

Input: Horizon T, Initial epoch length T_1

First phase, $i = 1$:

Initialize a single instance of OFUL(δ) (see Algorithm 2)

for times $t \in \{1, \cdots, T_1\}$ **do**

  Play the action given by the common OFUL

  Update OFUL's state by the observed rewards similar to Algorithm 2

end for

Let est $\leftarrow \widehat{\theta}_{T_1}$; the parameter estimate of Common OFUL at the end of phase 1

Subsequent Phases:

for phase $i \in \{2, \ldots, N\}$ **do**

  $\delta_i \leftarrow \frac{\delta}{2^{i-1}}$

  $T_i = T_1(\log T)^{i-1}$

  Initialize one (modified) ALB-Norm(δ) (see Algorithm 3) instance per agent

  for times $t \in \{T_i + 1, \ldots, T_{i+1}\}$ **do**

    Play arm by ALB-Norm (denoted as $\beta_{B_t,t}$) and receive reward y_t

    Every agent updates their ALB-Norm state with corrected reward

$$\tilde{y}_t = y_t - \langle \beta_{B_t,t}, \mathsf{est} \rangle$$

  end for

  $\widehat{\theta}_{T_i}$: parameter estimate after i-th epoch

  est $\leftarrow$ est $+ \widehat{\theta}_{T_i}$

end for

Algorithm 2 OFUL of (Chatterji et al., 2020)

Input: Parameters b, $\delta > 0$, number of rounds $\tilde{T}$

for $t = 1, 2, \ldots, \tilde{T}$ **do**

  Select the best arm estimate as

$$j_t = \text{argmax}_{i \in [K]} \left[\max_{\theta \in \mathcal{C}_{t-1}} \{\langle \alpha_{i,t}, \theta \rangle\} \right],$$

  where $\mathcal{C}_t$ is the confidence set with radius

$$\frac{b + \sqrt{d}}{\rho_{\min}\sqrt{t}} \log(K\tilde{T}/\delta)$$

  Play arm j_t, and update $\mathcal{C}_t$

end for

Algorithm 3 Adaptive Linear Bandit (norm)–ALB−Norm of (Ghosh et al., 2021b)

Input: The initial exploration period τ_1, intial phase length $T_1 := \lceil \sqrt{T} \rceil$, $\delta_1 > 0$, $\delta_s > 0$.

Select an arm at random, sample 2τ rewards

Obtain initial estimate (b_1) of $\|\theta^*\|$ according to Section 3.3 of (Ghosh et al., 2021b).

for epochs $i = 1, 2 \ldots, N$ **do**

  Play OFUL (Algorithm 2) with slack δ_i and norm estimate b_i until the end of epoch i (denoted by $\mathcal{E}_i$)

  At $t = \mathcal{E}_i$, refine estimate of $\|\theta^*\|$ as,

$$b_{i+1} = \max_{\theta \in \mathcal{C}_{\mathcal{E}_i}} \|\theta\|$$

  Set $T_{i+1} = 2T_i$

  $\delta_{i+1} = \frac{\delta_i}{2}$.

end for

ments. Using the fact that $\|\theta^* - \widehat{\theta}_{T_1} - \widehat{\theta}_{T_2}\|$ is small, we again use the norm adaptive algorithm ALB−norm to obtain smaller regret. Hence, the regret in this phase is proportional to $\|\theta^* - \widehat{\theta}_{T_1} - \widehat{\theta}_{T_2}\|$.

So, this successive estimation procedure continues upto the N-th epoch. At each epoch, we shift the reward by an inner product obtained of the estimate obtained from the previous round. The algorithm is detailed in Algorithm 1. Note that in the above algorithm, we use the estimate obtained in the previous epoch and judiciously use a norm adaptive (which adapts to the norm of the problem) algorithm. By judiciously choosing the time epochs, we show that the overall regret of LR−SCB can be reduced to $\mathcal{O}(\mathsf{polylog} T)$.

5. Regret Guarantee for LR−SCB

In this section, we provide the regret guarantee of LR−SCB. We stick to the notation of Section 3. Moreover, we select the time epochs in the following manner:

$$T_i = T_1(\log T)^{i-1}.$$

With this choice, the number of epochs is given by,

$$N = \mathcal{O}\left(\frac{\log(T/T_1)}{\log\log T}\right).$$

To ease notation, let us define

$$\Lambda = \left(\frac{1}{(\log\log T)} \log\left(\frac{\rho_{\min}^2 T}{d^2 \log^4(KT/\delta)\log(dT/\delta)}\right)\right)$$

and,

$$\mathfrak{T} = \log^3\left(\frac{Kd^2(\log T)(\log^4 KT/\delta)(\log dT/\delta)}{\rho_{\min}^2\,\delta}\right) \times \log^2\left(\frac{d^3(\log T)(\log^4 KT/\delta)(\log dT/\delta)}{\rho_{\min}^2\,\delta}\right)$$

We have the following theorem.

Theorem 5.1. *Suppose we play Algorithm 1 with initial phase length T_1 time and probability slack $\delta > 0$, where*

$$T_1 = C_1 \frac{d^2}{\rho_{\min}^2} \log^4(KT/\delta) \log(dT/\delta) \quad \textit{and}$$

$$d \geq C_1 \frac{\log T}{\log\log T} \log(K^2/\delta).$$

Then the regret of the player for a horizon of T satisfies

$$R(T) \leq C_2 \left[\left(\frac{d}{\rho_{\min}} \right)^{3/2} \Lambda^5 \, \mathfrak{T} \, \sqrt{\log T} \right] = \mathcal{O}\left(\left(\frac{d}{\rho_{\min}} \right)^{3/2} \mathsf{polylog}(T, K, d, \delta) \right)$$

with probability at least $1 - c\delta$, where c, C, C_1, C_2 are universal constants.

The proof is deferred to the Appendix. We make the following remarks:

Remark 5.2. The above theorem shows that the (instance independent) regret of stochastic contextual bandits is $\mathsf{polylog}(T)$. This is a huge improvement over the $\sqrt{T}$ regret presented in (Chu et al., 2011; Li et al., 2010; Chatterji et al., 2020). So, the stochastic contexts indeed help in regret reduction.

Remark 5.3. Note that the dependence on dimension d is worse in `LR-SCB` compared to SupLinUCB of (Chu et al., 2011) ($\mathcal{O}\left(\left(\frac{d}{\rho_{\min}} \right)^{3/2} \right)$ vs. $\mathcal{O}(\sqrt{d})$). Furthermore, one needs $d \geq \log(K^2)$ for the anti-concentration of the contexts to kick in, which was crucial in the analysis of the shifted OFUL.

Remark 5.4. We require the initial length $T_1 = \tilde{\mathcal{O}}(d^2/\rho_{\min}^2)$ for the norm adaptive algorithm, `ALB-norm` to work (see (Ghosh et al., 2021b)).

5.0.1. Special Case—Contexts are drawn from Uniform Distribution

Here we assume the contexts come from $\mathsf{Unif}[-c_0/\sqrt{d}, c_0/\sqrt{d}]^{\otimes d}$ for a constant c_0. For this we have $\rho_{\min} = c_1/d$, and hence the following result.

Corollary 5.5. *Suppose the initial phase length*

$$T_1 = \tilde{\mathcal{O}}(d^4) \;\; \textit{and} \;\; d \geq C_1 \frac{\log T}{\log\log T} \log(K^2/\delta).$$

Then, playing Algorithm 1 for T times incur a regret of

$$R(T) \leq \mathcal{O}\left(d^3 \; \mathsf{polylog}(T, K, d, \delta) \right),$$

with probability at least $1 - \delta$.

5.1. Proof Sketch

We now present a brief proof sketch of Theorem 5.1. The full proof is deferred to Appendix A. For simplicity and the clarity of exposition, we only focus on the dependence on time horizon T. We break the learning horizon in epochs of lengths $T_1, T_2, \ldots, T_N$.

Regret in Epoch 1: In the first epoch, we play the OFUL algorithm (Algorithm 2). Hence, for (Chatterji et al., 2020), we incur a regret of $\mathcal{O}(\sqrt{T_1})$.

Regret in Epoch 2: In the second epoch, we use the parameter estimate learned in the first epoch and accordingly modify the reward functions. Hence, the underlying parameter in second epoch is the shifted parameter. We leverage the analysis of a shifted OFUL to handle this. Moreover, note that since we are estimating θ^* in the first epoch, from (Chatterji et al., 2020), we have

$$\|\widehat{\theta}_{T_1} - \theta^*\| \leq \mathcal{O}(1/\sqrt{T_1})$$

In order to exploit the fact that the norm of the shifted parameter is small, we use a norm-adaptive algorithm, namely `ALB-norm`, in this round, whose regret is given by

$$\text{Reg}_{\text{epoch 2}} = \mathcal{O}(\|\widehat{\theta}_{T_1} - \theta^*\|)\sqrt{\frac{1}{T_2}} = \mathcal{O}(\sqrt{\frac{T_2}{T_1}})$$

Regret in Subsequent Epochs: We continue to shift the parameter by the estimate learnt from the previous epoch. For Epoch 3, we learn $\widehat{\theta}_{T_2}$, which is an estimate of the parameter $\theta^* - \widehat{\theta}_{T_1}$. Using the same `ALB-Norm`, the regret here is

$$\text{Reg}_{\text{epoch 3}} = \mathcal{O}(\|\widehat{\theta}_{T_2} - (\theta^* - \widehat{\theta}_{T_1})\|)\sqrt{\frac{1}{T_3}} = \mathcal{O}(\sqrt{\frac{T_3}{T_2}})$$

Total Regret: Combining the above expressions, the total regret is given by

$$R(T) \leq \mathcal{O}\left(\sqrt{\frac{1}{T_1}} + \sum_{i=1}^{N} \sqrt{\frac{T_i}{T_{i-1}}} \right)$$

Choice of T_i: We choose aggressively increasing epoch lengths. This is because, we get to exploit the estimation performance of previous epoch to the new one, and get low regret owing to norm adaptive algorithms. We select

$$T_i = T_1 (\log T)^{i-1}$$

and as a result, the total number of epochs is

$$N = \mathcal{O}\left(\frac{\log(T/T_1)}{\log\log T} \right)$$

Choice of T_1: We use the ALB-norm algorithm of (Ghosh et al., 2021b), which imposes a condition on T_1. It turns out (showed formally in Appendix A) we require

$$T_1 \geq \tilde{\mathcal{O}}(d^2/\rho^2_{\min}).$$

Hence, with the above choice of T_1 and combining the regret in different epochs, we obtain

$$R(T) \leq \mathcal{O}\left(\text{polylog}(T)\right),$$

which proves the theorem.

6. Shifted OFUL

In this section, we establish a relationship between the regret of the standard OFUL algorithm and the shift OFUL for linear contextual bandits, and show that shifts can not reduce the regret of OFUL. We crucially leverage the analysis of shifted OFUL in Algorithm 1. Beyond Algorithm 1, this analysis may be of independent interest.

We keep the problem setup same as Section 3. We define the shifted version of OFUL below.

Recall that the OFUL algortihm is used to make a decision of which action to take at time-step t, given the history of past actions $X_1, \cdots, X_{t-1}$ and observed rewards $Y_1, \cdots, Y_{t-1}$. The Γ shifted OFUL is an algorithm identical to OFUL that describes the action to take at time step t, based on the past actions $X_1, \cdots, X_{t-1}$ and the observed rewards $\widetilde{Y}_1^{(\Gamma)}, \cdots, \widetilde{Y}_{t-1}^{(\Gamma)}$, where for all $1 \leq s \leq t-1$, $\widetilde{Y}_s = Y_s - \langle X_s, \Gamma \rangle$.

Let us first recall the definition of regret for an un-shifted standard OFUL instance.

Definition 6.1 (OFUL)**.** For a linear contextual bandit instance with unknown parameter θ^*, and a sequence of (possibly random) actions $X_{1:T} := X_1, \cdots, X_T$, we denote the regret obtained upto round T as

$$R_T(X_{1:T}) := \sum_{t=1}^{T} \max_{1 \leq j \leq K} \langle \beta_{j,t} - X_t, \theta^* \rangle.$$

Using the same notation as above, we now define the regret of an instance of the Γ shifted system.

Definition 6.2 (Γ shifted OFUL)**.** For a linear contextual bandit system with unknown parameter θ^*, the modified set of rewards and a sequence of (possibly random) actions $X_{1:T} := X_1, \cdots, X_T$, we denote its regret upto time T as

$$R_T^{(\Gamma)}(X_{1:T}) := \sum_{t=1}^{T} \max_{1 \leq j \leq K} \langle \beta_{j,t} - X_t, \theta^* - \Gamma \rangle$$

We now show that the shifted OFUL algorithm incurs higher regret than that of unshifted one, with high probability. We have the following result.

Lemma 6.3. *Consider a linear contextual bandit instance with parameter θ^* with $||\theta^*|| \leq 1$ and the context vectors at each time are sampled independently from any (coordinate-wise) bounded distribution (i.e., $[-c/\sqrt{d}, c/\sqrt{d}]^{\otimes d}$) for a constant c. Let $\Gamma \in \mathbb{R}^d$ be such that $||\theta^* - \Gamma|| \leq \psi$ for a constant $\psi < \frac{1}{2\sqrt{2}}$, and $X_{1:T} = (X_1, \cdots, X_T)$ be the set of actions chosen by the Γ shifted OFUL. Then, with probability at-least $\left(1 - \binom{K}{2} e^{-c_1 d} - K e^{-c_2 d}\right)$,*

$$\mathcal{R}_T(X_{1:T}) \leq \mathcal{R}_T^{(\Gamma)}(X_{1:T}),$$

where the constants c_1 and c_2 depend on ψ.

Remark 6.4. The above lemma shows that for a deterministic Γ shift, provided $d \geq \Omega(\log K)$, the shifted system always suffers higher regret with probability at least $1 - c\exp(-c_1 d)$

6.0.1. Proof Sketch

The proof of the above Lemma is deferred in Appendix D. We now give a brief sketch here. To show the above, we first show the following using definitions and some basic facts in optimization literature.

Proposition 6.5. *Suppose for a linear contextual bandit instance with parameter θ^*, an algorithm plays the sequence of actions $X_1, \cdots, X_T$, then*

$$\mathcal{R}_T(X_{1:T}) \leq \mathcal{R}_T^{(\Gamma)}(X_{1:T}) + \sum_{t=1}^{T} \left(\langle X_t - \underset{\beta \in \{\beta_{1,t}, \cdots, \beta_{K,t}\}}{\text{argmax}} \langle \beta, \theta^* \rangle, \Gamma \rangle \right).$$

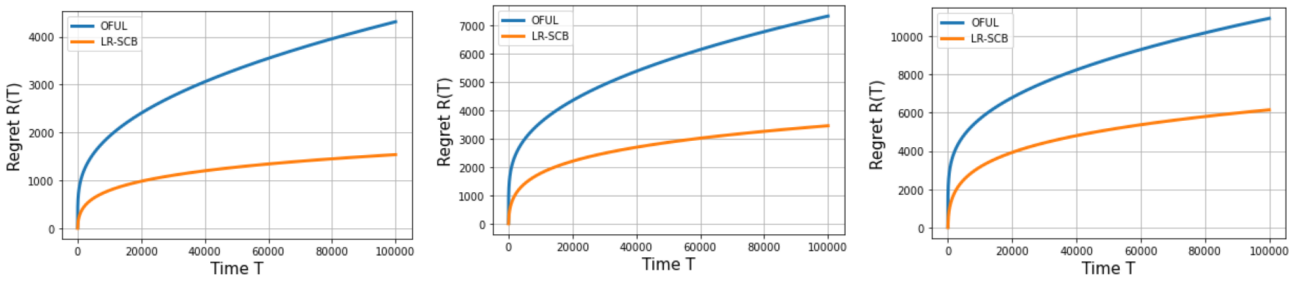
From the above, it is clear that provided,

$$\underset{\beta \in \{\beta_{1,t}, \cdots, \beta_{K,t}\}}{\text{argmax}} \langle \beta, \theta^* \rangle = \underset{\beta \in \{\beta_{1,t}, \cdots, \beta_{K,t}\}}{\text{argmax}} \langle \beta, \Gamma \rangle,$$

the second term in Proposition 6.5 is negative, and we have Lemma 6.3. We now concentrate on the probability under which the above mentioned event occurs. For this, we use the anti-concentration property of the coordinate-wise bounded (and hence sub-Gaussian) random variables, along with the fact that the contexts are drawn in an independent manner. Leveraging these, we obtain the probability of the above-mentioned event is at least $1 - \binom{K}{2} e^{-c_1 d} - K e^{-c_2 d}$, which proves the lemma.

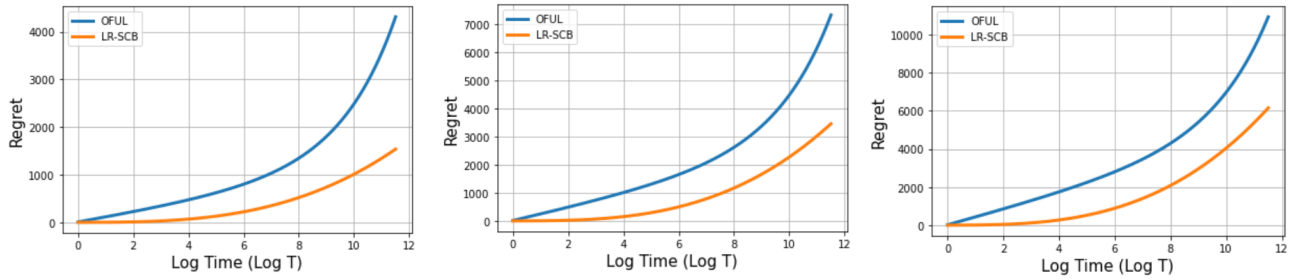
7. Simulations

In this section, we validate our theoretical findings of Section 5 via simulations. We assume that the contexts are drawn i.i.d from $\text{Unif}[-1/\sqrt{d}, 1/\sqrt{d}]^{\otimes d}$. We run Algorithm 1 with $K = 20$ arms with different dimension $d = \{20, 15, 30\}$. Moreover, we compare our results with that of the OFUL (Algorithm 2), and show the LR-SCB attanins much smaller regret compared to OFUL.



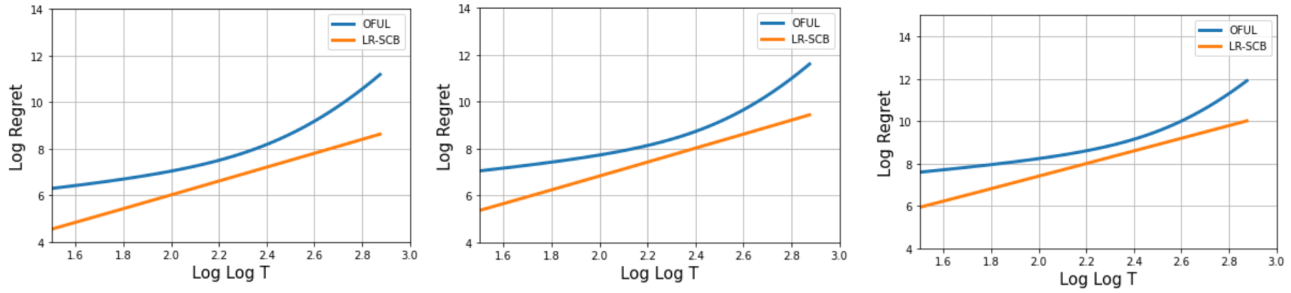
(a) $d = 20, K = 20$ (b) $d = 25, K = 20$ (c) $d = 30, K = 20$

Figure 1. Regret Scaling with respect to horizon T for OFUL and LR−SCB. The plots are produced by taking an average over 50 trials.



(a) $d = 20, K = 20$ (b) $d = 25, K = 20$ (c) $d = 30, K = 20$

Figure 2. Regret Scaling with respect to $\log T$ for OFUL and LR−SCB. Note that the regret of LR−SCB grows much slowly, compared to OFUL. The plots are produced by taking an average over 50 trials.



(a) $d = 20, K = 20$ (b) $d = 25, K = 20$ (c) $d = 30, K = 20$

Figure 3. Scaling of $\log R(T)$ with respect to $\log \log T$ for OFUL and LR−SCB. The linear increase of LR−SCB indicates a $\mathsf{polylog}(T)$ regret. The plots are produced by taking an average over 50 trials.

7.0.1. $R(T)$ vs. T:

We first plot the the variation of regret $R(T)$, with respect to the learning horizon T for OFUL as well as LR−SCB, for different dimension $d \in \{20, 25, 30\}$. It is shown in Figure 1. We observe that the regret of LR−SCB is much smaller than that of OFUL. This indeed validates our theoretical finding, since for OFUL, the regret $R_{OFUL}(T) = \mathcal{O}(\sqrt{T})$, and for LR−SCB, from Theorem 5.1, $R_{LR-SCB}(T) = \mathcal{O}(\mathsf{polylog}T)$. We run 50 instances, and take average over trials to obtain the plots in Figure 1.

7.0.2. $R(T)$ vs. $\log T$

To understand the regret scaling a bit better, we now plot the $R_{OFUL}(T)$ and $R_{LR-SCB}(T)$ with $\log T$. The plots are shown in Figure 2. We observe here that the regret scales quite aggressively for OFUL, while it increases at a much slower rate for LR−SCB.

Note that since, $R_{OFUL}(T) = \mathcal{O}(\sqrt{T})$, the plot of $R_{OFUL}(T)$ vs. $\log T$ is expected to grow at an exponential speed, which we can see from Figure 2 in all 3 cases. On the other hand, since $R_{LR-SCB}(T) = \mathcal{O}(\mathsf{polylog}T)$, the

$R_{OFUL}(T)$ vs. $\log T$ plot is expected to grow at a polynomial rate, which is evidenced by the slow rate of increase. Hence, Figure 2 clearly hints towards a $\mathsf{polylog}(T)$ regret of LR−SCB, which validates Theorem 5.1.

7.0.3. $\log R(T)$ vs. $\log\log T$

In order to further understand the regret scaling of LR−SCB, we plot $\log R(T)$ against $\log\log T$, for both OFUL and LR−SCB. The results are shown in Figure 3. Note that for LR−SCB, we obtain lines with slope slightly more than 2.

This clearly indicates a $\mathcal{O}(\mathsf{polylog}T)$ regret of LR−SCB. Recall that the regret of LR−SCB is $R_{LR-SCB}(T) = \mathcal{O}(\mathsf{polylog}T)$, and hence $\log R_{LR-SCB}$ is a linear function of $\log\log T$, which we evidence. Furthermore, this hints that the polynomial dependence on $\log T$ is close to a quadratic one. On the other hand, for OFUL, note that the log regret is not a straight line, and keeps on increasing. This implies that the regret of OFUL is not poly-logarithmic, which matches the known results. We emphasize that, it is quite non-trivial to capture the regret of OFUL and LR−SCB in $\log\log T$ scale. Hence, we ran the learning algorithms for $T = 5 \times 10^7$, to get the above mentioned results.

8. Conclusion and Future work

In this paper, we exploit the stochasticity of the contexts and obtain an instance-independent poly logarithmic regret bound for linear contextual bandits. Our analysis crucially relies on leveraging the norm adaptive learning algorithms, like ALB−norm. In this paper, we only obtain an upper bound, and hence a natural question arises about the tightness of the result. An immediate future work is to obtain an lower bound in the presence of stochastic context, and see whether our result is tight. Additionally, we want to understand the (structured) stochastic contextual bandit framework beyond linearity, and ask for similar guarantees. We keep these as our future endevors.

References

Abbasi-yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 24, pp. 2312–2320. Curran Associates, Inc., 2011.

Abe, N., Biermann, A. W., and Long, P. M. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293, 2003.

Agarwal, A., Dudík, M., Kale, S., Langford, J., and Schapire, R. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, pp. 19–26. PMLR, 2012.

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. Taming the monster: A fast and simple algorithm for contextual bandits. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1638–1646, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/agarwalb14.html.

Agarwal, A., Bird, S., Cozowicz, M., Hoang, L., Langford, J., Lee, S., Li, J., Melamed, D., Oshri, G., Ribas, O., et al. Making contextual decisions with low technical debt. *arXiv preprint arXiv:1606.03966*, 2016.

Agarwal, A., Luo, H., Neyshabur, B., and Schapire, R. E. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pp. 12–38. PMLR, 2017.

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

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 19–26. JMLR Workshop and Conference Proceedings, 2011.

Chatterji, N., Muthukumar, V., and Bartlett, P. Osom: A simultaneously optimal algorithm for multi-armed and linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 1844–1854. PMLR, 2020.

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.

Dimakopoulou, M., Zhou, Z., Athey, S., and Imbens, G. Balanced linear contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3445–3453, 2019.

Foster, D. and Rakhlin, A. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pp. 3199–3210. PMLR, 2020.

Foster, D. J., Krishnamurthy, A., and Luo, H. Model selection for contextual bandits, 2019.

Gentile, C., Li, S., and Zappella, G. Online clustering of bandits. In *International Conference on Machine Learning*, pp. 757–765. PMLR, 2014.

Gentile, C., Li, S., Kar, P., Karatzoglou, A., Zappella, G., and Etrue, E. On context-dependent clustering of bandits. In *International Conference on Machine Learning*, pp. 1253–1262. PMLR, 2017.

Ghosh, A., Chowdhury, S. R., and Ramchandran, K. Model selection with near optimal rates for reinforcement learning with general model classes. *arXiv preprint arXiv:2107.05849*, 2021a.

Ghosh, A., Sankararaman, A., and Kannan, R. Problem-complexity adaptive model selection for stochastic linear bandits. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1396–1404. PMLR, 13–15 Apr 2021b. URL http://proceedings.mlr.press/v130/ghosh21a.html.

Ghosh, A., Sankararaman, A., and Ramchandran, K. Collaborative learning and personalization in multi-agent stochastic linear bandits. *arXiv preprint arXiv:2106.08902*, 2021c.

Ghosh, A., Sankararaman, A., and Ramchandran, K. Model selection for generic contextual bandits. *arXiv preprint arXiv:2107.03455*, 2021d.

Krishnamurthy, S. K. and Athey, S. Optimal model selection in contextual bandits with many classes via offline oracles. *arXiv preprint arXiv:2106.06483*, 2021.

Lee, J., Pacchiano, A., Muthukumar, V., Kong, W., and Brunskill, E. Online model selection for reinforcement learning with function approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3340–3348. PMLR, 2021.

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.

Pacchiano, A., Dann, C., Gentile, C., and Bartlett, P. Regret bound balancing and elimination for model selection in bandits and rl. *arXiv preprint arXiv:2012.13045*, 2020.

Simchi-Levi, D. and Xu, Y. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 2021.

Tewari, A. and Murphy, S. A. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health - Sensors, Analytic Methods, and Applications*, 2017.

Zhou, D., Li, L., and Gu, Q. Neural contextual bandits with ucb-based exploration. In *International Conference on Machine Learning*, pp. 11492–11502. PMLR, 2020.

Zhou, Z., Xu, R., and Blanchet, J. Learning in generalized linear contextual bandits with stochastic delays. *Advances in Neural Information Processing Systems*, 32: 5197–5208, 2019.

Supplementary Material for "Logarithmic Regret for Stochastic Contextual Linear Bandits"

A. Proof of Theorem 5.1

Regret in Phase 1: We run the OFUL algorithm (shown in Algorithm 2 for T_1 time steps. Hence, in this phase, the center indeed learns the parameter θ^*. Let $\widehat{\theta}_{T_1}$ be the corresponding estimate. Provided, $T_1 > \tau_{\min}(\delta)$, from (Chatterji et al., 2020), we have,

$$\|\widehat{\theta}_{T_1} - \theta^*\| \leq \mathcal{O}\left(\sqrt{\frac{d}{\rho_{\min} T_1}}\right) \log(KT_1/\delta) \log(dT_1/\delta),$$

with probability at least $1-\delta$. The corresponding regret (call it R_{T_1}) is

$$R_{T_1} = \mathcal{O}\left(\sqrt{\frac{dT_1}{\rho_{\min}}}\right) \log(KT/\delta) \log(dT/\delta),$$

with probability at least $1-\delta$.

Regret in Phase 2: In this phase, we take advantange of the learned paameter, $\widehat{\theta}_{T_1}$. Here, the learning proceeds as the following: At each time t, out of K contexts, $\{\beta_{r,t}\}_{r=1}^K$, suppose the player chooses a context vector, $\beta_{r,t}$, (corresponding to the r-th arm). Thereafter, the player generates the reward $y_t = \langle \beta_{r,t}, \theta^* \rangle + \xi_{i,t}$. Subsequently, using the previous estimate, the player calculates the corrected reward

$$\tilde{y}_t = y_t - \langle \beta_{r,t}, \widehat{\theta}_{T_1} \rangle.$$

Note that the player has the information about $(\beta_{r,t}, \widehat{\theta}_{T_1})$ and so it can compute $\tilde{y}_t$. With this shift, the center basically learns the vector $\theta^- \widehat{\theta}_{T_1}$.

In this phase, we use a variation of the `ALB-norm` algorithm of (Ghosh et al., 2021b)[5]. The variation is reproduced in Section B. Note that the `ALB-norm` algorithm is a norm adaptive algorithm, which is particularly useful when the parameter norm is small. `ALB-norm` uses the OFUL algorithm of (Chatterji et al., 2020) repeatedly over doubling epochs. At the beginning of each epoch, it estimates the parameter norm, and runs OFUL with the norm estimate (see (Ghosh et al., 2021b, Algorithm 1)), and keeps on refining it. Hence, it is shown in (Ghosh et al., 2021b, Algorithm 1) that while estimating the parameter Ψ^*, with high probability, the regret of `ALB-norm` is

$$R_{\texttt{ALB-norm}} \leq \|\Psi^*\| \; R_{OFUL}.$$

We use the `ALB-Norm` with this shifted system. However, since `ALB-Norm` is equivalent to playing the OFUL algorithm on doubling epochs, it is sufficient to obtain the performance of a shifted OFUL system, and the same conclusion extends to `ALB-Norm` (see (Ghosh et al., 2021b)). In Appendix D, we present an analysis of shifted OFUL. In particular we show that shifts (by a fixed vector) can not reduce the regret (which is intuitive). Note that we learn $\widehat{\theta}_{T_1}$ in the previous phase, and fix it throughout this phase. Hence, conditioned on the observations of the first phase, $\hat{\theta}_{T_1}$ is a fixed (deterministic) vector. In particular, in Lemma D.9, it is shown that provided $d \geq C \log(K^2/\delta_2)$, we have $R_{OFUL} \leq R_{OFUL}^{shift}$ with probability at least $1-\delta_2$.

Hence, using Lemma D.9 of Appendix D, the regret in phase 2 (call it R_{T_2}) is given by

$$R_{T_2} \leq \mathcal{O}\left(\|\theta^* - \hat{\theta}_{T_1}\| \sqrt{\frac{dT_2}{\rho_{\min}}} \log(KT_2/\delta_2) \log(dT_2/\delta_2) \right),$$

[5] We reproduce the algorithm in Appendix B.

with probability at least $1 - c\delta_2$, provided $d \geq C \log(K^2/\delta_2)$. Substituting, we obtain

$$R_{T_2} \leq \mathcal{O}\left(\frac{d}{\rho_{\min}}\sqrt{\frac{T_2}{T_1}}\right)\log^2(KT_2/\delta_2)\log^2(dT_2/\delta_2)$$

with probability exceeding $1 - c\delta_2$.

Regret in Phase 3: At the end of phase 2, we obtain the estimate $\widehat{\theta}_{T_2}$. Note that this is an estimate of $\theta^* - \widehat{\theta}_{T_1}$. In Phase 3, the idea is to exploit this estimate. The intuition is similar to that of phase 2. Since $\widehat{\theta}_{T_2}$ is an estimate of $\theta^* - \widehat{\theta}_{T_1}$, the quantity $\|\theta^* - \widehat{\theta}_{T_1} - \widehat{\theta}_{T_2}\|$ will be small, an a norm-adaptive algorithm, like `ALB-norm` should exploit this fact.

In order to show this, we first show that, similar to the OFUL algorithm, it is possible for the `ALB-norm` algorithm to estimate the parameter of interest. In Appendix C, we show this formally. Intuitively, this makes sense, since `ALB-norm` is basically the OFUL algorithm of (Chatterji et al., 2020) applied repeatedly over doubling epochs. Since, the OFUL algorithm estimates the underlying parameter, in Section C, we show that `ALB-norm` also performs similar parameter estimation.

Furthermore, now the corrected regret is given by,

$$\tilde{y}_t = y_t - \langle \beta_{r,t}, \widehat{\theta}_{T_1}\rangle - \langle \beta_{r,t}, \widehat{\theta}_{T_2}\rangle.$$

In other words, we shift the center by an amount given corresponding to $\widehat{\theta}_{T_1} + \widehat{\theta}_{T_2}$. We use the same analysis in Section D to show that provided $d \geq C\log(K^2/\delta_3)$, we have $R_{OFUL} \leq R^{shift}_{OFUL}$. Hence, the regret of this phase is given by,

$$\begin{aligned} R_{T_3} &\leq \mathcal{O}\left(\|\theta^* - \hat{\theta}_{T_1} - \widehat{\theta}_{T_2}\|\sqrt{\frac{dT_3}{\rho_{\min}}}\log(KT_3/\delta_3)\log(dT_3/\delta_3)\right) \\ &\leq \mathcal{O}\left(\frac{d}{\rho_{\min}}\sqrt{\frac{T_3}{T_2}}\right)\log^2(KT_3/\delta_3)\log^2(dT_3/\delta_3) \end{aligned}$$

with probability at least $1 - c\delta_3$.

Subsequent Phases: For phase $i > 3$, the same argument holds, and the regret is given by,

$$R_{T_i} \leq \mathcal{O}\left(\frac{d}{\rho_{\min}}\sqrt{\frac{T_i}{T_{i-1}}}\right)\log^2(KT_i/\delta_i)\log^2(dT_i/\delta_i)$$

with probability at least $1 - c\delta_i$, provided $d \geq C\log(K^2/\delta_i)$.

Total Regret: We now characterize the total regret of the agent. Let us assume the number of phases is N. We have

$$\begin{aligned} R_T &= R_{T_1} + \ldots + R_{T_N} \\ &\leq \mathcal{O}(\sqrt{\frac{d}{\rho_{\min}}}\sqrt{T_1})\log(KT_1/\delta)\log(dT_1/\delta) + \sum_{i=2}^{N}\mathcal{O}\left(\frac{d}{\rho_{\min}}\sqrt{\frac{T_i}{T_{i-1}}}\right)\log^2(KT_i/\delta_i)\log^2(dT_i/\delta_i). \end{aligned}$$

Since we consider $\delta_i = \delta/2^{i-1}$, the above regret holds with probability at least

$$\begin{aligned} &1 - c(\delta_1 + \delta_2 + \ldots + \delta_N) \\ &\geq 1 - (\delta + \delta/2 + \delta/4 + \ldots) \\ &\geq 1 - 2c\delta, \end{aligned}$$

where c is an universal constant.

We now choose the length of phases as

$$T_i = T_1(\log T)^{i-1},$$

where T_1 is the initial length. With this, we obtain, the number of epochs, $N = \mathcal{O}\left(\frac{\log(T/T_1)}{\log\log T}\right)$. Subsequently, the overall regret is given by,

$$R_T \leq \mathcal{O}\left[(\sqrt{\frac{d}{\rho_{\min}}}\sqrt{T_1})\log(KT/\delta)\log(dT/\delta) + N\frac{d}{\rho_{\min}}\sqrt{\log T}\ N^2\log^2(KT_1(\log T)/\delta)\ N^2\log^2(dT_1(\log T)/\delta)\right],$$

where we substitute δ_i and upper bound the number of epochs by N. Substituting N, we obtain

$$R_T \leq \mathcal{O}\left[(\sqrt{\frac{d}{\rho_{\min}}}\sqrt{T_1})\log(KT_1/\delta)\log(dT_1/\delta) + \frac{d\sqrt{\log T}}{\rho_{\min}}\left(\frac{\log(T/T_1)}{\log\log T}\right)^5\log^2(KT_1(\log T)/\delta)\log^2(dT_1(\log T)/\delta)\right],$$

with probability at least $1 - c\delta$, provided

$$d \geq CN\ \log(K^2/\delta) \geq C\ \left(\frac{\log(T/T_1)}{\log\log T}\right)\log(K^2/\delta)$$

The next job is to choose the length of the first epoch T_1. For the norm adaptive algorithm, `ALB-norm` to work, one needs (from (Ghosh et al., 2021b, Theorem 1))

$$T_1 = C\max\left\{\frac{d^2}{\rho_{\min}^2}\log^4(KT/\delta), \tau_{\min}(\delta)^2\right\}$$

for a large enough universal constant C, where $\tau_{\min} = \left[\frac{16}{\rho_{\min}^2} + \frac{8}{3\rho_{\min}}\right]\log(\frac{2dT}{\delta})$. Hence, we need to choose

$$T_1 = C_1\frac{d^2}{\rho_{\min}^2}\log^4(KT/\delta)\log(dT/\delta).$$

To ease notation, let us define

$$\Lambda = \left(\frac{1}{(\log\log T)}\log\left(\frac{\rho_{\min}^2 T}{d^2\log^4(KT/\delta)\log(dT/\delta)}\right)\right)$$

and,

$$\mathfrak{T} = \log^3\left(\frac{Kd^2(\log T)(\log^4 KT/\delta)(\log dT/\delta)}{\rho_{\min}^2\,\delta}\right)\log^2\left(\frac{d^3(\log T)(\log^4 KT/\delta)(\log dT/\delta)}{\rho_{\min}^2\,\delta}\right)$$

With this, the overall regret is given by

$$\begin{aligned} R_T &\leq \mathcal{O}\left[\left(\frac{d}{\rho_{\min}}\right)^{3/2}\mathfrak{T} + \left(\frac{d}{\rho_{\min}}\right)\Lambda^5\,\mathfrak{T}\,\sqrt{\log T}\right] \\ &\leq \mathcal{O}\left[\left(\frac{d}{\rho_{\min}}\right)^{3/2}\Lambda^5\ \mathfrak{T}\ \sqrt{\log T}\right], \end{aligned}$$

with probability at least $1 - c\delta$. This requires,

$$d \geq C\ \left(\frac{\log(T/T_1)}{\log\log T}\right)\log(K^2/\delta).$$

Since, T_1 is a function of d, we choose a sufficient condition on d, which is given by

$$d \geq C\left(\frac{\log T}{\log\log T}\right)\log(K^2/\delta),$$

which concludes the proof.

B. Modified `ALB-Norm` from (Ghosh et al., 2021b)

In this section, we reproduce `ALB-Norm` from (Ghosh et al., 2021b), and prove a Corollary of the main theorem from (Ghosh et al., 2021b).

Corollary B.1 (Corollary of Theorem 1 from (Ghosh et al., 2021b))**.** *The regret of Algorithm 3 at the end of T time-steps satisfies with probability at-least $1 - 18\delta_1 - \delta_s$,*

$$R(T) \leq C\|\theta^*\|(\sqrt{K} + \sqrt{d})\sqrt{T}\log\left(\frac{KT}{\delta_1}\right),$$

where C is an universal constant.

The proof follows by recomputing Lemma 1 from (Ghosh et al., 2021b) as follows.

Lemma B.2. *If T is sufficiently large such that $\frac{2C\sigma\sqrt{d}}{T^{\frac{1}{4}}}\log\left(\frac{K\sqrt{T}}{\delta_1}\right) \leq 1$, then with probability at-least $1 - 8\delta_1 - \delta_s$, for all i large, $b_i \leq 2\|\theta^*\|$ holds, where b_i is defined in Line 11 of Algorithm 3.*

Proof of Lemma B.2. We start with Equation (8) of (Ghosh et al., 2021b). Reproducing Equation (8) by substituting $T_1 = \lceil\sqrt{T}\rceil$, with probability at-least $1 - 8\delta_1$, for all phases $i \geq 2$,

$$b_{i+1} \leq \|\theta^*\| + ip\frac{b_i}{2^{\frac{i-1}{2}}T^{\frac{1}{4}}} + iq\frac{\sqrt{d}}{2^{\frac{i-1}{2}}T^{\frac{1}{4}}}, \tag{3}$$

holds, where p and q are defined in (Ghosh et al., 2021b) as

$$p = \left(\frac{14\log\left(\frac{2K\sqrt{T}}{\delta_1}\right)}{\sqrt{\rho_{min}}}\right),$$
$$q = \left(\frac{2C\sigma\log\left(\frac{2K\sqrt{T}}{\delta_1}\right)}{\sqrt{\rho_{min}}}\right).$$

For all $i \geq 2$, $\frac{i}{2^{\frac{i-1}{2}}} \leq 2$. Thus, for all $i \geq 1$, Equation (3) can be rewritten as

$$\begin{aligned} b_{i+1} &\leq \|\theta^*\| + \frac{pb_i}{T^{\frac{1}{4}}} + \frac{q\sqrt{d}}{T^{\frac{1}{4}}}, \\ &\leq \|\theta^*\| + \frac{C\sigma\sqrt{d}}{T^{\frac{1}{4}}}\log\left(\frac{K\sqrt{T}}{\delta_1}\right)b_i. \end{aligned} \tag{4}$$

where $b_1 := 1$. We set this initial estimate as 1, since $\max_{i\in\{1,\cdots,N\}}\|\theta_i^*\| \leq 1$. We prove the lemma by induction that $b_i \leq 2\|\theta^*\|$.

Base case, $i = 1$ - We know from the initialization (Line 3 of Algorithm 3), that with probability at-least $1 - \delta_s$,

$$\begin{aligned} b_1 &\leq \|\theta^*\| + \sqrt{2}\sigma\sqrt{\frac{d}{\tau}\log\left(\frac{1}{\delta_s}\right)}, \\ &\leq 2\|\theta^*\|. \end{aligned}$$

where τ and δ_s are defined in Line 2 and input respectively of Algorithm 3.

Induction Step - Assume that for some $i \geq 1$, for all $1 \leq j \leq i$, $b_j \leq 2\|\theta^*\|$. Now, consider case $i+1$. From recursion in Equation (4), that

$$\begin{aligned} b_{i+1} &\leq \|\theta^*\| + \frac{C\sigma\sqrt{d}}{T^{\frac{1}{4}}} \log\left(\frac{K\sqrt{T}}{\delta_1}\right) b_i, \\ &\overset{(a)}{\leq} \|\theta^*\| \left(1 + \frac{2C\sigma\sqrt{d}}{T^{\frac{1}{4}}} \log\left(\frac{K\sqrt{T}}{\delta_1}\right)\right), \\ &\overset{(b)}{\leq} 2\|\theta^*\|. \end{aligned}$$

Step (a) follows from the induction hypothesis. Step (b) follows from the fact that T is large enough such that $\frac{2C\sigma\sqrt{d}}{T^{\frac{1}{4}}} \log\left(\frac{K\sqrt{T}}{\delta_1}\right) \leq 1$. This concludes the proof of Lemma. □

C. Parameter estimation for modified `ALB-norm`

In this section we show that, similar to the OFUL algorithm of (Chatterji et al., 2020), the modified `ALB-norm` algorithm described in the previous section, also estimated the underlying parameter while minimizing regret. We have the following result:

Proposition C.1. *Suppose we run the modified `ALB-norm` algorithm, with underlying parameter Ψ for $\mathcal{T}$ rounds (with the same stochastic context assumptions given in Section 3. The estimate returned by `ALB-norm` satisfies*

$$\|\widehat{\Psi} - \Psi\| \leq \mathcal{O}\left(\sqrt{\frac{d}{\rho_{\min}\mathcal{T}}}\right) \log(K\mathcal{T}/\delta) \log(d\mathcal{T}/\delta)),$$

with probability at least $1-\delta$.

Proof. As shown in Algorithm 3, the `ALB-norm` , algorithm works in doubling epochs. At each epoch, it runs the OFUL algorithm of (Chatterji et al., 2020) with a modified norm estimate. Let the doubling epochs be defined as $\{\mathcal{T}_1, \ldots, \mathcal{T}_N$, where N is the total number of epochs. Also, the parameter-estimate at the end of the last epoch is $\widehat{\Psi}$. Since, `ALB-norm` plays OFUL at the last epoch, we obtain,

$$\|\widehat{\Psi} - \Psi\| \leq \mathcal{O}\left(\sqrt{\frac{d}{\rho_{\min}\mathcal{T}_N}}\right) \log(K\mathcal{T}_N/\delta) \log(d\mathcal{T}_N/\delta))$$

with probability at least $1-\delta$. Now we have $\mathcal{T}_N \leq \mathcal{T}$ and,

$$\mathcal{T}_N + \mathcal{T}_{N-1} + \ldots + \mathcal{T}_1 = \mathcal{T}.$$

With the doubling epochs, we have

$$\begin{aligned} &\mathcal{T}_N + \mathcal{T}_N/2 + \ldots \geq \mathcal{T} \\ &\mathcal{T}_N\,(1 + 1/2 + \ldots) \geq \mathcal{T} \\ &\mathcal{T}_N \geq \mathcal{T}/2. \end{aligned}$$

Substituting the above, we have

$$\|\widehat{\Psi} - \Psi\| \leq \mathcal{O}\left(\sqrt{\frac{d}{\rho_{\min}\mathcal{T}}}\right) \log(K\mathcal{T}/\delta) \log(d\mathcal{T}/\delta))$$

with probability at least $1-\delta$, which concludes the proof. □

D. Shifted OFUL Regret

Here, we establish a relationship between the regret of the standard OFUL algorithm and the shift compensated algorithm. We define the shifted version of OFUL below.

Definition D.1. The OFUL algortihm is used to make a decision of which action to take at time-step t, given the history of past actions $X_1, \cdots, X_{t-1}$ and observed rewards $Y_1, \cdots, Y_{t-1}$. The Γ shifted OFUL is an algorithm identical to OFUL that describes the action to take at time step t, based on the past actions $X_1, \cdots, X_{t-1}$ and the observed rewards $\widetilde{Y}_1^{(\Gamma)}, \cdots, \widetilde{Y}_{t-1}^{(\Gamma)}$, where for all $1 \leq s \leq t-1$, $\widetilde{Y}_s = Y_s - \langle X_s, \Gamma \rangle$.

Definition D.2. For a linear bandit instance with unknown parameter θ^*, and a sequence of (possibly random) actions $X_{1:T} := X_1, \cdots, X_T$, denote by $\mathcal{R}_T(X_{1:T}) := \sum_{t=1}^T \max_{1 \leq j \leq K} \langle \beta_{j,t} - X_t, \theta^* \rangle$.

Definition D.3. For a linear bandit system with unknown parameter θ^*, and a sequence of (possibly random) actions $X_{1:T} := X_1, \cdots, X_T$, denote by $\mathcal{R}_T^{(\Gamma)}(X_{1:T}) := \sum_{t=1}^T \max_{1 \leq j \leq K} \langle \beta_{j,t} - X_t, \theta^* - \Gamma \rangle$.

Proposition D.4. *Suppose for a linear bandit instance with parameter θ^*, an algorithm plays the sequence of actions $X_1, \cdots, X_T$, then*

$$\mathcal{R}_T(X_{1:T}) \leq \mathcal{R}_T^{(\Gamma)}(X_{1:T}) + \sum_{t=1}^T \left(\langle X_t - \operatorname*{argmax}_{\beta \in \{\beta_{1,t}, \cdots, \beta_{K,t}\}} \langle \beta, \theta^* \rangle, \Gamma \rangle \right).$$

Proof. From the definition of $\mathcal{R}_T^{(\Gamma)}$, we can write the regret as

$$\begin{aligned} \mathcal{R}_T^{(\Gamma)}(X_{1:T}) &= \sum_{t=1}^T \max_{1 \leq j \leq K} \langle \beta_{j,t} - X_t, \theta^* + \Gamma \rangle, \\ &\overset{(a)}{\leq} \sum_{t=1}^T \max_{1 \leq j \leq K} \langle \beta_{j,t}, \theta^* \rangle + \langle \beta_t^*, \Gamma \rangle - \langle X_t, \theta^* \rangle - \langle X_t, \Gamma \rangle, \end{aligned} \tag{5}$$

where, $\beta_t^* := \operatorname{argmax}_{\beta \in \{\beta_{1,t}, \cdots, \beta_{K,t}\}} \langle \beta, \theta^* \rangle$. The inequality (a) follows from the following elementary fact.

Lemma D.5. *Let $\mathcal{X}$ be a compact set, and functions $f, g : \mathcal{X} \to \mathbb{R}$, such that $\sup_{x \in \mathcal{X}} |f(x)| < \infty$ and $\sup_{x \in \mathcal{X}} |g(x)| < \infty$. Then,*

$$\max_{x \in \mathcal{X}} (f(x) + g(x)) \geq \max_{x \in \mathcal{X}} f(x) + \min_{x \in \mathcal{X}} g(x).$$

that Rewriting Equation (5), we see that

$$\mathcal{R}_T^{(\Gamma)}(X_{1:T}) \leq \mathcal{R}_T + \sum_{t=1}^T \langle \beta_t^* - X_t, \Gamma \rangle,$$

and thus the proposition is proved. □

Corollary D.6. *Suppose for all time t, $\operatorname{argmax}_{\beta \in \{\beta_{1,t}, \cdots, \beta_{K,t}\}} \langle \beta, \theta^* \rangle = \operatorname{argmax}_{\beta \in \{\beta_{1,t}, \cdots, \beta_{K,t}\}} \langle \beta, \Gamma \rangle$. Then,*

$$\mathcal{R}_T(X_{1:T}) \leq \mathcal{R}_T^{(\Gamma)}(X_{1:T}).$$

Proof. From the hypothesis of the theorem, we can observe the following,

$$\begin{aligned} \sum_{t=1}^T \left(\langle X_t - \operatorname*{argmax}_{\beta \in \{\beta_{1,t}, \cdots, \beta_{K,t}\}} \langle \beta, \theta^* \rangle, \Gamma \rangle \right) &= \sum_{t=1}^T \left(\langle X_t - \operatorname*{argmax}_{\beta \in \{\beta_{1,t}, \cdots, \beta_{K,t}\}} \langle \beta, \Gamma \rangle, \Gamma \rangle \right), \\ &\leq 0. \end{aligned}$$

Plugging the above bound into Proposition D.4 completes the proof. □

D.0.1. HIGH PROBABILITY BOUND ON $\mathcal{R}_T^{(\Gamma)}$

Lemma D.7. *Suppose the K context vectors $\beta_1, \cdots, \beta_K$ are such that for all i, $||\beta_i|| \leq 2$ and for all $i \neq j$, $|\langle \beta_i - \beta_j, \theta^* \rangle| \geq 4||\theta^* - \Gamma||$, where θ^* is the unknown linear bandit parameter and Γ is a fixed vector. Then*

$$\operatorname*{argmax}_{1 \leq j \leq K} \langle \beta_j, \theta_i^* \rangle = \operatorname*{argmax}_{1 \leq j \leq K} \langle \beta_j, \Gamma \rangle.$$

Proof. We will prove the following more stronger statement. Let $i \neq j \in [K]$ be such that $\langle \theta^*, \beta_i \rangle \geq \langle \theta^*, \beta_j \rangle$. Then, under the hypothesis of the proposition statement, we have $\langle \theta^*, \beta_i - \beta_j \rangle \geq 4||\theta^* - \Gamma||$. Thus, the following chain holds,

$$\begin{aligned}
\langle \beta_i - \beta_j, \Gamma \rangle &= \langle \beta_i - \beta_j, \theta^* \rangle + \langle \beta_i - \beta_j, \Gamma - \theta^* \rangle, \\
&\geq 4||\theta^* - \Gamma|| + \langle \beta_i - \beta_j, \Gamma - \theta^* \rangle, \\
&\geq 4||\theta^* - \Gamma|| - ||\beta_i - \beta_j||||\Gamma - \theta^*||, \\
&\geq 0.
\end{aligned}$$

The first inequality follows from the hypothesis of the proposition statement, the second follows from Cauchy Schwartz inequality and the last follows from the fact that $||\beta_i - \beta_j|| \leq 2$. Thus, we have shown that under the hypothesis of the Proposition, the ordering of the coordinates whether by inner product with θ^* or with Γ remains unchanged. In particular, the argmax is identical. □

Lemma D.8. *Let θ^* be a fixed vector with $\|\theta^*\| \leq 1$, and $\Gamma \in \mathbb{R}^d$ be any arbitrary vector such that $||\theta^* - \Gamma|| \leq \psi$, for some constant ψ. Let $\beta_1, \cdots, \beta_K$ be i.i.d. vectors, supported on $[-c/\sqrt{d}, c/\sqrt{d}]^{\otimes d}$ for a constant c. Then,*

$$\mathbb{P}\left[\operatorname*{argmax}_{1 \leq j \leq K} \langle \beta_j, \theta_i^* \rangle = \operatorname*{argmax}_{1 \leq j \leq K} \langle \beta_j, \Gamma \rangle\right] \geq \left(1 - \binom{K}{2} e^{-\frac{d}{4}(1-8\psi^2)^2} - K e^{-\frac{\sqrt{5}-1}{2} d}\right).$$

Proof. Denote by the *Good event* $\mathcal{E} := \left\{\operatorname{argmax}_{1 \leq j \leq K} \langle \beta_j, \theta_i^* \rangle = \operatorname{argmax}_{1 \leq j \leq K} \langle \beta_j, \Gamma \rangle\right\}$ From Lemma D.7, we know that a sufficient condition for event $\mathcal{E}$ to hold is that for all $i \neq j$, we have $\left|\langle \theta^*, \beta_i - \beta_j \rangle\right| \geq 2||\theta^* - \Gamma||$ and for all i, $||\beta_i|| < 2$. Thus, from a simple union bound, we get

$$\begin{aligned}
\mathbb{P}[\mathcal{E}^c] &\leq \sum_{1 \leq i < j \leq K} \mathbb{P}\left[\left|\langle \theta^*, \beta_i - \beta_j \rangle\right| \leq 4||\theta^* - \Gamma||\right] + \sum_{i=1}^{K} \mathbb{P}[||\beta_i|| \geq 2], \\
&= \binom{K}{2} \mathbb{P}\left[\left|\langle \theta^*, \beta_1 - \beta_2 \rangle\right| \leq 4||\theta^* - \Gamma||\right] + K\mathbb{P}[||\beta_1|| \geq 2].
\end{aligned}$$

The second equality follows from the fact that $\beta_1, \cdots, \beta_K$ are i.i.d. Now, since $||\theta^*|| \leq 1$, we have from Cauchy Schwartz that, almost-surely, $\left|\langle \theta^*, \beta_1 - \beta_2 \rangle\right| \leq ||\beta_1 - \beta_2||$. Thus,

$$\begin{aligned}
\mathbb{P}\left[|\langle \theta^*, \beta_1 - \beta_2 \rangle| \leq 4||\theta^* - \Gamma||\right] &\leq \mathbb{P}[||\beta_1 - \beta_2|| \leq 4||\theta^* - \Gamma||], \\
&\leq \mathbb{P}[||\beta_1 - \beta_2|| \leq 4\psi], \\
&= \mathbb{P}[||\beta_1 - \beta_2||^2 \leq 16\psi^2], \\
&\overset{(a)}{\leq} e^{-\frac{c_1 d}{4}},
\end{aligned}$$

where the constant c_1 depends on ψ. The first inequality follows from Cauchy Schwartz, and the fact that $||\theta^*|| \leq 1$. The last inequality follows from the fact that, $\mathbb{E}\|\beta_1 - \beta_2\|^2 = c_2$ for a constant c_2, and since $\{\beta_1, \beta_2\}$ are coordinate-wise bounded, we use standard sub-Gaussian concentration to argue that $\|\beta_1 - \beta_2\|^2$ is close to its expectation. Finally, we obtain that

$$\mathbb{P}\left(\|\beta_1 - \beta_2\|^2 - \mathbb{E}\|\beta_1 - \beta_2\|^2 \leq -t\right) \leq \exp\left(-c_3\, dt^2\right).$$

Choosing t as a constant, we obtain (a).

Finally, we also need to ensure that the context vectors $\beta_1, \cdots, \beta_K$ have norms bounded by 2. This can also be similarly be bounded by the upper tail inequality as

$$\begin{aligned}\mathbb{P}[||\beta_1|| \geq 2] &= \mathbb{P}[d||\beta_1||^2 \geq 4d], \\ &\overset{(b)}{\leq} e^{-c_4 d}.\end{aligned}$$

for a constant c_4, where inequality (b) follows from the upper-tail concentration bound for sub-Gaussian random variables. Putting this all together concludes the proof. □

Lemma D.9. *Consider a linear bandit instance with parameter θ^* with $||\theta^*|| \leq 1$ and the context vectors at each time are sampled uniformly and independently from on a distribution with support $[-c/\sqrt{d}, c/\sqrt{d}]^{\otimes d}$ for a constant c, i.e., the contexts are i.i.d. across time and arms. Let $\Gamma \in \mathbb{R}^d$ be such that $||\theta^* - \Gamma|| \leq \psi$ for a constant $\psi < \frac{1}{2\sqrt{2}}$, and $X_{1:T} = (X_1, \cdots, X_T)$ be the set of actions chosen by the Γ shifted OFUL. Then, with probability at-least $\left(1 - \binom{K}{2} e^{-c_1 d} - K e^{-c_2 d}\right)$,*

$$\mathcal{R}_T(X_{1:T}) \leq \mathcal{R}_T^{(\Gamma)}(X_{1:T}),$$

where the constants c_1 and c_2 depend on ψ.

Proof. This follows by combining Lemma D.8 and D.7. □