A Simple Approach for Non-stationary Linear Bandits

Peng Zhao, Lijun Zhang, Yuan Jiang, Zhi-Hua Zhou
National Key Laboratory for Novel Software Technology,
Nanjing University, Nanjing 210023, China
{zhaop, zhanglj, jiangy, zhouzh}@lamda.nju.edu.cn

Abstract

This paper investigates the problem of non-stationary linear bandits, where the unknown regression parameter is evolving over time. Previous studies have adopted sophisticated mechanisms, such as sliding window or weighted penalty to achieve near-optimal dynamic regret. In this paper, we demonstrate that a simple restarted strategy is sufficient to attain the same regret guarantee. Specifically, we design an UCB-type algorithm to balance exploitation and exploration, and restart it periodically to handle the drift of unknown parameters. Let T be the time horizon, d be the dimension, and P_T be the path-length that measures the fluctuation of the evolving unknown parameter, our approach enjoys an $\widetilde{O}(d^{2/3}(1+P_T)^{1/3}T^{2/3})$ dynamic regret, which is nearly optimal, matching the $\Omega(d^{2/3}(1+P_T)^{1/3}T^{2/3})$ minimax lower bound up to logarithmic factors. Empirical studies also validate the efficacy of our approach.

1 Introduction

Multi-Armed Bandits (MAB) [Robbins, 1952] models the sequential decision-making with partial information, where the player requires to choose one of the K slot machines at each iteration in order to maximize the cumulative reward. MAB is a paradigmatic instance of the exploration versus exploitation trade-offs, which is fundamental in many areas of artificial intelligence, such as reinforcement learning [Sutton and Barto, 2018] and evolutionary algorithms [Črepinšek et al., 2013].

In many real-world decision-making problems, each arm is usually associated with certain side information. Therefore, researchers start to formulate structured bandits in which the reward distributions of each arm are connected by a common but unknown parameter. Particularly, stochastic linear bandits (SLB) has received much attention [Auer, 2002, Dani et al., 2007, Chu et al., 2011, Abbasi-Yadkori et al., 2011, Li et al., 2019]. In SLB, at iteration t, the player makes a decision X_t from a feasible set $\mathcal{X} \subseteq \mathbb{R}^d$, and then observes the reward r_t satisfying

$$\mathbb{E}[r_t|X_t] = X_t^{\mathrm{T}}\theta_*, \tag{1}$$

where θ_* is an unknown regression parameter. The goal of the player is to minimize the (pseudo) regret,

$$\text{Regret}_T = T\max_{\mathbf{x}\in\mathcal{X}} \mathbf{x}^T\theta_* - \sum_{t=1}^{T} X_t^{\mathrm{T}}\theta_*. \tag{2}$$

The stochastic linear bandits problem is well-studied in literatures. By exploiting the tool of upper confidence bounds, various approaches demonstrate an $\widetilde{O}(d\sqrt{T})$ regret [Dani et al., 2007, Abbasi-Yadkori et al., 2011],[1] which matches the $\Omega(d\sqrt{T})$ lower bound established by Dani et al. [2007], up to $\log T$ factors.

However, the observation model (1) assumes that the regression parameter θ_* is fixed, which is unfortunately hard to satisfy in real-life applications, because data are usually collected in non-stationary environments. For instance, in recommender systems the regression parameter models customers' interests, which could vary over time when customers look through product pages. Therefore, it is crucial to facilitate stochastic linear bandits with capability of handling non-stationarity.

To address above issue, Cheung et al. [2019a] proposed the *non-stationary* linear bandits model, which assumes the reward r_t satisfies

$$\mathbb{E}[r_t|X_t] = X_t^{\mathrm{T}}\theta_t,$$

[1] We adopt the notation of $\widetilde{O}$ to suppress logarithmic factors in the time horizon T.

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

where θ_t is the unknown regression parameter at iteration t. Different from the standard SLB setting in (1), non-stationary linear bandits allow the unknown parameter to vary over time, whose evolution is often called *path-length* defined as $P_T = \sum_{t=2}^{T} \|\theta_{t-1} - \theta_t\|_2$, which naturally measures the non-stationarity of environments. The player's goal is to minimize the following (pseudo) *dynamic* regret,

$$\text{D-Regret}_T = \sum_{t=1}^{T} \max_{\mathbf{x} \in \mathcal{X}} \mathbf{x}^T \theta_t - \sum_{t=1}^{T} X_t^{\mathrm{T}} \theta_t, \qquad (3)$$

namely, the cumulative regret against the optimal strategy that has full information of unknown parameters.

Recently, Cheung et al. [2019a] proposed an algorithm for non-stationary linear bandits by using the sliding window least square estimator to track the evolving parameters; while Russac et al. [2019] adopted the weighted least square estimator. They both achieve $\widetilde{O}(d^{2/3} P_T^{1/3} T^{2/3})$ dynamic regret, matching the $\Omega(d^{2/3} P_T^{1/3} T^{2/3})$ lower bound established by Cheung et al. [2019a], up to $\log T$ factors. Although these two strategies attain nearly rate-optimal guarantees, their algorithms and analyses are fairly complicated. Instead, we discover that a quite simple algorithm based on the *restarted strategy* (simply running an UCB-style algorithm and restarting it periodically), surprisingly, achieves the same dynamic regret guarantee and is more efficient.

Our proposed algorithm enjoys the following three advantages compared with previous studies.

- The proposed algorithm is very simple and thus easy to analyze, only exploiting the standard self-normalized concentration inequality for classical stochastic linear bandits. Our algorithm and analysis can be further extended to the non-stationary generalized linear bandits.
- Compared with WindowUCB, the sliding window least square based approach [Cheung et al., 2019a], our approach supports online update and enjoys a one-pass manner *without* storing historical data. Meanwhile, WindowUCB demands an $O(w)$ memory where w is the window length; by contrast, our approach only requires a *constant* memory.
- Compared with WeightUCB, the weighted least square based approach [Russac et al., 2019], our approach and analysis are much simpler, without involving other complicated deviation results. Additionally, WeightUCB maintains and manipulates the covariance matrix and its variant, and thus takes a longer running time.

Overall, our approach is more friendly to the resource-constrained learning scenarios due to its simplicity.

2 Related Work

Online learning in non-stationary environments has drawn considerable attention recently, in both full-information and bandits settings. We focus on related work in the bandits setting.

Non-stationary multi-armed bandits problem with abrupt changes was first studied by Auer [2002]. Denoted by K the number of arms and by L the number of distribution changes, Auer [2002] proposed EXP3.S, a variant of EXP3, which achieves an $\widetilde{O}(\sqrt{KLT})$ regret bound when L is known. The rate is minimax optimal up to $\log T$ factors. Later studies demonstrated that $\widetilde{O}(\sqrt{KLT})$ regret is attainable by sliding window and weighted penalty strategies [Garivier and Moulines, 2011], as well as the restarted strategy [Allesiardo et al., 2017]. All these algorithms require the number of changes L as the input parameter, which is undesired in practice. Recently, Auer et al. [2019] achieved a near-optimal rate $\widetilde{O}(\sqrt{KLT})$ without knowing prior knowledge of L. On the other hand, Besbes et al. [2019] studied the non-stationary MAB with slowly changing distributions, and proved an $\widetilde{O}((K \log K)^{1/3} V_T^{1/3} T^{2/3})$ dynamic regret, where $V_T = \sum_{t=2}^{T} \|\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t-1}\|_\infty$ is the total variation of changes in reward distributions.

Non-stationary linear bandits problem was first studied by Cheung et al. [2019a]. The authors established an $\Omega(d^{2/3} P_T^{1/3} T^{2/3})$ minimax lower bound, and then proposed the WindowUCB algorithm based on the sliding window least square, achieving an $\widetilde{O}(d^{2/3} P_T^{1/3} T^{2/3})$ near-optimal dynamic regret. Nevertheless, to implement the sliding window least square, WindowUCB needs to store historical data in a buffer. A natural replacement is the weighted least square, which supports online update and enjoys both nice empirical performance and sound theoretical guarantee [Guo et al., 1993, Zhao et al., 2019]. Based on the idea, Russac et al. [2019] proposed the WeightUCB algorithm and proved that the approach attains the same dynamic regret. Nevertheless, both algorithmic design and regret analysis of WeightUCB are fairly complicated. Besides, WeightUCB needs to maintain and manipulate covariance matrix and its variant (in the same scale), which leads to an evidently longer running time. Finally, both WindowUCB and WeightUCB require the unknown quantity P_T as an input. To avoid the limitation, Cheung et al. [2019a] developed the bandits-over-bandits mechanism as a meta algorithm and finally obtained an $\widetilde{O}(d^{2/3} T^{2/3} (\max\{P_T, d^{-1/2} T^{1/4}\})^{1/3})$ parameter-free regret guarantee.

In this work, we propose a simple algorithm based on the restarted strategy for non-stationary linear bandits, and achieve near-optimal dynamic regret. We note that using the restarted strategy for non-stationary environments is not new, which has been applied in various scenarios, including non-stationary online convex optimization [Besbes et al., 2015], MAB with abrupt changes [Allesiardo et al., 2017], and MAB with gradual changes [Besbes et al., 2019]. However, to the best of our knowledge, our work is the first time to apply the restarted strategy to non-stationary linear bandits and generalized linear bandits.

3 Our Approach

In this section, we describe the proposed algorithm and present the main theoretical result, a near-optimal $\widetilde{O}(d^{2/3}(1+P_T)^{1/3}T^{2/3})$ dynamic regret for non-stationary linear bandits.

3.1 Setting and Assumptions

Setting. In non-stationary (infinite-armed) linear bandits, at each iteration t, let $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^d$ be the contextual information and r_t be the reward, and the model is assumed to be linearly parameterized, i.e.,

$$r_t = \mathbf{x}_t^{\mathrm{T}}\theta_t + \eta_t, \tag{4}$$

where $\theta_t \in \mathbb{R}^d$ is the unknown parameter and η_t is the noise satisfying certain tail condition specified below.

Assumptions. We assume the noise η_t be conditionally R-sub-Gaussian with a fixed constant $R > 0$. That is, $\mathbb{E}[\eta_t \mid X_{1:t}, \eta_{1:t-1}] = 0$, and for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}[\exp(\lambda\eta_t) \mid X_{1:t}, \eta_{1:t-1}] \le \exp\left(\lambda^2 R^2/2\right),$$

The feasible set and unknown parameters are assumed to be bounded, i.e., $\forall \mathbf{x} \in \mathcal{X}$, $\|\mathbf{x}\|_2 \le L$, and $\|\theta_t\|_2 \le S$ holds for all $t \in [T]$. For convenience, we further assume $\langle \mathbf{x}, \theta_t\rangle \le 1$, but we will keep the dependence in L and S for better comprehension of the results.

3.2 RestartUCB Algorithm

RestartUCB algorithm has two key ingredients: upper confidence bounds for the exploration–exploitation trade-off, and the restarted strategy for handling the non-stationarity of environments.

Specifically, our proposed RestartUCB algorithm proceeds in epochs. At each iteration, we first estimate the unknown regression parameter from historical data within the epoch, and then construct upper confidence bounds of the expected reward for selecting the arm. Finally, we periodically restart the algorithm to be resilient to the drift of underlying parameter θ_t.

In the following, we first specify the estimator used in the RestartUCB algorithm, then investigate its estimate error to construct upper confidence bounds, and finally describe the restarted strategy.

3.2.1 Estimator

At iteration t, we adopt the regularized least square estimator by only exploiting data in the current epoch,

$$\widehat{\theta}_t = \arg\min_{\theta} \lambda\|\theta\|_2^2 + \sum_{s=t_0}^{t-1}(X_s^{\mathrm{T}}\theta - r_s)^2, \tag{5}$$

where t_0 is the starting point of the current epoch, and $\lambda > 0$ is the regularization coefficient. Clearly, $\widehat{\theta}_t$ admits a closed-form solution as

$$\widehat{\theta}_t = V_{t-1}^{-1}\left(\sum_{s=t_0}^{t-1} r_s X_s\right), \tag{6}$$

where $V_{t-1} = \lambda I + \sum_{s=t_0}^{t-1} X_s X_s^{\mathrm{T}}$. We remark that the estimator (6) (essentially, both the terms of V_{t-1} and $\sum_{s=t_0}^{t-1} r_s X_s$) can be updated online *without* storing historical data in the memory owing to the restarted strategy. Furthermore, it is known that (5) can be *exactly* solved by the recursive least square algorithm, whose solution is provably equivalent to the closed-form expression (6). This feature can further accelerate our approach in that it saves the computation of the inverse of covariance matrix V_{t-1}, which is arguably the most time-consuming step at each iteration.

By contrast, Cheung et al. [2019a] adopted the following sliding window least square estimator,

$$\widehat{\theta}_t^{\mathrm{sw}} = (V_{t-1}^{\mathrm{sw}})^{-1}\left(\sum_{s=1\vee(t-w)}^{t-1} r_s X_s\right), \tag{7}$$

where $V_{t-1}^{\mathrm{sw}} = \lambda I + \sum_{s=1\vee(t-w)}^{t-1} X_s X_s^{\mathrm{T}}$ is the covariance matrix formed by historical data in the sliding window and $w > 0$ is the window length. For online update, WindowUCB will remove the oldest data item in the window and then add the new item. So it requires to store the nearest w data items in the memory for future update, resulting in an $O(w)$ space complexity which cannot be regarded as a constant because the setting of w depends on the time horizon T.

3.2.2 Upper Confidence Bounds

Based on the estimator $\widehat{\theta}_t$ in (6), we further construct upper confidence bounds for the expected reward. To this end, it is required to investigate the estimate error. Inspired by the analysis of WindowUCB [Cheung et al., 2019a], we have the following result.

Lemma 1. *For any $t \in [T]$ and $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds for all $\mathbf{x} \in \mathcal{X}$,*

$$|\mathbf{x}^{\mathrm{T}}(\theta_t - \widehat{\theta}_t)| \le L\sum_{p=t_0}^{t-1}\|\theta_p - \theta_{p+1}\|_2 + \beta_t\|\mathbf{x}\|_{V_{t-1}^{-1}}, \tag{8}$$

where β_t is the radius of confidence region,

$$\beta_t = \sqrt{\lambda} S + R\sqrt{2\log\frac{1}{\delta} + d\log\left(1+\frac{(t-t_0)L^2}{\lambda d}\right)}. \quad (9)$$

The estimate error (8) essentially suggests an upper confidence bound of the expected reward $\mathbf{x}^{\mathrm{T}}\theta_t$. Hence, we adopt the principle of *optimism in the face of uncertainty* [Auer, 2002] and choose the arm that maximizes its upper confidence bound,

$$\begin{aligned} X_t &= \arg\max_{\mathbf{x}\in\mathcal{X}} \left\{\mathbf{x}^{\mathrm{T}}\widehat{\theta}_t + \mathsf{ub}(\mathbf{x})\right\} \\ &= \arg\max_{\mathbf{x}\in\mathcal{X}} \left\{\mathbf{x}^{\mathrm{T}}\widehat{\theta}_t + \beta_t\|\mathbf{x}\|_{V_{t-1}^{-1}}\right\}, \end{aligned} \quad (10)$$

where $\mathsf{ub}(\mathbf{x}) = L\sum_{p=t_0}^{t-1}\|\theta_p - \theta_{p+1}\|_2 + \beta_t\|\mathbf{x}\|_{V_{t-1}^{-1}}$.

So at iteration t, the algorithm first solves the estimator based on (6), then obtains the confidence radius β_t by (9), and finally pulls the arm X_t according to the selection criteria (10).

3.2.3 Restarted Strategy

To handle the changes of unknown regression parameters, RestartUCB algorithm proceeds in epochs and restarts the procedure every H iterations, as illustrated in Figure 1. In each epoch, RestartUCB performs the UCB-style algorithm as described in the last subsection. We summarize overall procedures in Algorithm 1.

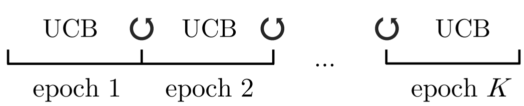


Figure 1: Illustration of RestartUCB algorithm.

Note that although the length of each epoch can be varied, we find that a fixed length is sufficient to achieve near-optimal theoretical guarantees.

3.3 Theoretical Guarantees

We show that RestartUCB algorithm enjoys a nearly optimal dynamic regret notwithstanding its simplicity.

First, we analyze the regret within each epoch (Theorem 1). Then, we sum over epochs to obtain the guarantee of the whole time horizon (Theorem 2).

Theorem 1. *For each epoch $\mathcal{E}$ whose size is H and any $\delta \in (0,1)$, with probability at least $1-2\delta$, the dynamic regret within the epoch is upper bounded by*

$$\textit{D-Regret}(\mathcal{E}) \le 2LH\mathcal{P}(\mathcal{E}) + 2\beta_H\sqrt{2dH\log\left(1+\frac{L^2H}{\lambda d}\right)},$$

Algorithm 1 RestartUCB

Input: time horizon T, epoch size H, confidence δ

Set epoch counter $j = 1$
while $j \le \lceil T/H \rceil$ **do**
  Set $\tau = (j-1)H$
  Initialize $X_\tau \in \mathcal{X}$
  $V_\tau = \lambda I_d$
  for $t = \tau+1, \ldots, \tau+H-1$ **do**
    Compute $\widehat{\theta}_t$ by (6) and β_t by (9)
    Select $X_t = \arg\max_{\mathbf{x}\in\mathcal{X}}\{\mathbf{x}^{\mathrm{T}}\widehat{\theta}_t + \beta_t\|\mathbf{x}\|_{V_{t-1}^{-1}}\}$
    Receive the reward r_t
    Update $V_t = V_{t-1} + X_t X_t^{\mathrm{T}}$
  end for
  Set $j = j+1$
end while

where $\beta_H = \sqrt{\lambda}S + R\sqrt{2\log\frac{1}{\delta} + d\log\left(1+\frac{HL^2}{\lambda d}\right)}$, and $\mathcal{P}(\mathcal{E})$ denotes the path-length within epoch $\mathcal{E}$, i.e., $\mathcal{P}(\mathcal{E}) = \sum_{t\in\mathcal{E}}\|\theta_{t-1} - \theta_t\|_2$.

By summing regret over epochs, we obtain dynamic regret over of the whole time horizon.

Theorem 2. *Algorithm 1* RestartUCB *enjoys the following dynamic regret guarantee,*

$$\textit{D-Regret}_T \le \widetilde{O}\big(HP_T + dT/\sqrt{H}\big). \quad (11)$$

By setting the epoch size $H = H^* = \lfloor (dT/P_T)^{2/3} \rfloor$, we achieve an $\widetilde{O}(d^{2/3}P_T^{1/3}T^{2/3})$ dynamic regret.

Remark 1. Cheung et al. [2019a] established an $\Omega(d^{2/3}P_T^{1/3}T^{2/3})$ minimax lower bound for the non-stationary linear bandits. Hence, the $\widetilde{O}(d^{2/3}P_T^{1/3}T^{2/3})$ dynamic regret exhibited in Theorem 2 is minimax optimal in all parameters up to $\log T$ factors.

Remark 2. As shown in Theorem 2, the setting of optimal epoch size H^* requires prior information of P_T, which is generally unavailable. We will discuss how to remove the undesired dependence in the next section.

4 Extensions

In this section, we first apply the restarted strategy to non-stationary generalized linear bandits, and then discuss how to adapt to the unknown path-length P_T.

4.1 Generalized Linear Bandits

Setting. Generalized linear bandits (GLB) assumes a *link function* $\mu : \mathbb{R} \mapsto \mathbb{R}$ such that $r_t = \mu(\mathbf{x}_t^{\mathrm{T}}\theta_t) + \eta_t$, where $\theta_t \in \mathbb{R}^d$ is the unknown parameter and can change over time. Evidently, linear and logistic models are two of special cases of the generalized linear model, with $\mu(x) = x$ and $\mu(x) = 1/(1+e^{-x})$, respectively.

For non-stationary GLB, dynamic regret is used as the performance measure, defined as

$$\text{D-Regret}_T = \sum_{t=1}^{T} \max_{\mathbf{x}\in\mathcal{X}} \mu(\mathbf{x}^{\mathrm{T}}\theta_t) - \mu(X_t^{\mathrm{T}}\theta_t). \tag{12}$$

Assumptions. We make the same assumptions with those of linear bandits as stated in Section 3.1, including tail conditions of noise, boundedness of feasible set, and boundedness of unknown regression parameters. In addition, following previous studies of GLB [Filippi et al., 2010, Li et al., 2017], we make two additional standard assumptions on the link function. Concretely, the link function is assumed to be k_μ-Lipschitz, and continuously differentiable with $c_\mu = \inf_{\{\theta,\mathbf{x}\in\mathcal{X}\}} \mu'(\theta^{\mathrm{T}}\mathbf{x}) > 0$. For simplicity, we do not impose the constraint on the parameter θ_t, which can be otherwise compensated by introducing additional projection step as done in the pioneering work of Filippi et al. [2010].

Estimator. The maximum quasi-likelihood estimator is typically adopted in GLB [Filippi et al., 2010, Li et al., 2017], where $\widehat{\theta}_t$ is set as the solution of $\sum_{s=t_0}^{t-1}\left(r_s - \mu(X_s^{\mathrm{T}}\theta)\right) X_s = 0$. Nevertheless, the estimator requires $\sum_{s=t_0}^{t-1} X_s X_s^{\mathrm{T}}$ to be invertible for all iterations in the regret analysis, which is a rather strong assumption. To address the issue, we solve the estimator $\widehat{\theta}_t$ by the following *regularized* estimation equation

$$\lambda c_\mu \theta + \sum_{s=t_0}^{t-1} \left(\mu(X_s^{\mathrm{T}}\theta) - r_s\right) X_s = 0, \tag{13}$$

where $\lambda > 0$ is the regularization coefficient. We have the following guarantee on the estimate error.

Lemma 2. *For any $t \in [T]$ and $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds for all $\mathbf{x} \in \mathcal{X}$,*

$$\begin{aligned} &|\mu(\mathbf{x}^{\mathrm{T}}\widehat{\theta}_t) - \mu(\mathbf{x}^{\mathrm{T}}\theta_t)| \\ &\leq \frac{k_\mu}{c_\mu}\left(k_\mu L \sum_{p=t_0}^{t-1} \|\theta_p - \theta_{p+1}\|_2 + \bar{\beta}_t \|\mathbf{x}\|_{V_{t-1}^{-1}}\right), \end{aligned}$$

where $\bar{\beta}_t$ is the radius of confidence region,

$$\bar{\beta}_t = c_\mu\sqrt{\lambda}S + R\sqrt{2\log\frac{1}{\delta} + d\log\left(1 + \frac{(t-t_0)L^2}{\lambda d}\right)}. \tag{14}$$

Based on Lemma 2, we can now specify the action selection criteria at iteration t as,

$$X_t = \arg\max_{\mathbf{x}\in\mathcal{X}} \left\{\mu(\mathbf{x}^{\mathrm{T}}\widehat{\theta}_t) + \frac{k_\mu}{c_\mu}\bar{\beta}_t\|\mathbf{x}\|_{V_t^{-1}}\right\}. \tag{15}$$

The algorithm for non-stationary generalized linear bandits (RestartGLB) is similar to that for linear bandits. At iteration t, RestartGLB algorithm first solves the estimator by (13), and then obtains the confidence radius $\bar{\beta}_t$ based on (14), and finally pulls the arm X_t according to (15).

Note that similar to the existing algorithm (based on the sliding window) for non-stationary GLB [Cheung et al., 2019b], our algorithm also requires to store the whole learning history to solve the estimation equation (13) at each iteration and thus is inefficient. Although there exist efficient algorithms for stationary GLB [Zhang et al., 2016, Jun et al., 2017], it remains open for non-stationary generalized linear bandits.

We have the following guarantee for RestartGLB.

Theorem 3. *The* RestartGLB *algorithm enjoys the dynamic regret of*

$$\textit{D-Regret}_T \leq \widetilde{O}\big(HP_T + dT/\sqrt{H}\big). \tag{16}$$

By setting the epoch size $H = H^* = \lfloor (dT/P_T)^{2/3} \rfloor$, we achieve an $\widetilde{O}(d^{2/3}P_T^{1/3}T^{2/3})$ dynamic regret.

The above dynamic regret is also minimax optimal for GLB up to logarithmic factors [Cheung et al., 2019a].

4.2 Adapting to Unknown Non-stationarity

Notice that in Theorem 2 and Theorem 3, the configuration of the optimal epoch size H^* requires knowledge of path-length P_T, which is generally unavailable. We compensate the lack of this information via the meta-expert framework studied in previous non-stationary bandits literatures [Agarwal et al., 2017, Cheung et al., 2019a, Zhao et al., 2020]. Specifically, we run the EXP3 algorithm [Auer et al., 2002] as a meta algorithm to adaptively choose the optimal epoch size. The method is referred to as *Bandits-over-Bandits* (BOB) [Cheung et al., 2019a], and we defer details to Appendix B.

RestartUCB algorithm together with BOB mechanism leads to the following dynamic regret without requiring the prior knowledge of the path-length P_T.

Theorem 4. RestartUCB *together with Bandits-over-Bandits mechanism enjoys the dynamic regret of*

$$\textit{D-Regret}_T \leq \widetilde{O}\Big(d^{\frac{2}{3}}T^{\frac{2}{3}}\big(\max\{P_T, d^{-\frac{1}{2}}T^{\frac{1}{4}}\}\big)^{\frac{1}{3}}\Big), \tag{17}$$

without requiring the path-length P_T ahead of time.

Remark 3. When the path-length P_T is sufficiently large ($P_T \geq d^{-\frac{1}{2}}T^{\frac{1}{4}}$), the attained dynamic regret in (17) becomes $\widetilde{O}(d^{2/3}P_T^{1/3}T^{2/3})$, demonstrating that in this case the approach achieves the minimax optimal dynamic regret guarantee without requiring prior knowledge of P_T. However, it remains open on how to obtain rate-optimal and parameter-free dynamic regret when the path-length P_T is small.

5 Analysis

In this section, we provide proofs of theoretical results presented in the previous two sections.

5.1 Analysis of Linear Bandits

We provide proofs of Lemma 1 and Theorems 1, 2.

Proof of Lemma 1. From the model assumption (4) and the estimator (6), we can verify that the estimate error can be decomposed as,

$$\widehat{\theta}_t - \theta_t = V_{t-1}^{-1}\left(\sum_{s=t_0}^{t-1} X_s X_s^{\mathrm{T}}(\theta_s - \theta_t) + \sum_{s=t_0}^{t-1} \eta_s X_s - \lambda\theta_t\right).$$

Therefore, by Cauchy-Schwartz inequality, we know that for any $\mathbf{x} \in \mathcal{X}$,

$$|\mathbf{x}^{\mathrm{T}}(\widehat{\theta}_t - \theta_t)| \leq \|\mathbf{x}\|_2 \cdot A_t + \|\mathbf{x}\|_{V_{t-1}^{-1}} \cdot B_t, \tag{18}$$

where

$$A_t = \left\|V_{t-1}^{-1}\left(\sum\nolimits_{s=t_0}^{t-1} X_s X_s^{\mathrm{T}}(\theta_s - \theta_t)\right)\right\|_2,$$
$$B_t = \left\|\sum\nolimits_{s=t_0}^{t-1} \eta_s X_s - \lambda\theta_t\right\|_{V_{t-1}^{-1}}.$$

These two terms can be bounded separately, summarized in the following lemma for a better presentation. We present the proof of Lemma 3 in Appendix A.

Lemma 3. *A_t and B_t can be upper bounded as follows.*

- $A_t \leq \sum_{p=t_0}^{t-1} \|\theta_p - \theta_{p+1}\|_2$;
- $B_t \leq \beta_t$, *where β_t is the confidence radius (9).*

Based on the inequality (18), Lemma 3, and the boundedness of the feasible set, we have

$$\langle \mathbf{x}, \widehat{\theta}_t - \theta_t\rangle \leq L\sum_{p=1}^{t}\|\theta_p - \theta_{p+1}\|_2 + \beta_t\|\mathbf{x}\|_{V_{t-1}^{-1}},$$

which competes the proof. □

Proof of Theorem 1. Due to Lemma 1 and the fact that $X_t^*, X_t \in \mathcal{X}$, each of the following holds with probability at least $1 - \delta$,

$$\langle X_t^*, \theta_t\rangle \leq \langle X_t^*, \widehat{\theta}_t\rangle + L\sum_{p=t_0}^{t-1}\|\theta_p - \theta_{p+1}\|_2 + \beta_t\|X_t^*\|_{V_{t-1}^{-1}},$$

$$\langle X_t, \theta_t\rangle \geq \langle X_t, \widehat{\theta}_t\rangle + L\sum_{p=t_0}^{t-1}\|\theta_p - \theta_{p+1}\|_2 + \beta_t\|X_t\|_{V_{t-1}^{-1}}.$$

By the union bound, the following holds with probability at least $1 - 2\delta$,

$$\begin{aligned}&\langle X_t^*, \theta_t\rangle - \langle X_t, \theta_t\rangle\\ &\leq \langle X_t^*, \widehat{\theta}_t\rangle - \langle X_t, \widehat{\theta}_t\rangle + 2L\sum_{p=t_0}^{t-1}\|\theta_p - \theta_{p+1}\|_2\\ &\qquad + \beta_t(\|X_t^*\|_{V_{t-1}^{-1}} + \|X_t\|_{V_{t-1}^{-1}})\\ &\leq 2L\sum_{p=t_0}^{t-1}\|\theta_p - \theta_{p+1}\|_2 + 2\beta_t\|X_t\|_{V_{t-1}^{-1}},\end{aligned}$$

where the last step comes from the following implication of the arm selection criteria (10),

$$\langle X_t^*, \widehat{\theta}_t\rangle + \beta_t\|X_t^*\|_{V_{t-1}^{-1}} \leq \langle X_t, \widehat{\theta}_t\rangle + \beta_t\|X_t\|_{V_{t-1}^{-1}}.$$

Hence, dynamic regret within epoch $\mathcal{E}$ is bounded by,

$$\begin{aligned}\text{D-Regret}(\mathcal{E}) &\leq \sum_{t\in\mathcal{E}} 2L\sum_{p=t_0}^{t-1}\|\theta_p - \theta_{p+1}\|_2 + 2\beta_t\|X_t\|_{V_{t-1}^{-1}}\\ &\leq 2LH\mathcal{P}(\mathcal{E}) + 2\beta_H\sqrt{2dH\log\left(1 + \frac{L^2H}{\lambda d}\right)},\end{aligned}$$

where the last inequality holds due to the standard elliptical potential lemma (Lemma 4), whose statement and proof are presented in Appendix C. □

Proof of Theorem 2. By taking the union bound over the dynamic regret of all $\lceil T/H\rceil$ epochs, we know that the following holds with probability at least $1 - 2/T$,

$$\begin{aligned}\text{D-Regret}_T &= \sum_{s=1}^{\lceil T/H\rceil}\text{D-Regret}(\mathcal{E}_s)\\ &\leq 2LHP_T + 2T\widetilde{\beta}_H\sqrt{\frac{2d}{H}\log\left(1 + \frac{L^2H}{\lambda d}\right)},\end{aligned}$$

where $\widetilde{\beta}_H = \sqrt{\lambda}S + R\sqrt{2\log(T\lceil\frac{T}{H}\rceil) + d\log\left(1 + \frac{HL^2}{\lambda d}\right)}$. Ignoring logarithmic factors, we finally obtain that

$$\text{D-Regret}_T \leq \widetilde{O}\big(HP_T + dT/\sqrt{H}\big).$$

By setting $H = H^* = \lfloor(dT/P_T)^{2/3}\rfloor$, we achieve an $\widetilde{O}(d^{2/3}P_T^{1/3}T^{2/3})$ near-optimal dynamic regret. □

5.2 Analysis of Generalized Linear Bandits

We provide proofs of Lemma 2 and Theorem 3.

Proof of Lemma 2. Define the function

$$g_t(\theta) = \lambda c_\mu \theta + \sum_{s=t_0}^{t-1}\mu(X_s^{\mathrm{T}}\theta)X_s, \tag{19}$$

then by the mean value theorem, we know that

$$g_t(\widehat{\theta}_t) - g_t(\theta_t) = G_t(\widehat{\theta}_t - \theta_t) \tag{20}$$

where $G_t = \int_0^1 \nabla g_t(s\theta_t + (1-s)\widehat{\theta}_t)\,\mathrm{d}s$. Notice that for any θ, the gradient of g_t is

$$\nabla g_t(\theta) = \lambda c_\mu I + \sum_{s=t_0}^{t-1} \mu'(X_s^{\mathrm{T}}\theta) X_s X_s^{\mathrm{T}} \succeq c_\mu V_{t-1},$$

which clearly implies $G_t \succeq c_\mu V_{t-1}$. From the function (19) and the estimation equation (13), we conclude that $g_t(\widehat{\theta}_t) - g_t(\theta_t)$ equals to

$$-\sum_{s=t_0}^{t-1} \big(\mu(X_s^{\mathrm{T}}\theta_t) - \mu(X_s^{\mathrm{T}}\theta_s)\big) X_s + \sum_{s=t_0}^{t-1} \eta_s X_s - \lambda c_\mu \theta_t.$$

Due to the Lipschitz continuity of the link function, $|\mu(\mathbf{x}^{\mathrm{T}}\widehat{\theta}_t) - \mu(\mathbf{x}^{\mathrm{T}}\theta_t)| \leq k_\mu |\langle \mathbf{x}, \widehat{\theta}_t - \theta_t\rangle|$. Meanwhile, from previous derivations, we have

$$\begin{aligned}
&|\mathbf{x}^{\mathrm{T}}(\widehat{\theta}_t - \theta_t)| \overset{(20)}{=} |\mathbf{x}^{\mathrm{T}} G_t^{-1}(g_t(\widehat{\theta}_t) - g_t(\theta_t))| \\
\leq\, & \underbrace{L \left\| G_t^{-1} \left(\sum_{s=t_0}^{t-1} \big(\mu(X_s^{\mathrm{T}}\theta_t) - \mu(X_s^{\mathrm{T}}\theta_s)\big) X_s \right) \right\|_2}_{\texttt{term (a)}} \\
&+ \underbrace{\left| \mathbf{x}^{\mathrm{T}} G_t^{-1} \left(\textstyle\sum_{s=t_0}^{t-1} \eta_s X_s \right) \right|}_{\texttt{term (b)}} + \underbrace{\left| \mathbf{x}^{\mathrm{T}} G_t^{-1} (\lambda c_\mu \theta_t) \right|}_{\texttt{term (c)}}
\end{aligned}$$

First, term (a) can be bounded as

$$\texttt{term (a)} \leq \frac{L k_\mu}{c_\mu} \sum_{p=t_0}^{t-1} \|\theta_p - \theta_{p+1}\|_2,$$

whose proof is basically same as that of Lemma 3 and can be found in Appendix H of Cheung et al. [2019b].

Then, term (b) can be upper bounded by the self-normalized concentration inequality [Abbasi-Yadkori et al., 2011, Theorem 1],

$$\texttt{term (b)} \leq R\sqrt{2\log\frac{1}{\delta} + d\log\left(1 + \frac{(t-t_0)L^2}{\lambda d}\right)} \|\mathbf{x}\|_{V_{t-1}^{-1}}.$$

Next, by noticing $G_t \succeq c_\mu V_{t-1}$, we obtain that

$$\texttt{term (c)} \leq \lambda \|\mathbf{x}\|_{V_{t-1}^{-1}} \|\theta_t\|_{V_{t-1}^{-1}} \leq \sqrt{\lambda} S \|\mathbf{x}\|_{V_{t-1}^{-1}}.$$

We compete the proof by combining upper bounds of all these three terms. $\square$

Proof of Theorem 3. Similar to the proof of Theorem 1, we know that with probability at least $1-2\delta$, dynamic regret within the epoch $\mathcal{E}$ (i.e., D-Regret($\mathcal{E}$)) is at most

$$\frac{2k_\mu^2}{c_\mu} LH\mathcal{P}(\mathcal{E}) + \frac{2k_\mu}{c_\mu}\bar{\beta}_H \sqrt{2dH\log\left(1 + \frac{L^2 H}{\lambda d}\right)},$$

where $\bar{\beta}_H$ is defined in (14).

By taking the union bound over all the epochs, we conclude that dynamic regret is bounded by

$$\frac{2k_\mu}{c_\mu}\left(k_\mu L H P_T + T\bar{\beta}_H \sqrt{\frac{2d}{H}\log\left(1 + \frac{L^2 H}{\lambda d}\right)} \right),$$

which is of order $\widetilde{O}\big(HP_T + dT/\sqrt{H}\big)$. $\square$

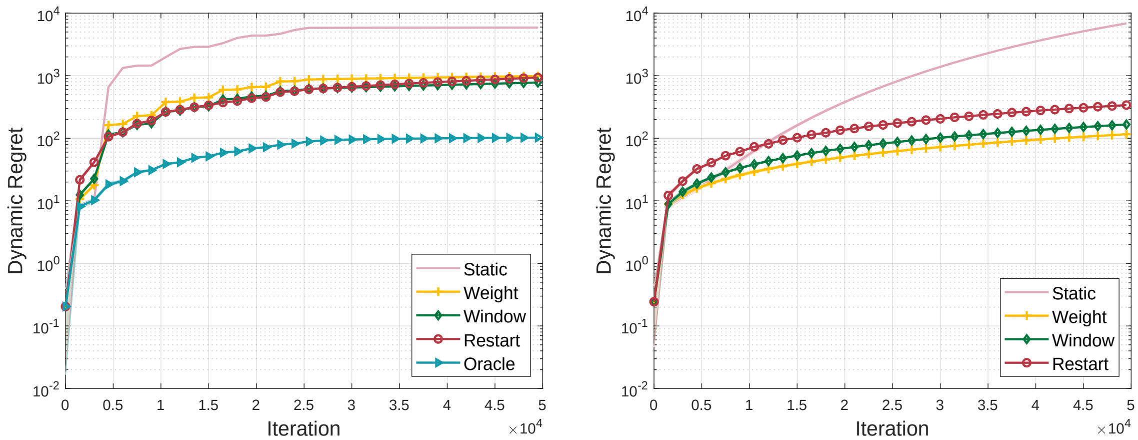
6 Empirical Studies

Despite the focus of this paper is on the theoretical aspect, we present empirical studies to further evaluate the proposed approach.

Contenders. We study two kinds of non-stationary environments: the underlying parameter is *abruptly changing* or *gradually changing*. Besides, We compare RestartUCB to (a) WindowUCB, based on the sliding window least square [Cheung et al., 2019a]; (b) WeightUCB, based on the weighted least square [Russac et al., 2019]; (c) StaticUCB, the algorithm designed for stationary linear bandits [Abbasi-Yadkori et al., 2011]. In the scenario of abrupt change, we additionally compare with OracleRestartUCB, which knows the exact information of change points and restarts the algorithm when reaching a change point.

Settings. In abruptly-changing environments, the unknown regression parameter θ_t is periodically set as $[1,0]$, $[-1,0]$, $[0,1]$, $[0,-1]$ in the first half of iterations, and $[1,0]$ for the remaining iterations. In gradually-changing environments, θ_t is moved from $[1,0]$ to $[-1,0]$ on the unit circle continuously. In both scenarios, we set $T = 50{,}000$ and number of arms $n = 20$. The feature is sampled from normal distribution $\mathcal{N}(0,1)$ and rescaled such that $L = 1$. The random noise is generated according to $\mathcal{N}(0, 0.1)$. Since the path-length P_T is available in the synthetic datasets, as suggested by the theory, we set the weight $\gamma = 1 - (dT/P_T)^{-2/3}$ for WeightUCB, the window size $w = \lfloor (dT/P_T)^{2/3} \rfloor$ for WindowUCB, and the epoch size $H = \lfloor (dT/P_T)^{2/3} \rfloor$ for RestartUCB. The simulation is repeated for 50 times, and we report the average and standard deviation.

Results. Figure 2 shows performance comparisons of different approaches, measured by the (pseudo-) dynamic regret, in logarithmic scale. In the *abruptly-changing environments*, OracleRestartUCB is definitely the best since it knows exact information of change points, and StaticUCB ranks the last as it does not take the non-stationarity into consideration. RestartUCB and WindowUCB have comparable performance, better than WeightUCB. In the *gradually-changing environments*, WeightUCB ranks the first, followed by



(a) abrupt change (b) gradual change

Figure 2: Comparisons of different approaches in terms of dynamic regret, in logarithmic scale.

WindowUCB and RestartUCB. Nevertheless, as will be shown later, WeightUCB takes a significantly longer running time than our approach.

Figure 3 reports the running time. We can see that time costs of RestartUCB, WindowUCB, and StaticUCB are basically the same, whereas WeightUCB requires a significantly longer running time, almost twice the cost of other contenders. The reason lies in the fact that WeightUCB algorithm involves the computation of the inverse of covariance matrix $V_t \in \mathbb{R}^{d\times d}$ and its variant $\widetilde{V}_t \in \mathbb{R}^{d\times d}$, while other three methods maintain and manipulate only one covariance matrix.

From empirical studies, we conclude that RestartUCB algorithm is more favored in abruptly-changing environments empirically, highly comparable to WindowUCB. We note that RestartUCB has an additional advantage over WindowUCB, RestartUCB supports the one-pass update without storing historical data, whereas WindowUCB has to maintain a buffer and thus needs to scan data multiple times owing to the sliding window strategy. On the other hand, compared with WeightUCB, our approach only maintains one covariance matrix and is thus simpler and faster. It is noteworthy that our approach can be further accelerated by the recursive least square, which will save the computation of the inverse of covariance matrix and can be particularly desired in high-dimensional problems.

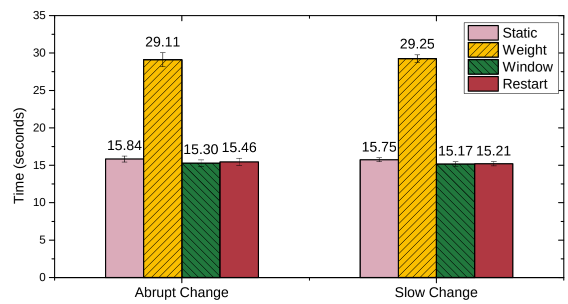


Figure 3: Comparisons in terms of running time.

7 Conclusion

In this paper, we study the problem of non-stationary linear bandits, where the unknown regression parameter θ_t is changing over time. We propose a simple algorithm based on the restarted strategy, which enjoys strong theoretical guarantees notwithstanding its simplicity. Concretely, our algorithm enjoys an $\widetilde{O}(d^{2/3}(1+P_T)^{1/3}T^{2/3})$ dynamic regret, and the rate is near-optimal, matching the minimax lower bound up to $\log T$ factors. The restarted strategy can be extended to the non-stationary generalized linear bandits and also achieves a near-optimal regret. Empirical studies validate the efficacy of the proposed approach, particularly in the abruptly-changing environments.

In the future, we would like to study how to design algorithms for non-stationary linear bandits that achieve rate-optimal dynamic regret without prior information. Moreover, as mentioned earlier, existing algorithms for non-stationary generalized linear bandits are inefficient in the sense that they require to store historical data in memory to compute the estimator, and we will explore more efficient algorithms for non-stationary GLB.

Acknowledgment

This research was supported by the National Science Foundation of China (61921006, 61673201), and the Collaborative Innovation Center of Novel Software Technology and Industrialization. We are grateful for the anonymous reviewers for their helpful comments.

References

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 2312–2320, 2011.

A. Agarwal, H. Luo, B. Neyshabur, and R. E. Schapire. Corralling a band of bandit algorithms. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, pages 12–38, 2017.

R. Allesiardo, R. Féraud, and O.-A. Maillard. The non-stationary stochastic multi-armed bandit problem. *International Journal of Data Science and Analytics*, 3(4):267–283, 2017.

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

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

P. Auer, P. Gajane, and R. Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Proceedings of the 32nd Conference On Learning Theory (COLT)*, volume 99, pages 138–158, 2019.

O. Besbes, Y. Gur, and A. J. Zeevi. Non-stationary stochastic optimization. *Operations Research*, 63(5): 1227–1244, 2015.

O. Besbes, Y. Gur, and A. J. Zeevi. Optimal exploration–exploitation in a multi-armed bandit problem with non-stationary rewards. *Stochastic Systems*, 9(4):319–337, 2019.

W. C. Cheung, D. Simchi-Levi, and R. Zhu. Learning to optimize under non-stationarity. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1079–1087, 2019a.

W. C. Cheung, D. Simchi-Levi, and R. Zhu. Hedging the drift: Learning to optimize under non-stationarity. *arXiv preprint*, arXiv:1903.01461, 2019b.

W. Chu, L. Li, L. Reyzin, and R. E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 208–214, 2011.

V. Dani, T. P. Hayes, and S. M. Kakade. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems 20 (NIPS)*, pages 345–352, 2007.

S. Filippi, O. Cappé, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems 23 (NIPS)*, pages 586–594, 2010.

A. Garivier and E. Moulines. On upper-confidence bound policies for switching bandit problems. In *Proceedings of the 22nd International Conference on Algorithmic Learning Theory (ALT)*, volume 6925, pages 174–188, 2011.

L. Guo, L. Ljung, and P. Priouret. Performance analysis of the forgetting factor RLS algorithm. *International Journal of Adaptive Control and Signal Processing*, 7(6):525–538, 1993.

K.-S. Jun, A. Bhargava, R. D. Nowak, and R. Willett. Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 99–109, 2017.

L. Li, Y. Lu, and D. Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 2071–2080, 2017.

Y. Li, Y. Wang, and Y. Zhou. Nearly minimax-optimal regret for linearly parameterized bandits. In *Proceedings of the 32nd Conference on Learning Theory (COLT)*, pages 2173–2174, 2019.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Y. Russac, C. Vernade, and O. Cappé. Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 12040–12049, 2019.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

M. Črepinšek, S.-H. Liu, and M. Mernik. Exploration and exploitation in evolutionary algorithms: A survey. *ACM Computing Surveys*, 45(3):35:1–35:33, 2013.

L. Zhang, T. Yang, R. Jin, Y. Xiao, and Z.-H. Zhou. Online stochastic linear optimization under one-bit feedback. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 392–401, 2016.

P. Zhao, X. Wang, S. Xie, L. Guo, and Z.-H. Zhou. Distribution-free one-pass learning. *IEEE Transaction on Knowledge and Data Engineering*, 2019.

P. Zhao, G. Wang, L. Zhang, and Z.-H. Zhou. Bandit convex optimization in non-stationary environments. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

A Proof of Lemma 3

Proof of Lemma 3. We first prove the upper bound of A_t. The essential proof is actually due to Cheung et al. [2019a] in analyzing sliding window based approach. For self-containedness, we restate here in the notations of our proposed restarted strategy.

$$
\begin{aligned}
& \left\| V_{t-1}^{-1} \left(\sum_{s=t_0}^{t-1} X_s X_s^{\mathrm{T}} (\theta_s - \theta_t) \right) \right\|_2 \\
= & \left\| V_{t-1}^{-1} \left(\sum_{s=t_0}^{t-1} X_s X_s^{\mathrm{T}} \Big(\sum_{p=s}^{t-1} (\theta_p - \theta_{p+1}) \Big) \right) \right\|_2 \\
= & \left\| V_{t-1}^{-1} \left(\sum_{p=t_0}^{t-1} \Big(\sum_{s=t_0}^{p} X_s X_s^{\mathrm{T}} (\theta_p - \theta_{p+1}) \Big) \right) \right\|_2 && (21) \\
\leq & \sum_{p=t_0}^{t-1} \left\| V_{t-1}^{-1} \Big(\sum_{s=t_0}^{p} X_s X_s^{\mathrm{T}} \Big) (\theta_p - \theta_{p+1}) \right\|_2 && (22) \\
\leq & \sum_{p=t_0}^{t-1} \lambda_{\max} \left(V_{t-1}^{-1} \Big(\sum_{s=t_0}^{p} X_s X_s^{\mathrm{T}} \Big) \right) \|\theta_p - \theta_{p+1}\|_2 && (23) \\
\leq & \sum_{p=t_0}^{t-1} \|\theta_p - \theta_{p+1}\|_2, && (24)
\end{aligned}
$$

where (21) holds by rearranging over the index pair of (s,p), (22) holds due to the triangle inequality, (23) and (24) can be obtained by the same argument in Appendix B of Cheung et al. [2019b]. We thus obtain the upper bound of A_t.

We proceed to prove the upper bound of B_t. From the self-normalized concentration inequality [Abbasi-Yadkori et al., 2011, Theorem 1], restated in Theorem 5 of Appendix C, we know that

$$
\begin{aligned}
& \left\| \sum_{s=t_0}^{t-1} \eta_s X_s \right\|_{V_{t-1}^{-1}} \\
\overset{(32)}{\leq} & \sqrt{2R^2 \log \left(\frac{\det(V_{t-1})^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right)} \\
\leq & R \sqrt{2 \log \frac{1}{\delta} + d \log \left(1 + \frac{(t-t_0)L^2}{d} \right)},
\end{aligned}
$$

where the last inequality is obtained from the analysis of the determinant, as shown in the proof of Lemma 4.

Meanwhile, since $V_{t-1} \succeq \lambda I$, we know that

$$
\|\lambda \theta_t\|_{V_{t-1}^{-1}}^2 \leq 1/\lambda_{\min}(V_{t-1}) \|\lambda \theta_t\|_2^2 \leq \frac{1}{\lambda} \|\lambda \theta_t\|_2^2 \leq \lambda S^2.
$$

Therefore, the upper bound of B_t can be immediately obtained by combining the above inequalities. □

B Bandit-over-Bandits Mechanism and Proof of Theorem 4

The RestartUCB algorithm requires prior information of the path-length P_T, which is generally unknown. Such a limitation can be avoided by utilizing the Bandits-over-bandits (BOB) mechanism, proposed by Cheung et al. [2019a] in designing parameter-free algorithm for non-stationary linear bandits based on sliding window least square estimator.

In the following, we first describe how to apply the BOB mechanism to eliminate the requirement of the unknown path-length in RestartUCB. Then, we present the proof of Theorem 4.

B.1 RestartUCB with BOB Mechanism

We name the RestartUCB algorithm with Bandit-over-Bandits mechanism as "RestartUCB-BOB", whose main idea is illustrated in Figure 4.

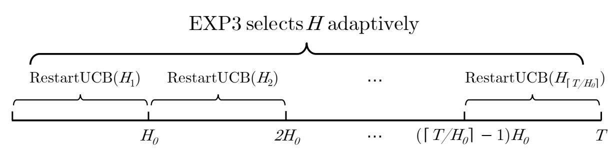


Figure 4: Illustration of Bandit-over-Bandits mechanism with application to RestartUCB algorithm.

From a high-level view, although the exact value of the optimal epoch size (or equivalently, the path-length P_T) is not clear, we can make some random guesses of its possible value, since the P_T is always bounded. Then, we can use a certain meta-algorithm to adaptively track the best epoch size, based on the returned reward returned. Specifically, The RestartUCB-BOB algorithm first sets an update period H_0, and then runs the RestartUCB with a particular epoch size in each period, and the epoch size will be adaptively adjusted by employing EXP3 [Auer et al., 2002] as the meta-algorithm. We refer the reader to Section 7.3 of Cheung et al. [2019b] for more descriptions of design motivations and algorithmic details.

In the configuration of RestartUCB-BOB, we set $H_0 = \lceil d\sqrt{T} \rceil$ and the pool of epoch sizes J as

$$
J = \{ H_i = \lfloor (d/(2S))^{2/3} \cdot 2^{i-1} \rfloor \mid i = 1, 2, \cdots, N \},
$$

where $N = \lceil \ln(d^{1/3} T^{1/2} (2S)^{2/3}) \rceil + 1$.

Denoted by $H_{\min}$ ($H_{\max}$) the minimal (maximal) epoch size in the pool J, we know that

$$
H_{\min} = \lfloor (d/(2S))^{2/3} \rfloor, H_{\max} = \lfloor d\sqrt{T} \rfloor \leq H_0. \quad (25)
$$

B.2 Proof of Theorem 4

Proof of Theorem 4. We begin with the following decomposition of the dynamic regret.

$$\begin{aligned}&\sum_{t=1}^{T}\langle X_t^*,\theta_t\rangle - \langle X_t,\theta_t\rangle\\ =&\underbrace{\sum_{t=1}^{T}\langle X_t^*,\theta_t\rangle - \sum_{i=1}^{\lceil T/H_0\rceil}\sum_{t=(i-1)H_0+1}^{iH_0}\langle X_t(H^\dagger),\theta_t\rangle}_{\texttt{term (i)}}\\ &+\underbrace{\sum_{i=1}^{\lceil T/H_0\rceil}\sum_{t=(i-1)H_0+1}^{iH_0}\langle X_t(H^\dagger),\theta_t\rangle - X_t(H_i),\theta_t\rangle}_{\texttt{term (ii)}},\end{aligned}$$

where $H^\dagger$ is the best epoch size to approximate the optimal epoch size H^* in the pool J, and $H^* = \lfloor (dT/(1+P_T))^{2/3}\rfloor$. Hence, it suffices to bound terms (i) and (ii). In the following, we consider two cases, either $(1+P_T)\geq d^{-1/2}T^{1/4}$ or $(1+P_T)< d^{-1/2}T^{1/4}$.

Case 1. when $(1+P_T)\geq d^{-1/2}T^{1/4}$.

In this case, it is easy to verify that $H^*\leq H_{\max}$ and we thus conclude that H^* lies in the the range of $[H_{\min},H_{\max}]$. Furthermore, from the configuration of the pool J, we confirm that there exists an epoch size $H^\dagger\in J$ such that $H^\dagger\leq H^*\leq 2H^\dagger$. So term (i) can be upper bounded by

$$\begin{aligned}\texttt{term (i)} &\leq \sum_{i=1}^{\lceil T/H_0\rceil}\widetilde{O}\Big(H^\dagger P_i + \frac{dH_0}{\sqrt{H^\dagger}}\Big) && (26)\\ &= \widetilde{O}\Big(H^\dagger P_T + \frac{dT}{\sqrt{H^\dagger}}\Big) && (27)\\ &\leq \widetilde{O}\Big(H^* P_T + \frac{dT}{\sqrt{2H^*}}\Big)\\ &= \widetilde{O}(d^{2/3}P_T^{1/3}T^{2/3}),\end{aligned}$$

where (26) is due to Theorem 2 and P_i denotes the path-length in the i-th update period. (27) follows by summing over all update periods, and the last inequality holds since the optimal epoch size H^* is provably in the range of $[H_{\min},H_{\max}]$ and satisfies $H^\dagger\leq H^*\leq 2H^\dagger$.

Next, we bound the term (ii),

$$\begin{aligned}\texttt{term (ii)} &\leq \widetilde{O}(\sqrt{H_0NT})\\ &\leq \widetilde{O}(d^{1/2}T^{3/4}) && (28)\\ &\leq \widetilde{O}(d^{2/3}T^{2/3}(1+P_T)^{1/3}),\end{aligned}$$

where the first inequality follows by the same argument as in the sliding window based approach [Cheung et al., 2019b, Proposition 1], building upon the regret analysis of EXP3. In addition, the last inequality holds due to the fact that $(1+P_T)\geq d^{-1/2}T^{1/4}$ implies

$$d^{1/2}T^{3/4} = d^{2/3}T^{2/3}d^{-1/3}T^{1/6} \leq d^{2/3}T^{2/3}(1+P_T)^{1/3}.$$

Hence, by combining the upper bounds of term (i) and term (ii), we know that the dynamic regret of RestartUCB-BOB is bounded by $\widetilde{O}(d^{2/3}T^{2/3}(1+P_T)^{1/3})$ under the condition of $(1+P_T)\geq d^{-1/2}T^{1/4}$.

Case 2. when $(1+P_T)< d^{-1/2}T^{1/4}$.

In this case, we cannot guarantee that the optimal epoch size H^* lies in the range of $[H_{\min},H_{\max}]$, so we set $H^\dagger$ as H_0,

$$\begin{aligned}\texttt{term (i)} &\leq \widetilde{O}\Big(H^\dagger P_T + \frac{dT}{\sqrt{H^\dagger}}\Big)\\ &\leq \widetilde{O}\Big(H_0 P_T + \frac{dT}{\sqrt{H_0}}\Big)\\ &= \widetilde{O}\Big(d\sqrt{T}P_T + d^{1/2}T^{3/4}\Big)\\ &\leq \widetilde{O}\Big(d^{1/2}T^{3/4}\Big)\end{aligned}$$

where the last inequality holds by exploiting the condition of $(1+P_T)\leq d^{-1/2}T^{1/4}$. The result in conjunction with the upper bound of term (ii) in (28) gives the $\widetilde{O}(d^{1/2}T^{3/4})$ dynamic regret under this condition.

Finally, note that the dynamic regret of above two cases can be rewritten in the following unified form,

$$\texttt{term (i)}+\texttt{term (ii)} \leq \widetilde{O}\Big(d^{\frac{2}{3}}T^{\frac{2}{3}}\big(\max\{P_T, d^{-\frac{1}{2}}T^{\frac{1}{4}}\}\big)^{\frac{1}{3}}\Big).$$

Hence, we complete the proof of Theorem 4. □

C Technical Lemmas

In this section, we provide several technical lemmas that frequently used in the proofs.

Theorem 5 (Self-Normalized Bound for Vector-Valued Martingales [Abbasi-Yadkori et al., 2011, Theorem 1]). *Let $\{F_t\}_{t=0}^\infty$ be a filtration. Let $\{\eta_t\}_{t=0}^\infty$ be a real-valued stochastic process such that η_t is F_t-measurable and conditionally R-sub-Gaussian for some $R>0$, namely,*

$$\forall\lambda\in\mathbb{R},\quad \mathbb{E}[\exp(\lambda\eta_t)|F_{t-1}]\leq \exp\left(\frac{\lambda^2R^2}{2}\right). \qquad (29)$$

Let $\{X_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued stochastic process such that X_t is F_{t-1}-measurable. Assume that V is a $d\times d$ positive definite matrix. For any $t\geq 0$, define

$$\bar{V}_t = V + \sum_{\tau=1}^{t}X_\tau X_\tau^{\mathrm{T}},\quad S_t = \sum_{\tau=1}^{t}\eta_\tau X_\tau. \qquad (30)$$

Then, for any $\delta > 0$, with probability at least $1-\delta$, for all $t \geq 0$,

$$\|S_t\|_{\bar{V}_t^{-1}}^2 \leq 2R^2 \log\left(\frac{\det(\bar{V}_t)^{1/2}\det(V)^{-1/2}}{\delta}\right). \quad (31)$$

Lemma 4 (Elliptical Potential Lemma)**.** *Suppose $U_0 = \lambda I$, $U_t = U_{t-1} + X_t X_t^{\mathrm{T}}$, and $\|X_t\|_2 \leq L$, then*

$$\sum_{t=1}^{T} \|U_{t-1}^{-\frac{1}{2}} X_t\|_2 \leq \sqrt{2dT \log\left(1 + \frac{L^2 T}{\lambda d}\right)}. \quad (32)$$

Proof. First, we have the following decomposition,

$$U_t = U_{t-1} + X_t X_t^{\mathrm{T}} = U_{t-1}^{\frac{1}{2}}(I + U_{t-1}^{-\frac{1}{2}} X_t X_t^{\mathrm{T}} U_{t-1}^{-\frac{1}{2}}) U_{t-1}^{\frac{1}{2}}.$$

Taking the determinant on both sides, we get

$$\det(U_t) = \det(U_{t-1})\det(I + U_{t-1}^{-\frac{1}{2}} X_t X_t^{\mathrm{T}} U_{t-1}^{-\frac{1}{2}}),$$

which in conjunction with Lemma 5 yields

$$\begin{aligned}\det(U_t) =& \det(U_{t-1})(1 + \|U_{t-1}^{-\frac{1}{2}} X_t\|_2^2) \\ \geq& \det(U_{t-1})\exp(\|U_{t-1}^{-\frac{1}{2}} X_t\|_2^2/2).\end{aligned}$$

Note that in the first inequality, we utilize the fact that $1 + x \geq \exp(x/2)$ holds for any $x \in [0,1]$. By taking advantage of the telescope structure, we have

$$\sum_{t=1}^{T} \|U_{t-1}^{-\frac{1}{2}} X_t\|_2^2 \leq 2\log\frac{\det(U_T)}{\det(U_0)} \leq 2d\log\left(1 + \frac{L^2 T}{\lambda d}\right),$$

where the last inequality follows from the fact that $\mathrm{Tr}(U_T) \leq \mathrm{Tr}(U_0) + L^2 T = \lambda d + L^2 T$, and thus $\det(U_T) \leq (\lambda + L^2 T/d)^d$.

Therefore, Cauchy-Schwartz inequality gives,

$$\begin{aligned}\sum_{t=1}^{T} \|U_{t-1}^{-\frac{1}{2}} X_t\|_2 \leq& \sqrt{T \sum_{t=1}^{T} \|U_{t-1}^{-\frac{1}{2}} X_t\|_2^2} \\ \leq& \sqrt{2dT \log\left(1 + \frac{L^2 T}{\lambda d}\right)}.\end{aligned}$$

□

Lemma 5.

$$\det(I + \mathbf{v}\mathbf{v}^{\mathrm{T}}) = 1 + \|\mathbf{v}\|_2^2. \quad (33)$$

Proof. Notice that

(i) $(I + \mathbf{v}\mathbf{v}^{\mathrm{T}})\mathbf{v} = (1 + \|\mathbf{v}\|_2^2)\mathbf{v}$, therefore, $\mathbf{v}$ is its eigenvector with $(1 + \|\mathbf{v}\|_2^2)$ as the eigenvalue;

(ii) $(I + \mathbf{v}\mathbf{v}^{\mathrm{T}})\mathbf{v}^{\perp} = \mathbf{v}^{\perp}$, therefore, $\mathbf{v}^{\perp} \perp \mathbf{v}$ is its eigenvector with 1 as the eigenvalue.

Consequently, $\det(I + \mathbf{v}\mathbf{v}^{\mathrm{T}}) = 1 + \|\mathbf{v}\|_2^2$. □