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 suf- cient to attain the same regret guarantee. Speci cally, 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 uctuation of the evolving unknown parameter, our approach enjoys an $\tilde{O}(d^{2/3}(1 + P_T)^{1/3}T^{2/3})$ dynamic regret, which is nearly optimal, matching the $(d^{2/3}(1+P_T)^{1/3}T^{2/3})$ minimax lower bound up to logarithmic factors. Empirical studies also validate the e cacy 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-o s, which is fundamental in many areas of arti cial intelligence, such as reinforcement learning [Sutton and Barto, 2018] and evolutionary algorithms [Crepinsek 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

$$\mathbb{E}[r_t|X_t] = X_t^{\mathrm{T}} \quad ; \tag{1}$$

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

$$\mathrm{Regret}_T = T \max_{x \in X} x^T \quad - \quad {}^{T=}$$

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

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 $\tilde{O}(d^{2/3}(1+P_T)^{1/3}T^{2/3})$ 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 $x_t \in \mathcal{X} \subseteq \mathbb{R}^d$ be the contextual information and r_t be the reward, and the model is assumed to be linearly parameterized, i.e.,

$$r_t = x_t^{\mathrm{T}}\theta_t + \eta_t, \tag{4}$$

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 $\lambda \in \mathbb{R}$,

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

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

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

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

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

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,

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

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 $\hat{\theta}_t$ 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* $x \in \mathcal{X}$,

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

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,

$$\begin{aligned} X_t &= \operatorname*{arg\,max}_{\mathbf{x}\in\mathcal{X}} \left\{\mathbf{x}^{\mathrm{T}}\hat{\theta}_t + \mathrm{ub}(\mathbf{x})\right\} \\ &= \operatorname*{arg\,max}_{\mathbf{x}\in\mathcal{X}} \left\{\mathbf{x}^{\mathrm{T}}\hat{\theta}_t + \beta_t\|\mathbf{x}\|_{V_t^{-1}}\right\}, \end{aligned} \quad (10)$$

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

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.

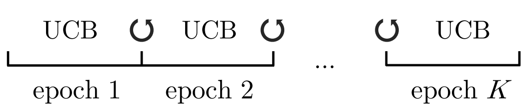


Figure 1: Illustration of RestartUCB algorithm.

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 $t = \tau+1, \dots, \tau+H-1$ **do**
    Compute $\hat{\theta}_t$ by (6) and β_t by (9)
    Select $X_t = \arg\max_{\mathbf{x}\in\mathcal{X}}\{\mathbf{x}^{\mathrm{T}}\hat{\theta}_t + \beta_t\|\mathbf{x}\|_{V_{t-1}^{-1}}\}$
    Receive the reward r_t
    Update $V_t = V_{t-1} + X_tX_t^{\mathrm{T}}$
  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,*

$$\textit{D-Regret}_T \le \tilde{\mathcal{O}}\left(HP_T + dT/\sqrt{H}\right). \quad (11)$$

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

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

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 $c_\mu = \inf_{\{\theta, \mathbf{x}\in\mathcal{X}\}} \mu'(\theta^\top \mathbf{x}) > 0$. 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 $\hat{\theta}_t$ is set as the solution of $\sum_{s=t_0}^{t-1} \big(r_s - \mu(X_s^\top \theta)\big) X_s = 0$. Nevertheless, the estimator requires $\sum_{s=t_0}^{t-1} X_s X_s^\top$ to be invertible for all iterations in the regret analysis, which is a rather strong assumption. To address the issue, we solve the estimator $\hat{\theta}_t$ by the following *regularized* estimation equation

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

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 $\mathbf{x} \in \mathcal{X}$,*

$$|\mu(\mathbf{x}^\top \hat{\theta}_t) - \mu(\mathbf{x}^\top \theta_t)| \le \frac{k_\mu}{c_\mu}\left(k_\mu L \sum_{p=t_0}^{t-1} \|\theta_p - \theta_{p+1}\|_2 + \beta_t \|\mathbf{x}\|_{V_{t-1}^{-1}}\right),$$

where β_t is the radius of confidence region,

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

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

$$X_t = \operatorname*{arg\,max}_{\mathbf{x}\in\mathcal{X}} \left\{\mu(\mathbf{x}^\top \hat{\theta}_t) + \frac{k_\mu}{c_\mu}\beta_t \|\mathbf{x}\|_{V_t^{-1}}\right\}. \tag{15}$$

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 β_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*

$$\text{D-Regret}_T \le \tilde{\mathcal{O}}\left(HP_T + dT/\sqrt{H}\right). \tag{16}$$

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

$$\text{D-Regret}_T \le \tilde{\mathcal{O}}\left(d^{\frac{2}{3}} T^{\frac{2}{3}} \left(\max\{P_T, d^{-\frac{1}{2}}T^{\frac{1}{4}}\}\right)^{\frac{1}{3}}\right), \tag{17}$$

without requiring the path-length P_T ahead of time.

Remark 3. When the path-length P_T is sufficiently large ($P_T \ge d^{-\frac{1}{2}}T^{\frac{1}{4}}$), the attained dynamic regret in (17) becomes $\tilde{\mathcal{O}}(d^{2/3}P_T^{1/3}T^{2/3})$, 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,

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

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

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

where

$$A_t = \left\| V_{t-1}^{-1} \sum_{s=t_0}^{t-1} X_s X_s^{\mathrm{T}}(\theta_s - \theta_t)\right\|_2,$$

$$B_t = \left\| \sum_{s=t_0}^{t-1} \eta_s X_s - \lambda\theta_t \right\|_{V_{t-1}^{-1}}.$$

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.*

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

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

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

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 $1-\delta$,

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

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

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

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

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

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

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

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

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 then by the mean value theorem, we know that

$$g_t(\hat\theta_t) - g_t(\theta_t) = G_t(\hat\theta_t - \theta_t) \tag{20}$$

where $G_t = \int_0^1 \nabla g_t(s\theta_t + (1-s)\hat\theta_t)\,ds$. Notice that for any θ, the gradient of g_t is

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

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

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

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

$$\begin{aligned}
|x^{\mathrm{T}}(\hat\theta_t - \theta_t)| &\overset{(20)}{=} \left|x^{\mathrm{T}}G_t^{-1}(g_t(\hat\theta_t) - g_t(\theta_t))\right| \\
&\le \underbrace{\left|L\, G_t^{-1}\sum_{s=t_0}^{t-1}\eta^{t-s-1}\left(\mu(X_s^{\mathrm{T}}\theta_t) - \mu(X_s^{\mathrm{T}}\theta_s)\right)X_s\right|}_{\text{term (a)}} \\
&\quad + \underbrace{\left|x^{\mathrm{T}}G_t^{-1}\sum_{s=t_0}^{t-1}\eta^{t-s-1}\eta_s X_s\right|}_{\text{term (b)}} + \underbrace{\left|x^{\mathrm{T}}G_t^{-1}(\lambda c_\mu\theta_t)\right|}_{\text{term (c)}}
\end{aligned}$$

First, term (a) can be bounded as

$$\text{term (a)} \le \frac{Lk_\mu}{c_\mu}\sum_{p=t_0}^{t-1}\|\theta_p - \theta_{p+1}\|_2,$$

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-normalize concentration inequality [Abbasi-Yadkori et al., 2011, Theorem 1],

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

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

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

We compete the proof by combining upper bounds of all these three terms. ☐

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

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

where β_H is defined in (14).

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

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

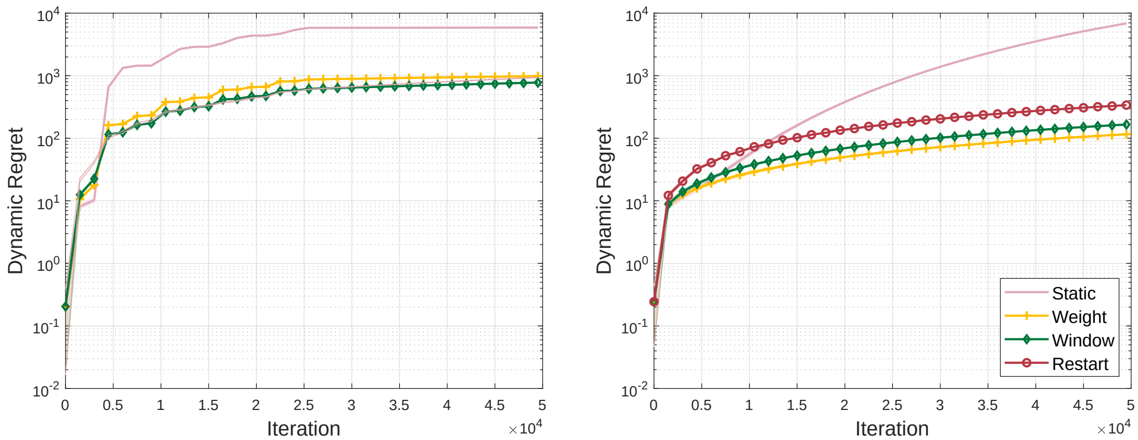
which is of order $\tilde{\mathcal{O}}\left(HP_T + dT/\sqrt{H}\right)$. ☐

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



(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 $\tilde{V}_t \in \mathbb{R}^{d \times d}$, 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.

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

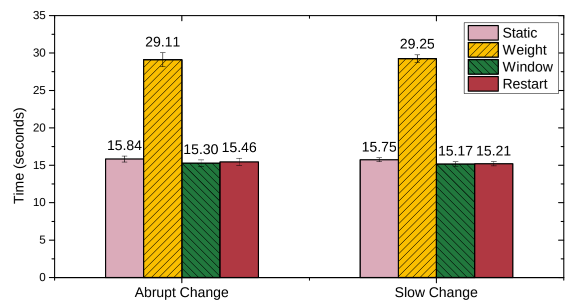


Figure 3: Comparisons in terms of running time.

$\widetilde{O}(d^{2/3}(1 + P_T)^{1/3}T^{2/3})$ 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. Pal, and C. Szepesvari. 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. Feraud, 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 con dence bounds for exploitation-exploration trade-o s. *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 payo 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. Cappe, A. Garivier, and C. Szepesvari. 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-con dence 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. Cappe. 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. Crepinsek, 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.