Non-stationary Linear Bandits Revisited

Peng Zhao ZHAOP@LAMDA.NJU.EDU.CN
Lijun Zhang ZHANGLJ@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China

Abstract

In this note, we revisit non-stationary linear bandits, a variant of stochastic linear bandits with a time-varying underlying regression parameter. Existing studies develop various algorithms and show that they enjoy an $\widetilde{\mathcal{O}}(T^{2/3}P_T^{1/3})$ dynamic regret, where T is the time horizon and P_T is the path-length that measures the fluctuation of the evolving unknown parameter. However, we discover that a serious technical flaw makes the argument ungrounded. We revisit the analysis and present a fix. Without modifying original algorithms, we can prove an $\widetilde{\mathcal{O}}(T^{3/4}P_T^{1/4})$ dynamic regret for these algorithms, slightly worse than the rate as was anticipated. We also show some impossibility results for the key quantity concerned in the regret analysis. Note that the above dynamic regret guarantee requires an oracle knowledge of the path-length P_T. Combining the bandit-over-bandit mechanism, we can also achieve the same guarantee in a parameter-free way.

1. Introduction

Non-stationary linear bandits (Cheung et al., 2019a) is a variant of classical stochastic linear bandits, in which the underlying regression parameter can change over time. Concretely speaking, 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 $\mathbb{E}[r_t|X_t] = X_t^\top \theta_t$, where θ_t is the unknown regression parameter at iteration t. Different from the standard stochastic setting, non-stationary linear bandits allows the unknown parameter to vary over time, whose evolution is often called path-length defined as $P_T = \sum_{t=2}^T \|\theta_{t-1} - \theta_t\|_2$, naturally measuring the non-stationarity of environments. The player's goal is to minimize the (pseudo) *dynamic regret*, defined as

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

namely, the cumulative regret against the optimal strategy that has full information of unknown parameters. The upper bound of dynamic regret depends on both horizon T and path-length P_T.

Related work. In the pioneering work (Cheung et al., 2019a), the authors propose the non-stationary linear bandits model and establish the minimax lower bound of $\Omega(T^{2/3}P_T^{1/3})$. On the upper bound sides, Cheung et al. (2019a) design the WindowUCB algorithm to deal with changing environments via the upper confidence bound strategy with sliding-window least square estimator, and they prove an $\widetilde{\mathcal{O}}(T^{2/3}P_T^{1/3})$ dynamic regret when the path-length P_T is known a priori. The $\widetilde{\mathcal{O}}(\cdot)$-notation suppresses logarithmic factors in T. When such prior information is unavailable, the authors further develop the bandit-over-bandit mechanism to attain a parameter-free algorithm without requiring knowledge of P_T, but with a suboptimal rate of $\widetilde{\mathcal{O}}(T^{2/3}(\max\{P_T, d^{-1/2}T^{1/4}\})^{1/3})$. In the subsequent studies, Russac et al. (2019) propose the WeightUCB algorithm based on the weighted least square estimator and also prove an $\widetilde{\mathcal{O}}(T^{2/3}P_T^{1/3})$ dynamic regret; Zhao et al. (2020) develop the RestartUCB algorithm and show that this simple restarted strategy is sufficient to achieve the same dynamic regret guarantee. There are also other variants based on perturbation methods (Kim and Tewari, 2020) and extensions to generalized linear bandits (Russac et al., 2020).

Our result. In this note, we reveal that the proof of a key lemma in previous analysis (Cheung et al., 2019b, Lemma 3) has serious technical flaw, which makes the final dynamic regret guarantee ungrounded. We provide a fix for the analysis and prove an $\widetilde{\mathcal{O}}(T^{3/4}P_T^{1/4})$ dynamic regret for the three mainstream algorithms (WindowUCB (Cheung et al., 2019a), WeightUCB (Russac et al., 2019), and RestartUCB (Zhao et al., 2020)). The attained dynamic regret bound is slightly worse than the $\widetilde{\mathcal{O}}(T^{2/3}P_T^{1/3})$ bound as was anticipated. We further show some impossibility results for the quantity concerned in the regret analysis. Note that the aforementioned dynamic regret guarantees require an oracle knowledge of the path-length P_T. It turns out that combining the bandit-over-bandit mechanism, this $\widetilde{\mathcal{O}}(T^{3/4}P_T^{1/4})$ dynamic regret bound is also achievable in a parameter-free way, namely, without requiring prior knowledge of P_T ahead of time.

2. Our Result

In this section, we revisit a key technical lemma that is commonly used for regret analysis in non-stationary linear bandits algorithms including WindowUCB, WeightUCB, and RestartUCB. For simplicity, we will focus on the RestartUCB algorithm, since the simple restarted strategy is sufficient to deliver the same dynamic regret guarantee as indicated by Zhao et al. (2020).

We first restate the problem setup and assumptions below, and then review the RestartUCB algorithm in Section 2.1. Next, we spot the technical flaws of original proofs in Section 2.2. We present new analysis for a fix and further discussion some impossibility results in Section 2.3. Finally, we give the overall dynamic regret analysis in Section 2.4.

Problem Setup. 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 that is linearly parameterized by

$$r_t = \mathbf{x}_t^\top \theta_t + \eta_t, \tag{2}$$

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

Assumptions. The noise η_t is conditionally R-sub-Gaussian with a fixed constant $R > 0$, i.e., $\mathbb{E}[\eta_t | X_{1:t}, \eta_{1:t-1}] = 0$ and $\forall \lambda \in \mathbb{R}$, $\mathbb{E}[\exp(\lambda\eta_t)|X_{1:t}, \eta_{1:t-1}] \leq \exp(\lambda^2 R^2/2)$. The feasible set and unknown parameters are bounded, i.e., $\forall \mathbf{x} \in \mathcal{X}$, $\|\mathbf{x}\|_2 \leq L$, and $\|\theta_t\|_2 \leq S$ holds for all $t \in [T]$.

2.1 RestartUCB Algorithm

The RestartUCB algorithm (Zhao et al., 2020) proceeds in epochs. At each iteration, the algorithm first estimates the unknown regression parameter from historical data within the epoch, and then constructs upper confidence bounds of the expected reward for selecting the arm. Finally, the algorithm is periodically restarted to be resilient to the drift of underlying parameter θ_t.

Algorithm 1 RestartUCB (Zhao et al., 2020)

Input: time horizon T, epoch size H, confidence δ, regularizer λ, scaling parameters S and L

Set epoch counter $j = 1$
while $j \leq \lceil T/H \rceil$ **do**
  Set $\tau = (j-1)H$
  Initialize $X_\tau \in \mathcal{X}$
  $V_\tau = \lambda I_d$, $S_\tau = \mathbf{0}$
  for $t = \tau+1, \ldots, \tau+H-1$ **do**
    Compute $\widehat{\theta}_t = V_{t-1}^{-1} S_t$, and β_t by (5) with $t_0 = \tau$
    Select $X_t = \arg\max_{\mathbf{x}\in\mathcal{X}} \{\mathbf{x}^\top \widehat{\theta}_t + \beta_t \|\mathbf{x}\|_{V_{t-1}^{-1}}\}$
    Receive the reward r_t
    Update $V_t = V_{t-1} + X_t X_t^\top$, $S_t = S_{t-1} + r_t X_t$
  end for
  Set $j = j+1$
end while

Specifically, at iteration t, RestartUCB adopts the regularized least square estimator by only exploiting data in the current epoch,

$$\widehat{\theta}_t = \arg\min_\theta \lambda \|\theta\|_2^2 + \sum_{s=t_0}^{t-1} (X_s^\top \theta - r_s)^2, \tag{3}$$

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 $\widehat{\theta}_t = V_{t-1}^{-1}\big(\sum_{s=t_0}^{t-1} r_s X_s\big)$, where $V_{t-1} = \lambda I + \sum_{s=t_0}^{t-1} X_s X_s^\top$. The algorithm adopts the principle of "optimism in the face of uncertainty" (Auer, 2002) and chooses the arm that maximizes its upper confidence bound,

$$X_t = \arg\max_{\mathbf{x}\in\mathcal{X}} \big\{\mathbf{x}^\top \widehat{\theta}_t + \beta_t \|\mathbf{x}\|_{V_{t-1}^{-1}}\big\}, \tag{4}$$

where β_t is the radius of confidence region set by

$$\beta_t = \sqrt{\lambda} S + R \sqrt{2 \log \frac{1}{\delta} + d \log \left(1 + \frac{(t-t_0)L^2}{\lambda d}\right)}. \tag{5}$$

The overall algorithm is summarized in Algorithm 1.

2.2 Technical flaws in previous analysis

Previous studies show an $\widetilde{\mathcal{O}}(T^{2/3} P_T^{1/3})$ dynamic regret for non-stationary linear bandits (including WindowUCB, WeightUCB, and RestartUCB, etc), however, the technical reasoning suffers from some gaps and makes the overall regret guarantee ungrounded. In the following, we will spot the flaws of their original proofs, and then present the new analysis for a fix in the next subsection.

Indeed, the flaw appears in a key technical lemma for regret analysis of WindowUCB, the pioneering study of non-stationary linear bandits (Cheung et al., 2019b, Lemma 3). The flaw is unfortunately inherited by the later studies, including WeightUCB (Russac et al., 2019, Theorem 2), RestartUCB (Zhao et al., 2020, Lemma 3), and perturbation based method (Kim and Tewari, 2020, Theorem 7). To be more concrete, Lemma 3 of Cheung et al. (2019b) (also see Lemma 3 of Zhao et al. (2020)) claims that for any $t \in [T]$,

$$\left\| V_{t-1}^{-1} \Big(\sum_{s=t_0}^{t-1} X_s X_s^\top (\theta_s - \theta_t) \Big) \right\|_2 \leq \sum_{p=t_0}^{t-1} \|\theta_p - \theta_{p+1}\|_2. \tag{6}$$

We restate their proof of the above claim (Cheung et al., 2019b, Appendix B) as follows:

$$\begin{aligned}
\left\| V_{t-1}^{-1} \Big(\sum_{s=t_0}^{t-1} X_s X_s^\top (\theta_s - \theta_t) \Big) \right\|_2 &= \left\| V_{t-1}^{-1} \Big(\sum_{s=t_0}^{t-1} X_s X_s^\top \Big(\sum_{p=s}^{t-1} (\theta_p - \theta_{p+1}) \Big) \Big) \right\|_2 \\
&= \left\| V_{t-1}^{-1} \Big(\sum_{p=t_0}^{t-1} \Big(\sum_{s=t_0}^{p} X_s X_s^\top (\theta_p - \theta_{p+1}) \Big) \Big) \right\|_2 \\
&\leq \sum_{p=t_0}^{t-1} \left\| V_{t-1}^{-1} \Big(\sum_{s=t_0}^{p} X_s X_s^\top \Big) (\theta_p - \theta_{p+1}) \right\|_2 \\
&\leq \sum_{p=t_0}^{t-1} \sigma_{\max} \left(V_{t-1}^{-1} \Big(\sum_{s=t_0}^{p} X_s X_s^\top \Big) \right) \|\theta_p - \theta_{p+1}\|_2 \\
&\leq \sum_{p=t_0}^{t-1} \|\theta_p - \theta_{p+1}\|_2,
\end{aligned} \tag{7}$$

where $\sigma_{\max}(\cdot)$ is the largest singular value. The key is the last step (7) but its proof is questionable: they need to show the following results holds universally for all $p \in \{t_0, \ldots, t-1\}$,

$$\sigma_{\max} \left(V_{t-1}^{-1} \Big(\sum_{s=t_0}^{p} X_s X_s^\top \Big) \right) \leq 1. \tag{8}$$

To this end, denoted by $A = \sum_{s=t_0}^{p} X_s X_s^\top$, the authors show that $V_{t-1}^{-1} A$ shares the same characteristics polynomial with $V_{t-1}^{-1/2} A V_{t-1}^{-1/2}$, namely, $\det(\eta I - V_{t-1}^{-1} A) = \det(\eta I - V_{t-1}^{-1/2} A V_{t-1}^{-1/2})$ holds for any η. Since $V_{t-1}^{-1/2} A V_{t-1}^{-1/2}$ is clearly symmetric positive semi-definite, they claim that

$$\mathbf{z}^\top V_{t-1}^{-1} A \mathbf{z} \geq 0 \tag{9}$$

also holds for $\mathbf{z} \in \mathcal{S}(1) = \{\mathbf{x} \mid \|\mathbf{x}\|_2 = 1\}$, which is crucial for their remaining proof.

$$\sigma_{\max} \left(V_{t-1}^{-1} \Big(\sum_{s=t_0}^{p} X_s X_s^\top \Big) \right) = \sup_{\mathbf{z} \in \mathcal{S}(1)} \mathbf{z}^\top V_{t-1}^{-1} \left(\sum_{s=t_0}^{p} X_s X_s^\top \right) \mathbf{z} \tag{10}$$

$$\overset{(9)}{\leq} \sup_{\mathbf{z} \in \mathcal{S}(1)} \left\{ \mathbf{z}^\top V_{t-1}^{-1} \Big(\sum_{s=t_0}^{p} X_s X_s^\top \Big) \mathbf{z} + \mathbf{z}^\top V_{t-1}^{-1} \Big(\sum_{s=p+1}^{t-1} X_s X_s^\top \Big) \mathbf{z} + \lambda \mathbf{z}^\top V_{t-1}^{-1} \mathbf{z} \right\} \tag{11}$$

$$= \sup_{\mathbf{z} \in \mathcal{S}(1)} \mathbf{z}^\top V_{t-1}^{-1} V_{t-1} \mathbf{z} = 1.$$

However, we identify that there are two issues in the above arguments, First, the step in (10) doubtful. For a matrix $M \in \mathbb{R}^{m\times n}$, we have $\|M\|_2 = \sup_{\|\mathbf{x}\|_2=1}\sup_{\|\mathbf{y}\|_2=1}|\mathbf{y}^\top M\mathbf{x}|$, while it is not warranted that $\|M\|_2 = \sup_{\|\mathbf{z}\|_2=1}|\mathbf{z}^\top M\mathbf{z}|$ which is seemingly important for the following arguments. Regardless of this first issue, the second issue about the claim (9) and the result in (11). We discover that the claim (9) is even more severe. We discover that the claim (9) is ungrounded (at least its current proof cannot stand for the correctness). The big caveat is that $V_{t-1}^{-1}A \in \mathbb{R}^{d\times d}$ is *not* guaranteed to be symmetric. The logic behind the claim is that, suppose $P, Q \in \mathbb{R}^{d\times d}$ are with the same characteristics polynomial, i.e., $\det(\eta I - Q) = \det(\eta I - P)$ holds for any η, and meanwhile P is symmetric positive semi-definite (which guarantees $\mathbf{z}^\top P\mathbf{z} \geq 0$ for any $\mathbf{z} \in \mathbb{R}^d$), then we can also have $\mathbf{z}^\top Q\mathbf{z} \geq 0$ for any $\mathbf{z} \in \mathbb{R}^d$. Unfortunately, the reasoning is not correct, and we give a simple counterexample. Let P be the 2-dim identity matrix $[1, 0; 0, 1]$, and $Q = [1, -10; 0, 1]$ is an *asymmetric* matrix, then clearly $\det(\eta I - P) = \det(\eta I - Q) = (\eta - 1)^2$ is true for any η; however, $\mathbf{z}^\top Q\mathbf{z} \geq 0$ does not hold in general, for example, $\mathbf{z}^\top Q\mathbf{z} = -8 < 0$ when $\mathbf{z} = (1, 1)^\top$.

To logic behind the claim is that, suppose $P, Q \in \mathbb{R}^{d\times d}$ are with the same characteristics polynomial, i.e., $\det(\eta I - Q) = \det(\eta I - P)$ holds for any η, and meanwhile P is symmetric positive semi-definite (which guarantees $\mathbf{z}^\top P\mathbf{z} \geq 0$ for any $\mathbf{z} \in \mathbb{R}^d$), then we can also have $\mathbf{z}^\top Q\mathbf{z} \geq 0$ for any $\mathbf{z} \in \mathbb{R}^d$. Unfortunately, the reasoning is not correct, and we give a simple counterexample. Let P be the 2-dim identity matrix $[1, 0; 0, 1]$, and $Q = [1, -10; 0, 1]$ is an *asymmetric* matrix, then clearly $\det(\eta I - P) = \det(\eta I - Q) = (\eta - 1)^2$ is true for any η; however, $\mathbf{z}^\top Q\mathbf{z} \geq 0$ does not hold in general, for example, $\mathbf{z}^\top Q\mathbf{z} = -8 < 0$ when $\mathbf{z} = (1, 1)^\top$.

2.3 New analysis

We provide a new analysis to fix the technical flaw and give a valid upper bound for the key quantity $\sigma_{\max}\left(V_{t-1}^{-1}\left(\sum_{s=t_0}^{p} X_s X_s^\top\right)\right)$. As shown below, there will be an extra $L^2\sqrt{dH/\lambda}$ coefficient in our result, comparing with the claim in (6) as was anticipated. At the end of this subsection, we will further show that the square-root dependence on restarting period H is dishearteningly necessary.

Lemma 1. *For any $t \in [T]$, we have*

$$\left\| V_{t-1}^{-1}\left(\sum_{s=t_0}^{t-1} X_s X_s^\top(\theta_s - \theta_t)\right)\right\|_2 \leq L\sqrt{\frac{dH}{\lambda}}\sum_{p=t_0}^{t-1}\|\theta_p - \theta_{p+1}\|_2.$$

Proof We continue arguments presented in (7). Denote by $\mathcal{S}(1) = \{\mathbf{x} \mid \|\mathbf{x}\|_2 = 1\}$ the unit sphere.

$$\begin{aligned}
\left\| V_{t-1}^{-1}\left(\sum_{s=t_0}^{p} X_s X_s^\top\right)\right\|_2 &= \sup_{\mathbf{z}\in\mathcal{S}(1)}\sup_{\widetilde{\mathbf{z}}\in\mathcal{S}(1)}\left|\mathbf{z}^\top V_{t-1}^{-1}\left(\sum_{s=t_0}^{p} X_s X_s^\top\right)\widetilde{\mathbf{z}}\right| \\
&= \left|\mathbf{z}_*^\top V_{t-1}^{-1}\left(\sum_{s=t_0}^{p} X_s X_s^\top\right)\widetilde{\mathbf{z}}_*\right| \qquad \text{(let } \mathbf{z}_*, \widetilde{\mathbf{z}}_* \text{ denote the optimizer)} \\
&\leq \|\mathbf{z}_*\|_{V_{t-1}^{-1}}\left\|\sum_{s=t_0}^{p} X_s(X_s^\top\widetilde{\mathbf{z}}_*)\right\|_{V_{t-1}^{-1}} \\
&\leq \|\mathbf{z}_*\|_{V_{t-1}^{-1}}\left\|\sum_{s=t_0}^{p} X_s\|X_s\|_2\|\widetilde{\mathbf{z}}_*\|_2\right\|_{V_{t-1}^{-1}}
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{L}{\sqrt{\lambda}} \left\| \sum_{s=t_0}^{p} X_s \right\|_{V_{t-1}^{-1}} \\
&\leq \frac{L}{\sqrt{\lambda}} \sum_{s=t_0}^{p} \|X_s\|_{V_{t-1}^{-1}} \\
&\leq \frac{L}{\sqrt{\lambda}} \sqrt{H} \sqrt{\sum_{s=t_0}^{p} \|X_s\|_{V_{t-1}^{-1}}^2} && \text{(by Cauchy-Schwarz inequality)} \\
&\leq L\sqrt{\frac{dH}{\lambda}}.
\end{aligned}
$$

In above, the first equation makes use of the property of the matrix 2-norm: for a matrix $M \in \mathbb{R}^{m\times n}$, $\|M\|_2 = \sup_{\|\mathbf{x}\|_2=1} \sup_{\|\mathbf{y}\|_2=1} |\mathbf{x}^\top M \mathbf{y}|$, whose proof can be found from the book (Meyer, 2000, Chapter 5, Eq. (5.2.9)). Moreover, we use the fact that for any $\mathbf{x}$, we have $\|\mathbf{x}\|_{V_{t-1}^{-1}} \leq \|\mathbf{x}\|_2/\sqrt{\lambda}$ as $V_{t-1} \succeq \lambda I$. Besides, the last step follows from the fact: for any $p \in \{t_0, \ldots, t-1\}$,

$$
\begin{aligned}
&\sum_{s=t_0}^{p} \|X_s\|_{V_{t-1}^{-1}}^2 = \sum_{s=t_0}^{p} \mathrm{Tr}(X_s^\top V_{t-1}^{-1} X_s) = \mathrm{Tr}\left(V_{t-1}^{-1} \sum_{s=t_0}^{p} X_s X_s^\top\right) \\
&\leq \mathrm{Tr}\left(V_{t-1}^{-1} \sum_{s=t_0}^{p} X_s X_s^\top\right) + \sum_{s=p+1}^{t-1} X_s^\top V_{t-1}^{-1} X_s + \lambda \sum_{i=1}^{d} \mathbf{e}_i^\top V_{t-1}^{-1} \mathbf{e}_i \qquad (12) \\
&= \mathrm{Tr}\left(V_{t-1}^{-1} \sum_{s=t_0}^{p} X_s X_s^\top\right) + \mathrm{Tr}\left(V_{t-1}^{-1} \sum_{s=p+1}^{t-1} X_s X_s^\top\right) + \mathrm{Tr}\left(V_{t-1}^{-1} \lambda \sum_{i=1}^{d} \mathbf{e}_i \mathbf{e}_i^\top\right) \\
&= \mathrm{Tr}(I_d) = d.
\end{aligned}
$$

Hence, we complete the proof. ■

Impossibility. We final note that the desirable claim (8) is actually impossible. In the following, we will construct a hard problem instance to show that the key quantity $\sigma_{\max}\big(V_{t-1}^{-1}\big(\sum_{s=t_0}^{p} X_s X_s^\top\big)\big)$ cannot be universally upper bounded by any constant without square-root dependence on H.

For notational convenience, we focus on the first restarting epoch, so the starting index $t_0 = 1$. Let $L = 1$ and $\lambda = 1$. We construct the feature as

$$
X_1 = \ldots = X_p = \left[\frac{1}{\sqrt{p}}, \frac{\sqrt{p-1}}{\sqrt{p}}\right]^\top, \text{ and } X_{p+1} = \ldots = X_H = \left[\frac{1}{\sqrt{H-p}}, \frac{\sqrt{H-p-1}}{\sqrt{H-p}}\right]^\top. \tag{13}
$$

Denote by $A = \sum_{s=1}^{p} X_s X_s^\top$ and $B = \sum_{s=p+1}^{H} X_s X_s^\top$, then the covariance matrix is $V_{t-1} = A + B + I_d$. Under such cases, considering the checkpoint of $p = \lfloor H/3 \rfloor$, we can prove that

$$
\|V_{t-1}^{-1} A\|_2 = \sigma_{\max}(V_{t-1}^{-1} A) \geq 0.0564\sqrt{H}. \tag{14}
$$

The proof involves some tedious calculations and is included in Appendix A. Moreover, we report some numerical results for validation: when $H = 3000$, $\sigma_{\max}(V_{t-1}^{-1} A) = 5.852$ and the theoretical lower bound is $0.0564\sqrt{H} = 3.087$; when $H = 30000$, $\sigma_{\max}(V_{t-1}^{-1}A) = 18.474$ and the theoretical lower bound is $0.0564\sqrt{H} = 9.763$.

2.4 Regret Analysis

Based on key technical lemma in Lemma 1, we have the following estimation error bound.

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

$$|\mathbf{x}^\top(\theta_t - \widehat{\theta}_t)| \leq L^2\sqrt{\frac{dH}{\lambda}}\sum_{p=t_0}^{t-1}\|\theta_p - \theta_{p+1}\|_2 + \beta_t\|\mathbf{x}\|_{V_{t-1}^{-1}}, \tag{15}$$

where β_t is the radius of confidence region same as defined in (5).

Comparing with the original result (Zhao et al., 2020, Lemma 1), the difference lies in the coefficient of the path-length $\sum_{p=t_0}^{t-1}\|\theta_p - \theta_{p+1}\|_2$. The original coefficient is 1, but its proof has serious flaws; the new analysis gives $L^2\sqrt{dH/\lambda}$, which is $\mathcal{O}(\sqrt{dH})$ worse than the anticipated one.

Based on Lemma 2, we can upper-bound the dynamic regret within each epoch (Theorem 3), and then sum over epochs to obtain the guarantee of the whole time horizon (Theorem 4).

Theorem 3. *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*

$$\text{D-Regret}(\mathcal{E}) \leq 2L^2\sqrt{\frac{d}{\lambda}}\cdot H^{\frac{3}{2}}\mathcal{P}(\mathcal{E}) + 2\beta_H\sqrt{2dH\log\left(1+\frac{L^2H}{\lambda d}\right)},$$

where $\beta_H = \sqrt{\lambda}S + R\sqrt{2\log(1/\delta) + d\log\left(1+\frac{HL^2}{\lambda d}\right)}$ is the confidence radius of the epoch, 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 4. *With probability at least $1 - 1/T$, the dynamic regret of RestartUCB (Algorithm 1) over the whole time horizon is upper bounded by*

$$\text{D-Regret}_T = \sum_{t=1}^{T}\max_{\mathbf{x}\in\mathcal{X}}\mathbf{x}^\top\theta_t - \sum_{t=1}^{T}X_t^\top\theta_t \leq \widetilde{\mathcal{O}}\left(d^{\frac{1}{2}}H^{\frac{3}{2}}P_T + dT/\sqrt{H}\right), \tag{16}$$

where $P_T = \sum_{t=2}^{T}\|\theta_{t-1} - \theta_t\|_2$ is the path-length, and H is the restarting period. We adopt the $\widetilde{\mathcal{O}}(\cdot)$-notation to suppress logarithmic factors in the time horizon T.

Furthermore, by setting the restarting period optimally as $H = \min\{\lfloor d^{1/4}(T/P_T)^{1/2}\rfloor, T\}$, RestartUCB achieves the following dynamic regret,

$$\text{D-Regret}_T \leq \begin{cases}\widetilde{\mathcal{O}}\left(d^{\frac{7}{8}}T^{\frac{3}{4}}P_T^{\frac{1}{4}}\right) & \text{when } P_T \geq \sqrt{d}/T, \\ \widetilde{\mathcal{O}}(d\sqrt{T}) & \text{when } P_T < \sqrt{d}/T.\end{cases} \tag{17}$$

The proofs of Theorem 3 and Theorem 4 are basically the same as analysis in previous work (Zhao et al., 2020, Section 5), so we omit the details here.

Notice that the optimal tuning in Theorem 4 requires the prior knowledge of path-length $P_T = \sum_{t=2}^{T} \|\theta_{t-1} - \theta_t\|_2$, which is generally unavailable. To (partially) address the issue, it is possible to use the Bandit-over-Bandit (BOB) mechanism (Cheung et al., 2019a) to compensate the lack of this information, with an $\mathcal{O}(T^{3/4})$ regret overhead (cf. the proof of Theorem 4 in (Zhao et al., 2020)). Combining the analysis and Theorem 3, we can prove that RestartUCB together with BOB mechanism leads to the following dynamic regret without requiring the prior knowledge of P_T.

Theorem 5. RESTARTUCB *together with the Bandit-over-Bandit (BOB) mechanism enjoys the dynamic regret of*

$$\text{D-Regret}_T \leq \widetilde{\mathcal{O}}\big(d^{\frac{7}{8}} T^{\frac{3}{4}} P_T^{\frac{1}{4}}\big), \tag{18}$$

without requiring the path-length P_T ahead of time.

Acknowledgment

The authors thank Jin-Hui Wu for many helpful discussions, especially on the impossibility result. We also acknowledge Yu-Hu Yan for carefully proofreading the paper.

References

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

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

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Hedging the drift: Learning to optimize under non-stationarity. *ArXiv preprint*, arXiv:1903.01461, 2019b.

Baekjin Kim and Ambuj Tewari. Randomized exploration for non-stationary stochastic linear bandits. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 71–80, 2020.

Carl D Meyer. *Matrix Analysis and Applied Linear Algebra*, volume 71. SIAM, 2000.

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

Yoan Russac, Olivier Cappé, and Aurélien Garivier. Algorithms for non-stationary generalized linear bandits. *ArXiv preprint*, arXiv:2003.10113, 2020.

Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. A simple approach for non-stationary linear bandits. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 746–755, 2020.

Appendix A. Proof of Impossibility Result

Proof For simplicity of notation, let $y = \sqrt{p-1}$ and $z = \sqrt{H-p-1}$. By the constructed example in (13), we have

$$A = \sum_{s=1}^{p} X_s X_s^\top = \begin{bmatrix} 1 & y \\ y & y^2 \end{bmatrix} \text{ and } B = \sum_{s=p+1}^{H} X_s X_s^\top = \begin{bmatrix} 1 & z \\ z & z^2 \end{bmatrix}.$$

For convenience, we will write the covariance matrix V_t simply V when no confusion can arise. So the concerned matrix $V^{-1}A$ can be calculated as

$$\begin{aligned} V^{-1}A &= \begin{bmatrix} 2+\lambda & y+z \\ y+z & y^2+z^2+\lambda \end{bmatrix}^{-1} \begin{bmatrix} 1 & y \\ y & y^2 \end{bmatrix} \\ &= \frac{1}{(2+\lambda)(y^2+z^2+\lambda)-(y+z)^2} \begin{bmatrix} y^2+z^2+\lambda & -(y+z) \\ -(y+z) & 2+\lambda \end{bmatrix} \begin{bmatrix} 1 & y \\ y & y^2 \end{bmatrix} \\ &= \frac{1}{(1+\lambda)(y^2+z^2)-2yz+(2+\lambda)\lambda} \begin{bmatrix} z^2-yz+\lambda & yz^2-y^2z+\lambda y \\ (1+\lambda)y-z & (1+\lambda)y^2-yz \end{bmatrix}. \end{aligned}$$

Denote by $s = (1+\lambda)(y^2+z^2) - 2yz + (2+\lambda)\lambda$, $\alpha = z^2 - yz + \lambda$, and $\beta = (1+\lambda)y - z$, we then have

$$V^{-1}A(V^{-1}A)^\top = \frac{1+y^2}{s^2} \begin{bmatrix} \alpha^2 & \alpha\beta \\ \alpha\beta & \beta^2 \end{bmatrix}.$$

The eigenvalues (we denote them by $\bar{\lambda}$, to distinguish the notation with the regularizer coefficient λ) of matrix $[\alpha^2, \alpha\beta; \alpha\beta, \beta^2]$ should satisfy $(\alpha^2 - \bar{\lambda})(\beta^2 - \bar{\lambda}) - \alpha^2\beta^2 = 0$. By solving the equation, we can obtain that

$$\bar{\lambda}_{\max} = \alpha^2 + \beta^2 = (z^2 - yz + \lambda)^2 + ((1+\lambda)y - z)^2 \geq (z^2 - yz + \lambda)^2.$$

When $\lambda = 1$ and $p = aH$ (here we assume aH is an integer for simplicity), we further have

$$\begin{aligned} \bar{\lambda}_{\max} &\geq (z^2 - yz + \lambda)^2 \\ &= \left((1-a)p - 1 - \sqrt{p-1}\sqrt{(1-a)p-1} + 1\right)^2 \\ &\geq \left((1-a)H - \sqrt{a(1-a)}H\right)^2 \\ &= (1-a)(\sqrt{1-a} - \sqrt{a})^2 H^2. \end{aligned} \tag{19}$$

Note that we require $a \in (0, 1/2)$ to make the second inequality hold. On the other hand,

$$\frac{1+y^2}{s^2} = \frac{p}{(2(y^2+z^2) - 2yz + 3)^2} \geq \frac{p}{(2(y^2+z^2)+4)^2} = \frac{a}{4H}. \tag{20}$$

Combining (19) and (20), we have

$$\sigma_{\max}(V^{-1}A) = \sqrt{\lambda_{\max}\left(V^{-1}A(V^{-1}A)^\top\right)} \geq \sqrt{\bar{\lambda}_{\max} \cdot \frac{1+y^2}{s^2}} \geq \sqrt{\frac{a'}{4}} \cdot \sqrt{H},$$

where $a' = (1-a)a(\sqrt{1-a} - \sqrt{a})^2$ is a universal constant. When choosing $a = 1/3$ as selected in the main paper, $a' = 0.0127$ and the lower bound is $\sigma_{\max}(V^{-1}A) \geq 0.0564\sqrt{H}$. ■