Revisiting Weighted Strategy for Non-stationary Parametric Bandits

Jing Wang, Peng Zhao, Zhi-Hua Zhou
National Key Laboratory for Novel Software Technology,
Nanjing University, Nanjing 210023, China
{wangjing, zhaop, zhouzh}@lamda.nju.edu.cn

Abstract

Non-stationary parametric bandits have attracted much attention recently. There are three principled ways to deal with non-stationarity, including sliding-window, weighted, and restart strategies. As many non-stationary environments exhibit gradual drifting patterns, the weighted strategy is commonly adopted in real-world applications. However, previous theoretical studies show that its analysis is more involved and the algorithms are either computationally less efficient or statistically suboptimal. This paper revisits the weighted strategy for non-stationary parametric bandits. In linear bandits (LB), we discover that this undesirable feature is due to an inadequate regret analysis, which results in an overly complex algorithm design. We propose a *refined analysis framework*, which simplifies the derivation and importantly produces a simpler weight-based algorithm that is as efficient as window/restart-based algorithms while retaining the same regret as previous studies. Furthermore, our new framework can be used to improve regret bounds of other parametric bandits, including Generalized Linear Bandits (GLB) and Self-Concordant Bandits (SCB). For example, we develop a simple weighted GLB algorithm with an $\widetilde{\mathcal{O}}(k_\mu^{5/4} c_\mu^{-3/4} d^{3/4} P_T^{1/4} T^{3/4})$ regret, improving the $\widetilde{\mathcal{O}}(k_\mu^{2} c_\mu^{-1} d^{9/10} P_T^{1/5} T^{4/5})$ bound in prior work, where k_μ and c_μ characterize the reward model's nonlinearity, P_T measures the non-stationarity, d and T denote the dimension and time horizon.

1 INTRODUCTION

Non-stationary parametric bandits model the sequential decision-making problems where the reward distributions of each arm are structured with an unknown *time-varying* parameter, which have been extensively studied in recent years [Cheung et al., 2019, Russac et al., 2019, Zhao et al., 2020, Russac et al., 2020, Faury et al., 2021, Russac et al., 2021, Wei and Luo, 2021, Deng et al., 2022, Liu et al., 2022] due to its significance in many real-world non-stationary online applications such as recommendation systems [Tomkins et al., 2021, Huleihel et al., 2021], as well as tight connection with theoretical foundation of reinforcement learning [Jin et al., 2020, Touati and Vincent, 2020].

Linear Bandits (LB) is a fundamental instance of parametric bandits, where the expected reward for pulling a certain arm at time t is the inner product between the arm's feature vector X_t and an unknown parameter θ_t, namely, $\mathbb{E}[r_t \mid X_t] = X_t^\top \theta_t$. Moreover, Generalized Linear Bandits (GLB) is introduced as a generalization of LB to model a broader range of reward functions such as binary rewards, where the expected reward obeys a generalized linear model as $\mathbb{E}[r_t \mid X_t] = \mu(X_t^\top \theta_t)$ with $\mu(\cdot)$ being an inverse link function. Notably, the non-stationary models allow the parameter θ_t in the above models to be time-varying. There are two typical non-stationarity measures to quantify the intensity of parameter changes: (i) in gradually drifting cases, path length $P_T = \sum_{t=2}^{T} \|\theta_{t-1} - \theta_t\|_2$ is used to measure the cumulative variations of the underlying parameters; and (ii) in piecewise-stationary cases, Γ_T denotes the number of parameter changes in T rounds.

To deal with non-stationarity, there are three principled ways: sliding-window, weighted, and restart strategies. For the sliding-window strategy, the learner maintains a time window that contains the most recent observed data to discard the outdated data. For the weighted strategy, the learner puts more weight on the most recent data and less weight on the old data to gradually forget the outdated data. For the restart strategy, the learner restarts the algorithm according to a certain period to discard the outdated data. The currently best-known result for non-stationary (generalized) linear bandits is by Wei and Luo [2021], who developed an optimal algorithm consisting of a non-stationarity detector and a base algorithm that performs well in near-stationary environments. Whenever the detector examines

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

Table 1: Comparisons of our dynamic regret bounds to previous best-known results for weight-based algorithms, under different non-stationary parametric bandits. Below, k_μ/c_μ denotes the degree of non-linearity and becomes 1 in LB case; d is the dimension, path length P_T and change number Γ_T are non-stationarity measures for drifting and piecewise-stationary cases, respectively.

Parametric Bandit Models	Previous Work	Our Results
Drifting LB	$\widetilde{\mathcal{O}}\big(d^{7/8}P_T^{1/4}T^{3/4}\big)$ [Russac et al., 2019]	$\widetilde{\mathcal{O}}\big(d^{3/4}P_T^{1/4}T^{3/4}\big)$ [Theorem 1]
Drifting GLB	$\widetilde{\mathcal{O}}\Big(\frac{k_\mu^2}{c_\mu}d^{9/10}P_T^{1/5}T^{4/5}\Big)$ [Faury et al., 2021]	$\widetilde{\mathcal{O}}\Big(\frac{k_\mu^{5/4}}{c_\mu^{3/4}}d^{3/4}P_T^{1/4}T^{3/4}\Big)$ [Theorem 2]
Drifting SCB	$\widetilde{\mathcal{O}}\Big(\frac{k_\mu^2}{c_\mu}d^{9/10}P_T^{1/5}T^{4/5}\Big)$ [Faury et al., 2021]	$\widetilde{\mathcal{O}}\Big(\frac{k_\mu^{5/4}}{c_\mu^{1/2}}d^{3/4}P_T^{1/4}T^{3/4}\Big)$ [Theorem 3]
Piecewise Stationary SCB	$\widetilde{\mathcal{O}}\Big(\frac{1}{c_\mu^{1/3}}d^{2/3}\Gamma_T^{1/3}T^{2/3}\Big)$ [Russac et al., 2021]	$\widetilde{\mathcal{O}}\big(d^{2/3}\Gamma_T^{1/3}T^{2/3}\big)$ [Theorem 4]

that the non-stationarity exceeds a certain limit, the algorithm will *restart* itself to resist the non-stationarity. In this sense, their algorithm can be regarded as an *adaptive restart-based algorithm*. Building on the RestartUCB algorithm [Zhao et al., 2020] and a carefully designed non-stationarity detector with multi-scale explorations, their algorithm can achieve an $\widetilde{\mathcal{O}}(\min\{\sqrt{\Gamma_T T}, P_T^{1/3}T^{2/3}\})$ optimal dynamic regret for both LB and GLB.

In real-world scenarios, the distribution change of environments often exhibits gradually drifting patterns [Crammer et al., 2010, Chiang et al., 2013, Gama et al., 2014], in such cases, a soft weighted strategy can be (empirically) more advantageous than a hard restart strategy to deal with the non-stationarity, as can be observed in bandits learning [Russac et al., 2019, Zhao et al., 2020, Deng et al., 2022], classification with concept drift [Anagnostopoulos et al., 2012, Zhao et al., 2021a], and adaptive system identification [Guo et al., 1993, Chu and Mak, 2017]. As a result, it will be highly attractive to design an *adaptive weight-based algorithm* for non-stationary parametric bandits, which imposes weights to discount the importance of past data, and the weights are set adaptively according to environments. Towards this end, we examine existing methods for non-stationary parametric bandits based on the weighted strategy, and (surprisingly) find that current results exhibit *unnatural* gaps compared to the other strategies, such as restart-based algorithms, as well as *unnatural* regret analysis transitions from GLB to LB.

Those unnatural phenomena motivate us to revisit the algorithm design and regret analysis of the weighted strategy for non-stationary parametric bandits [Russac et al., 2019, 2021, Faury et al., 2021]. Indeed, the key ingredient is the *estimation error* analysis for the weight-based estimator, which is usually decomposed into two parts — one is the *bias* part due to the parameter drift, and the other is the *variance* part due to the stochastic noise. Generally, the bias part is controlled by non-stationary strategies, and the variance part is handled by carefully designed concentration. Russac et al. [2019] provided the first analysis of the weight-based algorithm for LB, where they introduced a virtual window size in the analysis to control the bias in order to mimic the analysis of sliding-window strategy [Cheung et al., 2019]. Consequently, they have to use *different* local norms to control bias and variance parts, resulting in unexpected inefficiencies and complications. For LB, this leads to an algorithm D-LinUCB [Russac et al., 2019] requiring to maintain an extra covariance matrix, which is less efficient than the window-based and restart-based algorithms [Cheung et al., 2019, Zhao et al., 2020].

This analysis framework for weighted strategy introduces more severe issues in GLB, due to its more enriched and complicated structure. Specifically, Faury et al. [2021] studied the drifting GLB and designed a highly complex projection operation to control bias and variance parts following the way of Russac et al. [2019] to mimic sliding-window analysis, and finally attained an $\widetilde{\mathcal{O}}(d^{9/10}P_T^{1/5}T^{4/5})$ dynamic regret. Unfortunately, this *cannot* recover the $\widetilde{\mathcal{O}}(d^{7/8}P_T^{1/4}T^{3/4})$ bound enjoyed by the weight-based algorithm for drifting LB (a special case of GLB) [Russac et al., 2019]. Subsequently, Russac et al. [2021] investigated the non-stationary Self-Concordant Bandits (SCB), a subclass of GLB with many attractive structures. They can only conduct analysis under the piecewise-stationary setting, whereas failed in the more challenging drifting setting, due to technical difficulties in bounding bias using conventional analysis. As such, two open questions are proposed in their papers: (i) how to extend weight-based algorithms to drifting SCB; and (ii) how to replicate recent progress in stationary SCB [Abeille et al., 2021] to improve dependence on c_μ in non-stationary SCB.

Our Results. In this paper, we revisit the weighted strategy for the non-stationary parametric bandits. We discover that the earlier analysis framework may be inappropriate due to mimicking the sliding-window analysis, which demands bounding the bias and variance parts using *different* local norms. We present a *refined analysis framework* for the weighted strategy, in which a new analysis for the bias part is presented such that it is now allowed to use the *same* local norm to analyze both bias and variance parts. This refined analysis framework not only simplifies the regret analysis but also brings many benefits in algorithm designs, including improving efficiency for LB and resolving the projection problem brought by GLB and SCB. Table 1 summarizes our main results compared to earlier best-known results of weight-based algorithms. Specifically, based on our refined analysis framework, we achieve the following results: (i) for LB, our approach only needs to maintain one covariance instead of two and still enjoys the same regret as [Russac et al., 2019]; (ii) for GLB, our approach enjoys an $\widetilde{\mathcal{O}}(k_\mu^{5/4} c_\mu^{-3/4} d^{3/4} P_T^{1/4} T^{3/4})$ regret bound, whose order of d, P_T and T matches that in LB case; and (iii) for SCB, we achieve an $\widetilde{\mathcal{O}}(k_\mu^{5/4} c_\mu^{-1/2} d^{3/4} P_T^{1/4} T^{3/4})$ regret bound. In addition, for piecewise stationary SCB, our approach achieves an $\widetilde{\mathcal{O}}(d^{2/3} \Gamma_T^{1/3} T^{2/3})$ regret bound that can get rid of the influence of c_μ^{-1}, resolving the second open problem asked by Russac et al. [2021].

2 RELATED WORK

Linear Bandits. Non-stationary LB problem was first studied by Cheung et al. [2019]. They established an $\Omega(d^{2/3} P_T^{1/3} T^{2/3})$ minimax lower bound and then proposed SW-UCB algorithm based on the sliding-window strategy. Then Russac et al. [2019] proposed the D-LinUCB algorithm based on weighted strategy and Zhao et al. [2020] proposed the RestartUCB algorithm based on restart strategy. Note that the three works proved an $\widetilde{\mathcal{O}}(d^{2/3} P_T^{1/3} T^{2/3})$ regret bound, but there exists a subtle technical gap in the regret analysis as identified by Zhao and Zhang [2021]. After fixing the technical gap, all three aforementioned algorithms achieve an $\widetilde{\mathcal{O}}(d^{7/8} P_T^{1/4} T^{3/4})$ regret bound [Zhao and Zhang, 2021, Zhao et al., 2021b]. However, to achieve this result, all the three algorithms require the knowledge of the path length P_T as an input at the beginning of algorithmic implementation, which is undesired. To address so, Cheung et al. [2019] proposed the bandits-over-bandits (BOB) strategy as a meta-algorithm to learn the unknown parameter P_T which can be combined with the above algorithms to remove the requirement of this prior knowledge. Afterward, Wei and Luo [2021] proposed the MASTER algorithm with theoretically optimal $\widetilde{\mathcal{O}}(\min\{d\sqrt{\Gamma_T T}, dP_T^{1/3} T^{2/3}\})$ regret bound, also without requiring the non-stationarity level of environments (that is, Γ_T and P_T) in advance. Most recently, there also has been some new progress in the non-stationary (linear) bandits [Liu et al., 2022, Suk and Kpotufe, 2022, Abbasi-Yadkori et al., 2022, Clerici et al., 2023].

Generalized Linear Bandits. GLB problem was first introduced by Filippi et al. [2010]. They proposed GLM-UCB algorithm, achieving an $\widetilde{\mathcal{O}}(k_\mu c_\mu^{-1} d\sqrt{T})$ regret bound where k_μ, c_μ are the problem-dependent constants and k_μ/c_μ represents the nonlinearity of the generalized linear model. Faury et al. [2021] extended the stationary GLB to the drifting case, and proposed BVD-GLM-UCB algorithm with $\widetilde{\mathcal{O}}(k_\mu^2 c_\mu^{-1} d^{9/10} P_T^{1/5} T^{4/5})$ regret bound. Faury et al. [2020] studied a specific instance of GLB called Logistic Bandits (LogB), they first pointed out that under GLB setting, the problem-dependent constant $1/c_\mu$ could be very large in some cases like LogB, then they proposed the Logistic-UCB-1 algorithm with an $\widetilde{\mathcal{O}}(c_\mu^{-1/2} d\sqrt{T})$ regret bound and the Logistic-UCB-2 algorithm with an $\widetilde{\mathcal{O}}(d\sqrt{T} + c_\mu^{-1})$ regret bound. Subsequently, Abeille et al. [2021] established an $\Omega(d\sqrt{\mu'(X_*^\top \theta_*)T})$ regret lower bound for logistic bandits and provided an optimal algorithm OFULog. Russac et al. [2021] generalized the logistic bandits to self-concordant bandits and considered the piecewise-stationary case, their algorithm enjoys an $\widetilde{\mathcal{O}}(c_\mu^{-1/3} d^{2/3} \Gamma_T^{1/3} T^{2/3})$ regret bound. To deal with P_T-unknown cases, Faury et al. [2021] proposed a parameter-free algorithm by combining BVD-GLM-UCB with BOB strategy, but the final result is still suboptimal. Meanwhile, the optimal black-box algorithm [Wei and Luo, 2021] can also adaptively restart stationary algorithm GLM-UCB [Filippi et al., 2010] and achieve an optimal $\widetilde{\mathcal{O}}(\min\{k_\mu c_\mu^{-1}\sqrt{\Gamma_T T}, k_\mu^{4/3} c_\mu^{-1} dP_T^{1/3} T^{2/3}\})$ regret.

3 LINEAR BANDITS

In this section, we first introduce the problem setting of non-stationary LB, and then describe our LB-WeightUCB algorithm and its theoretical guarantee. Our algorithm has the same regret bound as the best-known weight-based algorithm [Russac et al., 2019] but is more efficient.

3.1 Problem Setting

At each round t, the learner chooses an arm X_t from a feasible set $\mathcal{X} \subseteq \mathbb{R}^d$ and receives a reward r_t such that

$$r_t = X_t^\top \theta_t + \eta_t, \tag{1}$$

where $\theta_t \in \mathbb{R}^d$ is the unknown time-varying parameter and η_t is the R-sub-Gaussian noise. The goal of the learner is to minimize the following (pseudo) *dynamic regret*:

$$R_T = \sum_{t=1}^{T} \max_{\mathbf{x}\in\mathcal{X}} \mathbf{x}^\top \theta_t - \sum_{t=1}^{T} X_t^\top \theta_t, \tag{2}$$

which is the cumulative regret against the optimal strategy that has full information of the unknown parameter. Here we consider the drifting case where we use path length $P_T = \sum_{t=2}^{T} \|\theta_{t-1} - \theta_t\|_2$ as the non-stationarity measure.

We work under the following standard boundedness assumption [Abbasi-Yadkori et al., 2011, Cheung et al., 2019, Russac et al., 2019, Zhao et al., 2020].

Assumption 1. The feasible set and unknown parameters are assumed to be bounded: $\forall \mathbf{x} \in \mathcal{X}$, $\|\mathbf{x}\|_2 \leq L$, and $\theta_t \in \Theta$ holds for all $t \in [T]$ where $\Theta \triangleq \{\theta \mid \|\theta\|_2 \leq S\}$.

3.2 Algorithm and Regret Guarantee

We propose the LB-WeightUCB algorithm, which attains the same guarantees as earlier methods while having higher efficiency. We first give the employed estimator and then derive its estimation error upper bound by our refined analysis framework, which is the key for algorithm design and regret analysis. Based on the estimation error bound, we propose our selection criterion and finally give the theoretical guarantee on its dynamic regret.

Estimator. We adopt the weighted regularized least square estimator *same* as D-LinUCB [Russac et al., 2019], the estimator $\widehat{\theta}_t$ is the solution to the following problem,

$$\min_{\theta} \lambda \|\theta\|_2^2 + \sum_{s=1}^{t-1} \gamma^{t-s-1} \left(X_s^\top \theta - r_s\right)^2, \tag{3}$$

where $\lambda > 0$ is the regularization coefficient and $\gamma \in (0,1)$ is the discounted factor. Clearly, $\widehat{\theta}_t$ admits a close-form solution $\widehat{\theta}_t = V_{t-1}^{-1}(\sum_{s=1}^{t-1}\gamma^{t-s-1} r_s X_s)$, where $V_t = \lambda I_d + \sum_{s=1}^{t}\gamma^{t-s}X_sX_s^\top, V_0 = \lambda I_d$ is the covariance matrix. Note that this close-form solution can be further transformed into a recursive formula [Haykin, 2002, Chapter 10.3]. This allows it to be updated online without the need to store historical data, which is another important computational advantage of the weighted strategy compared to the sliding-window strategy.

Upper Confidence Bounds. For estimator (3), we provide the following estimation error bound. Notably, this is *different* from the previous result [Russac et al., 2019, Appenidx B.3, second and third steps in Proof of Theorem 2], which is the key component being a more efficient algorithm and will be explained later.

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

$$\left|\mathbf{x}^\top(\widehat{\theta}_t - \theta_t)\right| \leq L^2\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{t-1}\gamma^{\frac{t-1}{2}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\left\|\theta_p - \theta_{p+1}\right\|_2 + \beta_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}}, \tag{4}$$

where β_t is the radius of confidence region set by

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

The proof of Lemma 1 is in Appendix A.2. Based on Lemma 1, we can specify the arm selection criterion as

$$X_t = \operatorname*{arg\,max}_{\mathbf{x}\in\mathcal{X}}\left\{\langle \mathbf{x}, \widehat{\theta}_t\rangle + \beta_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}}\right\}. \tag{6}$$

The overall algorithm is summarized in Algorithm 1. From the update procedure in Line 5 of Algorithm 1, we can observe that our algorithm needs to maintain a *single* covariance matrix $V_{t-1} \in \mathbb{R}^{d\times d}$. By contrast, the selection criterion of algorithm proposed in Russac et al. [2019] is like

$$X_t = \operatorname*{arg\,max}_{\mathbf{x}\in\mathcal{X}}\left\{\langle \mathbf{x}, \widehat{\theta}_t\rangle + \beta_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}}\right\}$$

, where β_{t-1}, V_{t-1}^{-1} are identical with those in our selection criterion (6) and $\widetilde{V}_{t-1} = \lambda I_d + \sum_{s=1}^{t-1}\gamma^{2(t-s-1)}X_sX_s^\top \in \mathbb{R}^{d\times d}$ is an extra covariance matrix. Thus, our algorithm is more efficient than their algorithm since it only needs to maintain one covariance matrix instead of two. This owes to the fact that our analysis of Lemma 1 only uses V_{t-1}^{-1} as the local norm to analyze both bias and variance parts, but the algorithm of Russac et al. [2019] requires to use l_2-norm and $V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}$-norm to control bias and variance parts, respectively. In Section 6, we provide a sketch of the analysis framework for Lemma 1 and a more detailed discussion is presented in Appendix A.1. Furthermore, we prove that our algorithm enjoys the same (even slightly better in d) regret as the algorithm of Russac et al. [2019].

Algorithm 1 LB-WeightUCB

Input: time horizon T, discounted factor γ, confidence δ, regularizer λ, parameters S, L and R

1: Set $V_0 = \lambda I_d$, $\widehat{\theta}_1 = \mathbf{0}$ and compute β_0 by (5)
2: **for** $t = 1, 2, ..., T$ **do**
3: Select $X_t = \arg\max_{\mathbf{x}\in\mathcal{X}}\{\langle\mathbf{x}, \widehat{\theta}_t\rangle + \beta_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}}\}$
4: Receive the reward r_t
5: Update $V_t = \gamma V_{t-1} + X_tX_t^\top + (1-\gamma)\lambda I_d$
6: Compute $\widehat{\theta}_{t+1}$ by (3) and β_t by (5)
7: **end for**

Theorem 1. *For all $\gamma \in (1/T, 1)$, $\lambda = d$, the dynamic regret of LB-WeightUCB (Algorithm 1) is bounded with probability at least $1 - 1/T$, by*

$$R_T \leq \widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^{3/2}}P_T + d(1-\gamma)^{1/2}T\right).$$

Furthermore, by setting the discounted factor optimally as $\gamma = 1 - \max\{1/T, \sqrt{P_T/(dT)}\}$, LB-WeightUCB ensures

$$R_T \leq \begin{cases} \widetilde{\mathcal{O}}\left(d^{3/4}P_T^{1/4}T^{3/4}\right) & \text{when } P_T \geq d/T, \\ \widetilde{\mathcal{O}}\left(d\sqrt{T}\right) & \text{when } P_T < d/T. \end{cases}$$

Compared to previous works [Cheung et al., 2019, Russac et al., 2019, Zhao et al., 2020], our approach improves from $\widetilde{\mathcal{O}}(d^{7/8}P_T^{1/4}T^{3/4})$ to $\widetilde{\mathcal{O}}(d^{3/4}P_T^{1/4}T^{3/4})$ when $P_T \geq d/T$. We remark that this improved dimensional dependence is simply owing to the more refined tuning of the discounted factor than the one used by [Russac et al., 2019], who did not take the dimension into the tuning. Their algorithm and regret bound can also benefit from the refined tuning. The proof of Theorem 1 is in Appendix A.3.

Further, notice that the optimal choice of discounted factor γ requires knowing P_T in advance. To achieve a parameter-free result for unknown P_T case, our algorithm can be combined with the BOB strategy [Cheung et al., 2019] and achieves an $\widetilde{\mathcal{O}}(d^{3/4}P_T^{1/4}T^{3/4})$ bound. However, this bound is not optimal, and it is possible to design an adaptive weight-based algorithm based on our result, in the spirit of Wei and Luo [2021], to further achieve an optimal dynamic regret without prior knowledge of P_T. This is very challenging since that at each round $t \in [T]$, we can only receive one data pair (X_t, r_t), which is not adequate for the learner to real-time update the discounted factor γ_t. At the same time, MASTER algorithm [Wei and Luo, 2021] can be considered as a special case of the adaptive weight-based algorithm since it only includes two circumstances: setting $\gamma_t = 0$ to restart at time t and setting $\gamma_t = 1$ to keep going. But for the adaptive weight-based algorithm, the choice of the discounted factor γ_t can be continuous in $[0, 1]$, which is more difficult than a binary decision. We leave this as an important open question for future study.

4 GENERALIZED LINEAR BANDITS

In this section, we apply the weighted strategy to drifting GLB. Compared to the best-known weight-based algorithm for drifting GLB [Faury et al., 2021], our algorithm is simpler and meanwhile has a better theoretical guarantee.

4.1 Problem Setting

GLB assumes an inverse link function $\mu : \mathbb{R} \to \mathbb{R}$ such that $r_t = \mu(X_t^\top \theta_t) + \eta_t$, where $\theta_t \in \mathbb{R}^d$ is the unknown parameter and can change over time. Similar to LB, we define *dynamic regret* for GLB as follows:

$$R_T = \sum_{t=1}^{T} \left(\max_{\mathbf{x} \in \mathcal{X}} \mu(\mathbf{x}^\top \theta_t) - \mu(X_t^\top \theta_t) \right). \tag{7}$$

Under the GLB setting, we make the same assumptions as those of LB, which include R-sub-Gaussian noise, boundedness of feasible set and unknown regression parameters (Assumption 1). In addition, we work under the standard boundedness assumption of the inverse link function [Filippi et al., 2010, Li et al., 2017, Faury et al., 2021].

Assumption 2. The inverse link function $\mu : \mathbb{R} \to \mathbb{R}$ is k_μ-Lipschitz, and continuously differentiable with

$$c_\mu \triangleq \inf_{\{\theta \in \Theta, \mathbf{x} \in \mathcal{X}\}} \mu'(\theta^\top \mathbf{x}) > 0, \quad \Theta = \{\theta \mid \|\theta\|_2 \le S\}.$$

Previous works [Zhao et al., 2020, Cheung et al., 2022] define a similar parameter $\widetilde{c}_\mu \triangleq \inf_{\{\theta \in \mathbb{R}^d, \mathbf{x} \in \mathcal{X}\}} \mu'(\theta^\top \mathbf{x}) > 0$ and obtain regret upper bound scaling with $1/\widetilde{c}_\mu$. Clearly, that $\widetilde{c}_\mu$ is smaller than our defined c_μ (and can be much smaller) as c_μ is defined on Θ while $\widetilde{c}_\mu$ is defined on $\mathbb{R}$. Therefore, $\widetilde{c}_\mu$ is less attractive to appear in the upper bound.

Algorithm 2 GLB-WeightUCB

Input: time horizon T, discounted factor γ, confidence δ, regularizer λ, inverse link function μ, parameters S, L and R

```
Set V_0 = λI_d, θ̂_1 = 0 and compute k_μ and c_μ
for t = 1, 2, ..., T do
    if ||θ̂_t||_2 ≤ S then
        let θ̃_t = θ̂_t
    else
        Do the projection and get θ̃_t by (9)
    end if
    Compute β̄_{t-1} by (11)
    Select X_t by (12)
    Receive the reward r_t
    Update V_t = γV_{t-1} + X_t X_t^T + (1 - γ)λI_d
    Compute θ̂_{t+1} according to (8)
end for
```

4.2 Algorithm and Regret Guarantee

We propose GLB-WeightUCB, which is a simpler algorithm with better theoretical guarantee compared to previous weight-based algorithm [Faury et al., 2021]. The key improvement is owing to our refined analysis framework, which is compatible with a simple projection step.

Estimator. At iteration t, we first adopt the quasi-maximum likelihood estimator (QMLE) without considering the projection onto the feasible domain. Specifically, the estimator $\widehat{\theta}_t$ is the solution of the following weighted regularized estimation equation:

$$\lambda c_\mu \theta + \sum_{s=1}^{t-1} \gamma^{t-s-1} \left(\mu(X_s^\top \theta) - r_s \right) X_s = 0. \tag{8}$$

Given that $\widehat{\theta}_t$ may not belong to the feasible set Θ and c_μ is defined over the parameter $\theta \in \Theta$, we need to perform the following projection step

$$\widetilde{\theta}_t = \mathop{\arg\min}_{\theta \in \Theta} \|g_t(\widehat{\theta}_t) - g_t(\theta)\|_{V_{t-1}^{-1}}, \tag{9}$$

where $V_t = \lambda I_d + \sum_{s=1}^{t} \gamma^{t-s} X_s X_s^\top$ and $g_t(\theta)$ is

$$g_t(\theta) \triangleq \lambda c_\mu \theta + \sum_{s=1}^{t-1} \gamma^{t-s-1} \mu(X_s^\top \theta) X_s. \tag{10}$$

However, previous work [Faury et al., 2021] cannot conduct the same simple projection in the drifting case as stationary GLB or piecewise-stationary GLB, since they use different local norms to measure the bias and variance parts separately for estimation error analysis. Consequently, they have to design a complicated projection to ensure that the bias and variance parts could be measured by different local norms (see [Faury et al., 2021, Section 4.1], and our restatements in Appendix B.1).

Our refined analysis framework is compatible with this projection operation, thanks to our analysis framework utilizing the same local norm for the bias and variance parts.

Upper Confidence Bounds. For estimator (8) with projection (9), we construct following estimation error bound.

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

$$\left|\mu(\mathbf{x}^\top \widetilde{\theta}_t) - \mu(\mathbf{x}^\top \theta_t)\right| \leq \frac{2k_\mu}{c_\mu}\left(\sum_{p=1}^{t-1} C(p) \left\|\theta_p - \theta_{p+1}\right\|_2 + \bar{\beta}_{t-1} \left\|\mathbf{x}\right\|_{V_{t-1}^{-1}}\right),$$

where $C(p) = k_\mu L^2 \sqrt{\frac{d}{\lambda}} \gamma^{\frac{t-1}{2}} \sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}$, and $\bar{\beta}_t$ is the radius of confidence region set by

$$\bar{\beta}_t = \sqrt{\lambda} c_\mu S + R\sqrt{2\log\frac{1}{\delta} + d\log\left(1 + \frac{L^2(1-\gamma^{2t})}{\lambda d(1-\gamma^2)}\right)}. \tag{11}$$

The proof of Lemma 2 is in Appendix B.2. Then, based on Lemma 2, we can specify the arm selection criterion as

$$X_t = \arg\max_{\mathbf{x}\in\mathcal{X}} \left\{\mu(\mathbf{x}^\top \widetilde{\theta}_t) + \frac{2k_\mu}{c_\mu} \bar{\beta}_{t-1} \left\|\mathbf{x}\right\|_{V_{t-1}^{-1}}\right\}. \tag{12}$$

The overall algorithm is summarized in Algorithm 2.

Notice that the estimation equation (8) and the confidence radius (11) are the same as those used in Algorithm 1 of Faury et al. [2021]. But importantly, the final (projected) estimators of the two approaches are significantly different. With a simpler projection operation and our refined analysis framework, we can immediately attain an improved regret guarantee for weight-based algorithm.

Theorem 2. *For all $\gamma \in (1/T, 1)$, $\lambda = d/c_\mu^2$, the regret of GLB-WeightUCB (Algorithm 2) is bounded with probability at least $1 - 1/T$, by*

$$R_T \leq \widetilde{\mathcal{O}}\left(k_\mu^2 \frac{1}{(1-\gamma)^{3/2}} P_T + \frac{k_\mu}{c_\mu} d(1-\gamma)^{1/2} T\right).$$

By optimally setting $\gamma = 1 - \max\{1/T, \sqrt{k_\mu c_\mu P_T/(dT)}\}$, GLB-WeightUCB achieves the following dynamic regret,

$$R_T \leq \begin{cases} \widetilde{\mathcal{O}}\left(\frac{k_\mu^{5/4}}{c_\mu^{3/4}} d^{3/4} P_T^{1/4} T^{3/4}\right) & \text{when } P_T \geq \frac{d}{k_\mu c_\mu T}, \\ \widetilde{\mathcal{O}}\left(\frac{k_\mu}{c_\mu} d\sqrt{T}\right) & \text{when } 0 \leq P_T < \frac{d}{k_\mu c_\mu T}. \end{cases}$$

Compared to BVD-GLM-UCB (the best-known weight-based algorithm for drifting GLB) [Faury et al., 2021], focusing on the dependence on d, P_T, and T, we can see that our approach improves the regret from $\widetilde{\mathcal{O}}(d^{9/10} P_T^{1/5} T^{4/5})$ to $\widetilde{\mathcal{O}}(d^{3/4} P_T^{1/4} T^{3/4})$. Furthermore, our result also improves their result upon the c_μ dependence from c_μ^{-1} to $c_\mu^{-3/4}$.

5 SELF-CONCORDANT BANDITS

This section studies Self-Concordant Bandits (SCB), an important subclass of GLB with many attractive structures.

5.1 Problem Setting

For SCB, the reward's distribution belongs to a canonical exponential family: $\mathbb{P}_\theta [r \mid \mathbf{x}] = \exp(r\mathbf{x}^\top\theta - b(\mathbf{x}^\top\theta) + c(r))$ where $b(\cdot)$ is a twice continuously differentiable function and $c(\cdot)$ is a real-valued function. Owing to the benign properties of exponential families, we have $\mathbb{E}[r \mid \mathbf{x}] = b'(\mathbf{x}^\top\theta)$ and $\text{Var}[r \mid \mathbf{x}] = b''(\mathbf{x}^\top\theta)$ where b' denotes the first derivative of the function b, and b'' denotes its second derivative. Then, we can introduce the (inverse) link function $\mu(\cdot) \triangleq b'(\cdot)$ such that at $t \in [T]$ the following holds

$$\mathbb{E}[r_t \mid X_t] = \mu(X_t^\top\theta_t), \text{Var}[r_t \mid X_t] = \mu'(X_t^\top\theta_t). \tag{13}$$

SCB requires the link function satisfy $|\mu''| \leq \mu'$, usually referred to general self-concordant property. We further introduce the notation $\eta_t = r_t - \mu(X_t^\top\theta_t)$ to denote the noise. SCB successfully models many important real-world applications and captures the reward structure. For example, choosing $\mu(x) = (1+e^{-x})^{-1}$ yields the Logistic Bandits (LogB), which is often adopted to model the binary-feedback reward in recommendation system [Zhang et al., 2016, Jun et al., 2017, Dong et al., 2019].

We make several standard assumptions same as LB and GLB, including boundedness of feasible set and unknown regression parameters (Assumption 1), and non-linearity measure on link function (Assumption 2). In addition, similar to Russac et al. [2021], we need assumptions on boundedness of reward, and for the convenience of analysis we let $L = 1$ which means $\|\mathbf{x}\|_2 \leq 1$ for all $\mathbf{x} \in \mathcal{X}$.

Assumption 3. The reward received at each round satisfies $0 \leq r_t \leq m$ for all $t \in [T]$ and some constant $m > 0$.

5.2 Algorithm and Regret Guarantee

We propose the SCB-WeightUCB algorithm. Compared to GLB, we use a new local norm for projection and regret analysis which is the key to improving the order of c_μ^{-1}.

Estimator. At iteration t, we first adopt the same maximum likelihood estimator as GLB which is defined in (8). Different from GLB, here we use a new local norm to perform the projection onto the feasible set Θ,

$$\widetilde{\theta}_t = \arg\min_{\theta\in\Theta} \left\|g_t(\widehat{\theta}_t) - g_t(\theta)\right\|_{H_t^{-1}(\theta)}, \tag{14}$$

where $g_t(\theta)$ is the same as (10) while $H_t(\theta)$ is defined as

$$H_t(\theta) \triangleq \lambda c_\mu I_d + \sum_{s=1}^{t-1} \gamma^{t-s-1} \mu'(X_s^\top\theta) X_s X_s^\top. \tag{15}$$

Notably, compared to V_t, $H_t(\theta)$ depends on the function curvature along the dynamics and thus can capture more *local* information. Combining this projection step with the weighted version self-normalized concentration as restated in Theorem 6 will remove a constant $c_\mu^{-1/2}$ in regret bound.

Upper Confidence Bound. For estimator (8) with projection (14), we construct following estimation error bound.

Lemma 3. *For any $\mathbf{x} \in \mathcal{X}$, $\gamma \in (0,1)$ and $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds for all $t \in [T]$*

$$\left|\mu(\mathbf{x}^\top \widetilde{\theta}_t) - \mu(\mathbf{x}^\top \theta_t)\right|$$
$$\leq \frac{\sqrt{4+8S}k_\mu}{\sqrt{c_\mu}} \left(\sum_{p=1}^{t-1} C(p) \left\|\theta_p - \theta_{p+1}\right\|_2 + \widetilde{\beta}_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}} \right),$$

where $C(p) = L^2 \sqrt{\frac{d}{\lambda} \frac{k_\mu}{\sqrt{c_\mu}}} \gamma^{\frac{t-1}{2}} \sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}$, and $\widetilde{\beta}_t$ is the radius of confidence region set by

$$\widetilde{\beta}_t = \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}} \left(\log \frac{1}{\delta} + d \log 2 \right) + \frac{dm}{\sqrt{\lambda c_\mu}} \log \left(1 + \frac{L^2 k_\mu (1-\gamma^{2t})}{\lambda c_\mu d (1-\gamma^2)} \right) + \sqrt{\lambda c_\mu} S. \tag{16}$$

The proof of Lemma 3 is in Appendix C.1. Based on Lemma 3, we can specify the arm selection criterion as

$$X_t = \arg\max_{\mathbf{x} \in \mathcal{X}} \left\{ \mu(\mathbf{x}^\top \widetilde{\theta}_t) + 2\sqrt{1+2S} \frac{k_\mu}{\sqrt{c_\mu}} \widetilde{\beta}_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}} \right\}. \tag{17}$$

Our algorithm for SCB (named SCB-WeightUCB) follows the same procedure of Algorithm 2, and the difference is that $\widetilde{\theta}_t$ is computed by (14), $\widetilde{\beta}_{t-1}$ is computed by (16) and X_t is computed by (17). Further, we have the following guarantee for SCB-WeightUCB algorithm.

Theorem 3. *For all $\gamma \in (1/T, 1)$, $\lambda = d\log(T)/c_\mu$, the dynamic regret of SCB-WeightUCB is bounded with probability at least $1 - 1/T$, by*

$$R_T \leq \widetilde{\mathcal{O}} \left(\frac{k_\mu^2}{\sqrt{c_\mu}} \frac{1}{(1-\gamma)^{3/2}} P_T + \frac{k_\mu}{\sqrt{c_\mu}} d (1-\gamma)^{1/2} T \right).$$

By setting $\gamma = 1 - \max\{1/T, \sqrt{k_\mu P_T/(dT)}\}$, we achieve

$$R_T \leq \begin{cases} \widetilde{\mathcal{O}} \left(\frac{k_\mu^{5/4}}{c_\mu^{1/2}} d^{3/4} P_T^{1/4} T^{3/4} \right) & \text{when } P_T \geq \frac{d}{k_\mu T}, \\ \widetilde{\mathcal{O}} \left(\frac{k_\mu}{c_\mu^{1/2}} d\sqrt{T} \right) & \text{when } 0 \leq P_T < \frac{d}{k_\mu T}. \end{cases}$$

Compared to GLB, we improve the order of c_μ from c_μ^{-1} to $c_\mu^{-1/2}$ by exploiting the self-concordant properties. At the same time, in near-stationary environments (P_T is small enough), our result can recover to the performance of LogUCB1 algorithm [Faury et al., 2020]. The proof of Theorem 3 is presented in Appendix C.2.

In addition, for the piecewise-stationary SCB, we propose SCB-PW-WeightUCB algorithm that gets rid of influence of c_μ and thus directly improves upon [Russac et al., 2021].

Theorem 4. *For all $\gamma \in (1/2, 1)$, $D = \log(T)/\log(1/\gamma)$ and $\lambda = d\log(T)/c_\mu$, the regret of SCB-PW-WeightUCB is bounded with probability at least $1 - 1/T$, by*

$$R_T \leq \widetilde{\mathcal{O}} \left(\frac{1}{1-\gamma} \Gamma_T + \frac{1}{\sqrt{1-\gamma}} + d\sqrt{(1-\gamma)} T \right).$$

By setting $\gamma = 1 - \max\{1/T, (\Gamma_T/(dT))^{2/3}\}$, we achieve

$$R_T \leq \begin{cases} \widetilde{\mathcal{O}} \left(d^{2/3} \Gamma_T^{1/3} T^{2/3} \right) & \text{when } \Gamma_T \geq d/\sqrt{T}, \\ \widetilde{\mathcal{O}} \left(d\sqrt{T} \right) & \text{when } 0 \leq \Gamma_T < d/\sqrt{T}. \end{cases}$$

The overall algorithm and analysis are in Appendix D.

6 REFINED ANALYSIS FRAMEWORK

This section presents a proof sketch for Lemma 1 (estimation error analysis for weighted linear bandits), which also serves as a description of our proposed analysis framework.

Proof Sketch. From the model assumption (1) and the estimator (3), the estimation error can be split into two parts,

$$\widehat{\theta}_t - \theta_t = \underbrace{V_{t-1}^{-1} \left(\sum_{s=1}^{t-1} \gamma^{t-s-1} X_s X_s^\top (\theta_s - \theta_t) \right)}_{\texttt{bias part}} + \underbrace{V_{t-1}^{-1} \left(\sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s - \lambda \theta_t \right)}_{\texttt{variance part}},$$

where the *bias* part is caused by the parameter drifting, and the *variance* part is due to the stochastic noise. Then, by the Cauchy-Schwarz inequality, for any $\mathbf{x} \in \mathcal{X}$,

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

where $A_t = \| \sum_{s=1}^{t-1} \gamma^{t-s-1} X_s X_s^\top (\theta_s - \theta_t) \|_{V_{t-1}^{-1}}$ and $B_t = \| \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s - \lambda \theta_t \|_{V_{t-1}^{-1}}$.

Choosing an appropriate local norm for (18) is the key to simplify and improve the estimation error analysis. We note that the previous analysis [Russac et al., 2019] had to use *different* local norms: using l_2-norm in the bias part, and $V_{t-1}^{-1} \widetilde{V}_{t-1} V_{t-1}^{-1}$-norm in the variance part, namely,

$$|\mathbf{x}^\top (\widehat{\theta}_t - \theta_t)| \leq \|\mathbf{x}\|_2 \, A_t' + \|\mathbf{x}\|_{V_{t-1}^{-1} \widetilde{V}_{t-1} V_{t-1}^{-1}} B_t', \tag{19}$$

where $A_t' = \|V_{t-1}^{-1} \sum_{s=1}^{t-1} \gamma^{t-s-1} X_s X_s^\top (\theta_s - \theta_t) \|_2$ and $B_t' = \| \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s - \lambda \theta_t \|_{\widetilde{V}_{t-1}^{-1}}$ and $\widetilde{V}_t = \lambda I_d + \sum_{s=1}^{t} \gamma^{2(t-s)} X_s X_s^\top$. Since the need for using sliding-window analysis to analyze the bias part, they have to use l_2-norm to get the format of A_t'. For the variance part, to use weighted version of self-normalized concentration (Theorem 5), they use $V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}$-norm to control $\mathbf{x}$ term so that B_t' term can be normed by $\widetilde{V}_{t-1}^{-1}$. As an improvement, we can directly use the *same* V_{t-1}^{-1}-norm to control both parts, which benefits from our new analysis for the bias part and modified analysis for the variance part.

Bias Part Analysis. The first step is to extract the variations of underlying parameters as follows,

$$A_t \leq L \sum_{p=1}^{t-1} \sum_{s=1}^{p} \gamma^{t-s-1} \|X_s\|_{V_{t-1}^{-1}} \|\theta_p - \theta_{p+1}\|_2.$$

Term $\sum_{s=1}^{p} \gamma^{t-s-1} \|X_s\|_{V_{t-1}^{-1}}$ should be able to further derive an expression about discounted factor γ, which can control the variation item. After some derivation, we get

$$A_t \leq L\sqrt{d} \sum_{p=1}^{t-1} \gamma^{\frac{t-1}{2}} \sqrt{\frac{\gamma^{-p}-1}{1-\gamma}} \|\theta_p - \theta_{p+1}\|_2 \,,$$

where the variation item is only controlled by the discounted factor γ instead of a virtual window size.

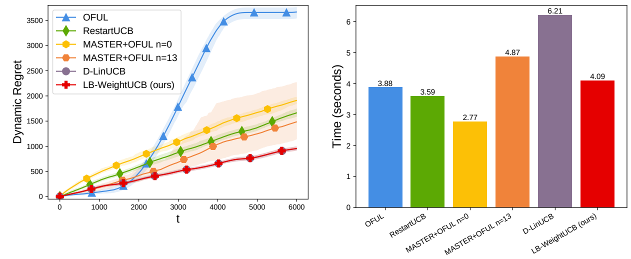
Variance Part Analysis. Based on the definition of V_t and $\widetilde{V}_t$, we can find that $V_t \succeq \widetilde{V}_t$, so we have $B_t \leq B_t' \leq \beta_{t-1}$ where β_{t-1} is the confidence radius (5) and the second inequality is by Theorem 5. So we keep the same confidence bound β_{t-1} while only need to compute V_{t-1}^{-1} instead of $V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}$ when doing the arm selection.

Combining the analysis for bias and variance parts, we can finish the proof of Lemma 1. □

Remark 1. The key step (18) in our analysis framework also resolves the projection issue in GLB. Specifically, after the projection step, the bias-variance decomposition can only be performed in V_{t-1}^{-1}-norm. To accommodate previous analysis (19), Faury et al. [2021] have to inject a highly complex projection operation in the algorithm, whereas our framework already satisfies this condition owing to the usage of the same V_{t-1}^{-1}-norm for the bias and variance parts.

7 EXPERIMENTS

In this section, we further empirically examine the performance of our proposed algorithms. We present two synthetic experiments on drifting LB and GLB, respectively. For each experiment, we set the dimension of the feature space to $d = 2$, the number of rounds to $T = 6000$, and the number of arms to $n = 50$. The features of each arm are sampled from the normal distribution $\mathcal{N}(0, 1)$ and subsequently rescaled to satisfy $L = 1$. We initialize the time-varying parameter θ_t to $[1, 0]$ and rotate it uniformly counterclockwise around the unit circle, completing one full revolution from 0 to 2π over the course of T rounds and returning to the starting point $[1, 0]$.



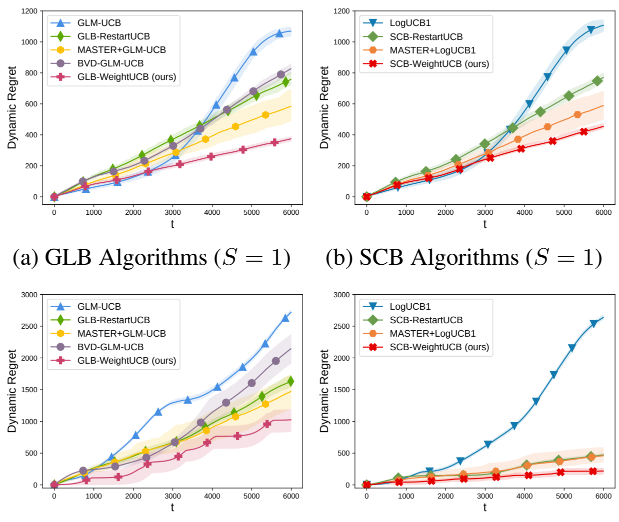
(a) LB: Cumulative regret (b) LB: Average running time

Figure 1: Experiments of linear bandits.

7.1 Linear Bandits

Setting. We consider the linear model $r_t = X_t^\top \theta_t + \eta_t$ where the random noise η_t is drawn from the normal distribution $\mathcal{N}(0, 1)$ at each time t independently. We compare the performance of our proposed LB-WeightUCB algorithm to: (a) the static algorithm OFUL [Abbasi-Yadkori et al., 2011]; (b) the restart-based algorithm RestartUCB [Zhao et al., 2020]; (c) the weight-based algorithm D-LinUCB [Russac et al., 2019]; and (d) the adaptive restart algorithm MASTER+OFUL [Wei and Luo, 2021]. Since P_T is computable, we set the discounted factor $\gamma = 1 - \max\{1/T, \sqrt{P_T/(dT)}\}$ for LB-WeightUCB and D-LinUCB, and set the window size w and restarting period H as $w = H = d^{1/4}\sqrt{T/(1 + P_T)}$. For MASTER, there is a parameter n representing the initial value of a multi-scale exploration parameter (see the input of Procedure 1 in [Wei and Luo, 2021]) and the origin MASTER algorithm lets it start from 0 (i.e., $n = 0, 1, ...$). However, a small initial value of n will lead to high-frequent restart and thus achieve poor performance. To address this issue, we experiment with a larger initial value of $n = 13$, which leads to greatly improved performance in our case.

Results. The experimental results are averaged over 20 independent trials. Figure 1a shows the cumulative dynamic regret performance, where the shaded area denotes the variance of the 20 independent trials of experimental results. Figure 1b reports the average time per run, with each run containing 6000 rounds. Our LB-WeightUCB algorithm performs as well as D-LinUCB but significantly more efficient, with over 1.5 times speedup. Figure 1a also shows that when equipped with a fine-tuned n, MASTER+OFUL ($n = 13$) performs better than RestartUCB, whereas a vanilla MASTER+OFUL ($n = 0$) performs worse due to overly active restarts at the beginning. However, a larger initial value of n results in greater time overhead, since at each restart, MASTER+OFUL needs to do Procedure 1 once, resulting in an $\mathcal{O}(n2^n)$ time complexity. More importantly, neither adaptive restart (MASTER+OFUL) nor periodical restart (RestartUCB) outperforms our weighted strategy in slowly-evolving environments.



(a) GLB Algorithms ($S = 1$) (b) SCB Algorithms ($S = 1$)

(c) GLB Algorithms ($S = 5$) (d) SCB Algorithms ($S = 5$)

Figure 2: Experiments of generalized linear bandits.

7.2 Generalized Linear Bandits

Setting. We employ the logistic model in GLB experiment, i.e., the reward satisfies $r_t \sim \text{Bernoulli}(\mu(X_t^\top \theta_t))$ with logistic function $\mu(x) = (1 + e^{-x})^{-1}$. We consider two cases of $S = 1$ and $S = 5$, respectively. We compare the performance of our proposed GLB-WeightUCB and SCB-WeightUCB algorithm to: (a) GLM-UCB, static algorithm for GLB [Filippi et al., 2010]; (b) LogUCB1, static algorithm for LogB [Faury et al., 2020]; (c) BVD-GLM-UCB, weight-based algorithm for GLB [Faury et al., 2021]; (d) GLB-RestartUCB, restart algorithm for GLB [Zhao et al., 2020]; (e) SCB-RestartUCB, restart algorihtm for SCB [Zhao et al., 2020]; (f) MASTER+GLM-UCB, adaptive restart algorithm for GLB [Wei and Luo, 2021]; and (g) MASTER+LogUCB1, adaptive restart algorithm for LogB [Wei and Luo, 2021]. We set discounted factor $\gamma = 1 - \max\{1/T, \sqrt{c_\mu P_T/(dT)}\}$ for GLB-WeightUCB, $\gamma = 1 - (P_T/(\sqrt{d}T))^{2/5}$ for BVD-GLM-UCB and $\gamma = 1 - \max\{1/T, \sqrt{P_T/(dT)}\}$ for SCB-WeightUCB. We set restarting period $H = d^{1/4}\sqrt{T/(1+P_T)}$ for both GLB-RestartUCB and SCB-RestartUCB. We set regularizer $\lambda = d$ for GLM-UCB, BVD-GLM-UCB, GLB-RestartUCB and MASTER+GLM-UCB, $\lambda = d/c_\mu^2$ for GLB-WeightUCB and $\lambda = d \log T/c_\mu$ for LogUCB1, SCB-RestartUCB, MASTER+LogUCB1 and SCB-WeightUCB. Note that for LogB, $k_\mu = 1/4 < 1$, so we don't need to control the order of k_μ. For the two MASTER algorithms, we set $n = 13$.

Results. We present the average cumulative dynamic regret results of our experiments on 20 independent trials in Figures 2. When S is small ($S = 1, c_\mu^{-1} \approx 5$), all of the weight-based algorithms outperform the static algorithms, and our GLB-WeightUCB and SCB-WeightUCB are better than BVD-GLM-UCB. When S is large ($S = 5, c_\mu^{-1} \approx 152$), SCB-WeightUCB significantly outperforms GLB-WeightUCB, demonstrating the importance of considering self-concordant property (recall that LogB is an instance of SCB). In contrast, the performance of BVD-GLM-UCB drops dramatically, as it does not take the c_μ^{-1} issue into account. Similar to LB, the experimental results of GLB also demonstrate the empirical advantage of the weighted strategy over (adaptive) restart strategy in slowly-evolving environments. Specifically, we observe that GLB-WeightUCB consistently outperforms MASTER+GLM-UCB, and SCB-WeightUCB consistently outperforms MASTER+LogUCB1.

8 CONCLUSION

This paper revisits the weight-based algorithms for three non-stationary parametric bandit models (LB, GLB, SCB). We identify that the inadequacies of the previous work are due to the inadequate analysis of the estimation error. We thus propose a refined analysis framework that enables the usage of the same local norm for both the bias and variance part in estimation error analysis. Our framework ensures more efficient algorithms for all three bandit models and improves the regret bounds for GLB and SCB settings.

The importance of our work lies in the fact that we have now made the weight-based algorithms for non-stationary LB/GLB/SCB as competitive as the restart-based algorithms, in terms of both computational efficiency and regret guarantee. Given that the weighted strategy is particularly appealing in gradually drifting scenarios that are commonly seen in real-world applications, it is essential to further design *adaptive* weight-based algorithms for non-stationary parametric bandits with optimal dynamic regret without requiring the knowledge of environmental non-stationarity, in the spirit of the currently best-known result achieved by adaptive restart strategy [Wei and Luo, 2021].

In this work, we employ $P_T = \sum_{t=2}^T \|\theta_{t-1} - \theta_t\|_2$ as a measure to capture the gradually changing environment. However, this metric may not be precise enough in capturing only the gradual changes in the environment, as it can also include other types of variations, such as abrupt changes. This might be able to serve as an explanation why weight-based algorithms do not exhibit a significant theoretical advantage, yet perform remarkably well in experiments on gradually changing environments compared to restart-based algorithms. To overcome this limitation, future research could explore more refined characterizations of gradual changes, drawing inspiration from the ideas behind Sobolev or Holder classes [Baby and Wang, 2019] or other information-theoretic tools [Liu et al., 2022].

Acknowledgements

This research was supported by NSFC (61921006), JiangsuSF (BK20220776) and Collaborative Innovation Center of Novel Software Technology and Industrialization. The authors greatly appreciate helpful comments and suggestions from anonymous reviewers and meta-reviewer. The authors also thank Yu-Hu Yan and Yu-Jie Zhang for many helpful discussions.

References

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

Yasin Abbasi-Yadkori, András György, and Nevena Lazic. A new look at dynamic regret for non-stationary stochastic bandits. *ArXiv preprint*, arXiv: 2201.06532, 2022.

Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3691–3699, 2021.

Christoforos Anagnostopoulos, Dimitris K Tasoulis, Niall M Adams, Nicos G Pavlidis, and David J Hand. Online linear and quadratic discriminant analysis with adaptive forgetting for streaming classification. *Statistical Analysis and Data Mining*, 5(2):139–166, 2012.

Dheeraj Baby and Yu-Xiang Wang. Online forecasting of total-variation-bounded sequences. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 11069–11079, 2019.

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, 2019.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Hedging the drift: Learning to optimize under nonstationarity. *Manag. Sci.*, 68(3):1696–1713, 2022.

Chao-Kai Chiang, Chia-Jung Lee, and Chi-Jen Lu. Beating bandits in gradually evolving worlds. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, pages 210–227, 2013.

Yijing Chu and Cheuk Ming Mak. A variable forgetting factor diffusion recursive least squares algorithm for distributed estimation. *Signal Processing*, 140:219–225, 2017.

Giulia Clerici, Pierre Laforgue, and Nicolò Cesa-Bianchi. Linear bandits with memory: from rotting to rising. *ArXiv preprint*, arXiv:2302.08345, 2023.

Koby Crammer, Yishay Mansour, Eyal Even-Dar, and Jennifer Wortman Vaughan. Regret minimization with concept drift. In *Proceedings of the 23rd Conference on Learning Theory (COLT)*, pages 168–180, 2010.

Yuntian Deng, Xingyu Zhou, Baekjin Kim, Ambuj Tewari, Abhishek Gupta, and Ness Shroff. Weighted Gaussian process bandits for non-stationary environments. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 6909–6932, 2022.

Shi Dong, Tengyu Ma, and Benjamin Van Roy. On the performance of thompson sampling on logistic bandits. In *Proceedings of the 32rd Conference on Learning Theory (COLT)*, pages 1158–1160, 2019.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 3052–3060, 2020.

Louis Faury, Yoan Russac, Marc Abeille, and Clément Calauzènes. Regret bounds for generalized linear bandits under parameter drift. *ArXiv preprint*, arXiv:2103.05750, 2021.

Sarah Filippi, Olivier Cappé, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems 23 (NIPS)*, pages 586–594, 2010.

João Gama, Indre Zliobaite, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4): 44:1–44:37, 2014.

Lei Guo, Lennart Ljung, and Pierre Priouret. Performance analysis of the forgetting factor RLS algorithm. *International Journal of Adaptive Control and Signal Processing*, 7(6):525–537, 1993.

Simon S Haykin. *Adaptive filter theory*. Pearson Education India, 2002.

Wasim Huleihel, Soumyabrata Pal, and Ofer Shayevitz. Learning user preferences in non-stationary environments. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1432–1440, 2021.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of 33rd Conference on Learning Theory (COLT)*, pages 2137–2143, 2020.

Kwang-Sung Jun, Aniruddha Bhargava, Robert D. Nowak, and Rebecca Willett. Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 99–109, 2017.

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

Yueyang Liu, Benjamin Van Roy, and Kuang Xu. Nonstationary bandit learning via predictive sampling. *ArXiv preprint*, arXiv:2205.01970, 2022.

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.

Yoan Russac, Louis Faury, Olivier Cappé, and Aurélien Garivier. Self-concordant analysis of generalized linear bandits with forgetting. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 658–666, 2021.

Joe Suk and Samory Kpotufe. Tracking most significant arm switches in bandits. In *Proceedings of the 35th Conference on Learning Theory (COLT)*, pages 2160–2182, 2022.

Sabina Tomkins, Peng Liao, Predrag V. Klasnja, and Susan A. Murphy. Intelligentpooling: practical Thompson sampling for mhealth. *Machine Learning*, 110(9):2685–2727, 2021.

Ahmed Touati and Pascal Vincent. Efficient learning in non-stationary linear markov decision processes. *ArXiv preprint*, arXiv:2010.12870, 2020.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, pages 4300–4354, 2021.

Lijun Zhang, Tianbao Yang, Rong Jin, Yichi Xiao, and Zhi-Hua Zhou. Online stochastic linear optimization under one-bit feedback. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 392–401, 2016.

Peng Zhao and Lijun Zhang. Non-stationary linear bandits revisited. *ArXiv preprint*, arXiv:2103.05324, 2021.

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.

Peng Zhao, Xinqiang Wang, Siyu Xie, Lei Guo, and Zhi-Hua Zhou. Distribution-free one-pass learning. *IEEE Transaction on Knowledge and Data Engineering*, 33: 951–963, 2021a.

Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. A simple approach for non-stationary linear bandits. *ArXiv preprint*, arXiv:2103.05324, 2021b.

A Analysis of LB-WeightUCB

In this section, we provide the analysis for LB-WeightUCB algorithm. In Appendix A.1, we review the D-LinUCB algorithm proposed by Russac et al. [2019] and restate their estimation error analysis. In Appendix A.2, we present our own estimation error analysis for the proposed LB-WeightUCB algorithm, which is captured in Lemma 1. Finally, in Appendix A.3, we provide a proof for our dynamic regret bound, as stated in Theorem 1.

A.1 Review Estimation Error Analysis of D-LinUCB Algorithm

In this part, we review the previous estimation error analysis of the D-LinUCB algorithm [Russac et al., 2019] who has the same estimator as ours (3). The first step is to divide the estimation error into the bias and variance parts, where the bias part represents the error caused by parameter drift and the variance part represents the error caused by stochastic noise. Based on the reward model assumption and the estimator (same as eq (1) and eq (3)), the estimation error of D-LinUCB algorithm can be decomposed as

$$\begin{aligned}
\widehat{\theta}_t - \theta_t &= V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\gamma^{t-s-1} r_s X_s\right) - \theta_t \\
&= V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\gamma^{t-s-1}\left(X_s^\top\theta_s + \eta_s\right)X_s\right) - V_{t-1}^{-1}\left(\lambda I_d + \sum_{s=1}^{t-1}\gamma^{t-s-1}X_sX_s^\top\right)\theta_t \\
&= V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\gamma^{t-s-1}X_sX_s^\top\theta_s + \sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_sX_s\right) - V_{t-1}^{-1}\left(\lambda I_d + \sum_{s=1}^{t-1}\gamma^{t-s-1}X_sX_s^\top\right)\theta_t \\
&= \underbrace{V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\gamma^{t-s-1}X_sX_s^\top\left(\theta_s-\theta_t\right)\right)}_{\texttt{bias part}} + \underbrace{V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_sX_s - \lambda\theta_t\right)}_{\texttt{variance part}}.
\end{aligned} \tag{20}$$

Afterward, Russac et al. [2019] use different local norms (we will explain the reason of using different local norms later) for the bias and variance parts as follows,

$$|\mathbf{x}^\top(\widehat{\theta}_t - \theta_t)| \le \|\mathbf{x}\|_2 A_t' + \|\mathbf{x}\|_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}} B_t', \tag{21}$$

where $\widetilde{V}_t = \lambda I_d + \sum_{s=1}^{t}\gamma^{2(t-s)}X_sX_s^\top$ and

$$A_t' = \left\|V_{t-1}^{-1}\sum_{s=1}^{t-1}\gamma^{t-s-1}X_sX_s^\top\left(\theta_s-\theta_t\right)\right\|_2, \quad B_t' = \left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_sX_s - \lambda\theta_t\right\|_{\widetilde{V}_{t-1}^{-1}}.$$

For the bias part, Russac et al. [2019] divide it into two parts on the timeline by introducing a virtual window size D,

$$A_t' \le \underbrace{\left\|\sum_{s=t-D}^{t-1}V_{t-1}^{-1}\gamma^{t-s-1}X_sX_s^\top\left(\theta_s-\theta_t\right)\right\|_2}_{\texttt{virtual window}} + \underbrace{\left\|\sum_{s=1}^{t-D-1}V_{t-1}^{-1}\gamma^{t-s-1}X_sX_s^\top\left(\theta_s-\theta_t\right)\right\|_2}_{\texttt{small term}},$$

The first term can be considered as a virtual window containing the most recent data obtained after time $t-D$, and can be directly analyzed by the analysis of SW-UCB [Cheung et al., 2019] since it corresponds to the bias part of the estimation error of window strategy and this is why they use l_2-norm for bias part. The second term reflects the influence formed by the outdated data obtained before time $t-D$. Since γ^{t-s-1} will be very small when $s \le t-D-1$, this small term is dominated by the first virtual window term which means the bias part is actually controlled by the virtual window size D.

For the variance part, Russac et al. [2019] extend the previous self-normalized concentration [Abbasi-Yadkori et al., 2011, Theorem 1] to the weighted version which is restated in Theorem 5. This concentration requires to use $\widetilde{V}_t$ as the local norm. To this end, Russac et al. [2019] split the variance part as

$$\left|\mathbf{x}^\top V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_sX_s - \lambda\theta_t\right)\right| \le \|\mathbf{x}\|_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}}\left\|V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_sX_s - \lambda\theta_t\right)\right\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}},$$

where

$$\left\|V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_s X_s-\lambda\theta_t\right)\right\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}}=\left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_s X_s-\lambda\theta_t\right\|_{\widetilde{V}_{t-1}^{-1}}\leq\left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_s X_s\right\|_{\widetilde{V}_{t-1}^{-1}}+\sqrt{\lambda}S.$$

Then term $\|\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_s X_s\|_{\widetilde{V}_{t-1}^{-1}}$ can be bounded by Theorem 5. Finally, based on this analysis, D-LinUCB needs to use the following action selection criterion which only depends on the variance part since the bias part doesn't contain $\mathbf{x}$,

$$X_t=\operatorname*{arg\,max}_{\mathbf{x}\in\mathcal{X}}\left\{\langle\mathbf{x},\widehat{\theta}_t\rangle+\beta_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}}\right\},$$

where β_{t-1} is the upper bound of B_t' which is the same as (5). From this selection criterion, it can be seen that D-LinUCB needs to maintain two covariance matrices, namely, V_t and $\widetilde{V}_t$ at round t during the algorithm running.

In the next section, we present our proof for the estimation error upper bound. The difference between our analysis and D-LinUCB's analysis mainly starts at step (21), which is the key step of the analysis and our new analysis framework allows us to employ the *same* local norm for both bias and variance parts.

A.2 Proof of Lemma 1

Proof. Using the same derivation in (20), the estimation error of LB-WeightUCB algorithm can also be decomposed as

$$\widehat{\theta}_t-\theta_t=\underbrace{V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\gamma^{t-s-1}X_sX_s^\top(\theta_s-\theta_t)\right)}_{\texttt{bias part}}+\underbrace{V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_sX_s-\lambda\theta_t\right)}_{\texttt{variance part}}.$$

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

$$\left|\mathbf{x}^\top\left(\widehat{\theta}_t-\theta_t\right)\right|\leq\|\mathbf{x}\|_{V_{t-1}^{-1}}(A_t+B_t),\tag{22}$$

where

$$A_t=\left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}X_sX_s^\top(\theta_s-\theta_t)\right\|_{V_{t-1}^{-1}},\quad B_t=\left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_sX_s-\lambda\theta_t\right\|_{V_{t-1}^{-1}}.$$

The above two terms can be bounded separately, as summarized in the following two lemmas,

Lemma 4. *For any $t\in[T]$, we have*

$$\left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}X_sX_s^\top(\theta_s-\theta_t)\right\|_{V_{t-1}^{-1}}\leq L\sqrt{d}\sum_{p=1}^{t-1}\gamma^{\frac{t-1}{2}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\|\theta_p-\theta_{p+1}\|_2.$$

Lemma 5. *For any $\delta\in(0,1)$, with probability at least $1-\delta$, the following holds for all $t\in[T]$,*

$$\left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_sX_s-\lambda\theta_t\right\|_{V_{t-1}^{-1}}\leq\sqrt{\lambda}S+R\sqrt{2\log\frac{1}{\delta}+d\log\left(1+\frac{L^2(1-\gamma^{2t-2})}{\lambda d(1-\gamma^2)}\right)},$$

Based on the inequality (22), Lemma 4, Lemma 5, and the boundedness assumption of the feasible set, for any $\mathbf{x}\in\mathcal{X}$, $\gamma\in(0,1)$ and $\delta\in(0,1)$, with probability at least $1-\delta$, the following holds for all $t\in[T]$,

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

where $\beta_t\triangleq\sqrt{\lambda}S+R\sqrt{2\log\frac{1}{\delta}+d\log\left(1+\frac{L^2(1-\gamma^{2t})}{\lambda d(1-\gamma^2)}\right)}$ is the confidence radius used in LB-WeightUCB. Hence we complete the proof. □

Proof of Lemma 4. The first step is to extract the variations of the parameter θ_t as follows,

$$\begin{aligned}\left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}X_sX_s^\top(\theta_s-\theta_t)\right\|_{V_{t-1}^{-1}} &= \left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}X_sX_s^\top\sum_{p=s}^{t-1}(\theta_p-\theta_{p+1})\right\|_{V_{t-1}^{-1}}\\ &= \left\|\sum_{p=1}^{t-1}\sum_{s=1}^{p}\gamma^{t-s-1}X_sX_s^\top(\theta_p-\theta_{p+1})\right\|_{V_{t-1}^{-1}}\\ &\leq \sum_{p=1}^{t-1}\left\|\sum_{s=1}^{p}\gamma^{t-s-1}X_s\|X_s\|_2\|\theta_p-\theta_{p+1}\|_2\right\|_{V_{t-1}^{-1}}\\ &\leq L\sum_{p=1}^{t-1}\sum_{s=1}^{p}\gamma^{t-s-1}\|X_s\|_{V_{t-1}^{-1}}\|\theta_p-\theta_{p+1}\|_2,\end{aligned}$$

and term $\sum_{s=1}^{p}\gamma^{t-s-1}\|X_s\|_{V_{t-1}^{-1}}$ can be able to further derive an expression about discounted factor γ as follows,

$$\sum_{s=1}^{p}\gamma^{t-s-1}\|X_s\|_{V_{t-1}^{-1}} \leq \gamma^{\frac{t-1}{2}}\sqrt{\sum_{s=1}^{p}\gamma^{-s}}\sqrt{\sum_{s=1}^{p}\gamma^{t-s-1}\|X_s\|^2_{V_{t-1}^{-1}}} \leq \sqrt{d}\gamma^{\frac{t-1}{2}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}. \tag{23}$$

In above, we use the fact that for any $\mathbf{x}$, $\|\mathbf{x}\|_{V_{t-1}^{-1}} \leq \|\mathbf{x}\|_2/\sqrt{\lambda}$ since $V_{t-1}\succeq\lambda I_d$. The second last step holds by the Cauchy-Schwarz inequality. Besides, the last step follows the fact,

$$\forall p\in[t-1],\quad \sum_{s=1}^{p}\gamma^{t-s-1}\|X_s\|^2_{V_{t-1}^{-1}}\leq d, \tag{24}$$

which can be proven by the following argument.

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

Hence, we complete the proof. □

Proof of Lemma 5. Let $\widetilde{V}_t \triangleq \lambda I_d + \sum_{s=1}^{t}\gamma^{2(t-s)}X_sX_s^\top$,

$$\left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_sX_s-\lambda\theta_t\right\|_{V_{t-1}^{-1}} \leq \left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_sX_s\right\|_{V_{t-1}^{-1}}+\|\lambda\theta_t\|_{V_{t-1}^{-1}} \leq \left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_sX_s\right\|_{\widetilde{V}_{t-1}^{-1}}+\sqrt{\lambda}S.$$

Recall that $V_t=\lambda I_d+\sum_{s=1}^{t}\gamma^{t-s}X_sX_s^\top$, so the last inequality comes from

$$V_t=\lambda I_d+\sum_{s=1}^{t}\gamma^{t-s}X_sX_s^\top \succeq \lambda I_d+\sum_{s=1}^{t}\gamma^{2(t-s)}X_sX_s^\top=\widetilde{V}_t.$$

We emphasize that the $\widetilde{V}_t$ is introduced into analysis *only*, which is actually *not* required in our algorithmic implementation. From the weighted version maximal deviation inequality [Russac et al., 2019, Theorem 1], restated in Theorem 5, we can get the bound for the first term $\|\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_sX_s\|_{\widetilde{V}_{t-1}^{-1}}$ as below by just let $w_s=\gamma^{t-s-1}$, $\mu_t=\lambda$,

$$\left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_sX_s\right\|_{\widetilde{V}_{t-1}^{-1}} \leq R\sqrt{2\log\frac{1}{\delta}+d\log\left(1+\frac{L^2\sum_{s=1}^{t-1}\gamma^{2(t-s-1)}}{\lambda d}\right)} \leq R\sqrt{2\log\frac{1}{\delta}+d\log\left(1+\frac{L^2(1-\gamma^{2t-2})}{\lambda d(1-\gamma^2)}\right)},$$

which completes the proof. □

A.3 Proof of Theorem 1

Proof. Let $X_t^* \triangleq \arg\max_{\mathbf{x}\in\mathcal{X}} \mathbf{x}^\top \theta_t$. Due to Lemma 1 and the fact that $X_t^*, X_t \in \mathcal{X}$, each of the following holds with probability at least $1-\delta$,

$$\forall t \in [T], X_t^{*\top}\theta_t \leq X_t^{*\top}\widehat{\theta}_t + L^2\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{t-1}\gamma^{\frac{t-1}{2}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\,\|\theta_p-\theta_{p+1}\|_2 + \beta_{t-1}\|X_t^*\|_{V_{t-1}^{-1}}$$

$$\forall t \in [T], X_t^{\top}\theta_t \geq X_t^{\top}\widehat{\theta}_t - L^2\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{t-1}\gamma^{\frac{t-1}{2}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\,\|\theta_p-\theta_{p+1}\|_2 - \beta_{t-1}\|X_t\|_{V_{t-1}^{-1}}.$$

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

$$\begin{aligned}\forall t \in [T], X_t^{*\top}\theta_t - X_t^\top\theta_t &\leq X_t^{*\top}\widehat{\theta}_t - X_t^\top\widehat{\theta}_t + 2L^2\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{t-1}\gamma^{\frac{t-1}{2}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\,\|\theta_p-\theta_{p+1}\|_2 + \beta_{t-1}(\|X_t^*\|_{V_{t-1}^{-1}} + \|X_t\|_{V_{t-1}^{-1}})\\ &\leq 2L^2\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{t-1}\gamma^{\frac{t-1}{2}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\,\|\theta_p-\theta_{p+1}\|_2 + 2\beta_{t-1}\|X_t\|_{V_{t-1}^{-1}},\end{aligned}$$

where the last step comes from the arm selection criterion (6) such that

$$X_t^{*\top}\widehat{\theta}_t + \beta_{t-1}\|X_t^*\|_{V_{t-1}^{-1}} \leq X_t^\top\widehat{\theta}_t + \beta_{t-1}\|X_t\|_{V_{t-1}^{-1}}.$$

Hence, the following dynamic regret bound holds with probability at least $1-2\delta$ and can be divided into two parts,

$$R_T = \sum_{t=1}^T\left(X_t^{*\top}\theta_t - X_t^\top\theta_t\right) \leq \underbrace{2L^2\sqrt{\frac{d}{\lambda}}\sum_{t=1}^T\sum_{p=1}^{t-1}\gamma^{\frac{t-1}{2}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\,\|\theta_p-\theta_{p+1}\|_2}_{\texttt{bias part}} + \underbrace{2\beta_T\sum_{t=1}^T\|X_t\|_{V_{t-1}^{-1}}}_{\texttt{variance part}},$$

where $\beta_T = \sqrt{\lambda}S + R\sqrt{2\log\frac{1}{\delta} + d\log\left(1+\frac{L^2(1-\gamma^{2T})}{\lambda d(1-\gamma^2)}\right)}$ is the confidence radius.

Now we derive the upper bound for the bias and variance parts separately.

Bias Part. For the bias part, we need to extract path length P_T and show the control of the discounted factor γ on P_T.

$$\begin{aligned}2L^2\sqrt{\frac{d}{\lambda}}\sum_{t=1}^T\sum_{p=1}^{t-1}\gamma^{\frac{t-1}{2}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\,\|\theta_p-\theta_{p+1}\|_2 &= 2L^2\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{T-1}\sum_{t=p+1}^{T}\gamma^{\frac{t-1}{2}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\,\|\theta_p-\theta_{p+1}\|_2\\ &= 2L^2\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{T-1}\frac{\gamma^{\frac{p}{2}}-\gamma^{\frac{T}{2}}}{1-\gamma^{\frac{1}{2}}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\,\|\theta_p-\theta_{p+1}\|_2\\ &\leq 2L^2\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{T-1}\frac{\gamma^{\frac{p}{2}}-\gamma^{\frac{T}{2}}}{(1-\gamma^{\frac{1}{2}})\frac{1+\gamma^{\frac{1}{2}}}{2}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\,\|\theta_p-\theta_{p+1}\|_2\\ &\leq 4L^2\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{T-1}\frac{\gamma^{\frac{p}{2}}\gamma^{-\frac{p}{2}}}{(1-\gamma)^{3/2}}\,\|\theta_p-\theta_{p+1}\|_2\\ &= 4L^2\sqrt{\frac{d}{\lambda}}\frac{1}{(1-\gamma)^{3/2}}P_T.\end{aligned}$$

So for the bias part, we have

$$2L^2\sqrt{\frac{d}{\lambda}}\sum_{t=1}^T\sum_{p=1}^{t-1}\gamma^{\frac{t-1}{2}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\,\|\theta_p-\theta_{p+1}\|_2 \leq 4L^2\sqrt{\frac{d}{\lambda}}\frac{1}{(1-\gamma)^{3/2}}P_T. \tag{25}$$

Variance Part. First, use the Cauchy-Schwarz inequality, we know that $2\beta_T\sum_{t=1}^T\|X_t\|_{V_{t-1}^{-1}} \leq 2\beta_T\sqrt{T\sum_{t=1}^T\|X_t\|^2_{V_{t-1}^{-1}}}$.

Then by Lemma 11 (potential lemma), we have the following upper bound:

$$2\beta_T \sum_{t=1}^{T} \|X_t\|_{V_{t-1}^{-1}} \leq 2\beta_T \sqrt{2\max\{1, L^2/\lambda\} dT} \sqrt{T \log \frac{1}{\gamma} + \log\left(1 + \frac{L^2}{\lambda d(1-\gamma)}\right)}. \tag{26}$$

Combining the upper bounds of the bias and variance parts and with confidence level $\delta = 1/(2T)$, by union bound we have the following dynamic regret bound with probability at least $1 - 1/T$,

$$R_T \leq 4L^2 \sqrt{\frac{d}{\lambda}} \frac{1}{(1-\gamma)^{3/2}} P_T + 2\beta_T \sqrt{2\max\{1, L^2/\lambda\} dT} \sqrt{T \log \frac{1}{\gamma} + \log\left(1 + \frac{L^2}{\lambda d(1-\gamma)}\right)}.$$

where $\beta_T = \sqrt{\lambda} S + R\sqrt{2\log T + 2\log 2 + d\log\left(1 + \frac{L^2(1-\gamma^{2T})}{\lambda d(1-\gamma^2)}\right)}$. Since that there has a term $T\sqrt{\log(1/\gamma)}$ in the regret bound, we cannot let γ close to 0, so we set $\gamma \geq 1/T$ and have $\log(1/\gamma) \leq C(1-\gamma)$, where $C = \log T/(1 - 1/T)$.

Then, ignoring logarithmic factors in time horizon T, and let $\lambda = d$, we finally obtain

$$R_T \leq \widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^{3/2}} P_T + d(1-\gamma)^{1/2} T\right).$$

When $P_T < d/T$ (which corresponds a small amount of non-stationarity), we simply set $\gamma = 1 - 1/T$ and achieve an $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret bound. Besides, when coming to the non-degenerated case ($P_T \geq d/T$), We set the discounted factor optimally as $1 - \gamma = \sqrt{P_T/(dT)}$ and attain an $\widetilde{\mathcal{O}}(d^{3/4} P_T^{1/4} T^{3/4})$ dynamic regret bound, which completes the proof. □

B Analysis of GLB-WeightUCB

In this section, we provide the analysis for GLB-WeightUCB algorithm. In Appendix B.1, we review the projection issue of GLB and restate the BVD-GLM-UCB algorithm of Faury et al. [2021]. In Appendix B.2, we present the proof of the estimation error upper bound of our GLB-WeightUCB algorithm (namely, Lemma 2). Finally, in Appendix B.3, we provide the proof of dynamic regret upper bound as stated in Theorem 2.

B.1 Review Projection Step of BVD-GLM-UCB Algorithm

As mentioned in Section 4.2, the main difficulty of GLB is that the result of MLE or QMLE estimator $\widehat{\theta}_t$ may not belong to the feasible set Θ and c_μ is defined over the parameter $\theta \in \Theta$. Under stationary environments, Filippi et al. [2010] overcame this difficulty by introducing a projection step as

$$\widetilde{\theta}_t = \underset{\theta \in \Theta}{\arg\min} \|g_t(\widehat{\theta}_t) - g_t(\theta)\|_{V_{t-1}^{-1}}, \tag{27}$$

where $V_t = \lambda I_d + \sum_{s=1}^{t} X_s X_s^\top$ and $g_t(\theta) = \lambda c_\mu \theta + \sum_{s=1}^{t-1} \mu(X_s^\top \theta) X_s$ are the static version (by setting $\gamma = 1$). Based on the QMLE, we know that

$$g_t(\widehat{\theta}_t) = \lambda c_\mu \widehat{\theta}_t + \sum_{s=1}^{t-1} \mu(X_s^\top \widehat{\theta}_t) X_s = \sum_{s=1}^{t-1} r_s X_s, \tag{28}$$

and then by the mean value theorem, we know that

$$g_t(\theta_1) - g_t(\theta_2) = G_t(\theta_1, \theta_2)(\theta_1 - \theta_2), \tag{29}$$

where $G_t(\theta_1, \theta_2) \triangleq \int_0^1 \nabla g_t(s\theta_2 + (1-s)\theta_1) \mathrm{d}s \in \mathbb{R}^{d \times d}$. Notice that for any $\theta \in \Theta$, the gradient of g_t satisfies

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

which clearly implies $\forall \theta_1, \theta_2 \in \Theta, G_t(\theta_1, \theta_2) \succeq c_\mu V_{t-1}$.

Algorithm 3 BVD-GLM-UCB [Faury et al., 2021]

Input: time horizon T, discounted factor γ, confidence δ, regularizer λ, inverse link function μ, parameters S, L and R

1: Set $V_0 = \lambda I_d$, $\widehat{\theta}_1 = \mathbf{0}$ and compute k_μ and c_μ
2: **for** $t = 1, 2, ..., T$ **do**
3: Find θ_t^p by solving $\theta_t^p \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \left\| g_t(\theta) - g_t(\widehat{\theta}_t) \right\|_{V_t^{-2}} \text{ s.t } \Theta \cap \mathcal{E}_t^\delta(\theta) \neq \emptyset \right\}$
4: Select $\widetilde{\theta}_t \in \Theta \cap \mathcal{E}_t^\delta(\theta_t^p)$ where $\mathcal{E}_t^\delta(\theta) := \left\{ \theta' \in \mathbb{R}^d \,\middle|\, \|g_t(\theta') - g_t(\theta)\|_{\widetilde{V}_t^{-1}} \leq \bar{\beta}_t(\delta) \right\}$
5: Compute $\bar{\beta}_{t-1}$ by $\bar{\beta}_t = \sqrt{\lambda} c_\mu S + R\sqrt{2\log\frac{1}{\delta} + d\log\left(1 + \frac{L^2(1-\gamma^{2t})}{\lambda d(1-\gamma^2)}\right)}$
6: Select X_t by $X_t = \arg\max_{\mathbf{x} \in \mathcal{X}} \left\{ \mu(\mathbf{x}^\top \widetilde{\theta}_t) + \frac{2k_\mu}{c_\mu} \bar{\beta}_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}} \right\}$
7: Receive the reward r_t
8: Update $V_t = \gamma V_{t-1} + X_t X_t^\top + (1-\gamma)\lambda I_d$, $\widetilde{V}_t = \gamma^2 V_{t-1} + X_t X_t^\top + (1-\gamma^2)\lambda I_d$
9: Compute $\widehat{\theta}_{t+1}$ according to $\lambda c_\mu \theta + \sum_{s=1}^t \gamma^{t-s} \left(\mu(X_s^\top \theta) - r_s\right) X_s = 0$
10: **end for**

By this projection step, Filippi et al. [2010] can analyze the estimation error like,

$$\begin{aligned}
|\mu(\mathbf{x}^\top \widetilde{\theta}_t) - \mu(\mathbf{x}^\top \theta_t)| &\leq k_\mu |\mathbf{x}^\top (\widetilde{\theta}_t - \theta_t)| \\
&= k_\mu |\mathbf{x}^\top G_t^{-1}(\theta_t, \widetilde{\theta}_t)(g_t(\widetilde{\theta}_t) - g_t(\theta_t))| \\
&\leq k_\mu \|\mathbf{x}\|_{G_t^{-1}(\theta_t, \widetilde{\theta}_t)} \|g_t(\widetilde{\theta}_t) - g_t(\theta_t)\|_{G_t^{-1}(\theta_t, \widetilde{\theta}_t)} \\
&\leq \frac{k_\mu}{c_\mu} \|\mathbf{x}\|_{V_{t-1}^{-1}} \|g_t(\widetilde{\theta}_t) - g_t(\theta_t)\|_{V_{t-1}^{-1}} \\
&\leq \frac{2k_\mu}{c_\mu} \|\mathbf{x}\|_{V_{t-1}^{-1}} \|g_t(\widehat{\theta}_t) - g_t(\theta_t)\|_{V_{t-1}^{-1}},
\end{aligned}$$

where the last step comes from the projection step. After doing the projection step, term $g_t(\widehat{\theta}_t) - g_t(\theta_t)$ is the estimation error of the MLE without projection. Notice that in piecewise-stationary case, Russac et al. [2021] can also use this projection step. Faury et al. [2021] believe that these two previous works could use this projection operation mainly due to their stationary or piecewise-stationary setting. They mention that for the drifting case, the estimation error is always divided into the bias (tracking error) and variance (learning error) part, and this simple projection operation ignores the bias part which needs to be generalized to adapt to the two sources of deviation. In the analysis, the problem is that after the projection step estimation error term $g_t(\widehat{\theta}_t) - g_t(\theta_t)$ need to be separate into the bias part and variance parts, and Faury et al. [2021] need to use l_2-norm for bias part and V_{t-1}^{-1} for variance part. But the whole estimation error is already normed by V_{t-1}^{-1}, which means they cannot use the previous analysis of the window strategy for the bias part.

To this end, Faury et al. [2021] propose the BVD-GLM-UCB algorithm for drifting generalized linear bandits, as restated in Algorithm 3, where a new projection step is devised to solve this problem. Specifically, at each round t, the first step is to construct the confidence set $\mathcal{E}_t^\delta(\theta)$ which represents the influence of the stochastic noise.

$$\mathcal{E}_t^\delta(\theta) := \left\{ \theta' \in \mathbb{R}^d \,\middle|\, \|g_t(\theta') - g_t(\theta)\|_{\widetilde{V}_t^{-1}} \leq \bar{\beta}_t(\delta) \right\}. \tag{30}$$

The second step is to find a confidence set $\mathcal{E}_t^\delta(\theta_t^p)$ that intersects with the feasible set, and the gap between θ_t^p and $\widehat{\theta}_t$ represents the influence of parameter drift.

$$\theta_t^p \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \left\| g_t(\theta) - g_t(\widehat{\theta}_t) \right\|_{V_t^{-2}} \text{ s.t } \Theta \cap \mathcal{E}_t^\delta(\theta) \neq \emptyset \right\}. \tag{31}$$

After obtaining the solution θ_t^p via computing the optimization problem (31), the third step is to select $\widetilde{\theta}_t$ from $\Theta \cap \mathcal{E}_t^\delta(\theta_t^p)$. Based on this projection step, Faury et al. [2021] can separate the bias and variance parts before projection as follows,

$$\begin{aligned}
|\mu(\mathbf{x}^\top \widetilde{\theta}_t) - \mu(\mathbf{x}^\top \theta_t)| &\leq k_\mu |\mathbf{x}^\top (\widetilde{\theta}_t - \theta_t)| \\
&= k_\mu |\mathbf{x}^\top G_t^{-1}(\theta_t, \widetilde{\theta}_t)(g_t(\widetilde{\theta}_t) - g_t(\theta_t))| \qquad (32)
\end{aligned}$$

$$\leq k_\mu|\mathbf{x}^\top G_t^{-1}(\theta_t,\widetilde{\theta}_t)(g_t(\widetilde{\theta}_t)-g_t(\theta_t^p)+g_t(\theta_t^p)-g_t(\widehat{\theta}_t)+g_t(\widehat{\theta}_t)-g_t(\bar{\theta}_t)+g_t(\bar{\theta}_t)-g_t(\theta_t))| \tag{33}$$

$$\leq \underbrace{k_\mu|\mathbf{x}^\top G_t^{-1}(\theta_t,\widetilde{\theta}_t)(g_t(\widetilde{\theta}_t)-g_t(\theta_t^p)+g_t(\widehat{\theta}_t)-g_t(\bar{\theta}_t)|}_{\texttt{bias part}} \tag{34}$$

$$+\underbrace{k_\mu|\mathbf{x}^\top G_t^{-1}(\theta_t,\widetilde{\theta}_t)(g_t(\theta_t^p)-g_t(\widehat{\theta}_t)+g_t(\bar{\theta}_t)-g_t(\theta_t))|}_{\texttt{variance part}}. \tag{35}$$

Their bias-variance decomposition motivates the choice of *different* local norms for bounding bias and variance parts in their algorithm and analysis. Notably, due to the complications of the projection step (see (30) and (31)), the overall algorithm is fairly complicated and less attractive for practical implementations, and moreover, it needs to maintain two covariance matrices V_t and $\widetilde{V}_t$ (due to the constructed confidence region (30)) at each round t during the algorithm running. In the next section, we will show that the simple projection used in the stationary GLB (27) can be sufficient for coping with the drifting GLB via our refined analysis framework.

B.2 Proof of Lemma 2

Proof. Base on the estimator equation (8), we know that

$$g_t(\widehat{\theta}_t)=\lambda c_\mu\widehat{\theta}_t+\sum_{s=1}^{t-1}\gamma^{t-s-1}\mu(X_s^\top\widehat{\theta}_t)X_s=\sum_{s=1}^{t-1}\gamma^{t-s-1}r_sX_s, \tag{36}$$

and then by the mean value theorem, we know that

$$g_t(\theta_1)-g_t(\theta_2)=G_t(\theta_1,\theta_2)(\theta_1-\theta_2), \tag{37}$$

where $G_t(\theta_1,\theta_2)\triangleq\int_0^1\nabla g_t(s\theta_2+(1-s)\theta_1)\mathrm{d}s\in\mathbb{R}^{d\times d}$. Notice that for any $\theta\in\Theta$, the gradient of g_t is

$$\nabla g_t(\theta)=\lambda c_\mu I_d+\sum_{s=1}^{t-1}\gamma^{t-s-1}\mu'(X_s^\top\theta)X_sX_s^\top\succeq c_\mu V_{t-1},$$

which clearly implies $\forall\theta_1,\theta_2\in\Theta, G_t(\theta_1,\theta_2)\succeq c_\mu V_{t-1}$.

By Assumption 2, the mean value theorem (37) on g_t and the projection (9), we have

$$\begin{aligned}|\mu(\mathbf{x}^\top\widetilde{\theta}_t)-\mu(\mathbf{x}^\top\theta_t)| &\leq k_\mu|\mathbf{x}^\top(\widetilde{\theta}_t-\theta_t)|\\ &=k_\mu|\mathbf{x}^\top G_t^{-1}(\theta_t,\widetilde{\theta}_t)(g_t(\widetilde{\theta}_t)-g_t(\theta_t))|\\ &\leq k_\mu\|\mathbf{x}\|_{G_t^{-1}(\theta_t,\widetilde{\theta}_t)}\|g_t(\widetilde{\theta}_t)-g_t(\theta_t)\|_{G_t^{-1}(\theta_t,\widetilde{\theta}_t)}\\ &\leq\frac{k_\mu}{c_\mu}\|\mathbf{x}\|_{V_{t-1}^{-1}}\|g_t(\widetilde{\theta}_t)-g_t(\theta_t)\|_{V_{t-1}^{-1}}\\ &\leq\frac{2k_\mu}{c_\mu}\|\mathbf{x}\|_{V_{t-1}^{-1}}\|g_t(\widehat{\theta}_t)-g_t(\theta_t)\|_{V_{t-1}^{-1}},\end{aligned}$$

then based on the model assumption, the function g_t (10) and $g_t(\widehat{\theta}_t)$ (36), we have,

$$\begin{aligned}g_t(\theta_t)-g_t(\widehat{\theta}_t)&=\lambda c_\mu\theta_t+\sum_{s=1}^{t-1}\gamma^{t-s-1}\mu(X_s^\top\theta_t)X_s-\sum_{s=1}^{t-1}\gamma^{t-s-1}r_sX_s\\ &=\lambda c_\mu\theta_t+\sum_{s=1}^{t-1}\gamma^{t-s-1}\mu(X_s^\top\theta_t)X_s-\sum_{s=1}^{t-1}\gamma^{t-s-1}(\mu(X_s^\top\theta_s)+\eta_s)X_s \qquad (38)\\ &=\underbrace{\sum_{s=1}^{t-1}\gamma^{t-s-1}(\mu(X_s^\top\theta_t)-\mu(X_s^\top\theta_s))X_s}_{\texttt{bias part}}+\underbrace{\lambda c_\mu\theta_t-\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_sX_s}_{\texttt{variance part}}. \qquad (39)\end{aligned}$$

Then, by the Cauchy-Schwarz inequality, we know that for any $\mathbf{x}\in\mathcal{X}$,

$$\left|\mu(\mathbf{x}^\top\widetilde{\theta}_t)-\mu(\mathbf{x}^\top\theta_t)\right|\leq\frac{2k_\mu}{c_\mu}\|\mathbf{x}\|_{V_{t-1}^{-1}}\left(C_t+D_t\right), \tag{40}$$

where

$$C_t = \left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} (\mu(X_s^\top \theta_t) - \mu(X_s^\top \theta_s)) X_s \right\|_{V_{t-1}^{-1}}, \quad D_t = \left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s - \lambda c_\mu \theta_t \right\|_{V_{t-1}^{-1}}.$$

This two terms can be bounded separately, as summarized in the following lemmas.

Lemma 6. *For any $t \in [T]$, we have*

$$\left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} (\mu(X_s^\top \theta_t) - \mu(X_s^\top \theta_s)) X_s \right\|_{V_{t-1}^{-1}} \leq L k_\mu \sqrt{d} \sum_{p=1}^{t-1} \gamma^{\frac{t-1}{2}} \sqrt{\frac{\gamma^{-p} - 1}{1-\gamma}} \left\| \theta_p - \theta_{p+1} \right\|_2. \tag{41}$$

Lemma 7. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds for all $t \in [T]$,*

$$\left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s - \lambda c_\mu \theta_t \right\|_{V_{t-1}^{-1}} \leq \sqrt{\lambda} c_\mu S + R \sqrt{2 \log \frac{1}{\delta} + d \log \left(1 + \frac{L^2 (1-\gamma^{2t-2})}{\lambda d (1-\gamma^2)}\right)}. \tag{42}$$

Based on the inequality (40), Lemma 6, Lemma 7, and the boundedness assumption of the feasible set, we have for any $\mathbf{x} \in \mathcal{X}$, $\gamma \in (0,1)$, $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds for all $t \in [T]$,

$$\begin{aligned}
\left| \mu(\mathbf{x}^\top \widetilde{\theta}_t) - \mu(\mathbf{x}^\top \theta_t) \right| &\leq \frac{2k_\mu}{c_\mu} \|\mathbf{x}\|_{V_{t-1}^{-1}} \left(L k_\mu \sqrt{d} \sum_{p=1}^{t-1} \gamma^{\frac{t-1}{2}} \sqrt{\frac{\gamma^{-p}-1}{1-\gamma}} \left\| \theta_p - \theta_{p+1} \right\|_2 + \bar{\beta}_{t-1} \right) \\
&\leq \frac{2k_\mu}{c_\mu} \left(L^2 k_\mu \sqrt{\frac{d}{\lambda}} \sum_{p=1}^{t-1} \gamma^{\frac{t-1}{2}} \sqrt{\frac{\gamma^{-p}-1}{1-\gamma}} \left\| \theta_p - \theta_{p+1} \right\|_2 + \bar{\beta}_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}} \right),
\end{aligned}$$

where $\bar{\beta}_t \triangleq \sqrt{\lambda} c_\mu S + R \sqrt{2 \log \frac{1}{\delta} + d \log \left(1 + \frac{L^2(1-\gamma^{2t})}{\lambda d (1-\gamma^2)}\right)}$ is the confidence radius used in GLB-WeightUCB. Hence we complete the proof. □

Proof of Lemma 6. Here we need to extract the variations of the time-varying parameter θ_t

$$\begin{aligned}
\left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} (\mu(X_s^\top \theta_t) - \mu(X_s^\top \theta_s)) X_s \right\|_{V_{t-1}^{-1}} &\leq \left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} \sum_{p=s}^{t-1} (\mu(X_s^\top \theta_p) - \mu(X_s^\top \theta_{p+1})) X_s \right\|_{V_{t-1}^{-1}} \\
&= \left\| \sum_{p=1}^{t-1} \sum_{s=1}^{p} \gamma^{t-s-1} \alpha(X_s, \theta_p, \theta_{p+1}) (X_s^\top \theta_p - X_s^\top \theta_{p+1}) X_s \right\|_{V_{t-1}^{-1}} \\
&= \left\| \sum_{p=1}^{t-1} \sum_{s=1}^{p} \gamma^{t-s-1} \alpha(X_s, \theta_p, \theta_{p+1}) X_s X_s^\top (\theta_p - \theta_{p+1}) \right\|_{V_{t-1}^{-1}} \\
&\leq \sum_{p=1}^{t-1} \left\| \sum_{s=1}^{p} \gamma^{t-s-1} \alpha(X_s, \theta_p, \theta_{p+1}) X_s \|X_s\|_2 \|\theta_p - \theta_{p+1}\|_2 \right\|_{V_{t-1}^{-1}} \\
&\leq L \sum_{p=1}^{t-1} \sum_{s=1}^{p} \gamma^{t-s-1} |\alpha(X_s, \theta_p, \theta_{p+1})| \, \|X_s\|_{V_{t-1}^{-1}} \, \|\theta_p - \theta_{p+1}\|_2 \\
&\leq L k_\mu \sum_{p=1}^{t-1} \sum_{s=1}^{p} \gamma^{t-s-1} \, \|X_s\|_{V_{t-1}^{-1}} \, \|\theta_p - \theta_{p+1}\|_2.
\end{aligned}$$

where the forth equation is due to the mean value theorem where $\alpha(\mathbf{x}, \theta_1, \theta_2) = \int_0^1 \mu'(v \mathbf{x}^\top \theta_2 + (1-v) x^\top \theta_1) \mathrm{d}v$:

$$\mu(X_s^\top \theta_p) - \mu(X_s^\top \theta_{p+1}) = \alpha(X_s, \theta_p, \theta_{p+1})(X_s^\top \theta_p - X_s^\top \theta_{p+1}).$$

Next, the derivation for the bound of term $\sum_{s=1}^{p} \gamma^{t-s-1} \|X_s\|_{V_{t-1}^{-1}}$ is the same as the inequality (23) in A.2, hence we complete the proof. □

Proof of Lemma 7. Same as the linear case, we need to use $\widetilde{V}_t = \lambda I_d + \sum_{s=1}^{t} \gamma^{2(t-s)} X_s X_s^\top$.

$$
\begin{aligned}
D_t &= \left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s - \lambda c_\mu \theta_t \right\|_{V_{t-1}^{-1}} \\
&\leq \left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s \right\|_{V_{t-1}^{-1}} + \|\lambda c_\mu \theta_t\|_{V_{t-1}^{-1}} \\
&\leq \left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s \right\|_{\widetilde{V}_{t-1}^{-1}} + \sqrt{\lambda} c_\mu S \\
&\leq R\sqrt{2\log\frac{1}{\delta} + d\log\left(1 + \frac{L^2(1-\gamma^{2t-2})}{\lambda d(1-\gamma^2)}\right)} + \sqrt{\lambda} c_\mu S.
\end{aligned}
$$

Again, we emphasize that the $\widetilde{V}_t$ is introduced into analysis *only*. The proof here is the same as the proof of Lemma 5 in A.2, the only difference is an extra c_μ in the second term. □

B.3 Proof of Theorem 2

Proof. Let $X_t^* \triangleq \arg\max_{\mathbf{x}\in\mathcal{X}} \mu(\mathbf{x}^\top \theta_t)$. Due to Lemma 2 and the fact that $X_t^*, X_t \in \mathcal{X}$, each of the following holds with probability at least $1-\delta$,

$$
\forall t \in [T], \mu(X_t^{*\top}\theta_t) \leq \mu(X_t^{*\top}\widetilde{\theta}_t) + \frac{2k_\mu}{c_\mu}\left(L^2 k_\mu \sqrt{\frac{d}{\lambda}} \sum_{p=1}^{t-1} \gamma^{\frac{t-1}{2}} \sqrt{\frac{\gamma^{-p}-1}{1-\gamma}} \|\theta_p - \theta_{p+1}\|_2 + \bar{\beta}_{t-1}\|X_t^*\|_{V_{t-1}^{-1}}\right),
$$

$$
\forall t \in [T], \mu(X_t^{\top}\theta_t) \geq \mu(X_t^{\top}\widetilde{\theta}_t) - \frac{2k_\mu}{c_\mu}\left(L^2 k_\mu \sqrt{\frac{d}{\lambda}} \sum_{p=1}^{t-1} \gamma^{\frac{t-1}{2}} \sqrt{\frac{\gamma^{-p}-1}{1-\gamma}} \|\theta_p - \theta_{p+1}\|_2 + \bar{\beta}_{t-1}\|X_t\|_{V_{t-1}^{-1}}\right).
$$

By the union bound, the following holds with probability at least $1-2\delta$: $\forall t \in [T]$

$$
\begin{aligned}
&\mu(X_t^{*\top}\theta_t) - \mu(X_t^{\top}\theta_t) \\
\leq\ & \mu(X_t^{*\top}\widetilde{\theta}_t) - \mu(X_t^{\top}\widetilde{\theta}_t) + \frac{4L^2k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}} \sum_{p=1}^{t-1} \gamma^{\frac{t-1}{2}} \sqrt{\frac{\gamma^{-p}-1}{1-\gamma}} \|\theta_p - \theta_{p+1}\|_2 + \frac{2k_\mu}{c_\mu}\left(\bar{\beta}_{t-1}\|X_t^*\|_{V_{t-1}^{-1}} + \bar{\beta}_{t-1}\|X_t\|_{V_{t-1}^{-1}}\right) \\
\leq\ & \frac{4L^2k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}} \sum_{p=1}^{t-1} \gamma^{\frac{t-1}{2}} \sqrt{\frac{\gamma^{-p}-1}{1-\gamma}} \|\theta_p - \theta_{p+1}\|_2 + \frac{4k_\mu}{c_\mu}\bar{\beta}_{t-1}\|X_t\|_{V_{t-1}^{-1}},
\end{aligned}
$$

where the last step comes from the arm selection criterion (12) such that

$$
\mu(X_t^{*\top}\widetilde{\theta}_t) + \frac{2k_\mu}{c_\mu}\bar{\beta}_{t-1}\|X_t^*\|_{V_{t-1}^{-1}} \leq \mu(X_t^{\top}\widetilde{\theta}_t) + \frac{2k_\mu}{c_\mu}\bar{\beta}_{t-1}\|X_t\|_{V_{t-1}^{-1}}.
$$

Hence the following dynamic regret bound holds with probability at least $1-2\delta$ and can be divided into two parts,

$$
\begin{aligned}
R_T &= \sum_{t=1}^{T} \max_{\mathbf{x}\in\mathcal{X}} \mu(\mathbf{x}^\top\theta_t) - \mu(X_t^\top\theta_t) \\
&\leq \underbrace{\frac{4L^2k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}} \sum_{t=1}^{T}\sum_{p=1}^{t-1} \gamma^{\frac{t-1}{2}} \sqrt{\frac{\gamma^{-p}-1}{1-\gamma}} \|\theta_p - \theta_{p+1}\|_2}_{\texttt{bias part}} + \underbrace{\frac{4k_\mu}{c_\mu}\bar{\beta}_T \sum_{t=1}^{T} \|X_t\|_{V_{t-1}^{-1}}}_{\texttt{variance part}}.
\end{aligned}
\tag{43}
$$

where $\bar{\beta}_t = \sqrt{\lambda}c_\mu S + R\sqrt{2\log\frac{1}{\delta} + d\log\left(1 + \frac{L^2(1-\gamma^{2t})}{\lambda d(1-\gamma^2)}\right)}$ is the confidence radius.

Now we derive the upper bound for these two parts separately.

Algorithm 4 SCB-WeightUCB

Input: time horizon T, discounted factor γ, confidence δ, regularizer λ, inverse link function μ, parameters S, L and m

Set $V_0 = \lambda I_d$, $\widehat{\theta}_1 = \mathbf{0}$ and compute k_μ and c_μ
for $t = 1, 2, ..., T$ **do**
  if $\|\widehat{\theta}_t\|_2 \leq S$ **then**
    let $\widetilde{\theta}_t = \widehat{\theta}_t$
  else
    Do the projection and get $\widetilde{\theta}_t$ by (14)
  end if
  Compute $\widetilde{\beta}_{t-1}$ by (16)
  Select X_t by (17)
  Receive the reward r_t
  Update $V_t = \gamma V_{t-1} + X_t X_t^\top + (1-\gamma)\lambda I_d$
  Compute $\widehat{\theta}_{t+1}$ according to (8)
end for

Bias Part. Similar to the proof of inequality (25), we have

$$\frac{4L^2k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\sum_{t=1}^{T}\sum_{p=1}^{t-1}\gamma^{\frac{t-1}{2}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\|\theta_p - \theta_{p+1}\|_2 \leq \frac{8L^2k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\frac{1}{(1-\gamma)^{3/2}}P_T.$$

Variance Part. Similar to the proof of inequality (26), we have

$$\frac{4k_\mu}{c_\mu}\bar{\beta}_T\sqrt{T}\sqrt{\sum_{t=1}^{T}\|X_t\|^2_{V_{t-1}^{-1}}} \leq \frac{4k_\mu}{c_\mu}\bar{\beta}_T\sqrt{2\max\{1, L^2/\lambda\}dT}\sqrt{T\log\frac{1}{\gamma} + \log\left(1 + \frac{L^2}{\lambda d(1-\gamma)}\right)}.$$

Combine the upper bound for the bias and variance parts, and let $\delta = 1/(2T^2)$, we have the following regret bound with probability at least $1 - 1/T$,

$$R_T \leq \frac{8L^2k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\frac{1}{(1-\gamma)^{3/2}}P_T + \frac{4k_\mu}{c_\mu}\bar{\beta}_T\sqrt{2\max\{1, L^2/\lambda\}dT}\sqrt{T\log\frac{1}{\gamma} + \log\left(1 + \frac{L^2}{\lambda d(1-\gamma)}\right)}.$$

where $\bar{\beta}_t = \sqrt{\lambda}c_\mu S + R\sqrt{4\log T + 2\log 2 + d\log\left(1 + \frac{L^2(1-\gamma^{2t})}{\lambda d(1-\gamma^2)}\right)}$. We set $\gamma \geq 1/T$ and $\lambda = d/c_\mu^2$, and obtain that,

$$R_T \leq \widetilde{\mathcal{O}}\left(k_\mu^2\frac{1}{(1-\gamma)^{3/2}}P_T + \frac{k_\mu}{c_\mu}d(1-\gamma)^{1/2}T\right).$$

When $P_T < d/(k_\mu c_\mu T)$, we set $\gamma = 1 - 1/T$ and achieve an $\widetilde{\mathcal{O}}(k_\mu c_\mu^{-1}d\sqrt{T})$ regret bound. When $P_T \geq d/(k_\mu c_\mu T)$, We set γ optimally as $1 - \gamma = \sqrt{k_\mu c_\mu P_T/(dT)}$ and attain an $\widetilde{\mathcal{O}}(k_\mu^{5/4}c_\mu^{-3/4}d^{3/4}P_T^{1/4}T^{3/4})$ regret bound. Notice that, if $k_\mu < 1$, we just let $1 - \gamma = \sqrt{c_\mu P_T/(dT)}$ and the regret bound becomes $\widetilde{\mathcal{O}}(k_\mu^2 c_\mu^{-3/4}d^{3/4}P_T^{1/4}T^{3/4})$. □

C Analysis of SCB-WeightUCB

In this section, we first present SCB-WeightUCB algorithm in Algorithm 4, Then, in Appendix C.1 we present the proof of the estimation error upper bound of our SCB-WeightUCB algorithm (Lemma 3). Finally, in Appendix C.2, we provide the proof of dynamic regret upper bound (Theorem 3).

C.1 Proof of Lemma 3

Proof. Base on the estimator equation (8), we know that

$$g_t(\widehat{\theta}_t) = \lambda c_\mu \widehat{\theta}_t + \sum_{s=1}^{t-1}\gamma^{t-s-1}\mu(X_s^\top\widehat{\theta}_t)X_s = \sum_{s=1}^{t-1}\gamma^{t-s-1}r_s X_s, \tag{44}$$

and then by the mean value theorem, we know that

$$g_t(\theta_1) - g_t(\theta_2) = G_t(\theta_1, \theta_2)(\theta_1 - \theta_2), \tag{45}$$

where $G_t(\theta_1, \theta_2) \triangleq \int_0^1 \nabla g_t(s\theta_2 + (1-s)\theta_1)\mathrm{d}s \in \mathbb{R}^{d\times d}$. Notice that for any $\theta \in \Theta$, the gradient of g_t is

$$\nabla g_t(\theta) = \lambda c_\mu I_d + \sum_{s=1}^{t-1} \gamma^{t-s-1}\mu'(X_s^\top \theta) X_s X_s^\top \succeq c_\mu V_{t-1},$$

which clearly implies $\forall \theta_1, \theta_2 \in \Theta, G_t(\theta_1, \theta_2) \succeq c_\mu V_{t-1}$ and $\forall \theta, H_t(\theta) \succeq c_\mu V_{t-1}$, where $H_t(\theta)$ is defined as

$$H_t(\theta) \triangleq \lambda c_\mu I_d + \sum_{s=1}^{t-1} \gamma^{t-s-1}\mu'(X_s^\top \theta) X_s X_s^\top. \tag{46}$$

By Assumption 2, the mean value theorem (37) on g_t, the projection (14) and Lemma 14, we have

$$\begin{aligned}
|\mu(\mathbf{x}^\top \widetilde{\theta}_t) - \mu(\mathbf{x}^\top \theta_t)| &\leq k_\mu |\mathbf{x}^\top (\widetilde{\theta}_t - \theta_t)| \\
&= k_\mu |\mathbf{x}^\top G_t^{-1}(\theta_t, \widetilde{\theta}_t)(g_t(\widetilde{\theta}_t) - g_t(\theta_t))| \\
&\leq k_\mu \|\mathbf{x}\|_{G_t^{-1}(\theta_t, \widetilde{\theta}_t)} \|g_t(\widetilde{\theta}_t) - g_t(\theta_t)\|_{G_t^{-1}(\theta_t, \widetilde{\theta}_t)} \\
&\leq k_\mu \|\mathbf{x}\|_{G_t^{-1}(\theta_t, \widetilde{\theta}_t)} \left(\|g_t(\widetilde{\theta}_t) - g_t(\widehat{\theta}_t)\|_{G_t^{-1}(\theta_t, \widetilde{\theta}_t)} + \|g_t(\widehat{\theta}_t) - g_t(\theta_t)\|_{G_t^{-1}(\theta_t, \widetilde{\theta}_t)} \right) \\
&\leq \sqrt{1+2S} k_\mu \|\mathbf{x}\|_{G_t^{-1}(\theta_t, \widetilde{\theta}_t)} \left(\|g_t(\widetilde{\theta}_t) - g_t(\widehat{\theta}_t)\|_{H_t^{-1}(\widetilde{\theta}_t)} + \|g_t(\widehat{\theta}_t) - g_t(\theta_t)\|_{H_t^{-1}(\theta_t)} \right) \\
&\leq 2\sqrt{1+2S} \frac{k_\mu}{\sqrt{c_\mu}} \|\mathbf{x}\|_{V_{t-1}^{-1}} \|g_t(\widehat{\theta}_t) - g_t(\theta_t)\|_{H_t^{-1}(\theta_t)},
\end{aligned}$$

then based on the model assumption (13), the function g_t (10) and the $g_t(\widehat{\theta}_t)$ (44), we have,

$$\begin{aligned}
g_t(\theta_t) - g_t(\widehat{\theta}_t) &= \lambda c_\mu \theta_t + \sum_{s=1}^{t-1} \gamma^{t-s-1} \mu(X_s^\top \theta_t) X_s - \sum_{s=1}^{t-1} \gamma^{t-s-1} r_s X_s \\
&= \lambda c_\mu \theta_t + \sum_{s=1}^{t-1} \gamma^{t-s-1} \mu(X_s^\top \theta_t) X_s - \sum_{s=1}^{t-1} \gamma^{t-s-1} (\mu(X_s^\top \theta_s) + \eta_s) X_s \\
&= \sum_{s=1}^{t-1} \gamma^{t-s-1} (\mu(X_s^\top \theta_t) - \mu(X_s^\top \theta_s)) X_s + \lambda c_\mu \theta_t - \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s,
\end{aligned}$$

then, by Cauchy-Schwarz inequality, we have

$$\left| \mu(\mathbf{x}^\top \widetilde{\theta}_t) - \mu(\mathbf{x}^\top \theta_t) \right| \leq 2\sqrt{1+2S} \frac{k_\mu}{\sqrt{c_\mu}} \|\mathbf{x}\|_{V_{t-1}^{-1}} (E_t + F_t), \tag{47}$$

where

$$E_t = \left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} (\mu(X_s^\top \theta_t) - \mu(X_s^\top \theta_s)) X_s \right\|_{H_t^{-1}(\theta_t)}, \quad F_t = \left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s - \lambda c_\mu \theta_t \right\|_{H_t^{-1}(\theta_t)}.$$

This two terms can be bounded separately.

Lemma 8. *For any $t \in [T]$, we have*

$$\left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} (\mu(X_s^\top \theta_t) - \mu(X_s^\top \theta_s)) X_s \right\|_{H_t^{-1}(\theta_t)} \leq L \frac{k_\mu}{\sqrt{c_\mu}} \sqrt{d} \sum_{p=1}^{t-1} \gamma^{\frac{t-1}{2}} \sqrt{\frac{\gamma^{-p} - 1}{1-\gamma}} \|\theta_p - \theta_{p+1}\|_2. \tag{48}$$

Lemma 9. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have for all $t \in [T]$,*

$$\left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_s X_s - \lambda c_\mu \theta_t\right\|_{H_t^{-1}(\theta_t)} \le \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}}\log\frac{1}{\delta} + \frac{dm}{\sqrt{\lambda c_\mu}}\log\left(1+\frac{L^2 k_\mu(1-\gamma^{2t-2})}{\lambda c_\mu d(1-\gamma^2)}\right) + \frac{2m}{\sqrt{\lambda c_\mu}} d\log(2) + \sqrt{\lambda c_\mu} S, \tag{49}$$

Based on the inequality (47), Lemma 6 and Lemma 7, and the boundedness assumption of the feasible set, we have for any $\mathbf{x} \in \mathcal{X}$, $\delta \in (0,1)$, with probability at least $1-\delta$, we have for all $t \in [T]$,

$$\begin{aligned}\left|\mu(\mathbf{x}^\top \widetilde{\theta}_t) - \mu(\mathbf{x}^\top \theta_t)\right| &\le 2\sqrt{1+2S}\frac{k_\mu}{\sqrt{c_\mu}}\|\mathbf{x}\|_{V_{t-1}^{-1}}\left(L\frac{k_\mu}{\sqrt{c_\mu}}\sqrt{d}\sum_{p=1}^{t-1}\gamma^{\frac{t-1}{2}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\|\theta_p-\theta_{p+1}\|_2 + \widetilde{\beta}_{t-1}\right),\\ &\le 2\sqrt{1+2S}\frac{k_\mu}{\sqrt{c_\mu}}\left(L^2\frac{k_\mu}{\sqrt{\lambda c_\mu}}\sqrt{d}\sum_{p=1}^{t-1}\gamma^{\frac{t-1}{2}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\|\theta_p-\theta_{p+1}\|_2 + \widetilde{\beta}_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}}\right),\end{aligned}$$

where $\widetilde{\beta}_t \triangleq \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}}\log\frac{1}{\delta} + \frac{dm}{\sqrt{\lambda c_\mu}}\log\left(1+\frac{L^2k_\mu(1-\gamma^{2t})}{\lambda c_\mu d(1-\gamma^2)}\right) + \frac{2m}{\sqrt{\lambda c_\mu}}d\log(2) + \sqrt{\lambda c_\mu}S$ is the confidence radius used in SCB-WeightUCB. Hence we completes the proof. □

Proof of Lemma 8. Since $\forall\theta, H_t(\theta) \succeq c_\mu V_{t-1}$, we have

$$\left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}(\mu(X_s^\top\theta_t)-\mu(X_s^\top\theta_s))X_s\right\|_{H_t^{-1}(\theta_t)} \le \frac{1}{\sqrt{c_\mu}}\left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}(\mu(X_s^\top\theta_t)-\mu(X_s^\top\theta_s))X_s\right\|_{V_{t-1}^{-1}}.$$

Then use Lemma 6 and we complete the proof. □

Proof of Lemma 9. Let $\widetilde{H}_t(\theta) \triangleq \lambda c_\mu \gamma^{-2(t-1)} I_d + \sum_{s=1}^{t-1}\gamma^{-2s}\mu'(X_s^\top\theta)X_sX_s^\top$ which is *only* used in the analysis.

$$\begin{aligned}F_t = \left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_s X_s - \lambda c_\mu\theta_t\right\|_{H_t^{-1}(\theta_t)} &\le \left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_s X_s\right\|_{H_t^{-1}(\theta_t)} + \|\lambda c_\mu \theta_t\|_{H_t^{-1}(\theta_t)}\\ &\le \left\|\sum_{s=1}^{t-1}\gamma^{-s}\eta_s X_s\right\|_{\widetilde{H}_t^{-1}(\theta_t)} + \sqrt{\lambda c_\mu}S.\end{aligned}$$

Recall that $H_t(\theta) = \lambda c_\mu I_d + \sum_{s=1}^{t-1}\gamma^{t-s-1}\mu'(X_s^\top\theta)X_sX_s^\top$, so the last inequality comes from

$$\gamma^{-2(t-1)}H_t(\theta) = \lambda c_\mu\gamma^{-2(t-1)}I_d + \sum_{s=1}^{t-1}\gamma^{-t-s+1}\mu'(X_s^\top\theta)X_sX_s^\top \succeq \lambda c_\mu\gamma^{-2(t-1)}I_d + \sum_{s=1}^{t-1}\gamma^{-2s}\mu'(X_s^\top\theta)X_sX_s^\top = \widetilde{H}_t^{-1}(\theta). \tag{50}$$

From the weighted version concentration inequality [Russac et al., 2021, Theorem 3], restated in Theorem 6, we can get the bound for the first term $\|\sum_{s=1}^{t-1}\gamma^{-s}\eta_sX_s\|_{\widetilde{H}_t^{-1}(\theta_t)}$. First by the model assumption (13), we know that $\sigma_t^2 = \mathbb{E}[\eta_t^2|\mathcal{F}_t] = \mathrm{Var}[r_t|\mathcal{F}_t] = \mu''(X_t\theta_t)$, then just let $w_t = \gamma^{-t}, \lambda_t = \lambda c_\mu\gamma^{-2t}$ and we have,

$$\left\|\sum_{s=1}^{t-1}\gamma^{-s}\eta_sX_s\right\|_{\widetilde{H}_t^{-1}(\theta_t)} \le \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}}\log\left(\frac{\det(\widetilde{H}_t)^{1/2}}{\delta(\lambda c_\mu\gamma^{-2(t-1)})^{d/2}}\right) + \frac{2m}{\sqrt{\lambda c_\mu}}d\log(2).$$

Then use Lemma 12 and let $w_{t,s} = \gamma^{-2s}\mu'(X_s^\top\theta_t), \lambda_t = \lambda c_\mu\gamma^{-2(t-1)}$, we get the upper bound for $\det(\widetilde{H}_t)$,

$$\det(\widetilde{H}_t) \le \left(\lambda c_\mu\gamma^{-2(t-1)} + \frac{L^2k_\mu\sum_{s=1}^{t-1}\gamma^{-2s}}{d}\right)^d,$$

then,

$$\begin{aligned}\left\|\sum_{s=1}^{t-1}\gamma^{-s}\eta_s X_s\right\|_{\widetilde{H}_t^{-1}(\theta_t)} &\leq \frac{\sqrt{\lambda c_\mu}}{2m}+\frac{2m}{\sqrt{\lambda c_\mu}}\log\frac{1}{\delta}+\frac{dm}{\sqrt{\lambda c_\mu}}\log\left(\frac{\lambda c_\mu\gamma^{-2(t-1)}+\frac{L^2k_\mu\sum_{s=1}^{t-1}\gamma^{-2s}}{d}}{\lambda c_\mu\gamma^{-2(t-1)}}\right)+\frac{2m}{\sqrt{\lambda c_\mu}}d\log(2)\\ &\leq \frac{\sqrt{\lambda c_\mu}}{2m}+\frac{2m}{\sqrt{\lambda c_\mu}}\log\frac{1}{\delta}+\frac{dm}{\sqrt{\lambda c_\mu}}\log\left(1+\frac{L^2k_\mu(1-\gamma^{2t-2})}{\lambda c_\mu d(1-\gamma^2)}\right)+\frac{2m}{\sqrt{\lambda c_\mu}}d\log(2).\end{aligned}$$

Therefore, we get the upper bound for F_t term. □

C.2 Proof of Theorem 3

Proof. Let $X_t^* \triangleq \arg\max_{\mathbf{x}\in\mathcal{X}} \mu(\mathbf{x}^\top\theta_t)$. Due to Lemma 2 and the fact that $X_t^*, X_t \in \mathcal{X}$, each of the following holds with probability at least $1-\delta$,

$$\forall t\in[T], \mu(X_t^{*\top}\theta_t) \leq \mu(X_t^{*\top}\widetilde{\theta}_t)+2\sqrt{1+2S}\frac{k_\mu}{\sqrt{c_\mu}}\left(L^2\frac{k_\mu}{\sqrt{\lambda c_\mu}}\sqrt{d}\sum_{p=1}^{t-1}\gamma^{\frac{t-1}{2}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\|\theta_p-\theta_{p+1}\|_2+\widetilde{\beta}_{t-1}\|X_t^*\|_{V_{t-1}^{-1}}\right),$$

$$\forall t\in[T], \mu(X_t^{\top}\theta_t) \geq \mu(X_t^{\top}\widetilde{\theta}_t)-2\sqrt{1+2S}\frac{k_\mu}{\sqrt{c_\mu}}\left(L^2\frac{k_\mu}{\sqrt{\lambda c_\mu}}\sqrt{d}\sum_{p=1}^{t-1}\gamma^{\frac{t-1}{2}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\|\theta_p-\theta_{p+1}\|_2+\widetilde{\beta}_{t-1}\|X_t\|_{V_{t-1}^{-1}}\right).$$

By the union bound, the following holds with probability at least $1-2\delta$: $\forall t\in[T]$

$$\begin{aligned}&\mu(X_t^{*\top}\theta_t)-\mu(X_t^\top\theta_t)\\ &\leq \mu(X_t^{*\top}\widetilde{\theta}_t)-\mu(X_t^\top\widetilde{\theta}_t)+2\sqrt{1+2S}\Bigg(\frac{2L^2k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{t-1}\gamma^{\frac{t-1}{2}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\|\theta_p-\theta_{p+1}\|_2\\ &\qquad\qquad+\frac{k_\mu}{\sqrt{c_\mu}}\left(\widetilde{\beta}_{t-1}\|X_t^*\|_{V_{t-1}^{-1}}+\widetilde{\beta}_{t-1}\|X_t\|_{V_{t-1}^{-1}}\right)\Bigg)\\ &\leq \frac{4\sqrt{1+2S}L^2k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{t-1}\gamma^{\frac{t-1}{2}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\|\theta_p-\theta_{p+1}\|_2+\frac{4\sqrt{1+2S}k_\mu}{\sqrt{c_\mu}}\widetilde{\beta}_{t-1}\|X_t\|_{V_{t-1}^{-1}},\end{aligned}$$

where the last step comes from the arm selection criterion (17) such that

$$\mu(X_t^{*\top}\widetilde{\theta}_t)+2\sqrt{1+2S}\frac{k_\mu}{\sqrt{c_\mu}}\widetilde{\beta}_{t-1}\|X_t^*\|_{V_{t-1}^{-1}} \leq \mu(X_t^{\top}\widetilde{\theta}_t)+2\sqrt{1+2S}\frac{k_\mu}{\sqrt{c_\mu}}\widetilde{\beta}_{t-1}\|X_t\|_{V_{t-1}^{-1}}.$$

Hence, the following dynamic regret bound holds with probability at least $1-2\delta$ and can be divided into two parts,

$$\begin{aligned}R_T &= \sum_{t=1}^T \mu(X_t^{*\top}\theta_t)-\mu(X_t^\top\theta_t)\\ &\leq \underbrace{\frac{4\sqrt{1+2S}L^2k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\sum_{t=1}^T\sum_{p=1}^{t-1}\gamma^{\frac{t-1}{2}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\|\theta_p-\theta_{p+1}\|_2}_{\texttt{bias part}}+\underbrace{\frac{4\sqrt{1+2S}k_\mu}{\sqrt{c_\mu}}\widetilde{\beta}_T\sum_{t=1}^T\|X_t\|_{V_{t-1}^{-1}}}_{\texttt{variance part}}.\end{aligned}$$

where $\widetilde{\beta}_t=\frac{\sqrt{\lambda c_\mu}}{2m}+\frac{2m}{\sqrt{\lambda c_\mu}}\log\frac{1}{\delta}+\frac{dm}{\sqrt{\lambda c_\mu}}\log\left(1+\frac{L^2k_\mu(1-\gamma^{2t})}{\lambda c_\mu d(1-\gamma^2)}\right)+\frac{2m}{\sqrt{\lambda c_\mu}}d\log(2)+\sqrt{\lambda c_\mu}S$ is the confidence radius.

Now we derive the upper bound for these two parts separately.

Bias Part. Similar to the proof of inequality (25), we have

$$\frac{4\sqrt{1+2S}L^2k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\sum_{t=1}^T\sum_{p=1}^{t-1}\gamma^{\frac{t-1}{2}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\|\theta_p-\theta_{p+1}\|_2 \leq \frac{8\sqrt{1+2S}L^2k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\frac{1}{(1-\gamma)^{3/2}}P_T.$$

Variance Part. First use the Cauchy-Schwarz inequality, we know that

$$\frac{4\sqrt{1+2S}k_\mu}{\sqrt{c_\mu}}\widetilde{\beta}_T\sum_{t=1}^T\|X_t\|_{V_{t-1}^{-1}} \leq \frac{4\sqrt{1+2S}k_\mu}{\sqrt{c_\mu}}\widetilde{\beta}_T\sqrt{T}\sqrt{\sum_{t=1}^T\|X_t\|^2_{V_{t-1}^{-1}}}.$$

Then for the term $\sqrt{\sum_{t=1}^T\|X_t\|^2_{V_{t-1}^{-1}}}$, we can directly use the Lemma 11 to bound it,

$$\frac{4\sqrt{1+2S}k_\mu}{\sqrt{c_\mu}}\widetilde{\beta}_T\sqrt{T}\sqrt{\sum_{t=1}^T\|X_t\|^2_{V_{t-1}^{-1}}} \leq \frac{4\sqrt{1+2S}k_\mu}{\sqrt{c_\mu}}\widetilde{\beta}_T\sqrt{2\max\{1,L^2/\lambda\}dT}\sqrt{T\log\frac{1}{\gamma}+\log\left(1+\frac{L^2}{\lambda d(1-\gamma)}\right)}.$$

Combining the upper bound for the bias and variance parts, and letting $\delta = 1/(2T)$, we have the following regret bound with probability at least $1-1/T$,

$$R_T \leq \frac{8\sqrt{1+2S}L^2k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\frac{1}{(1-\gamma)^{3/2}}P_T + \frac{4\sqrt{1+2S}k_\mu}{\sqrt{c_\mu}}\widetilde{\beta}_T\sqrt{2\max\{1,L^2/\lambda\}dT}\sqrt{T\log\frac{1}{\gamma}+\log\left(1+\frac{L^2}{\lambda d(1-\gamma)}\right)}.$$

where $\widetilde{\beta}_t = \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}}\log(2T) + \frac{dm}{\sqrt{\lambda c_\mu}}\log\left(1+\frac{L^2k_\mu(1-\gamma^{2t})}{\lambda c_\mu d(1-\gamma^2)}\right) + \frac{2m}{\sqrt{\lambda c_\mu}}d\log(2) + \sqrt{\lambda c_\mu}S$. Since that there is a $T\sqrt{\log(1/\gamma)}$ term in the regret bound, which means that we cannot let γ close to 0, so we set $\gamma \geq 1/T$, then we have $\log(1/\gamma) \leq C(1-\gamma)$, where $C = \log T/(1-1/T)$. Then, ignoring logarithmic factors in time horizon T, and let $\lambda = d\log(T)/c_\mu$, we finally obtain that,

$$R_T \leq \widetilde{\mathcal{O}}\left(\frac{k_\mu^2}{\sqrt{c_\mu}}\frac{1}{(1-\gamma)^{3/2}}P_T + \frac{k_\mu}{\sqrt{c_\mu}}d(1-\gamma)^{1/2}T\right).$$

When $P_T < d/(k_\mu T)$ (which corresponds a small amount of non-stationarity), we simply set $\gamma = 1-1/T$ and achieve an $\widetilde{\mathcal{O}}(k_\mu c_\mu^{-1/2}d\sqrt{T})$ regret bound. Besides, when coming to the non-degenerated case of $P_T \geq d/(k_\mu T)$, We set the discounted factor optimally as $1-\gamma = \sqrt{k_\mu P_T/(dT)}$ and attain an $\widetilde{\mathcal{O}}(k_\mu^{5/4}c_\mu^{-1/2}d^{3/4}P_T^{1/4}T^{3/4})$ regret bound, which completes the proof. □

D Piecewise-Stationary SCB

In this section, we study SCB under piecewise-stationary environment and our work is a direct improvement over [Russac et al., 2021]. Next, we will first propose our SCB-PW-WeightUCB algorithm, and then, present the analysis of the confidence set. Finally, we give the proof of the dynamic regret upper bound.

D.1 SCB-PW-WeightUCB Algorithm

Inspired by Abeille et al. [2021], we make a direct improvement over Russac et al. [2021]. Just like Russac et al. [2021], for $D \geq 1$, define $\mathcal{T}(D) = \{1 \leq t \leq T,$ such that $\theta_s = \theta_t$ for $t-D \leq s \leq t-1\}$. $t \in \mathcal{T}(D)$ when t is at least D steps away from the previous closest changing point. But the difference is that Russac et al. [2021] considers D as an analysis parameter, and we treat D as a tunable algorithm parameter. Notice that, the D here is *not* a virtual window size, but the algorithm's estimate of how durable the environment is stationary.

Estimator. At iteration t, we adopt the same maximum likelihood estimator as in the drifting case as defined in (8).

Confidence Set. We further construct confidence set for the real θ_t. For $\delta \in (0,1)$, we define,

$$\mathcal{C}_t(\delta) \triangleq \left\{\theta\in\Theta \,\middle|\, \|g_t(\theta) - g_t(\widehat{\theta}_t)\|_{H_t^{-1}(\theta)} \leq \rho_t\right\},$$

where $\rho_t = \frac{2L^2Sk_\mu}{\sqrt{\lambda c_\mu}}\frac{\gamma^D}{1-\gamma} + \frac{Lm}{\sqrt{\lambda c_\mu}}\frac{\gamma^D}{1-\gamma} + \breve{\beta}_t$ and $\breve{\beta}_t = \frac{dm}{\sqrt{\lambda c_\mu}}\log\left(1+\frac{L^2k_\mu(1-\gamma^{2D})}{\lambda c_\mu d(1-\gamma)}\right) + \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}}\log\frac{1}{\delta} + \frac{2m}{\sqrt{\lambda c_\mu}}d\log(2) + \sqrt{\lambda c_\mu}S$.

Algorithm 5 SCB-PW-WeightUCB

Input: time horizon T, discounted factor γ, confidence δ, regularizer λ, inverse link function μ, parameters S, L and m, changing confidence D

1: Set $\widehat{\theta}_0 = \mathbf{0}$ and compute k_μ and c_μ
2: **for** $t = 1, 2, 3, ..., T$ **do**
3: Compute $(X_t, \widetilde{\theta}_t) = \arg\max_{\mathbf{x}\in\mathcal{X},\theta\in\mathcal{C}_t(\delta)} \mu(\mathbf{x}^\top\theta)$
4: Select X_t and receive the reward r_t
5: Compute $\widehat{\theta}_{t+1}$ according to (8)
6: **end for**

Lemma 10. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have $\forall t \in \mathcal{T}(D), \theta_t \in \mathcal{C}_t(\delta)$.*

$$\mathcal{C}_t(\delta) = \left\{\theta \in \Theta \mid \|g_t(\theta) - g_t(\widehat{\theta}_t)\|_{H_t^{-1}(\theta)} \le \frac{2L^2 S k_\mu}{\sqrt{\lambda c_\mu}}\frac{\gamma^D}{1-\gamma} + \frac{Lm}{\sqrt{\lambda c_\mu}}\frac{\gamma^D}{1-\gamma} + \breve{\beta}_t\right\},$$

where $\breve{\beta}_t = \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}}\log\frac{1}{\delta} + \frac{dm}{\sqrt{\lambda c_\mu}}\log\left(1 + \frac{L^2 k_\mu(1-\gamma^{2D})}{\lambda c_\mu d(1-\gamma)}\right) + \frac{2m}{\sqrt{\lambda c_\mu}} d\log(2) + \sqrt{\lambda c_\mu} S.$

The proof of Lemma 10 is presented in Appendix D.2.

Selection Criteria. Algorithms discussed earlier for drifting cases are using bonus-based selection criteria. But here we use a parameter-based selection criterion as follows,

$$(X_t, \widetilde{\theta}_t) = \underset{\mathbf{x}\in\mathcal{X},\theta\in\mathcal{C}_t(\delta)}{\arg\max}\ \mu(\mathbf{x}^\top\theta). \tag{51}$$

The main difference between parameter-based and bonus-based selection criteria is discussed in Section 3.2 of Abeille et al. [2021]. The overall algorithm is summarized in Algorithm 5.

D.2 Proof of Lemma 10

Proof. Based on the model assumption (13), the function g_t (10) and the $g_t(\widehat{\theta}_t)$ (44), we have,

$$\begin{aligned}
g_t(\theta_t) - g_t(\widehat{\theta}_t) &= \lambda c_\mu \theta_t + \sum_{s=1}^{t-1}\gamma^{t-s-1}\mu(X_s^\top\theta_t)X_s - \sum_{s=1}^{t-1}\gamma^{t-s-1} r_s X_s \\
&= \lambda c_\mu \theta_t + \sum_{s=1}^{t-1}\gamma^{t-s-1}\mu(X_s^\top\theta_t)X_s - \sum_{s=1}^{t-1}\gamma^{t-s-1}(\mu(X_s^\top\theta_s) + \eta_s)X_s \\
&= \sum_{s=1}^{t-1}\gamma^{t-s-1}(\mu(X_s^\top\theta_t) - \mu(X_s^\top\theta_s))X_s + \lambda c_\mu\theta_t - \sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_s X_s.
\end{aligned}$$

Then,

$$\begin{aligned}
\|g_t(\theta_t) - g_t(\widehat{\theta}_t)\|_{H_t^{-1}(\theta_t)} &= \left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}(\mu(X_s^\top\theta_t) - \mu(X_s^\top\theta_s))X_s + \lambda c_\mu\theta_t - \sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_s X_s\right\|_{H_t^{-1}(\theta_t)} \\
&\le \left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}(\mu(X_s^\top\theta_t) - \mu(X_s^\top\theta_s))X_s\right\|_{H_t^{-1}(\theta_t)} + \left\|\lambda c_\mu\theta_t - \sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_s X_s\right\|_{H_t^{-1}(\theta_t)} \\
&\le \underbrace{\left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}(\mu(X_s^\top\theta_t) - \mu(X_s^\top\theta_s))X_s\right\|_{H_t^{-1}(\theta_t)}}_{\texttt{term (a)}} + \underbrace{\left\|\sum_{s=1}^{t-D-1}\gamma^{t-s-1}\eta_s X_s\right\|_{H_t^{-1}(\theta_t)}}_{\texttt{term (b)}} \\
&\quad + \underbrace{\left\|\sum_{s=t-D}^{t-1}\gamma^{t-s-1}\eta_s X_s - \lambda c_\mu\theta_t\right\|_{H_t^{-1}(\theta_t)}}_{\texttt{term (c)}}.
\end{aligned}$$

Term (a). Since $t \in \mathcal{T}(D)$, we have

$$\begin{aligned}\left\|\sum_{s=1}^{t-1}\gamma^{t-s-1}(\mu(X_s^\top\theta_t)-\mu(X_s^\top\theta_s))X_s\right\|_{H_t^{-1}(\theta_t)} &= \left\|\sum_{s=1}^{t-D-1}\gamma^{t-s-1}(\mu(X_s^\top\theta_t)-\mu(X_s^\top\theta_s))X_s\right\|_{H_t^{-1}(\theta_t)}\\ &\leq \left\|\sum_{s=1}^{t-D-1}\gamma^{t-s-1}k_\mu X_s^\top(\theta_t-\theta_s)X_s\right\|_{H_t^{-1}(\theta_t)}\\ &\leq \sum_{s=1}^{t-D-1}\gamma^{t-s-1}k_\mu\|X_s\|_2\|(\theta_t-\theta_s)\|_2\,\|X_s\|_{H_t^{-1}(\theta_t)}\\ &\leq \frac{2L^2Sk_\mu}{\sqrt{\lambda c_\mu}}\frac{\gamma^D}{1-\gamma}.\end{aligned}$$

Term (b).

$$\left\|\sum_{s=1}^{t-D-1}\gamma^{t-s-1}\eta_s X_s\right\|_{H_t^{-1}(\theta_t)} \leq \sum_{s=1}^{t-D-1}\gamma^{t-s-1}m\,\|X_s\|_{H_t^{-1}(\theta_t)} \leq \frac{Lm}{\sqrt{\lambda c_\mu}}\sum_{s=1}^{t-D-1}\gamma^{t-s-1} \leq \frac{Lm}{\sqrt{\lambda c_\mu}}\frac{\gamma^D}{1-\gamma}.$$

Term (c). Let $\widetilde{H}_{t-D:t}(\theta) = \lambda c_\mu \gamma^{-2(t-1)} I_d + \sum_{s=t-D}^{t-1}\gamma^{-2s}\mu'(X_s^\top\theta)X_sX_s^\top$

$$\begin{aligned}\left\|\sum_{s=t-D}^{t-1}\gamma^{t-s-1}\eta_s X_s - \lambda c_\mu\theta_t\right\|_{H_t^{-1}(\theta_t)} &\leq \left\|\sum_{s=t-D}^{t-1}\gamma^{t-s-1}\eta_s X_s\right\|_{H_t^{-1}(\theta_t)} + \sqrt{\lambda c_\mu}S\\ &\leq \left\|\sum_{s=t-D}^{t-1}\gamma^{-s}\eta_s X_s\right\|_{\widetilde{H}_t^{-1}(\theta_t)} + \sqrt{\lambda c_\mu}S\\ &\leq \left\|\sum_{s=t-D}^{t-1}\gamma^{-s}\eta_s X_s\right\|_{\widetilde{H}_{t-D:t}^{-1}(\theta_t)} + \sqrt{\lambda c_\mu}S.\end{aligned}$$

We already proof that $\gamma^{-2(t-1)}H_t(\theta) \succeq \widetilde{H}_t^{-1}(\theta)$ in (50), and obviously $\widetilde{H}_t(\theta) \succeq \widetilde{H}_{t-D:t}(\theta)$. Next, we need to bound the term $\|\sum_{s=t-D}^{t-1}\gamma^{-s}\eta_s X_s\|_{\widetilde{H}_{t-D:t}^{-1}(\theta_t)}$ using self-normalization bound [Russac et al., 2021, Theorem 3], restated in Theorem 6 by let $w_t = \gamma^{-t}, \lambda_t = \lambda c_\mu \gamma^{-2t}$, then we have

$$\begin{aligned}&\left\|\sum_{s=t-D}^{t-1}\gamma^{-s}\eta_s X_s\right\|_{\widetilde{H}_{t-D:t}^{-1}(\theta_t)}\\ \leq&\frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}}\log\left(\frac{\det\left(\widetilde{H}_{t-D:t}\right)^{1/2}}{\delta(\lambda c_\mu\gamma^{-2(t-1)})^{d/2}}\right) + \frac{2m}{\sqrt{\lambda c_\mu}}d\log(2)\\ \leq&\frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}}\log\frac{1}{\delta} + \frac{dm}{\sqrt{\lambda c_\mu}}\log\left(\frac{\lambda c_\mu\gamma^{-2(t-1)} + \frac{L^2k_\mu\sum_{s=t-D}^{t}\gamma^{-2s}}{d}}{\lambda c_\mu\gamma^{-2(t-1)}}\right) + \frac{2m}{\sqrt{\lambda c_\mu}}d\log(2)\\ \leq&\frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}}\log\frac{1}{\delta} + \frac{dm}{\sqrt{\lambda c_\mu}}\log\left(1 + \frac{L^2k_\mu(1-\gamma^{2D})}{\lambda c_\mu d(1-\gamma)}\right) + \frac{2m}{\sqrt{\lambda c_\mu}}d\log(2).\end{aligned}$$

Let $\breve{\beta}_t \triangleq \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}}\log\frac{1}{\delta} + \frac{dm}{\sqrt{\lambda c_\mu}}\log\left(1+\frac{L^2k_\mu(1-\gamma^{2D})}{\lambda c_\mu d(1-\gamma)}\right) + \frac{2m}{\sqrt{\lambda c_\mu}}d\log(2) + \sqrt{\lambda c_\mu}S$, finally we have,

$$\|g_t(\theta_t) - g_t(\widehat{\theta}_t)\|_{H_t^{-1}(\theta_t)} \leq \frac{2L^2Sk_\mu}{\sqrt{\lambda c_\mu}}\frac{\gamma^D}{1-\gamma} + \frac{Lm}{\sqrt{\lambda c_\mu}}\frac{\gamma^D}{1-\gamma} + \breve{\beta}_t,$$

which completes the proof. □

D.3 Proof of Theorem 4

Proof. Let $R_t = \mu(X_t^{*\top}\theta_t) - \mu(X_t^\top\theta_t)$

$$R_T = \sum_{t=1}^T R_t = \sum_{t\notin\mathcal{T}(D)} R_t + \sum_{t\in\mathcal{T}(D)} R_t = \Gamma_T D + \sum_{t\in\mathcal{T}(D)} R_t.$$

For $t \in \mathcal{T}(D)$, by selection criterion (51),

$$\begin{aligned}
R_t &= \mu(X_t^{*\top}\theta_t) - \mu(X_t^\top\theta_t) \\
&\le \mu(X_t^\top\widetilde{\theta}_t) - \mu(X_t^\top\widehat{\theta}_t) + \mu(X_t^\top\widehat{\theta}_t) - \mu(X_t^\top\theta_t) \\
&\le \alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t)\left|X_t^\top\left(\widetilde{\theta}_t - \widehat{\theta}_t\right)\right| + \alpha(X_t,\theta_t,\widehat{\theta}_t)\left|X_t^\top\left(\theta_t - \widehat{\theta}_t\right)\right| \\
&\le \sqrt{1+2S}\Big(\alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t)\left\|X_t\right\|_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}\left\|g_t(\widetilde{\theta}_t) - g_t(\widehat{\theta}_t)\right\|_{H_t^{-1}(\widetilde{\theta}_t)} \\
&\qquad + \alpha(X_t,\theta_t,\widehat{\theta}_t)\left\|X_t\right\|_{G_t^{-1}(\theta_t,\widehat{\theta}_t)}\left\|g_t(\theta_t) - g_t(\widehat{\theta}_t)\right\|_{H_t^{-1}(\theta_t)}\Big).
\end{aligned}$$

where $\alpha(\mathbf{x},\theta_1,\theta_2) = \int_0^1 \mu'(v\mathbf{x}^\top\theta_2 + (1-v)x^\top\theta_1)\mathrm{d}v$, and the last second inequality comes from the mean value theorem $\mu(\mathbf{x}^\top\theta_1) - \mu(\mathbf{x}^\top\theta_2) = \alpha(\mathbf{x},\theta_1,\theta_2)(\mathbf{x}^\top\theta_1 - \mathbf{x}^\top\theta_2)$. Since that $\widetilde{\theta}_t \in \mathcal{C}_t(\delta)$ and with probability at least $1-\delta$, $\forall t\in[T]$, $\theta_t \in \mathcal{C}_t(\delta)$, and by union bound, the following dynamic regret bound hold with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{t\in\mathcal{T}(D)} R_t &\le \sum_{t\in\mathcal{T}(D)} \sqrt{1+2S}\left(\alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t)\left\|X_t\right\|_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}\rho_t + \alpha(X_t,\theta_t,\widehat{\theta}_t)\left\|X_t\right\|_{G_t^{-1}(\theta_t,\widehat{\theta}_t)}\rho_t\right) \\
&\le \sqrt{1+2S}\rho_T\left(\sum_{t\in\mathcal{T}(D)} \alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t)\left\|X_t\right\|_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)} + \sum_{t\in\mathcal{T}(D)} \alpha(X_t,\theta_t,\widehat{\theta}_t)\left\|X_t\right\|_{G_t^{-1}(\theta_t,\widehat{\theta}_t)}\right).
\end{aligned}$$

Now we try to derive the upper bound for term $\sum_{t\in\mathcal{T}(D)} \alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t)\left\|X_t\right\|_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}$.

Based on the definition of g_t (10), we have

$$\begin{aligned}
g_t(\theta_1) - g_t(\theta_2) &= \lambda c_\mu(\theta_1 - \theta_2) + \sum_{s=1}^{t-1}\gamma^{t-s-1}(\mu(X_s^\top\theta_1) - \mu(X_s^\top\theta_2))X_s \\
&= \lambda c_\mu(\theta_1 - \theta_2) + \sum_{s=1}^{t-1}\gamma^{t-s-1}\alpha(X_s,\theta_1,\theta_2)X_s^\top X_s(\theta_1 - \theta_2) \\
&= \left(\lambda c_\mu + \sum_{s=1}^{t-1}\gamma^{t-s-1}\alpha(X_s,\theta_1,\theta_2)X_s^\top X_s\right)(\theta_1 - \theta_2).
\end{aligned} \tag{52}$$

Then based on the definition of G_t (45), we know $G_t(\theta_1,\theta_2) = \lambda c_\mu + \sum_{s=1}^{t-1}\gamma^{t-s-1}\alpha(X_s,\theta_1,\theta_2)X_s^\top X_s$. which means $G_t(\widetilde{\theta}_t,\widehat{\theta}_t) = \lambda c_\mu I_d + \sum_{s=1}^{t-1}\gamma^{t-s-1}\alpha(X_s,\widetilde{\theta}_t,\widehat{\theta}_t)X_s X_s^\top$, if we let $\widetilde{X}_s = \sqrt{\alpha(X_s,\widetilde{\theta}_t,\widehat{\theta}_t)}X_s$, then

$$\begin{aligned}
\sum_{t\in\mathcal{T}(D)} \alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t)\left\|X_t\right\|_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)} &\le \sqrt{\sum_{t=1}^T \alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t)}\sqrt{\sum_{t=1}^T \alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t)\left\|X_t\right\|^2_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}} \\
&\le \sqrt{k_\mu T}\sqrt{\sum_{t=1}^T \left\|\widetilde{X}_t\right\|^2_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}}.
\end{aligned} \tag{53}$$

Then for the term $\sqrt{\sum_{t=1}^T \|\widetilde{X}_t\|^2_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}}$, we can directly use the Lemma 11 to bound it,

$$\sqrt{k_\mu T}\sqrt{\sum_{t=1}^T \left\|\widetilde{X}_t\right\|^2_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}} \le \sqrt{2k_\mu \max\{1, L^2k_\mu/(\lambda c_\mu)\}dT}\sqrt{T\log\frac{1}{\gamma} + \log\left(1 + \frac{L^2k_\mu}{\lambda c_\mu d(1-\gamma)}\right)}.$$

We can bound term $\sum_{t\in\mathcal{T}(D)} \alpha(X_t,\theta_t,\widehat{\theta}_t)\,\|X_t\|_{G_t^{-1}(\theta_t,\widehat{\theta}_t)}$ in the same way and get,

$$\sum_{t\in\mathcal{T}(D)} \alpha(X_t,\theta_t,\widehat{\theta}_t)\,\|X_t\|_{G_t^{-1}(\theta_t,\widehat{\theta}_t)} \leq \sqrt{2k_\mu \max\{1, L^2k_\mu/(\lambda c_\mu)\}dT}\sqrt{T\log\frac{1}{\gamma} + \log\left(1+\frac{L^2k_\mu}{\lambda c_\mu d(1-\gamma)}\right)}.$$

Combine these two bound and let $\delta = 1/T$, we have the following regret bound with probability at least $1-1/T$,

$$R_T \leq \Gamma_T D + 2\sqrt{1+2S}\rho_T\sqrt{2k_\mu \max\{1, L^2k_\mu/(\lambda c_\mu)\}dT}\sqrt{T\log\frac{1}{\gamma} + \log\left(1+\frac{L^2k_\mu}{\lambda c_\mu d(1-\gamma)}\right)},$$

where $\rho_t = \frac{2L^2Sk_\mu}{\sqrt{\lambda c_\mu}}\frac{\gamma^D}{1-\gamma} + \frac{Lm}{\sqrt{\lambda c_\mu}}\frac{\gamma^D}{1-\gamma} + \breve{\beta}_t$ and $\breve{\beta}_t = \frac{dm}{\sqrt{\lambda c_\mu}}\log\left(1+\frac{L^2k_\mu(1-\gamma^{2D})}{\lambda c_\mu d(1-\gamma)}\right) + \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}}\log(T) + \frac{2m}{\sqrt{\lambda c_\mu}}d\log(2) + \sqrt{\lambda c_\mu}S$. Since that there is a $T\sqrt{\log(1/\gamma)}$ term in the regret bound, which means that we cannot let γ close to 0, so we set $\gamma \geq 1/2$, then we have $\log(1/\gamma) \leq 2\log(2)(1-\gamma)$. Then, we set $D = \log(T)/\log(1/\gamma)$, noticing that $0 < 1/\gamma - 1 < 1$ and using $\log(1+x) \geq x/2$ for $0 < x < 1$, we have

$$\log\frac{1}{\gamma} = \log(1+1/\gamma-1) \geq \frac{1-\gamma}{2\gamma}.$$

Therefore, we have $D \leq \frac{2\gamma\log(T)}{1-\gamma}$. Then, ignoring logarithmic factors in time horizon T, and let $\lambda = d\log(T)/c_\mu$, we finally obtain that,

$$\begin{aligned} R_T &\leq \widetilde{\mathcal{O}}\left(\frac{1}{1-\gamma}\Gamma_T + \left(\frac{1}{\sqrt{d}}\frac{1}{1-\gamma}\frac{1}{T} + \sqrt{d}\right)\sqrt{d(1-\gamma)T}\right) \\ &\leq \widetilde{\mathcal{O}}\left(\frac{1}{1-\gamma}\Gamma_T + \frac{1}{\sqrt{1-\gamma}} + d\sqrt{(1-\gamma)T}\right). \end{aligned}$$

When $\Gamma_T < d/\sqrt{T}$ (which corresponds a small amount of non-stationarity), we simply set $\gamma = 1-1/T$ and achieve an $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret bound. Besides, when coming to the non-degenerated case of $\Gamma_T > d/\sqrt{T}$, We set the discounted factor optimally as $1-\gamma = (\Gamma_T/(dT))^{2/3}$ and attain an $\widetilde{\mathcal{O}}(d^{2/3}\Gamma_T^{1/3}T^{2/3})$ dynamic regret bound, which completes the proof. □

E Thechnical Lemmas

In this section, we provide several useful lemmas, mainly about weighted version self-normalized concentration, weighted version potential lemma and some derivatives of self-concordant property.

E.1 Weighted Version Self-normalized Concentration

Theorem 5 (Weighted Version Self-Normalized Bound for Vector-Valued Martingales [Russac et al., 2019, Theorem 1])**.** *Let $\{\mathcal{F}_t\}_{t=0}^\infty$ be a filtration, $\{\eta_t\}_{t=0}^\infty$ be a real-valued stochastic process such that η_t is $\mathcal{F}_t$-measurable and η_t is conditionally R-sub-Gaussian for some $R \geq 0$, such that*

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

Let $\{X_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued stochastic process such that X_t is $\mathcal{F}_{t-1}$-measurable. For any $t \geq 0$, define

$$\widetilde{V}_t = \mu_t I_d + \sum_{s=1}^t w_s^2 X_s X_s^\top, \qquad S_t = \sum_{s=1}^t w_s\eta_s X_s.$$

where $\forall s \geq 0, t\geq 0, w_s, \mu_t > 0$. Then, for any $\delta > 0$, with probability at least $1-\delta$, we have

$$\forall t\geq 0, \|S_t\|_{\widetilde{V}_t^{-1}} \leq R\sqrt{2\log\frac{1}{\delta} + d\log\left(1+\frac{L^2\sum_{s=1}^t w_s^2}{d\mu_t}\right)}.$$

Theorem 6 (Theorem 3 of Russac et al. [2021])**.** *Let t be a fixed time instant. Let $\{\mathcal{F}_t\}_{t=0}^\infty$ be a filtration. Let $\{X_t\}_{t=0}^\infty$ be a stochastic process on $\mathbb{R}^d$ such that X_t is $\mathcal{F}_{t-1}$ measurable and $\|X_t\|_2 \leq 1$. Let $\{\eta_t\}_{t=0}^\infty$ be a martingale difference sequence such that η_t is $\mathcal{F}_t$ measurable. Assume that the weights are non-decreasing, strictly positive and the time horizon* *is known. Furthermore, assume that conditionally on $\mathcal{F}_t$ we have $|\eta_t| \leq m$ a.s. Let $\{\lambda_t\}_{u=0}^{\infty}$ be a deterministic sequence of regularization terms and denote $\sigma_t^2 = \mathbb{E}\left[\eta_t^2 \mid \mathcal{F}_t\right]$. Let $\widetilde{H}_t = \sum_{s=1}^{t-1} w_s^2 \sigma_s^2 X_s X_s^\top + \lambda_{t-1}\mathbf{I}_d$ and $S_t = \sum_{s=1}^{t-1} w_s \eta_s X_s$, then for any $\delta \in (0,1]$, with probability at least $1-\delta$,*

$$\forall t \geq 0, \|S_t\|_{\widetilde{H}_t^{-1}} \leq \frac{\sqrt{\lambda_{t-1}}}{2mw_{t-1}} + \frac{2mw_{t-1}}{\sqrt{\lambda_{t-1}}} \log\left(\frac{\det\left(\widetilde{H}_t\right)^{1/2}}{\delta \lambda_{t-1}^{d/2}}\right) + \frac{2mw_{t-1}}{\sqrt{\lambda_{t-1}}} d \log(2).$$

E.2 Weighted Version Potential Lemma

Lemma 11 (Weighted Version Potential Lemma [Faury et al., 2021, Lemma 8])**.** *Suppose $V_t = \sum_{s=1}^t \gamma^{t-s} X_s X_s^\top + \lambda I_d, V_0 = \lambda I_d, \gamma \in (0,1]$ and $\|X_t\|_2 \leq L$ for all $t \geq 1$, then the following inequality holds,*

$$\sum_{t=1}^T \|X_t\|_{V_{t-1}^{-1}}^2 \leq 2\max\{1, L^2/\lambda\} d \left(T \log\frac{1}{\gamma} + \log\left(1 + \frac{L^2}{\lambda d(1-\gamma)}\right)\right).$$

Lemma 12 (Determinant inequality)**.** *Let $V_t = \sum_{s=1}^t w_{t,s} X_s X_s^\top + \lambda_t I_d, V_0 = \lambda_0 I_d$. Assume $\|\mathbf{x}\|_2 \leq L$ and we have,*

$$\det(V_t) \leq \left(\lambda_t + \frac{L^2 \sum_{s=1}^t w_{t,s}}{d}\right)^d.$$

Proof. Now we have $V_t = \sum_{s=1}^t w_{t,s} X_s X_s^\top + \lambda_t I_d$, take the trace on both sides, and get the upper bound of $\mathrm{Tr}(V_t)$

$$\mathrm{Tr}(V_t) = \mathrm{Tr}(\lambda_t I_d) + \sum_{s=1}^t w_{t,s} \mathrm{Tr}\left(X_s X_s^\top\right) = \lambda_t d + \sum_{s=1}^t w_{t,s} \|X_s\|_2^2 \leq \lambda_t d + L^2 \sum_{s=1}^t w_{t,s}. \tag{54}$$

Base on the definition of determinant and the upper bound of $\mathrm{Tr}(V_t)$ (54), we can get the upper bound for $\det(V_t)$,

$$\det(V_t) = \prod_{i=1}^d \lambda_i \leq \left(\frac{\sum_{i=1}^d \lambda_i}{d}\right)^d = \left(\frac{\mathrm{Tr}(V_t)}{d}\right)^d \leq \left(\lambda_t + \frac{L^2 \sum_{s=1}^t w_{t,s}}{d}\right)^d.$$

□

E.3 Self-Concordant Properties

Based on the generalized self-concordant property of the (inverse) link function $\mu(\cdot)$, we have the following lemma, which will be later used to derive Lemma 14.

Lemma 13 (Lemma 9 of Faury et al. [2020])**.** *For any $z_1, z_2 \in \mathbb{R}$, we have the following inequality:*

$$\mu'(z_1)\frac{1-\exp(-|z_1-z_2|)}{|z_1-z_2|} \leq \int_0^1 \mu'(z_1 + v(z_2-z_1))\mathrm{d}v \leq \mu'(z_1)\frac{\exp(|z_1-z_2|)-1}{|z_1-z_2|}.$$

Furthermore, $\int_0^1 \mu'(z_1 + v(z_2-z_1))\mathrm{d}v \geq \mu'(z_1)(1+|z_1-z_2|)^{-1}$.

The following lemma provides a weighted version of Lemma 10 of Faury et al. [2020] which can be easily proven.

Lemma 14. *With G_t defined in (45) and H_t defined in (46), the following inequalities hold*

$$\forall \theta_1, \theta_2 \in \Theta, \quad G_t(\theta_1, \theta_2) \geq (1+2S)^{-1} H_t(\theta_1), \quad G_t(\theta_1, \theta_2) \geq (1+2S)^{-1} H_t(\theta_2).$$