Revisiting Weighted Strategy for Non-stationary Parametric Bandits and MDPs

Jing Wang, Peng Zhao, *Member, IEEE,* and Zhi-Hua Zhou, *Fellow, IEEE*

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. Moreover, we extend our framework to non-stationary Markov Decision Processes (MDPs) with function approximation, focusing on Linear Mixture MDP and Multinomial Logit (MNL) Mixture MDP. For both classes, we propose algorithms based on the weighted strategy and establish dynamic regret guarantees using our analysis framework.

Index Terms—dynamic regret, non-stationary bandits, discounted factor, online MDPs, function approximation.

I. 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 [1]–[11] due to their significance in many real-world non-stationary online applications such as recommendation systems [12], [13]. This line of work also has a tight connection with the theoretical foundation of Reinforcement Learning (RL), particularly in the context of episodic Markov Decision Processes (MDPs) with function approximation [14]–[17]. In

these settings, parametric bandits are frequently employed to model both reward and transition dynamics across episodes. Moreover, when these underlying dynamics exhibit *non-stationary* behavior across different episodes, non-stationary parametric bandit techniques naturally extend to capture the non-stationary dynamics of rewards and transitions [18], [19].

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 (e.g. 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. Furthermore, LB and GLB have fundamental applications in Markov Decision Processes (MDPs) with function approximation. As a representative instance, the Linear Mixture MDP adopts LB to model both reward functions and transition dynamics. Building on this, the Multinomial Logit (MNL) Mixture MDP was introduced to address the limitation of linear functions to model probabilities. By employing the MNL bandit (a special case of GLB), it effectively models transition probabilities and ensures valid distributions. Notably, the non-stationary models allow the parameter θ_t in the above models to be time-varying; therefore, we use dynamic regret [20], [21] to evaluate the algorithm's performance. 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 assigns more weight to the most recent data and less to older data, gradually forgetting 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 and episodic MDPs with linear function approximation is by [8], who developed a minimax optimal algorithm consisting of a non-stationarity detector and a base algorithm that performs well in near-stationary environments. Whenever the detector ex-

J. Wang, P. Zhao and Z.-H. Zhou are with National Key Laboratory for Novel Software Technology and the School of Artificial Intelligence, Nanjing University, Nanjing 210023, China. (e-mail: {wangjing,zhaop,zhouzh}@lamda.nju.edu.cn). This paper was presented in part at the Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS), 2023.

TABLE I: Comparisons of our dynamic regret bounds to the previous best-known results for weight-based algorithms, under different non-stationary bandit and MDP settings. Below, k_μ/c_μ characterize the non-linearity in GLB/SCB (reducing to 1 for LB) and κ denotes the non-linearity in MNL Mixture MDP; d is the dimension, H is the length of an episode in the MDP setting, path length P_T and the change number Γ_T are non-stationarity measures for drifting and piecewise-stationary cases, respectively, and total path length Δ measures the non-stationarity in the drifting MDP case.

Settings	Previous Work	Our Results
Drifting LB	$\widetilde{\mathcal{O}}\big(d^{7/8}P_T^{1/4}T^{3/4}\big)$ [2]	$\widetilde{\mathcal{O}}\big(d^{3/4}P_T^{1/4}T^{3/4}\big)$ [Theorem 1]
Drifting GLB	$\widetilde{\mathcal{O}}\big(\frac{k_\mu^2}{c_\mu}d^{9/10}P_T^{1/5}T^{4/5}\big)$ [6]	$\widetilde{\mathcal{O}}\big(\frac{k_\mu^{5/4}}{c_\mu^{3/4}}d^{3/4}P_T^{1/4}T^{3/4}\big)$ [Theorem 2]
Drifting SCB	$\widetilde{\mathcal{O}}\big(\frac{k_\mu^2}{c_\mu}d^{9/10}P_T^{1/5}T^{4/5}\big)$ [6]	$\widetilde{\mathcal{O}}\big(\frac{k_\mu^{5/4}}{c_\mu^{1/2}}d^{3/4}P_T^{1/4}T^{3/4}\big)$ [Theorem 3]
Piecewise Stationary SCB	$\widetilde{\mathcal{O}}\big(\frac{1}{c_\mu^{1/3}}d^{2/3}\Gamma_T^{1/3}T^{2/3}\big)$ [7]	$\widetilde{\mathcal{O}}\big(d^{2/3}\Gamma_T^{1/3}T^{2/3}\big)$ [Theorem 4]
Drifting Linear Mixture MDP	$\widetilde{\mathcal{O}}\big(Hd\Delta^{1/4}T^{3/4}\big)$ [18]	$\widetilde{\mathcal{O}}\big(Hd\Delta^{1/4}T^{3/4}\big)$ [Theorem 5]
Drifting MNL Mixture MDP	—	$\widetilde{\mathcal{O}}\big(\kappa^{-1}Hd\Delta^{1/4}T^{3/4}\big)$ [Theorem 6]

amines that the non-stationarity exceeds a certain limit, the algorithm will *restart* itself to handle the non-stationarity. In this sense, their algorithm can be regarded as an *adaptive restart-based algorithm*. Building on the RestartUCB algorithm [3] 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 and $\widetilde{\mathcal{O}}(\min\{\sqrt{\Gamma_T T}, \Delta^{1/3}T^{2/3}\})$ dynamic regret for Episodic MDPs where Δ represents the total path length which includes the cumulative parameters variations of both reward function and transition function.

In real-world scenarios, the distributional change of environments often exhibits gradually drifting patterns [22]–[27], 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 [2], [3], [9], classification with concept drift [28], [29], and adaptive system identification [30], [31]. 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 [2], [6], [7]. 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. [2] provided the first analysis of a weight-based algorithm for linear bandits (LB). In their bias analysis, they introduced a

virtual window size in the analysis to control the bias in order to mimic the analysis of a sliding-window strategy [1]. For the variance analysis, [2] developed a weighted version of the self-normalized concentration inequality, which required a specially designed local norm. This introduced additional analytical complexity, since the previously studied sliding-window [1] and restart strategy [3] could directly apply the standard self-normalized concentration inequality [32]. Consequently, they have to use *different* local norms to control bias and variance parts, resulting in unexpected inefficiencies of algorithm design and complications of analysis. For LB, this leads to an algorithm D-LinUCB [2] requiring the maintenance of an extra covariance matrix as the local norm for the weighted version self-normalized concentration, which is less efficient than the window and restart-based algorithms [1], [3].

This analysis framework for weighted strategy introduces more severe issues in GLB, due to its more enriched and complicated structure. Specifically, [6] studied the drifting GLB and designed a highly complex projection operation to control bias and variance parts following the way of [2] 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) [2]. Subsequently, [7] 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 they failed in the more challenging drifting setting, due to technical difficulties in bounding bias using conventional analysis. Moreover, since the weighted version of the self-normalized concentration for LB [2] could not be extended to the SCB setting, they further redesigned a weighted version specifically for SCB, building on the self-normalized concentration for stationary SCB [33], which introduced substantial additional complexity into the 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 [34] to improve dependence on c_μ in non-stationary SCB.

Our Results. In this paper, we revisit the weighted strategy for non-stationary parametric bandits and MDPs. We discover that the earlier analysis framework for the weighted strategy may be inappropriate due to its reliance on mimicking the sliding-window analysis and the specifically designed weighted version of self-normalized concentration, which requires bounding the bias and variance components using *different* local norms, and designing new weighted versions of this concentration tool for every new setting further introduces significant and unnecessary analytical complexity. As a result, there is no unified analysis framework that can be applied directly across different settings. To address this, we propose a *refined analysis framework* for the weighted strategy. In our framework, a new analysis for the bias part is presented, while the variance part analysis only relies on the standard self-normalized concentration [32] without the need for an additional weighted version and enables the use of a *single* local norm to analyze both the bias and variance components. This refinement simplifies the analysis of the weighted strategy and makes the approach more broadly applicable to other decision-making settings. It also brings several benefits to algorithm design, including improved efficiency for LB and a resolution to the projection issue encountered in GLB and SCB. Furthermore, our analysis framework is not limited to the bandit setting and can be extended to online Markov decision processes (MDPs) scenarios. In this paper, we extend our results to two fundamental classes of MDPs: (i) non-stationary linear mixture MDPs, and (ii) non-stationary multinomial logit (MNL) mixture MDPs. Table I summarizes our main results compared with the best-known results for weight-based algorithms. Specifically, based on our refined analysis framework, we achieve: (i) for LB, our approach only needs to maintain one covariance instead of two and still enjoys the same regret as [2]; (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; (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, and 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 [7]; (iv) for Linear Mixture MDP, we achieve an $\widetilde{\mathcal{O}}(H d \Delta^{1/4} T^{3/4})$ regret bound that enjoys the same regret as [18] that was achieved by the restarted strategy; and (v) for MNL Mixture MDP, we establish the first dynamic regret bound of $\widetilde{\mathcal{O}}(H d \Delta^{1/4} T^{3/4})$ in the literature.

Compared with our earlier conference version [11], this extended version presents additional results, along with a simpler, clearer analysis and refined presentation. Firstly, this extended version further simplifies the analysis compared to our conference version [11]. The earlier approach [2] relied on three key components: an extra covariance matrix, a weighted self-normalized concentration inequality, and a weighted potential lemma. In our conference version [11], we removed the need to maintain an additional covariance matrix. In this extended version, we take it a step further by showing that the standard self-normalized concentration inequality is sufficient for analyzing the weighted strategy. As

a result, the only essential component for the weighted strategy analysis is the weighted potential lemma. The maintenance of two covariance matrices and the use of weighted self-normalized concentration, as done in previous works [2], [11], are unnecessary. This simplification makes our analysis and algorithm both much simpler and more general. Secondly, this simplification makes our approach much more scalable and easier to extend to other bandit settings. Both earlier work [7] and our conference version [11] required designing a new weighted self-normalized concentration inequality when adapting the method to a new setting (e.g., SCB), which limited their generality. Our refined analysis removes this need, allowing the same framework to be applied across different problems without requiring problem-specific weighted concentration results. Thirdly, we extend our results to two fundamental settings of online MDPs with function approximation: linear mixture MDPs and multinomial logit (MNL) mixture MDPs. Notably, we provide the first dynamic regret guarantee for MNL mixture MDPs, demonstrating both the effectiveness and the broad applicability of our refined analytical framework for weighted strategy.

II. RELATED WORK

Linear Bandits. The non-stationary LB problem was first studied by [1]. 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 [2] proposed the D-LinUCB algorithm based on a weighted strategy, and [3] proposed the RestartUCB algorithm based on a 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 [35]. 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 [35], [36]. However, to achieve this result, all 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 this, [1] 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, [8] proposed the MASTER algorithm with theoretically optimal $\widetilde{\mathcal{O}}(\min\{d\sqrt{\Gamma_T T}, d P_T^{1/3} T^{2/3}\})$ regret bound, also without requiring the non-stationarity level of environments (that is, Γ_T and P_T) in advance, but requires fixed arm set assumption. Most recently, there has also been some new progress in the non-stationary (linear) bandits [10], [37]–[39]. Furthermore, [40, Remark 4] bypassed the aforementioned technical gap by restarting adversarial LB algorithms. However, it is important to note that this only applies to LB and requires fixed arm set assumption and known P_T.

Generalized Linear Bandits. The GLB problem was first introduced by [41]. 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. [6] 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. [33] studied a specific instance of GLB called Logistic Bandits (LogB). They first pointed out that under the 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, [34] established an $\Omega(d\sqrt{\mu'(X_*^\top \theta_*)T})$ regret lower bound for logistic bandits and provided an optimal algorithm OFULog. [7] 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, [6] proposed a parameter-free algorithm by combining BVD-GLM-UCB with the BOB strategy, but the final result is still suboptimal. Meanwhile, the black-box algorithm [8] can adaptively restart the stationary algorithm GLM-UCB [41] and achieve an $\widetilde{\mathcal{O}}(\min k_\mu c_\mu^{-1} \sqrt{\Gamma_T T}, k_\mu^{4/3} c_\mu^{-1} d P_T^{1/3} T^{2/3})$ regret, which matches the lower bound for non-stationary LB in terms of P_T and T, and therefore optimal for non-stationary GLBs, since LB is a special case of GLB (i.e. $\mu(x) = x$). Recently, there has been notable progress in GLB regarding its efficiency and regret optimality in terms of non-linearity. Readers can refer to [42] for the latest advancements. Nonetheless, these results focus on the static regret setting.

MDP with Function Approximation. Reinforcement learning with function approximation has attracted significant attention recently [16]–[18], [43], [44], with two fundamental approaches: linear function approximation and generalized linear function approximation. Among these, Linear Mixture MDP was first introduced by [43], [44], which is a representative model for linear function approximation. They proposed the UCRL-VTR algorithm, achieving a regret bound of $\widetilde{\mathcal{O}}(d\sqrt{H^3 T})$, where H is the episode horizon. Building on this, [18] extended the stationary Linear Mixture MDP to drifting case, they establish $\Omega(d^{5/6} \Delta^{1/3} H^{2/3} T^{2/3})$ minimax lower bound for non-stationary linear mixture MDPs, and then proposed the SW-LSVI-UCB algorithm, which achieves a regret bound of $\widetilde{\mathcal{O}}(Hd\Delta^{1/4} T^{3/4})$, where Δ quantifies the cumulative variation of the underlying parameters. To better capture the probabilistic nature of transition dynamics, [16] explored a class of generalized function approximation models and introduced the MNL Mixture MDP, which leverages the MNL function to model transitions. They proposed the UCRL-MNL algorithm, achieving a regret bound of $\widetilde{\mathcal{O}}(\kappa^{-1} d\sqrt{H^3 T})$, where κ represents the nonlinearity of the MNL model, H is the episode horizon, and K is the total number of episodes. Later, [17] further achieved an $\widetilde{\mathcal{O}}(d\sqrt{H^3 T} + \kappa^{-1} d^2 H^2)$ regret bound for MNL Mixture MDP.

III. LINEAR BANDIT

In this section, we first introduce the problem setting of non-stationary LB, and describe our LB-WeightUCB algorithm and its theoretical guarantee. Then we present a proof sketch of Lemma 1 to illustrate our proposed analysis framework in detail. Notably, our algorithm achieves the same regret bound as the best-known weight-based algorithm [2] without relying

on a specially designed weighted version of self-normalized concentration and can be more efficient.

A. 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

$$r_t = X_t^\top \theta_t + \eta_t, \qquad (1)$$

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*:

$$\text{D-REG}_T = \sum_{t=1}^T \max_{\mathbf{x} \in \mathcal{X}} \mathbf{x}^\top \theta_t - \sum_{t=1}^T X_t^\top \theta_t, \qquad (2)$$

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. Notice that in this paper we focus on a fixed arm set $\mathcal{X}$. A time-varying arm set $\mathcal{X}_t$ does not introduce any additional difficulty for our weight-based algorithm. The only difference is that the optimal comparator in the dynamic regret (2) would need to be updated to $\max_{\mathbf{x} \in \mathcal{X}_t} \mathbf{x}^\top \theta_t$, and the arm selection step (6) would be performed over the time-varying arm set $\mathcal{X}_t$ instead of $\mathcal{X}$. This does not affect the analysis. For simplicity, we stick to the fixed arm set setting in this paper.

We work under the following standard boundedness assumption [1]–[3], [32].

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 $\Theta \triangleq \{\theta \mid \|\theta\|_2 \leq S\}$.

B. Algorithm and Regret Guarantee

We propose the LB-WeightUCB algorithm, which attains the same regret guarantees as previous methods while enjoying better 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 a weighted regularized least square estimator similar to D-LinUCB [2], the estimator $\widehat{\theta}_t$ is the solution to the following problem,

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

where $\lambda > 0$ is the regularization coefficient and $\forall t \in [T], s \in [t-1], w_{t-1,s}$ is the weighted factor. To deal with non-stationarity, we set $w_{t,s} = \gamma^{t-s}$, where $\gamma \in (0,1)$ is the discounted factor. This approach assigns lower weights to older data while giving higher weights to more recent data, thereby better adapting to changes over time. Clearly, $\widehat{\theta}_t$ admits a closed-form solution $\widehat{\theta}_t = V_{t-1}^{-1}(\sum_{s=1}^{t-1} w_{t-1,s} r_s X_s)$,

---

Algorithm 1 LB-WeightUCB

Require: 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 $X_t = \arg\max_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \mathbf{x}, \widehat{\theta}_t \rangle + \beta_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}} \right\}$
4:      Receive the reward r_t
5:      Update $V_t = \gamma V_{t-1} + X_t X_t^\top + (1-\gamma)\lambda I_d$
6:      Compute $\widehat{\theta}_{t+1}$ by (3) and β_t by (5) with $w_{t,s} = \gamma^{t-s}$
7: **end for**

---

where $V_t = \lambda I_d + \sum_{s=1}^{t} w_{t,s} X_s X_s^\top, V_0 = \lambda I_d$ is the covariance matrix. Note that this closed-form solution can be further transformed into a recursive formula such that $V_t = \gamma V_{t-1} + X_t X_t^\top + (1-\gamma)\lambda I_d$ where we set $w_{t,s} = \gamma^{t-s}$. This allows it to be updated online without storing historical data, which is another important computational advantage of the weighted strategy over 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 [2, Appendix B.3, second and third steps in Proof of Theorem 2]. This difference is key to our algorithm's improved efficiency, as we discuss 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* $t \in [T]$

$$\left| \mathbf{x}^\top (\widehat{\theta}_t - \theta_t) \right| \tag{4}$$

$$\leq L^2 \sqrt{\frac{d}{\lambda}} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \|\theta_p - \theta_{p+1}\|_2 + \beta_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}},$$

where β_t *is the radius of the confidence region set by*

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

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

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

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 the algorithm proposed in [2] is

$$X_t = \arg\max_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \mathbf{x}, \widehat{\theta}_t \rangle + \beta_{t-1} \|\mathbf{x}\|_{V_{t-1}'^{-1} \widetilde{V}_{t-1} V_{t-1}'^{-1}} \right\},$$

where β_{t-1} is similar to those in our selection criterion (6), $V_{t-1}' = \lambda I_d + \sum_{s=1}^{t-1} \gamma^{t-s-1} X_s X_s^\top \in \mathbb{R}^{d \times d}$, and $\widetilde{V}_{t-1} = \lambda I_d + \sum_{s=1}^{t-1} \gamma^{2(t-s-1)} X_s X_s^\top \in \mathbb{R}^{d \times d}$ 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 [2] 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 III-C, we provide a sketch of the analysis framework for Lemma 1, and a more detailed discussion is presented in Appendix A-A. Furthermore, we prove that our algorithm enjoys the same (even slightly better in d) regret as the algorithm of [2].

Theorem 1. *Let the weighted factor* $w_{t,s} = \gamma^{t-s}$, *where* $\gamma \in (1/T, 1)$, *and let* $\lambda = d$, *the dynamic regret of LB-WeightUCB (Algorithm 1) is bounded with probability at least* $1 - 1/T$, *by*

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

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

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

Compared to previous works [1]–[3], 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 [2], who did not take the dimension into the tuning. Their algorithm and regret can also benefit from the refined tuning. The proof of Theorem 1 is in Appendix A-C.

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 [1] and achieves an $\widetilde{\mathcal{O}}(d^{3/4} P_T^{1/4} T^{3/4})$ bound. We provide the BOB version of LB-WeightUCB and detailed regret analysis in Appendix H. 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 [8], to further achieve an optimal dynamic regret without prior knowledge of P_T. This is very challenging since 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 [8] 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. However, 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. Additionally, we note that our approach can handle time-varying arm set settings, whereas MASTER relies on the fixed arm set assumption. It remains unclear whether optimal regret can be achieved under time-varying arm set.

C. Refined analysis framework

In this section, we present 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,

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

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

$$|\mathbf{x}^\top(\widehat{\theta}_t - \theta_t)| \le \|\mathbf{x}\|_{V_{t-1}^{-1}} (A_t + B_t), \tag{7}$$

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

Choosing an appropriate local norm for (7) is the key to simplifying and improving the estimation error analysis. Note that the previous analysis [2] 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,

$$|\mathbf{x}^\top(\widehat{\theta}_t - \theta_t)| \le \|\mathbf{x}\|_2 A_t' + \|\mathbf{x}\|_{V_{t-1}'^{-1}\widetilde{V}_{t-1}V_{t-1}'^{-1}} B_t', \tag{8}$$

where we have $A_t' = \|V_{t-1}'^{-1}\sum_{s=1}^{t-1}\gamma^{t-s-1} X_s X_s^\top (\theta_s - \theta_t)\|_2$, $B_t' = \|\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_s X_s - \lambda\theta_t\|_{\widetilde{V}_{t-1}^{-1}}$ and $V_t' = \lambda I_d + \sum_{s=1}^{t-1}\gamma^{t-s} X_s X_s^\top$, $\widetilde{V}_t = \lambda I_d + \sum_{s=1}^{t}\gamma^{2(t-s)} X_s X_s^\top$. Due to 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, they use the $V_{t-1}'^{-1}\widetilde{V}_{t-1}V_{t-1}'^{-1}$-norm to control $\mathbf{x}$ term so that B_t' term can be normed by $\widetilde{V}_{t-1}^{-1}$.

As an improvement, we 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 key step of bias part analysis is to extract the variations of underlying parameters as follows,

$$A_t \le L\sum_{p=1}^{t-1}\sum_{s=1}^{p} w_{t-1,s} \|X_s\|_{V_{t-1}^{-1}} \|\theta_p - \theta_{p+1}\|_2$$
$$\le L\sqrt{d}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}} \|\theta_p - \theta_{p+1}\|_2.$$

Based on that, we can obtain an upper bound for bias part related to the path length $P_T = \sum_{t=2}^{T}\|\theta_{t-1} - \theta_t\|_2$. A precise proof for the above argument can be found in Lemma 7.

Variance Part Analysis. The key lies in analyzing the following self-normalized term with weighted factor $w_{t-1,s}$,

$$B_t \le \left\|\sum_{s=1}^{t-1} w_{t-1,s}\eta_s X_s\right\|_{V_{t-1}^{-1}} + \sqrt{\lambda}S$$

$$= \left\|\sum_{s=1}^{t-1}\widetilde{\eta}_s\widetilde{X}_s\right\|_{V_{t-1}^{-1}} + \sqrt{\lambda}S.$$

Here, notice that $V_t = \lambda I_d + \sum_{s=1}^{t}\widetilde{X}_s\widetilde{X}_s^\top$, where we define $\widetilde{\eta}_s \triangleq \sqrt{w_{t-1,s}}\eta_s$ and $\widetilde{X}_s \triangleq \sqrt{w_{t-1,s}}X_s$. Notably, for all $t \in [T]$ and $s \in [t-1]$, it holds that $|w_{t-1,s}| \le 1$, which ensures that $\widetilde{\eta}_s$ remains R-sub-Gaussian, and a precise argument can be found in Lemma 8. Consequently, we can directly apply the self-normalized concentration (Theorem 7) to control the variance term, without requiring the weighted version of the self-normalized concentration proposed in Theorem 1 of [2].

Combining the analysis for bias and variance parts, we can finish the proof of Lemma 1. $\qquad\square$

The estimation error analysis for weighted strategies involves first decomposing the estimation error into bias and variance parts, then analyzing them separately. [2] used different local norms to decompose estimation errors, mimicking sliding window analysis for the bias term and specifically designing a weighted version of self-normalized concentration for the variance term. Our refined analysis framework shows that such complexity is unnecessary. Bias and variance can be decomposed using the same local norm, with a dedicated bias analysis for the weighted strategy, and the variance term no longer requires specially designed concentrations or additional local norms. With the estimation error bound, we proceed to the regret analysis, where we need to use a weighted potential lemma to bound the regret.

Weighted Potential Lemma. Term $\|\mathbf{x}\|_{V_{t-1}^{-1}}$ in (7) induces a summation term $\sum_{t=1}^{T}\|X_t\|_{V_{t-1}^{-1}}$ in the variance part of regret analysis. Since $V_{t-1} = \lambda I_d + \sum_{s=1}^{t-1} w_{t-1,s} X_s X_s^\top$ incorporates the weighted factors, we cannot directly apply the standard potential lemma. Instead, we need to use a weighted potential lemma (see Lemma 9) for regret analysis with weighted factor $w_{t,s} = \gamma^{t-s}$, such that

$$\sum_{t=1}^{T}\|X_t\|_{V_{t-1}^{-1}} = \widetilde{\mathcal{O}}\left(T\sqrt{d\log\frac{1}{\gamma}}\right). \tag{9}$$

The next is to choose the discounted factor γ appropriately, so that the bias and variance terms in the regret bound are well balanced. A smaller γ corresponds to faster forgetting, which helps reduce the bias caused by non-stationarity. However, a smaller γ will also increase the variance by the key term $\mathcal{O}(\sqrt{\log\frac{1}{\gamma}})$ shown in (9). For details on how to optimally select γ to balance these two parts, please refer to Appendix A-C.

To summarize, for non-stationary LB analysis, the weighted strategy is as simple as the restarted or sliding-window strategies, *with only the difference being the requirement of the weighted potential lemma* for regret analysis, without the need for more complicated deviation results.

Remark 1. The key step (7) 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 (8), [6] has 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.

IV. GENERALIZED LINEAR BANDIT

In this section, we apply the weighted strategy to drifting GLB. Compared to the best-known weight-based algorithm for drifting GLB [6], our algorithm is simpler and meanwhile has a better dynamic regret. Additionally, we consider a key class of GLB, known as the Self-Concordant Bandit (SCB), and further improve the theoretical guarantees in this setting.

A. 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:

$$\text{D-REG}_T = \sum_{t=1}^{T} \left(\max_{\mathbf{x} \in \mathcal{X}} \mu(\mathbf{x}^\top \theta_t) - \mu(X_t^\top \theta_t) \right). \quad (10)$$

Under GLB, 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 [6], [41], [45].

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

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

Previous works [3], [40] 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, $\widetilde{c}_\mu$ 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 regret upper bound.

B. Algorithm and Regret Guarantee

We propose GLB-WeightUCB, which is a simpler algorithm with better theoretical guarantee compared to previous weight-based algorithm [6]. 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 equation:

$$\lambda c_\mu \theta + \sum_{s=1}^{t-1} w_{t-1,s} \left(\mu(X_s^\top \theta) - r_s \right) X_s = 0. \quad (11)$$

Similar to the Estimator (3), we set $w_{t,s} = \gamma^{t-s}$, where $\gamma \in (0,1)$ is the discounted factor. 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

$$\widetilde{\theta}_t = \arg\min_{\theta \in \Theta} \|g_t(\widehat{\theta}_t) - g_t(\theta)\|_{V_{t-1}^{-1}}, \quad (12)$$

Algorithm 2 GLB-WeightUCB

Require: time horizon T, discounted factor γ, confidence δ, regularizer λ, link function μ, parameters S, L and R
1: Set $V_0 = \lambda I_d$, $\widehat{\theta}_1 = \mathbf{0}$, compute k_μ, c_μ and $\bar{\beta}_0$ by (14)
2: **for** $t = 1, 2, ..., T$ **do**
3:    **if** $\|\widehat{\theta}_t\|_2 \leq S$ **then**
4:       let $\theta_t = \widehat{\theta}_t$
5:    **else**
6:       Do the projection and get $\widetilde{\theta}_t$ by (12)
7:    **end if**
8:    Select X_t by (15)
9:    Receive the reward r_t
10:   Update $V_t = \gamma V_{t-1} + X_t X_t^\top + (1-\gamma)\lambda I_d$
11:   Compute $\widehat{\theta}_{t+1}$ according to (11) with $w_{t,s} = \gamma^{t-s}$
12:   Compute $\bar{\beta}_t$ by (14) with $w_{t,s} = \gamma^{t-s}$
13: **end for**

---

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

$$g_t(\theta) \triangleq \lambda c_\mu \theta + \sum_{s=1}^{t-1} w_{t-1,s} \mu(X_s^\top \theta) X_s. \quad (13)$$

However, previous work [6] 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 [6, Section 4.1], and our restatements in Appendix B-A).

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 (11) with projection (12), 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 $t \in [T]$

$$\left| \mu(\mathbf{x}^\top \widetilde{\theta}_t) - \mu(\mathbf{x}^\top \theta_t) \right|$$

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

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

$$\sqrt{\lambda} c_\mu S + R \sqrt{2 \log \frac{1}{\delta} + d \log \left(1 + \frac{L^2 \sum_{s=1}^{t} w_{t,s}}{\lambda d} \right)}. \quad (14)$$

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

$$X_t = \arg\max_{\mathbf{x} \in \mathcal{X}} \left\{ \mu(\mathbf{x}^\top \widetilde{\theta}_t) + \frac{2k_\mu}{c_\mu} \bar{\beta}_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}} \right\}. \quad (15)$$

The overall algorithm is summarized in Algorithm 2.

Notice that the estimation equation (11) and the confidence radius (14) are the same as those used in Algorithm 1 of [6].

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. *Let the weighted factor $w_{t,s} = \gamma^{t-s}$, where $\gamma \in (1/T, 1)$, and let $\lambda = d/c_\mu^2$, the regret of GLB-WeightUCB (Algorithm 2) is bounded with probability at least $1 - 1/T$ by*

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

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

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

Compared to BVD-GLM-UCB (the best-known weight-based algorithm for drifting GLB) [6], 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 $c_\mu^{-3/4}$.

C. Self-Concordant Bandits

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

$$\mathbb{E}[r_t \mid X_t] = \mu(X_t^\top \theta_t), \text{Var}[r_t \mid X_t] = \mu'(X_t^\top \theta_t). \quad (16)$$

SCB requires the link function satisfy $|\mu''| \leq \mu'$, usually referred to as 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 [46]–[48].

We make several standard assumptions same as LB and GLB, including boundedness of feasible set and unknown parameters (Assumption 1), and non-linearity measure on the (inverse) link function (Assumption 2). In addition, similar to [7], 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 $m > 0$.

Algorithm. 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}. At iteration t, we first adopt the same maximum likelihood estimator as GLB which is defined in (11). Different from GLB, here we use a new local norm to perform the projection onto the feasible set Θ,

$$\widetilde{\theta}_t = \arg\min_{\theta \in \Theta} \left\| g_t(\widehat{\theta}_t) - g_t(\theta) \right\|_{H_t^{-1}(\theta)}, \quad (17)$$

where $g_t(\theta)$ is the same as (13) while $H_t(\theta)$ is defined as

$$H_t(\theta) \triangleq \lambda c_\mu I_d + \sum_{s=1}^{t-1} w_{t-1,s} \mu'(X_s^\top \theta) X_s X_s^\top. \quad (18)$$

Notably, compared to V_t, $H_t(\theta)$ depends on the function curvature along the dynamics and thus captures more *local* information. Combining this projection step with the standard self-normalized concentration restated in Theorem 8 removes a constant $c_\mu^{-1/2}$ in the regret. For estimator (11) with projection (17), we construct the 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 $t \in [T]$*

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

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

$$\widetilde{\beta}_t = \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}} \left(\log \frac{1}{\delta} + d \log 2 \right)$$
$$+ \frac{dm}{\sqrt{\lambda c_\mu}} \log \left(1 + \frac{L^2 k_\mu \sum_{s=1}^t w_{t,s}}{\lambda c_\mu d} \right) + \sqrt{\lambda c_\mu} S. \quad (19)$$

The proof of Lemma 3 is in Appendix C-A. Based on Lemma 3, we can select the arm X_t as

$$\arg\max_{\mathbf{x} \in \mathcal{X}} \left\{ \mu(\mathbf{x}^\top \widetilde{\theta}_t) + 2\sqrt{1 + 2S} \frac{k_\mu}{\sqrt{c_\mu}} \widetilde{\beta}_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}} \right\}. \quad (20)$$

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 (17), $\widetilde{\beta}_{t-1}$ is computed by (19) and X_t is computed by (20). 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*

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

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

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

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. In near-stationary environments (P_T is small enough), our result can recover to the performance of LogUCB1 algorithm [33]. Proof of Theorem 3 is presented in Appendix C-B.

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 [7].

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*

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

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

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

The overall algorithm and analysis are in Appendix D.

V. Linear Mixture MDP

In this section, we apply the weighted strategy to non-stationary linear mixture MDPs, and describe our WeightUCRL algorithm and its theoretical guarantee. Our algorithm achieves the same regret bound as the previous restart-based algorithm [18].

A. Problem Setting

We build upon the previously established definition of episodic non-stationary MDPs [18] and provide the corresponding learning protocol.

Episodic Non-stationary MDPs. An episodic MDP is defined by a tuple $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$, where $\mathcal{S}$ is the state space; $\mathcal{A}$ is the action space; H is the length of each episode; $\mathbb{P} = \{\mathbb{P}_h^k\}_{h\in[H], k\in[K]}$, where $\mathbb{P}_h^k : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ is the transition probability at h-th step of k-th episode; and $r = \{r_h^k\}_{h\in[H], k\in[K]}$, where $r_h^k : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the reward function at h-th step of k-th episode. A policy is defined as $\pi = \{\pi_h^k\}_{h\in[H], k\in[K]}$, where each $\pi_h^k : \mathcal{S} \to \Delta(\mathcal{A})$ is a function that maps a state s to distributions over action space $\mathcal{A}$ at stage h of the k-th episode.

Learning Protocol. At the beginning of each episode k, the learner chooses a policy $\pi^k = \{\pi_h^k\}_{h=1}^H$. At each stage $h \in [H]$, starting from the initial stage s_1^k, the learner observes the state s_h^k, chooses an action a_h^k sampled from $\pi_h^k(s_h^k)$, obtains reward $r_h^k(s_h^k, a_h^k)$ and transitions to the next state $s_{h+1}^k \sim \mathbb{P}_h^k(\cdot \mid s_h^k, a_h^k)$ for $h \in [H]$. The episode ends when s_{H+1}^k is reached; when this happens, no action is taken, and the reward is equal to zero. For any policy $\pi = \{\pi_h^k\}_{h\in[H], k\in[K]}$ and

$(s,a) \in \mathcal{S} \times \mathcal{A}$, we define the action-value function $Q_h^{k,\pi}$ and value function $V_h^{k,\pi}$ as

$$Q_h^{k,\pi}(s,a) = \mathbb{E}\left[\sum_{h'=h}^H r_{h'}^k\left(s_{h'}^k, \pi_{h'}^k\left(s_{h'}^k\right)\right) \,\middle|\, s_h^k = s, a_h^k = a \right]$$

$$V_h^{k,\pi}(s) = \mathbb{E}_{a\sim\pi_h^k(\cdot \mid s)}\left[Q_h^{k,\pi}(s,a) \right].$$

We define $\forall V : \mathcal{S} \to \mathbb{R}$, $\left[\mathbb{P}_h^k V\right](s,a) = \mathbb{E}_{s'\sim\mathbb{P}_h^k(\cdot \mid s,a)}V(s')$, and the Bellman equation for policy π is given by

$$Q_h^{k,\pi}(s,a) = r_h^k(s,a) + \left[\mathbb{P}_h^k V_{h+1}^{k,\pi}\right](s,a)$$

$$V_h^{k,\pi}(s) = \mathbb{E}_{a\sim\pi_h^k(\cdot \mid s)}\left[Q_h^{k,\pi}(s,a) \right], \quad V_{H+1}^{k,\pi} = 0.$$

The learner's goal is to minimize the following dynamic regret,

$$\text{D-Reg}_T = \sum_{k=1}^K V_1^{k,\pi_*^k}\left(s_1^k\right) - \sum_{k=1}^K V_1^{k,\pi^k}\left(s_1^k\right), \quad (21)$$

where we denote $T \triangleq H \cdot K$, for consistency with the bandits notations. The dynamic regret measures the difference between the learner's policy and the optimal policy, namely, $\pi_*^k = \arg\max_\pi V_1^{k,\pi}(s_1^k)$.

Non-stationary Linear Mixture MDP. An MDP instance $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$ is a linear mixture MDP if there exist known feature maps $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and $\psi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ and for any $k \in [K], h \in [H]$, there exist unknown vectors $\theta_h^k \in \mathbb{R}^d$ and $\mathbf{w}_h^k \in \mathbb{R}^d$ such that

$$r_h^k(s,a) = \langle \phi(s,a), \theta_h^k \rangle$$
$$\mathbb{P}_h^k(s' \mid s,a) = \langle \psi(s' \mid s,a), \mathbf{w}_h^k \rangle, \quad (22)$$

here we consider the drifting case where we use path length $P_T^\theta = \sum_{k=2}^K \sum_{h=1}^H \left\| \theta_h^{k-1} - \theta_h^k \right\|_2$ to measure the non-stationarity of θ_h^k and $P_T^{\mathbf{w}} = \sum_{k=2}^K \sum_{h=1}^H \left\| \mathbf{w}_h^{k-1} - \mathbf{w}_h^k \right\|_2$ to measure the non-stationarity of $\mathbf{w}_h^k$, and we define $\Delta = P_T^\theta + P_T^{\mathbf{w}}$ as the total path length. We work under the following standard boundedness assumption [18].

Assumption 4. The feasible set and unknown parameters are assumed to be bounded: $\forall s \in \mathcal{S}, a \in \mathcal{A}, \|\phi(s,a)\| \leq L_\phi$; for any bounded function $V : \mathcal{S} \to [0,1]$ and $\forall s \in \mathcal{S}, a \in \mathcal{A}, \|\psi_V(s,a)\|_2 \leq L_\psi$, where $\psi_V(s,a) \triangleq \sum_{s'\in\mathcal{S}} \psi(s'|s,a)V(s')$; $\theta_h^k \in \Theta$ holds for all $k \in [K], h \in [H]$ where $\Theta \triangleq \{\theta \mid \|\theta\|_2 \leq S_\theta\}$ and $\mathbf{w}_h^k \in \mathcal{W}$ holds for all $k \in [K], h \in [H]$ where $\mathcal{W} \triangleq \{\mathbf{w} \mid \|\mathbf{w}\|_2 \leq S_{\mathbf{w}}\}$.

B. Algorithm and Regret Guarantee

We propose the WeightUCRL algorithm in this section. We first give the employed reward estimator and transition 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.

Reward Estimator. At h-th stage of k-th episode, we use the following weighted least square estimator $\widehat{\theta}_h^k$ to estimate the unknown parameter θ_h^k, which is the minimizer of

$$\frac{\lambda_\theta}{2}\|\theta\|_2^2 + \sum_{j=1}^{k-1} w_{k-1,j}\left(r_h^j(s_h^j,a_h^j) - \phi(s_h^j,a_h^j)^\top\theta\right)^2,$$

where $\lambda_\theta > 0$ is the regularization coefficient and $w_{k-1,j}$ is the weighted factor. Similar to the Estimator (3), we set $w_{k,j} = \gamma^{k-j}$, where $\gamma \in (0,1)$ is the discounted factor. Clearly, $\widehat{\theta}_h^k$ has a closed-form solution:

$$\widehat{\theta}_h^k = \left(\Lambda_h^{k-1}\right)^{-1}\left(\sum_{j=1}^{k-1} w_{k-1,j} r_h^j(s_h^j,a_h^j)\phi(s_h^j,a_h^j)\right), \quad (23)$$

where $\Lambda_h^k = \lambda_\theta I_d + \sum_{j=1}^k w_{k,j}\phi(s_h^j,a_h^j)\phi(s_h^j,a_h^j)^\top$.

Transition Estimator. First notice that, based on model assumption (22), we have

$$\left[\mathbb{P}_h^k V_{h+1}^k\right](s_h^k,a_h^k) = \sum_{s'}\left\langle\psi(s'\mid s_h^k,a_h^k),\mathbf{w}_h^k\right\rangle V_{h+1}^k(s')$$
$$= \left\langle\sum_{s'}\psi(s'\mid s_h^k,a_h^k)V_{h+1}^k(s'),\mathbf{w}_h^k\right\rangle$$
$$= \left\langle\psi_{h+1}^k(s_h^k,a_h^k),\mathbf{w}_h^k\right\rangle, \quad (24)$$

where we denote $\psi_{h+1}^k(s,a) \triangleq \psi_{V_{h+1}^k}(s,a)$ for simplicity. Based on Assumption 4, we know that $\forall s \in \mathcal{S}, a \in \mathcal{A}, \left\|\psi_{h+1}^k(s,a)\right\|_2 \leq HL_\psi$. We adopt the following weighted least square estimator $\widehat{\mathbf{w}}_h^k$ to estimate the unknown parameter $\mathbf{w}_h^k$, which is the minimizer of

$$\sum_{j=1}^{k-1}\alpha_{k-1,j}\left(\langle\psi_{h+1}^j(s_h^j,a_h^j),\mathbf{w}\rangle - V_{h+1}^j(s_{h+1}^j)\right)^2 + \frac{\lambda_\mathbf{w}}{2}\|\mathbf{w}\|_2^2,$$

where $\lambda_\mathbf{w} > 0$ is the regularization coefficient and $\alpha_{k-1,j}$ is the weighted factor, we set $\alpha_{k,j} = \gamma^{k-j}$, where $\gamma \in (0,1)$ is the discounted factor. For this least square estimator, we have a closed-form solution for $\widehat{\mathbf{w}}_h^k$:

$$\left(\Sigma_h^{k-1}\right)^{-1}\left(\sum_{j=1}^{k-1}\alpha_{k-1,j}V_{h+1}^j(s_{h+1}^j)\psi_{h+1}^j(s_h^j,a_h^j)\right), \quad (25)$$

where $\Sigma_h^k = \lambda_\mathbf{w} I_d + \sum_{j=1}^k \alpha_{k,j}\psi_{h+1}^j(s_h^j,a_h^j)\psi_{h+1}^j(s_h^j,a_h^j)^\top$.

Upper Confidence Bounds. For estimator (23) and (25), we provide the estimation error bounds, respectively.

Lemma 4. *For any $s \in \mathcal{S}, a \in \mathcal{A}$, the following holds for all $k \in [K], h \in [H]$,*

$$\left|\phi(s,a)^\top\left(\widehat{\theta}_h^k - \theta_h^k\right)\right| \leq \Gamma_{h,\theta}^{k-1} + \beta_\theta\|\phi(s,a)\|_{(\Lambda_h^{k-1})^{-1}},$$

where $\Gamma_{h,\theta}^{k-1} \triangleq L_\phi^2\sqrt{\frac{d}{\lambda_\theta}}\sum_{p=1}^{k-1}\sqrt{\sum_{j=1}^p w_{k-1,j}}\left\|\theta_h^p - \theta_h^{p+1}\right\|_2,$ $\beta_\theta \triangleq \sqrt{\lambda_\theta}S_\theta.$

Lemma 5. *For any $s \in \mathcal{S}, a \in \mathcal{A}$, and $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds for all $k \in [K], h \in [H]$*

$$\left|\psi_{h+1}^k(s,a)^\top\left(\widehat{\mathbf{w}}_h^k - \mathbf{w}_h^k\right)\right|$$

Algorithm 3 WeightUCRL

Require: episode number K, time horizon H, discounted factor γ, confidence δ, regularizer $\lambda_\theta, \lambda_\mathbf{w}$, parameters $S_\theta, S_\mathbf{w}, L_\phi, L_\psi$
1: Initialize $\left\{\pi_h^0\right\}_{h=1}^H$ as uniform distribution policies, $\left\{Q_h^0\right\}_{h=1}^H$ as zero functions.
2: Set $\forall h \in [H], \Lambda_h^0 = \lambda_\theta I_d, \Sigma_h^0 = \lambda_\mathbf{w} I_d$ for $k = 1$
3: **for** $k = 1, 2, ..., K$ **do**
4:     Receive the initial state s_1^k
5:     Initialize V_{H+1}^k as zero function
6:     **for** $h = H, H-1, ..., 1$ **do**
7:         $\psi_{h+1}^k(\cdot,\cdot) = \sum_{s'}\psi(s'\mid\cdot,\cdot)V_{h+1}^k(s')$
8:         Update $\Lambda_h^k = \gamma\Lambda_h^{k-1} + \phi(s_h^k,a_h^k)\phi(s_h^k,a_h^k)^\top + (1-\gamma)\lambda_\theta I_d$
9:         Update $\Sigma_h^k = \psi_{h+1}^k(s_h^k,a_h^k)\psi_{h+1}^k(s_h^k,a_h^k)^\top + \gamma\Sigma_h^{k-1} + (1-\gamma)\lambda_\mathbf{w} I_d$
10:         Compute $\widehat{\theta}_h^k$ by (23) and $\widehat{\mathbf{w}}_h^k$ by (25)
11:         Compute optimistic value function $Q_h^k(s,a)$ and $V_h^k(s)$ by (26)
12:     **end for**
13:     **for** $h = 1, ..., H$ **do**
14:         Choose policy as $\pi_h^k(s) = \arg\max_{a\in\mathcal{A}} Q_h^k(s,a)$
15:         Take action $a_h^k \sim \pi_h^k(s_h^k)$, then observe the reward $r_h^k(s_h^k,a_h^k)$ and receive the next state s_{h+1}^k
16:     **end for**
17: **end for**

$$\leq \Gamma_{h,\mathbf{w}}^{k-1} + \beta_\mathbf{w}^{k-1}\left\|\psi_{h+1}^k(s,a)\right\|_{(\Sigma_h^{k-1})^{-1}},$$

$\Gamma_{h,\mathbf{w}}^{k-1} \triangleq H^2 L_\psi^2\sqrt{\frac{d}{\lambda_\mathbf{w}}}\sum_{p=1}^{k-1}\sqrt{\sum_{j=1}^p\alpha_{k-1,j}}\left\|\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right\|_2,$ *and $\beta_\mathbf{w}^k$ is the radius of confidence region defined by*

$$H\sqrt{\frac{1}{2}\log\frac{1}{\delta} + \frac{d}{4}\log\left(1 + \frac{H^2 L_\psi^2\sum_{j=1}^k\alpha_{k,j}}{\lambda_\mathbf{w} d}\right)} + \sqrt{\lambda_\mathbf{w}}S_\mathbf{w}.$$

Proof of Lemma 4 and Lemma 5 are in Appendix E-A, E-B.

Arm Selection. Based on Lemma 4 and Lemma 5, we define the optimistic value function $Q_h^k(s,a)$ and $V_h^k(s)$ as follows,

$$Q_h^k(s,a) \triangleq \min\left\{H, \phi(s,a)^\top\widehat{\theta}_h^k + \beta_\theta\|\phi(s,a)\|_{(\Lambda_h^{k-1})^{-1}}\right.$$
$$\left. + \psi_{h+1}^k(s,a)^\top\widehat{\mathbf{w}}_h^k + \beta_\mathbf{w}^{k-1}\left\|\psi_{h+1}^k(s,a)\right\|_{(\Sigma_h^{k-1})^{-1}}\right\}$$
$$V_h^k(s) \triangleq \max_{a\in\mathcal{A}} Q_h^k(s,a) = \mathbb{E}_{a\sim\pi_h^k(\cdot\mid s)}\left[Q_h^k(s,a)\right]. \quad (26)$$

At state s_h^k, we can specify the action selection criterion of our policy $\pi_h^k(s_h^k)$ as $a_h^k = \arg\max_{a\in\mathcal{A}} Q_h^k(s_h^k,a)$. The overall algorithm is summarized in Algorithm 3. We show that our algorithm enjoys the following regret guarantee.

Theorem 5. *Let $T = KH$, $\delta = 1/(4T)$, $\lambda_\theta = d$, and $\lambda_\mathbf{w} = H^2 d$, $\forall k, j \in [K], w_{k,j} = \alpha_{k,j} = \gamma^{k-j}, \gamma \in (1/K, 1)$, the*

dynamic regret D-REG$_T$ *is bounded with probability at least* $1 - 1/T$ *by*

$$\mathcal{O}\left(Hd\left(\frac{1}{(1-\gamma)^{3/2}}\Delta + HK\sqrt{1-\gamma}\right) + H^{3/2}d\sqrt{HK}\right),$$

Furthermore, by setting the discounted factor optimally as $\gamma = 1 - \max\left\{1/K, \sqrt{\Delta/T}\right\}$, *we have*

$$\text{D-REG}_T \leq \begin{cases} \widetilde{\mathcal{O}}\left(Hd\Delta^{1/4}T^{3/4}\right) & \text{when } \Delta \geq H/K, \\ \widetilde{\mathcal{O}}\left(dH^{3/2}\sqrt{T}\right) & \text{when } \Delta < H/K. \end{cases}$$

The proof of Theorem 5 is in Appendix E-C. Compared to previous work [18], our results achieve the same order regret guarantees for dynamic regret in non-stationary environments. Furthermore, in near-stationary settings, our results recover the theoretical guarantees established for stationary environments [43], [44].

VI. MULTINOMIAL LOGIT MIXTURE MDP

In this section, we explore another class of MDPs, known as the Multinomial Logit (MNL) Mixture MDP, under the non-stationary setting. We introduce the MNL-WeightUCRL algorithm, which applies the weighted strategy to non-stationary MNL Mixture MDPs, and provide the first theoretical guarantee for non-stationary MNL Mixture MDP.

A. Problem Setting

To address the limitation that linear function approximation cannot guarantee valid distribution, a new class called MNL mixture MDPs has been proposed recently [16], [17]. The MNL mixture MDPs share the same episodic non-stationary MDPs structure and learning protocol as the linear mixture MDP, with the objective of minimizing dynamic regret (21). The key distinction lies in its modeling assumptions for the transition probabilities. Below, we present the formal definition of MNL Mixture MDPs.

Definition 1 (Reachable States). *For any* $(k, h, s, a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$, *we define the "reachable states" as the set of states that can be reached from state* s *taking action* a *at stage* h *of* k-*th episode within a single transition, i.e.,* $\mathcal{S}_h^k(s, a) \triangleq \{s' \in \mathcal{S} \mid \mathbb{P}_h^k(s' \mid s, a) > 0\}$. *Also, we define* $S_h^k(s, a) \triangleq \left|\mathcal{S}_h^k(s, a)\right|$ *and further define* $U \triangleq \max_{(k,h,s,a)} S_h^k(s, a)$ *as the maximum number of reachable states.*

Non-stationary MNL Mixture MDP. $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$ is a MNL mixture MDP if there exist known feature maps $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and $\psi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ and for any $k \in [K], h \in [H]$, there exist unknown vectors $\theta_h^k \in \mathbb{R}^d$ and $\mathbf{w}_h^k \in \mathbb{R}^d$ such that

$$r_h^k(s, a) = \langle \phi(s, a), \theta_h^k \rangle$$
$$\mathbb{P}_h^k(s' \mid s, a) = \frac{\exp\left(\psi(s' \mid s, a)^\top \mathbf{w}_h^k\right)}{\sum_{\widetilde{s} \in \mathcal{S}_h^k(s,a)} \exp\left(\psi(\widetilde{s} \mid s, a)^\top \mathbf{w}_h^k\right)}. \quad (27)$$

We work under the following standard assumptions of MNL Mixture MDP [16], [17].

Assumption 5. The feasible set and unknown parameters are assumed to be bounded: $\forall s \in \mathcal{S}, a \in \mathcal{A}, \|\phi(s, a)\| \leq L_\phi$; $\forall s's \in \mathcal{S}, a \in \mathcal{A}, \|\psi(s' \mid s, a)\| \leq L_\psi$; $\theta_h^k \in \Theta$ holds for all $k \in [K], h \in [H]$ where $\Theta \triangleq \{\theta \mid \|\theta\|_2 \leq S_\theta\}$ and $\mathbf{w}_h^k \in \mathcal{W}$ holds for all $k \in [K], h \in [H]$ where $\mathcal{W} \triangleq \{\mathbf{w} \mid \|\mathbf{w}\|_2 \leq S_\mathbf{w}\}$.

Assumption 6. There exists $0 < \kappa < 1$ such that for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and $s', s'' \in \mathcal{S}_h^k(s, a)$, it holds that $\inf_{\mathbf{w} \in \mathcal{W}} p_{s,a}^{s'}(\mathbf{w}) p_{s,a}^{s''}(\mathbf{w}) \geq \kappa$.

B. Algorithm and Regret Guarantee

We propose the MNL-WeightUCRL algorithm, which obtains the first dynamic regret guarantee for non-stationary MNL Mixture MDPs. We begin by presenting the estimator used in our approach and deriving its estimation error upper bound by our refined analysis framework, which is the key for algorithm design and regret analysis. Building on the estimation error bound, we propose our selection criterion and finally give the theoretical guarantee on its dynamic regret.

Reward Estimator. Since we use the same linear function as (22) to model the reward function, here we still use estimator (23) to estimate the unknown parameter θ_h^k.

Transition Estimator. For the trajectory $\{(s_h^k, a_h^k)\}_{h=1}^H$ at episode k, we define the variable: $y_h^k \in \{0, 1\}^{S_h^k}$, where $y_h^k(s') = \mathbb{1}(s_{h+1}^k = s')$ for $s' \in \mathcal{S}_h^k \triangleq \mathcal{S}_h^k(s_h^k, a_h^k)$ and $S_h^k \triangleq |\mathcal{S}_h^k|$. Furthermore, we denote $p_h^k(\psi(s' \mid s, a)^\top \mathbf{w}) = \frac{\exp(\psi(s' \mid s,a)^\top \mathbf{w})}{\sum_{\widetilde{s} \in \mathcal{S}_h^k(s,a)} \exp(\psi(\widetilde{s} \mid s,a)^\top \mathbf{w})}$. We define $\bar{\psi}_h^k(s') \triangleq \psi(s' \mid s_h^k, a_h^k)$. Then y_h^k is a sample from the multinomial distribution:

$$\text{multinomial}\left(1, \left[p_h^k(\bar{\psi}_h^k(s_1)^\top \mathbf{w}), \ldots, p_h^k(\bar{\psi}_h^k(s_{N_h^k})^\top \mathbf{w})\right]\right).$$

We use the following weighted maximum likelihood estimation (MLE) $\widehat{\mathbf{w}}_h^k$ to estimate the unknown parameter $\mathbf{w}_h^k$, which is the minimizer of

$$\frac{\lambda_\mathbf{w}\kappa}{2}\|\mathbf{w}\|_2^2 + \sum_{j=1}^{k-1} \alpha_{k-1,j} \sum_{s' \in \mathcal{S}_h^j} -y_h^j(s') \log p_h^k(\bar{\psi}_h^j(s')^\top \mathbf{w}),$$

where $\lambda_\mathbf{w} > 0$ is the regularization coefficient and $\alpha_{k-1,j}$ is the weighted factor. We set $\alpha_{k,j} = \gamma^{k-j}$, where $\gamma \in (0, 1)$ is the discounted factor. Specifically the estimator $\widehat{\mathbf{w}}_h^k$ is the solution of the following equation:

$$\sum_{j=1}^{k-1} \alpha_{k-1,j} \sum_{s' \in \mathcal{S}_h^j} \left(p_h^j(\bar{\psi}_h^j(s')^\top \mathbf{w}) - y_h^j(s')\right) \bar{\psi}_h^j(s')$$
$$+ \lambda_\mathbf{w}\kappa \mathbf{w} = 0. \quad (28)$$

Given that $\widehat{\mathbf{w}}_h^k$ may not belong to the feasible set $\mathcal{W}$ and κ is defined over the parameter $\mathbf{w} \in \mathcal{W}$, we need to perform the following projection step

$$\widetilde{\mathbf{w}}_h^k = \arg\min_{\mathbf{w} \in \mathcal{W}} \left\|g_h^k(\widehat{\mathbf{w}}_h^k) - g_h^k(\mathbf{w})\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}, \quad (29)$$

Algorithm 4 MNL-WeightUCRL

Require: episode number K, time horizon H, discounted factor γ, confidence δ, regularizer $\lambda_\theta, \lambda_{\mathbf{w}}$, parameters $S_\theta, S_{\mathbf{w}}, L_\phi, L_\psi$
1: Initialize $\{\pi_h^0\}_{h=1}^H$ as uniform distribution policies, $\{\bar{Q}_h^0\}_{h=1}^H$ as zero functions.
2: **for** $k = 1, 2, ..., K$ **do**
3:    Receive the initial state s_1^k
4:    Initialize $\bar{V}_{H+1}^k$ as zero function
5:    Set $\forall h \in [H], \Lambda_h^0 = \lambda_\theta I_d, \bar{\Sigma}_h^0 = \lambda_{\mathbf{w}} I_d$
6:    **for** $h = H, H-1, ..., 1$ **do**
7:       Update $\Lambda_h^k = \gamma \Lambda_h^{k-1} + \phi(s_h^k, a_h^k)\phi(s_h^k, a_h^k)^\top + (1 - \gamma)\lambda_\theta I_d$
8:       Update $\bar{\Sigma}_h^k = \sum_{s' \in \mathcal{S}_h^k} \bar{\psi}_h^j(s'|s_h^k, a_h^k)\bar{\psi}_h^j(s'|s_h^k, a_h^k)^\top + \gamma \bar{\Sigma}_h^{k-1} + (1 - \gamma)\lambda_{\mathbf{w}} I_d$
9:       Compute $\widehat{\theta}_h^k$ by (23) and $\widehat{\mathbf{w}}_h^k$ by (28)
10:      **if** $\|\widehat{\mathbf{w}}_h^k\|_2 \leq S_{\mathbf{w}}$ **then**
11:        let $\widetilde{\mathbf{w}}_h^k = \widehat{\mathbf{w}}_h^k$
12:      **else**
13:        Do the projection and get $\widetilde{\mathbf{w}}_h^k$ by (29)
14:      **end if**
15:      Compute optimistic value function $\bar{Q}_h^k(s, a)$ and $\bar{V}_h^k(s)$ by (30)
16:    **end for**
17:    **for** $h = 1, ..., H$ **do**
18:       Choose policy as $\pi_h^k(s) = \arg\max_{a \in \mathcal{A}} \bar{Q}_h^k(s, a)$
19:       Take action $a_h^k \sim \pi_h^k(s_h^k)$, then observe the reward $r_h^k(s_h^k, a_h^k)$ and receive the next state s_{h+1}^k
20:    **end for**
21: **end for**

where $\bar{\Sigma}_h^k = \lambda_{\mathbf{w}} I_d + \sum_{j=1}^k \alpha_{k,j} \sum_{s' \in \mathcal{S}_h^j} \bar{\psi}_h^j(s')\bar{\psi}_h^j(s')^\top$ and $g_h^k(\mathbf{w})$ is defined as

$$g_h^k(\mathbf{w}) \triangleq \lambda_{\mathbf{w}} \kappa \mathbf{w} + \sum_{j=1}^{k-1} \alpha_{k-1,j} \sum_{s' \in \mathcal{S}_h^j} p_h^j(\bar{\psi}_h^j(s')^\top \mathbf{w})\bar{\psi}_h^j(s').$$

Upper Confidence Bounds. Estimator (23) can directly apply Lemma 4 for UCB construction. And for estimator (28), we provide the following estimation error bounds. For simplicity, we define for any function $V : \mathcal{S} \to \mathbb{R}$, $[\mathbb{P}_h^k V](s, a) = \sum_{s' \in \mathcal{S}_h^k} p_h^k(\psi(s' \mid s, a)^\top \mathbf{w}_h^k)V(s')$, $[\widetilde{\mathbb{P}}_h^k V](s, a) = \sum_{s' \in \mathcal{S}_h^k} p_h^k(\psi(s' \mid s, a)^\top \widetilde{\mathbf{w}}_h^k)V(s')$.

Lemma 6. *For any $\mathbf{x} \in \mathcal{X}$, and $\delta \in (0, 1)$, $\forall k, j \in [K], \alpha_{k,j} \leq 1$, with probability at least $1 - \delta$, the following holds for all $k \in [K], h \in [H]$*

$$\left| \left[\widetilde{\mathbb{P}}_h^k V\right](s, a) - \left[\mathbb{P}_h^k V\right](s, a) \right|$$
$$\leq \frac{H}{\kappa} \left(\Gamma_{h,\mathbf{w}}^{k-1} + \bar{\beta}_{\mathbf{w}}^{k-1} \max_{s' \in \mathcal{S}_h^k} \|\psi(s' \mid s, a)\|_{(\bar{\Sigma}_h^{k-1})^{-1}} \right),$$

$\Gamma_{h,\mathbf{w}}^{k-1} \triangleq L_\psi^2 \sqrt{\frac{d}{\lambda_{\mathbf{w}}}} \sum_{p=1}^{k-1} \sqrt{\sum_{j=1}^p \alpha_{k-1,j}} \left\| \mathbf{w}_h^p - \mathbf{w}_h^{p+1} \right\|_2$, *and*

$\bar{\beta}_{\mathbf{w}}^k$ *is the radius of confidence region defined by*

$$\sqrt{\frac{1}{2}\log\frac{1}{\delta} + \frac{d}{4}\log\left(1 + \frac{UL_\psi^2 \sum_{j=1}^k \alpha_{k,j}}{\lambda_{\mathbf{w}} d}\right)} + \sqrt{\lambda_{\mathbf{w}}}\kappa S_{\mathbf{w}}.$$

Proof of Lemma 6 is in Appendix F-A.

Action Selection. Based on Lemma 4 and Lemma 6, we construct the optimistic value function $\bar{Q}_h^k(s, a)$, $\bar{V}_h^k(s)$ and the action selection criteria as follow,

$$\bar{Q}_h^k(s, a) = \min\left\{ H, \phi(s, a)^\top \widehat{\theta}_h^k + \beta_\theta \|\phi(s, a)\|_{(\Lambda_h^{k-1})^{-1}} \right.$$
$$\left. + [\widetilde{\mathbb{P}}_h^k \bar{V}_{h+1}^k](s, a) + \frac{H}{\kappa}\bar{\beta}_{\mathbf{w}}^{k-1} \max_{s' \in \mathcal{S}_h^k} \|\bar{\psi}_h^k(s')\|_{(\bar{\Sigma}_h^{k-1})^{-1}} \right\}$$
$$\bar{V}_h^k(s) = \max_{a \in \mathcal{A}} \bar{Q}_h^k(s, a) = \mathbb{E}_{a \sim \pi_h^k(\cdot \mid s)}\left[\bar{Q}_h^k(s, a)\right]$$
$$\pi_h^k(s) = \arg\max_{a \in \mathcal{A}} \bar{Q}_h^k(s, a). \tag{30}$$

The overall algorithm is summarized in Algorithm 4. We show that our algorithm has the following regret guarantee.

Theorem 6. *Let $\delta = 1/(4T)$, $\lambda_\theta = d$, and $\lambda_{\mathbf{w}} = d$, $\forall k, j \in [K], w_{k,j} = \alpha_{k,j} = \gamma^{k-j}$, $\gamma \in (1/K, 1)$, the dynamic regret* D-REG$_T$ *is bounded with probability at least $1 - 1/T$, by*

$$\mathcal{O}\left(\frac{Hd}{\kappa}\left(\frac{1}{(1-\gamma)^{3/2}}\Delta + HK\sqrt{1-\gamma} \right) + H^{3/2}d\sqrt{HK} \right).$$

Furthermore, by setting the discounted factor optimally as $\gamma = 1 - \max\left\{1/K, \sqrt{\Delta/T}\right\}$, we have

$$\text{D-REG}_T \leq \begin{cases} \widetilde{\mathcal{O}}\left(\kappa^{-1}Hd\Delta^{1/4}T^{3/4}\right) & \text{when } \Delta \geq H/K, \\ \widetilde{\mathcal{O}}\left(\kappa^{-1}dH^{3/2}\sqrt{T}\right) & \text{when } \Delta < H/K. \end{cases}$$

Proof of Theorem 6 is in Appendix F-B.

VII. 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. 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 [32]; (b) the restart-based algorithm RestartUCB [3]; (c) the weight-based algorithm D-LinUCB [2]; and (d) the adaptive restart algorithm
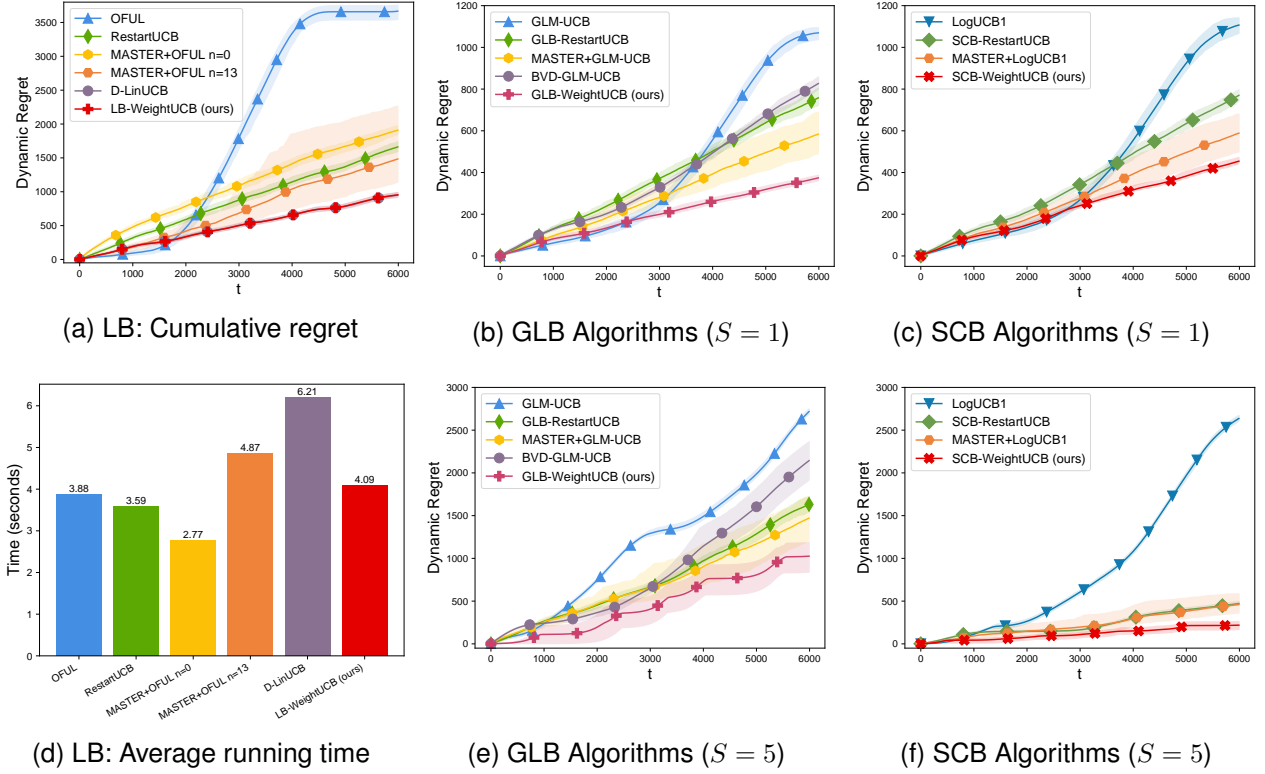
(a) LB: Cumulative regret (b) GLB Algorithms ($S = 1$) (c) SCB Algorithms ($S = 1$)

(d) LB: Average running time (e) GLB Algorithms ($S = 5$) (f) SCB Algorithms ($S = 5$)

Fig. 1: Experiments of generalized linear bandits.

MASTER+OFUL [8]. Among these algorithms, RestartUCB and D-LinUCB both require prior knowledge of P_T, whereas MASTER+OFUL does not. Although MASTER+OFUL operates under weaker requirements, we still include it in our comparison because it achieves theoretically optimal regret with a faster convergence rate, making it an important benchmark. 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 [8]), and the origin MASTER algorithm lets it start from 0 (i.e., $n = 0, 1, ...$). However, a small initial value of n leads to frequent restarts and thus poor performance. To this end, we experiment with a larger initial value of $n = 13$, which yields substantially improved performance in our case.

Results. The experimental results are averaged over 20 independent trials. Fig. 1a shows the cumulative dynamic regret performance, where the shaded area denotes the variance of experimental results. Fig. 1d reports the average time per run, with each run containing 6000 rounds. Our LB-WeightUCB algorithm performs as well as D-LinUCB but is significantly more efficient, with over 1.5 times speedup. Fig. 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.

B. Generalized Linear Bandits

Setting. We employ the logistic model in the 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 [41]; (b) LogUCB1, static algorithm for LogB [33]; (c) BVD-GLM-UCB, weight-based algorithm for GLB [6]; (d) GLB-RestartUCB, restart algorithm for GLB [3]; (e) SCB-RestartUCB, restart algorithm for SCB [3]; (f) MASTER+GLM-UCB, adaptive restart algorithm for GLB [8]; and (g) MASTER+LogUCB1, adaptive restart algorithm for LogB [8]. 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 two MASTER algorithms, we set $n = 13$.

Results. We present the average cumulative dynamic regret results of our experiments on 20 independent trials in Fig. 1. 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 the 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.

VIII. CONCLUSION

This paper revisits the weight-based algorithms for three non-stationary parametric bandit models (LB, GLB, SCB) and two non-stationary MDP settings (Linear Mixture MDP, MNL Mixture MDP). 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 parts in estimation error analysis. Our framework ensures more efficient algorithms for all three bandit models and two RL models, improves the regret bounds for GLB and SCB settings, and establishes the first dynamic regret bound for MNL Mixture MDP.

The importance of our work lies in the fact that we have now made the weight-based algorithms for non-stationary parametric bandits and MDPs as competitive as the restart-based algorithms, in terms of both computational efficiency and regret guarantee. Note that the current window-based, restart-based, and weight-based algorithms can only achieve a regret bound of $\widetilde{\mathcal{O}}(P_T^{1/4} T^{3/4})$, which does not match the optimal rate $\widetilde{\mathcal{O}}(P_T^{1/3} T^{2/3})$ attained by the MASTER algorithm, an adaptive restart strategy [8]. In the spirit of this best-known result, it is essential to design *adaptive* weight-based algorithms that can achieve the optimal dynamic regret bound without requiring prior knowledge of the environment's non-stationarity, given that weighted strategies are particularly effective in gradually drifting environments, which are commonly encountered in real-world applications. The current lower bound of $\Omega(P_T^{1/3} T^{2/3})$ is established under the fixed arm set assumption [40]. The MASTER algorithm [8] matches this rate with the same assumption, making $\Theta(P_T^{1/3} T^{2/3})$ the minimax optimal rate for the fixed arm set case. However, the minimax rate remains open for time-varying arm sets.

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 and restless changes [49], [50]. This might be able to explain 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 [51] or other information-theoretic tools [10].

APPENDIX A
ANALYSIS OF LB-WEIGHTUCB

In this section, we provide the analysis for LB-WeightUCB algorithm. In Appendix A-A, we review the D-LinUCB algorithm proposed by [2] and restate their estimation error analysis. In Appendix A-B, we present our own estimation error analysis for the proposed LB-WeightUCB algorithm, which is captured in Lemma 1. Finally, in Appendix A-C, we provide an analysis of dynamic regret, as stated in Theorem 1.

A. Review Estimation Error Analysis of D-LinUCB Algorithm

In this part, we review the previous estimation error analysis of the D-LinUCB algorithm [2], which 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 estimator (same as (1) and (3)), the estimation error of D-LinUCB algorithm can be decomposed as

$$
\begin{aligned}
\widehat{\theta}_t - \theta_t &= V'^{-1}_{t-1} \left(\sum_{s=1}^{t-1} \gamma^{t-s-1} r_s X_s \right) - \theta_t \\
&= V'^{-1}_{t-1} \left(\sum_{s=1}^{t-1} \gamma^{t-s-1} \left(X_s^\top \theta_s + \eta_s \right) X_s \right) \\
&\quad - V'^{-1}_{t-1} \left(\lambda I_d + \sum_{s=1}^{t-1} \gamma^{t-s-1} X_s X_s^\top \right) \theta_t \\
&= V'^{-1}_{t-1} \left(\sum_{s=1}^{t-1} \gamma^{t-s-1} X_s X_s^\top \theta_s + \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s \right) \\
&\quad - V'^{-1}_{t-1} \left(\lambda I_d + \sum_{s=1}^{t-1} \gamma^{t-s-1} X_s X_s^\top \right) \theta_t \\
&= \underbrace{V'^{-1}_{t-1} \left(\sum_{s=1}^{t-1} \gamma^{t-s-1} X_s X_s^\top \left(\theta_s - \theta_t \right) \right)}_{\texttt{bias part}} \\
&\quad + \underbrace{V'^{-1}_{t-1} \left(\sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s - \lambda \theta_t \right)}_{\texttt{variance part}}, \quad\quad (31)
\end{aligned}
$$

where $V'_t = \lambda I_d + \sum_{s=1}^{t-1} \gamma^{t-s} X_s X_s^\top$. Afterward, [2] uses different local norms (we will explain the reason for using different local norms later) for the bias and variance parts as

$$|\mathbf{x}^\top(\widehat{\theta}_t - \theta_t)| \leq \|\mathbf{x}\|_2 A'_t + \|\mathbf{x}\|_{V'^{-1}_{t-1}\widetilde{V}_{t-1}V'^{-1}_{t-1}} B'_t, \quad (32)$$

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

$$A'_t = \left\| V'^{-1}_{t-1} \sum_{s=1}^{t-1} \gamma^{t-s-1} X_s X_s^\top (\theta_s - \theta_t) \right\|_2$$

$$B'_t = \left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s - \lambda \theta_t \right\|_{\widetilde{V}^{-1}_{t-1}}.$$

For the bias part, [2] divide it into two parts on the timeline by introducing a virtual window size D,

$$A'_t \leq \underbrace{\left\| \sum_{s=t-D}^{t-1} V'^{-1}_{t-1} \gamma^{t-s-1} X_s X_s^\top (\theta_s - \theta_t) \right\|_2}_{\text{virtual window}}$$
$$+ \underbrace{\left\| \sum_{s=1}^{t-D-1} V'^{-1}_{t-1} \gamma^{t-s-1} X_s X_s^\top (\theta_s - \theta_t) \right\|_2}_{\text{small term}},$$

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 [1] since it corresponds to the bias part of the estimation error of the window strategy, and this is why they use l_2-norm for the 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 $s \leq t - D - 1$, 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, [2] extend the previous self-normalized concentration [32, Theorem 1] to the weighted version. This concentration requires the use of $\widetilde{V}_t$ as the local norm. To this end, [2] split the variance part as

$$\left| \mathbf{x}^\top V'^{-1}_{t-1} \left(\sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s - \lambda \theta_t \right) \right| \leq \|\mathbf{x}\|_{V'^{-1}_{t-1}\widetilde{V}_{t-1}V'^{-1}_{t-1}} C'_t,$$

where

$$C'_t = \left\| V'^{-1}_{t-1} \left(\sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s - \lambda \theta_t \right) \right\|_{V'_{t-1}\widetilde{V}^{-1}_{t-1}V'_{t-1}}$$
$$= \left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s - \lambda \theta_t \right\|_{\widetilde{V}^{-1}_{t-1}}$$
$$\leq \left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s \right\|_{\widetilde{V}^{-1}_{t-1}} + \sqrt{\lambda} S.$$

Then term $\| \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s \|_{\widetilde{V}^{-1}_{t-1}}$ can be bounded by the weighted version self-normalized concentration. 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 $\mathbf{x}$,

$$X_t = \arg\max_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \mathbf{x}, \widehat{\theta}_t \rangle + \beta_{t-1} \|\mathbf{x}\|_{V'^{-1}_{t-1}\widetilde{V}_{t-1}V'^{-1}_{t-1}} \right\},$$

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 (32), which is the key step of analysis, and our new analysis framework allows us to employ *same* local norm for both bias and variance parts.

B. Proof of Lemma 1

Proof. Using the same derivation in (31), the estimation error of LB-WeightUCB algorithm can also be decomposed as

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

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

$$\left| \mathbf{x}^\top \left(\widehat{\theta}_t - \theta_t \right) \right| \leq \|\mathbf{x}\|_{V^{-1}_{t-1}} (A_t + B_t), \quad (33)$$

where

$$A_t = \left\| \sum_{s=1}^{t-1} w_{t-1,s} X_s X_s^\top (\theta_s - \theta_t) \right\|_{V^{-1}_{t-1}}$$

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

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

Lemma 7. *For any $t \in [T]$, we have*

$$\left\| \sum_{s=1}^{t-1} w_{t-1,s} X_s X_s^\top (\theta_s - \theta_t) \right\|_{V^{-1}_{t-1}}$$
$$\leq L\sqrt{d} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \|\theta_p - \theta_{p+1}\|_2.$$

Lemma 8. *For any $\delta \in (0,1)$, with probability at least $1 - \delta$, the following holds for all $t \in [T]$,*

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

Based on the inequality (33), Lemma 7, Lemma 8, the boundedness assumption of the feasible set and the fact that for any $\mathbf{x}$, $\|\mathbf{x}\|_{V_{t-1}^{-1}} \leq \|\mathbf{x}\|_2 / \sqrt{\lambda}$ since $V_{t-1} \succeq \lambda I_d$, 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 $t \in [T]$,

$$|\mathbf{x}^\top(\widehat{\theta}_t - \theta_t)|$$

$$\leq L^2 \sqrt{\frac{d}{\lambda}} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \|\theta_p - \theta_{p+1}\|_2 + \beta_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}},$$

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

Proof of Lemma 7. The first step is to extract the variations of the parameter θ_t as follows,

$$\left\| \sum_{s=1}^{t-1} w_{t-1,s} X_s X_s^\top (\theta_s - \theta_t) \right\|_{V_{t-1}^{-1}}$$

$$= \left\| \sum_{s=1}^{t-1} w_{t-1,s} X_s X_s^\top \sum_{p=s}^{t-1} (\theta_p - \theta_{p+1}) \right\|_{V_{t-1}^{-1}}$$

$$= \left\| \sum_{p=1}^{t-1} \sum_{s=1}^{p} w_{t-1,s} X_s X_s^\top (\theta_p - \theta_{p+1}) \right\|_{V_{t-1}^{-1}}$$

$$\leq \sum_{p=1}^{t-1} \left\| \sum_{s=1}^{p} w_{t-1,s} X_s \|X_s\|_2 \|\theta_p - \theta_{p+1}\|_2 \right\|_{V_{t-1}^{-1}}$$

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

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

$$\sum_{s=1}^{p} w_{t-1,s} \|X_s\|_{V_{t-1}^{-1}} \leq \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \sqrt{\sum_{s=1}^{p} w_{t-1,s} \|X_s\|_{V_{t-1}^{-1}}^2}$$

$$\leq \sqrt{d} \sqrt{\sum_{s=1}^{p} w_{t-1,s}}. \tag{34}$$

In above, the second last step holds by the Cauchy-Schwarz inequality. Besides, the last step follows the Lemma 25 by letting $A_s = \sqrt{w_{t-1,s}} X_s$ and $U_{t-1} = V_{t-1}$. Hence, we complete the proof. $\square$

Proof of Lemma 8.

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

$$\leq \left\| \sum_{s=1}^{t-1} w_{t-1,s} \eta_s X_s \right\|_{V_{t-1}^{-1}} + \sqrt{\lambda} S.$$

We define $\widetilde{\eta}_s \triangleq \sqrt{w_{t-1,s}} \eta_s$ and $\widetilde{X}_s \triangleq \sqrt{w_{t-1,s}} X_s$, and notice that $\forall t \in [T], s \in [t-1], |w_{t-1,s}| \leq 1$, then $\widetilde{\eta}_s$ is still R-sub-Gaussian, then by Theorem 7, we have

$$\left\| \sum_{s=1}^{t-1} w_{t-1,s} \eta_s X_s \right\|_{V_{t-1}^{-1}}$$

$$\leq \sqrt{2R^2 \log \left(\frac{\det(V_{t-1})^{\frac{1}{2}} \det(V_0)^{-\frac{1}{2}}}{\delta} \right)}.$$

Then, based on Lemma 26 and $\det(V_0) = \lambda^d$, we have

$$\left\| \sum_{s=1}^{t-1} w_{t-1,s} \eta_s X_s \right\|_{V_{t-1}^{-1}}$$

$$\leq R \sqrt{2 \log \frac{1}{\delta} + d \log \left(1 + \frac{L^2 \sum_{s=1}^{t-1} w_{t-1,s}}{d\lambda}\right)}.$$

which completes the proof. $\square$

C. 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 $1 - \delta$,

$$\forall t \in [T], X_t^{*\top} \theta_t \leq X_t^{*\top} \widehat{\theta}_t + \beta_{t-1} \|X_t^*\|_{V_{t-1}^{-1}}$$

$$+ L^2 \sqrt{\frac{d}{\lambda}} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \|\theta_p - \theta_{p+1}\|_2,$$

$$\forall t \in [T], X_t^\top \theta_t \geq X_t^\top \widehat{\theta}_t - \beta_{t-1} \|X_t\|_{V_{t-1}^{-1}}$$

$$- L^2 \sqrt{\frac{d}{\lambda}} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \|\theta_p - \theta_{p+1}\|_2.$$

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

$$X_t^{*\top} \theta_t - X_t^\top \theta_t$$

$$\leq X_t^{*\top} \widehat{\theta}_t - X_t^\top \widehat{\theta}_t + 2L^2 \sqrt{\frac{d}{\lambda}} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \|\theta_p - \theta_{p+1}\|_2$$

$$+ \beta_{t-1}(\|X_t^*\|_{V_{t-1}^{-1}} + \|X_t\|_{V_{t-1}^{-1}})$$

$$\leq 2L^2 \sqrt{\frac{d}{\lambda}} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \|\theta_p - \theta_{p+1}\|_2 + 2\beta_{t-1} \|X_t\|_{V_{t-1}^{-1}},$$

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,

$$
\text{D-REG}_T = \sum_{t=1}^{T} \left(X_t^{*\top} \theta_t - X_t^\top \theta_t \right)
$$

$$
\leq \underbrace{2L^2 \sqrt{\frac{d}{\lambda} \sum_{t=1}^{T} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \|\theta_p - \theta_{p+1}\|_2}}_{\text{bias part}}
$$

$$
+ \underbrace{2\beta_T \sum_{t=1}^{T} \|X_t\|_{V_{t-1}^{-1}}}_{\text{variance part}},
$$

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. Notice that $w_{t-1,s} = \gamma^{t-s-1}$ with $\gamma \in (1/T, 1)$. For the bias part, we need to extract the path length P_T and show the control of the discounted factor γ on P_T.

$$
2L^2 \sqrt{\frac{d}{\lambda}} \sum_{t=1}^{T} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \|\theta_p - \theta_{p+1}\|_2
$$

$$
= 2L^2 \sqrt{\frac{d}{\lambda}} \sum_{p=1}^{T-1} \sum_{t=p+1}^{T} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \|\theta_p - \theta_{p+1}\|_2
$$

$$
= 2L^2 \sqrt{\frac{d}{\lambda}} \sum_{p=1}^{T-1} \sum_{t=p+1}^{T} \gamma^{\frac{t-1}{2}} \sqrt{\sum_{s=1}^{p} \gamma^{-s}} \|\theta_p - \theta_{p+1}\|_2
$$

$$
= 2L^2 \sqrt{\frac{d}{\lambda}} \sum_{p=1}^{T-1} \frac{\gamma^{\frac{p}{2}} - \gamma^{\frac{T}{2}}}{1 - \gamma^{\frac{1}{2}}} \sqrt{\frac{\gamma^{-p} - 1}{1 - \gamma}} \|\theta_p - \theta_{p+1}\|_2
$$

$$
\leq 2L^2 \sqrt{\frac{d}{\lambda}} \sum_{p=1}^{T-1} \frac{\gamma^{\frac{p}{2}} \gamma^{-\frac{p}{2}}}{(1 - \gamma^{\frac{1}{2}}) \frac{1 + \gamma^{\frac{1}{2}}}{2} \sqrt{1 - \gamma}} \|\theta_p - \theta_{p+1}\|_2
$$

$$
\leq 4L^2 \sqrt{\frac{d}{\lambda}} \frac{1}{(1 - \gamma)^{3/2}} P_T.
$$

So for the bias part, we have

$$
2L^2 \sqrt{\frac{d}{\lambda}} \sum_{t=1}^{T} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \|\theta_p - \theta_{p+1}\|_2
$$

$$
\leq 4L^2 \sqrt{\frac{d}{\lambda}} \frac{1}{(1 - \gamma)^{3/2}} P_T. \tag{35}
$$

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

by Lemma 9 (weighted potential lemma) with $w_{t,t} = \gamma^{t-t} = 1, c = w_{t,s}/w_{t,s-1} = \gamma$, we have the following upper bound:

$$
2\beta_T \sum_{t=1}^{T} \|X_t\|_{V_{t-1}^{-1}} \leq 2\beta_T \sqrt{2 \max\{1, \frac{L^2}{\lambda}\} dT}
$$

$$
\cdot \sqrt{T \log \frac{1}{\gamma} + \log\left(1 + \frac{L^2 \sum_{s=1}^{T} w_{T,s}}{d\lambda}\right)}. \tag{36}
$$

Notice that the main differences between weighted LB and standard LB in analysis are the need for path length analysis and the use of the weighted potential lemma. Further we have $\sum_{s=1}^{t} w_{t,s} = \frac{1-\gamma^t}{1-\gamma} \leq \frac{1}{1-\gamma}$. 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 $1 - 1/T$,

$$
\text{D-REG}_T \leq 4L^2 \sqrt{\frac{d}{\lambda}} \frac{1}{(1 - \gamma)^{3/2}} P_T + 2\beta_T \sqrt{2 \max\left\{1, \frac{L^2}{\lambda}\right\} dT}
$$

$$
\cdot \sqrt{T \log \frac{1}{\gamma} + \log\left(1 + \frac{L^2}{\lambda d(1 - \gamma)}\right)},
$$

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

Since the regret bound contains a term $T\sqrt{\log(1/\gamma)}$, 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

$$
\text{D-REG}_T \leq \widetilde{\mathcal{O}} \left(\frac{1}{(1-\gamma)^{3/2}} P_T + d(1-\gamma)^{1/2} T \right). \tag{37}
$$

When $P_T < d/T$ (which corresponds to 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 $\widetilde{\mathcal{O}}(d^{3/4} P_T^{1/4} T^{3/4})$ regret bound, which completes the proof. $\square$

Lemma 9 (Weighted Version Potential Lemma). *Suppose* $V_t = \sum_{s=1}^{t} w_{t,s} X_s X_s^\top + \lambda I_d, V_0 = \lambda I_d$, *the weight satisfies that,* $\forall t \in [T], s \in [t-1], w_{t,s}/w_{t-1,s} = c \leq 1$, $\forall t, s \in [T], w_{t,s} \in (0,1), w_{t,t} = 1$ *and* $\|X_t\|_2 \leq L$ *for all* $t \geq 1$*, then the following inequality holds,*

$$
\sum_{t=1}^{T} \|w_{t,t} X_t\|_{V_{t-1}^{-1}}^2
$$

$$
\leq 2 \max\left\{1, \frac{L^2}{\lambda}\right\} d\left(T \log \frac{1}{c} + \log\left(1 + \frac{L^2 \sum_{s=1}^{T} w_{T,s}}{d\lambda}\right)\right).
$$

Proof of Lemma 9.

$$
V_t = \sum_{s=1}^{t} w_{t,s} X_s X_s^\top + \lambda I_d
$$

$$
= \sum_{s=1}^{t-1} w_{t,s} X_s X_s^\top + w_{t,t} X_t X_t^\top + \lambda I_d
$$

$$= c \sum_{s=1}^{t-1} w_{t-1,s} X_s X_s^\top + X_t X_t^\top + \lambda I_d \qquad (w_{t,t}=1)$$

$$\succeq c \left(\sum_{s=1}^{t-1} w_{t-1,s} X_s X_s^\top + X_t X_t^\top + \lambda I_d \right), \qquad (c<1)$$

$$= c \left(V_{t-1} + X_t X_t^\top \right)$$

$$= c V_{t-1}^{1/2} \left(I_d + V_{t-1}^{-1/2} X_t X_t^\top V_{t-1}^{-1/2} \right) V_{t-1}^{1/2},$$

Taking the determinant on both sides and we obtain,

$$\det(V_t) \ge \det(c V_{t-1}) \det\left(I_d + V_{t-1}^{-1/2} X_t X_t^\top V_{t-1}^{-1/2} \right)$$

$$\det(V_t) \ge c^d \det(V_{t-1}) \left(1 + \|X_t\|_{V_{t-1}^{-1}}^2 \right)$$

$$\log \det(V_t) \ge d \log c + \log \det(V_{t-1}) + \log \left(1 + \|X_t\|_{V_{t-1}^{-1}}^2 \right)$$

$$d \log \frac{1}{c} + \log \frac{\det(V_t)}{det(V_{t-1})} \ge \log \left(1 + \|X_t\|_{V_{t-1}^{-1}}^2 \right)$$

Then summing from 1 to T, and telescoping we have,

$$dT \log \frac{1}{c} + \log \left(\frac{\det(V_T)}{\det(V_0)} \right)$$

$$\ge \sum_{t=1}^T \log \left(1 + \|X_t\|_{V_{t-1}^{-1}}^2 \right)$$

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

$$\ge \frac{1}{2 \max\{1, L^2/\lambda\}} \sum_{t=1}^T \|X_t\|_{V_{t-1}^{-1}}^2$$

So we have

$$\sum_{t=1}^T \|X_t\|_{V_{t-1}^{-1}}^2$$

$$\le 2 \max \left\{ 1, \frac{L^2}{\lambda} \right\} \cdot \left(dT \log \frac{1}{c} + \log \left(\frac{\det(V_T)}{\det(V_0)} \right) \right).$$

Finally, by using Lemma 26 and the fact $\det(V_0) = \lambda^d$, we complete the proof. $\qquad\square$

APPENDIX B
ANALYSIS OF GLB-WEIGHTUCB

In this section, we provide analysis for GLB-WeightUCB algorithm. In Appendix B-A, we review the projection issue of GLB and restate the BVD-GLM-UCB algorithm of [6]. In Appendix B-B, we present the proof of the estimation error upper bound of our GLB-WeightUCB algorithm (namely, Lemma 2). Finally, in Appendix B-C, we provide the proof of dynamic regret upper bound as stated in Theorem 2.

A. Review Projection Step of BVD-GLM-UCB Algorithm

As mentioned in Section IV-B, 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, [41] overcame this difficulty by introducing a projection step as

$$\widetilde{\theta}_t = \arg\min_{\theta \in \Theta} \|g_t(\widehat{\theta}_t) - g_t(\theta)\|_{V_{t-1}^{-1}}, \qquad (38)$$

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

$$g_t(\widehat{\theta}_t) = \lambda c_\mu \widehat{\theta}_t + \sum_{s=1}^{t-1} \mu(X_s^\top \widehat{\theta}_t) X_s = \sum_{s=1}^{t-1} r_s X_s, \qquad (39)$$

and then by the mean value theorem, we know that

$$g_t(\theta_1) - g_t(\theta_2) = G_t(\theta_1, \theta_2)(\theta_1 - \theta_2), \qquad (40)$$

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 $\forall \theta_1, \theta_2 \in \Theta, G_t(\theta_1, \theta_2) \succeq c_\mu V_{t-1}$. By this projection step, [41] can analyze the estimation error like,

$$|\mu(\mathbf{x}^\top \widetilde{\theta}_t) - \mu(\mathbf{x}^\top \theta_t)| \le k_\mu |\mathbf{x}^\top (\widetilde{\theta}_t - \theta_t)|$$

$$= k_\mu |\mathbf{x}^\top G_t^{-1}(\theta_t, \widetilde{\theta}_t)(g_t(\widetilde{\theta}_t) - g_t(\theta_t))|$$

$$\le k_\mu \|\mathbf{x}\|_{G_t^{-1}(\theta_t, \widetilde{\theta}_t)} \|g_t(\widetilde{\theta}_t) - g_t(\theta_t)\|_{G_t^{-1}(\theta_t, \widetilde{\theta}_t)}$$

$$\le \frac{k_\mu}{c_\mu} \|\mathbf{x}\|_{V_{t-1}^{-1}} \|g_t(\widetilde{\theta}_t) - g_t(\theta_t)\|_{V_{t-1}^{-1}}$$

$$\le \frac{2k_\mu}{c_\mu} \|\mathbf{x}\|_{V_{t-1}^{-1}} \|g_t(\widehat{\theta}_t) - g_t(\theta_t)\|_{V_{t-1}^{-1}},$$

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, [7] can also use this projection step. [6] 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 [6] 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, [6] propose the BVD-GLM-UCB algorithm for drifting generalized linear bandits, as restated in Algorithm 5, 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.

$$\mathcal{E}_t^\delta(\theta) := \left\{ \theta' \in \mathbb{R}^d \,\middle|\, \|g_t(\theta') - g_t(\theta)\|_{\widetilde{V}_t^{-1}} \le \bar{\beta}_t(\delta) \right\}. \quad (41)$$

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.

$$\theta_t^p \in \arg\min_{\theta \in \mathbb{R}^d} \left\| g_t(\theta) - g_t(\widehat{\theta}_t) \right\|_{V_t^{-2}}$$

$$\text{s.t } \Theta \cap \mathcal{E}_t^\delta(\theta) \ne \emptyset \qquad (42)$$

Algorithm 5 BVD-GLM-UCB [6]

Require: 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:    Solving $\theta_t^p \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \left\| g_t(\theta) - g_t(\widehat{\theta}_t) \right\|_{V_t^{-2}} \right.$ s.t $\left. \Theta \cap \mathcal{E}_t^\delta(\theta) \neq \emptyset \right\}$
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:    Get $\bar{\beta}_{t-1} = R\sqrt{2\log\frac{1}{\delta} + d\log\left(1 + \frac{L^2(1-\gamma^{2t-2})}{\lambda d(1-\gamma^2)}\right)} + \sqrt{\lambda}c_\mu S$
6:    Get $X_t = \arg\max_{\mathbf{x} \in \mathcal{X}} \left\{ \mu(\mathbf{x}^\top \widetilde{\theta}_t) + \frac{2k_\mu}{c_\mu}\bar{\beta}_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}} \right\}$
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 $\widehat{\theta}_{t+1}$ by $\sum_{s=1}^t \gamma^{t-s}\left(\mu(X_s^\top \theta) - r_s\right)X_s + \lambda c_\mu \theta = 0$
10: **end for**

After obtaining the solution θ_t^p via computing the optimization problem (42), the third step is to select $\widetilde{\theta}_t$ from $\Theta \cap \mathcal{E}_t^\delta(\theta_t^p)$. Based on this projection step, [6] can separate the bias and variance parts before projection as follows,

$$
\begin{aligned}
&|\mu(\mathbf{x}^\top \widetilde{\theta}_t) - \mu(\mathbf{x}^\top \theta_t)| \leq k_\mu |\mathbf{x}^\top (\widetilde{\theta}_t - \theta_t)| \\
&= k_\mu |\mathbf{x}^\top G_t^{-1}(\theta_t, \widetilde{\theta}_t)(g_t(\widetilde{\theta}_t) - g_t(\theta_t))| \\
&\leq k_\mu |\mathbf{x}^\top G_t^{-1}(\theta_t, \widetilde{\theta}_t)(g_t(\widetilde{\theta}_t) - g_t(\theta_t^p) + g_t(\theta_t^p) - g_t(\widehat{\theta}_t) \\
&\quad + g_t(\widehat{\theta}_t) - g_t(\bar{\theta}_t) + g_t(\bar{\theta}_t) - g_t(\theta_t))| \\
&\leq \underbrace{k_\mu |\mathbf{x}^\top G_t^{-1}(\theta_t, \widetilde{\theta}_t)(g_t(\widetilde{\theta}_t) - g_t(\theta_t^p) + g_t(\widehat{\theta}_t) - g_t(\bar{\theta}_t))|}_{\texttt{bias part}} \\
&\quad + \underbrace{k_\mu |\mathbf{x}^\top G_t^{-1}(\theta_t, \widetilde{\theta}_t)(g_t(\theta_t^p) - g_t(\widehat{\theta}_t) + g_t(\bar{\theta}_t) - g_t(\theta_t))|}_{\texttt{variance part}}.
\end{aligned}
$$

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 (41) and (42)), 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 (41)) at each round t during the algorithm running. In the next section, we will show that the simple projection used in the stationary GLB (38) can be sufficient for coping with drifting GLB via our refined analysis framework.

B. Proof of Lemma 2

Proof. Based on the estimator equation (11), we know that

$$
\begin{aligned}
g_t(\widehat{\theta}_t) &= \lambda c_\mu \widehat{\theta}_t + \sum_{s=1}^{t-1} w_{t-1,s}\mu(X_s^\top \widehat{\theta}_t)X_s \\
&= \sum_{s=1}^{t-1} w_{t-1,s} r_s X_s,
\end{aligned} \tag{43}
$$

and then by the mean value theorem, we know that

$$
g_t(\theta_1) - g_t(\theta_2) = G_t(\theta_1, \theta_2)(\theta_1 - \theta_2), \tag{44}
$$

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

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

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

By Assumption 2, the mean value theorem (44) on g_t and the projection (12), 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 (13) and $g_t(\widehat{\theta}_t)$ (43), we have,

$$
\begin{aligned}
&g_t(\theta_t) - g_t(\widehat{\theta}_t) \\
&= \lambda c_\mu \theta_t + \sum_{s=1}^{t-1} w_{t-1,s}\mu(X_s^\top \theta_t)X_s - \sum_{s=1}^{t-1} w_{t-1,s} r_s X_s \\
&= \lambda c_\mu \theta_t + \sum_{s=1}^{t-1} w_{t-1,s}\mu(X_s^\top \theta_t)X_s \\
&\quad - \sum_{s=1}^{t-1} w_{t-1,s}(\mu(X_s^\top \theta_s) + \eta_s)X_s \\
&= \underbrace{\sum_{s=1}^{t-1} w_{t-1,s}(\mu(X_s^\top \theta_t) - \mu(X_s^\top \theta_s))X_s}_{\texttt{bias part}} \\
&\quad + \underbrace{\lambda c_\mu \theta_t - \sum_{s=1}^{t-1} w_{t-1,s}\eta_s X_s}_{\texttt{variance part}}.
\end{aligned}
$$

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

$$
\left|\mu(\mathbf{x}^\top \widetilde{\theta}_t) - \mu(\mathbf{x}^\top \theta_t)\right| \leq \frac{2k_\mu}{c_\mu}\|\mathbf{x}\|_{V_{t-1}^{-1}}(C_t + D_t), \tag{45}
$$

where

$$C_t = \left\|\sum_{s=1}^{t-1} w_{t-1,s}(\mu(X_s^\top\theta_t) - \mu(X_s^\top\theta_s))X_s\right\|_{V_{t-1}^{-1}}$$

$$D_t = \left\|\sum_{s=1}^{t-1} w_{t-1,s}\eta_s X_s - \lambda c_\mu\theta_t\right\|_{V_{t-1}^{-1}}.$$

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

Lemma 10. *For any $t \in [T]$, we have*

$$\left\|\sum_{s=1}^{t-1} w_{t-1,s}(\mu(X_s^\top\theta_t) - \mu(X_s^\top\theta_s))X_s\right\|_{V_{t-1}^{-1}}$$

$$\leq Lk_\mu\sqrt{d}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}\|\theta_p - \theta_{p+1}\|_2}. \tag{46}$$

Lemma 11. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds for all $t \in [T]$,*

$$\left\|\sum_{s=1}^{t-1} w_{t-1,s}\eta_s X_s - \lambda c_\mu\theta_t\right\|_{V_{t-1}^{-1}} \leq \sqrt{\lambda}c_\mu S$$

$$+ R\sqrt{2\log\frac{1}{\delta} + d\log\left(1 + \frac{L^2\sum_{s=1}^{t-1} w_{t-1,s}}{\lambda d}\right)}. \tag{47}$$

Based on the inequality (45), Lemma 10, Lemma 11, 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 $t \in [T]$,

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

$$\cdot\left(Lk_\mu\sqrt{d}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}\|\theta_p - \theta_{p+1}\|_2} + \bar{\beta}_{t-1}\right)$$

$$\leq \frac{2k_\mu}{c_\mu}\left(L^2 k_\mu\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}\|\theta_p - \theta_{p+1}\|_2}\right.$$

$$\left. + \bar{\beta}_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}}\right),$$

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

Proof of Lemma 10. Here we need to extract the variations of the time-varying parameter θ_t

$$\left\|\sum_{s=1}^{t-1} w_{t-1,s}(\mu(X_s^\top\theta_t) - \mu(X_s^\top\theta_s))X_s\right\|_{V_{t-1}^{-1}}$$

$$\leq \left\|\sum_{s=1}^{t-1} w_{t-1,s}\sum_{p=s}^{t-1}(\mu(X_s^\top\theta_p) - \mu(X_s^\top\theta_{p+1}))X_s\right\|_{V_{t-1}^{-1}}$$

$$= \left\|\sum_{p=1}^{t-1}\sum_{s=1}^{p} w_{t-1,s}\alpha(X_s, \theta_p, \theta_{p+1})(X_s^\top\theta_p - X_s^\top\theta_{p+1})X_s\right\|_{V_{t-1}^{-1}}$$

$$\leq \sum_{p=1}^{t-1}\left\|\sum_{s=1}^{p} w_{t-1,s}\alpha(X_s, \theta_p, \theta_{p+1})X_s\|X_s\|_2\|\theta_p - \theta_{p+1}\|_2\right\|_{V_{t-1}^{-1}}$$

$$\leq L\sum_{p=1}^{t-1}\sum_{s=1}^{p} w_{t-1,s}|\alpha(X_s, \theta_p, \theta_{p+1})|\,\|X_s\|_{V_{t-1}^{-1}}\,\|\theta_p - \theta_{p+1}\|_2$$

$$\leq Lk_\mu\sum_{p=1}^{t-1}\sum_{s=1}^{p} w_{t-1,s}\,\|X_s\|_{V_{t-1}^{-1}}\,\|\theta_p - \theta_{p+1}\|_2.$$

where the fourth 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 of bound of term $\sum_{s=1}^{p} w_{t-1,s}\,\|X_s\|_{V_{t-1}^{-1}}$ is the same as the inequality (34) in Appendix A-B, hence we complete the proof. $\square$

Proof of Lemma 11. Same as the linear case, we define $\widetilde{\eta}_s \triangleq \sqrt{w_{t-1,s}}\eta_s$ and $\widetilde{X}_s \triangleq \sqrt{w_{t-1,s}}X_s$, and notice that $\forall t \in [T], s \in [t-1], |w_{t-1,s}| \leq 1$, then $\widetilde{\eta}_s$ is still R-sub-Gaussian, then by Theorem 7, we have

$$D_t = \left\|\sum_{s=1}^{t-1} w_{t-1,s}\eta_s X_s - \lambda c_\mu\theta_t\right\|_{V_{t-1}^{-1}}$$

$$\leq \left\|\sum_{s=1}^{t-1} w_{t-1,s}\eta_s X_s\right\|_{V_{t-1}^{-1}} + \|\lambda c_\mu\theta_t\|_{V_{t-1}^{-1}}$$

$$\leq \left\|\sum_{s=1}^{t-1} \widetilde{\eta}_s\widetilde{X}_s\right\|_{V_{t-1}^{-1}} + \sqrt{\lambda}c_\mu S$$

$$\leq R\sqrt{2\log\frac{1}{\delta} + d\log\left(1 + \frac{L^2\sum_{s=1}^{t-1} w_{t-1,s}}{d\lambda}\right)} + \sqrt{\lambda}c_\mu S.$$

The proof here is the same as the proof of Lemma 8 in A-B, the only difference is an extra c_μ in the second term. $\square$

C. 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 $1-\delta$,

$$\forall t \in [T], \mu(X_t^{*\top}\theta_t) \leq \mu(X_t^{*\top}\widetilde{\theta}_t)$$

$$+ \frac{2k_\mu}{c_\mu}\left(L^2 k_\mu\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}\|\theta_p - \theta_{p+1}\|_2}\right.$$

$$\left. + \bar{\beta}_{t-1}\|X_t^*\|_{V_{t-1}^{-1}}\right),$$

$$\forall t \in [T], \mu(X_t^\top\theta_t) \geq \mu(X_t^\top\widetilde{\theta}_t)$$

$$- \frac{2k_\mu}{c_\mu}\left(L^2 k_\mu\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}\|\theta_p - \theta_{p+1}\|_2}\right.$$

$$\left. + \bar{\beta}_{t-1}\|X_t\|_{V_{t-1}^{-1}}\right).$$

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

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

where the last step comes from the arm selection criterion (15) 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,

$$\text{D-REG}_T = \sum_{t=1}^{T}\max_{\mathbf{x}\in\mathcal{X}}\mu(\mathbf{x}^\top\theta_t) - \mu(X_t^\top\theta_t)$$
$$\leq \underbrace{\frac{4L^2k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\sum_{t=1}^{T}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p}w_{t-1,s}}\|\theta_p - \theta_{p+1}\|_2}_{\text{bias part}}$$
$$+ \underbrace{\frac{4k_\mu}{c_\mu}\bar{\beta}_T\sum_{t=1}^{T}\|X_t\|_{V_{t-1}^{-1}}}_{\text{variance part}}.$$

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.

Bias Part. Similar to the proof of inequality (35), we have

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

Variance Part. Similar to the proof of inequality (36), let $C_T^{\text{GLB}} \triangleq \frac{4k_\mu}{c_\mu}\bar{\beta}_T\sqrt{2\max\{1, L^2/\lambda\}dT}$ we have

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

Algorithm 6 SCB-WeightUCB

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

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:    **if** $\|\widehat{\theta}_t\|_2 \leq S$ **then**
4:       let $\widetilde{\theta}_t = \widehat{\theta}_t$
5:    **else**
6:       Do the projection and get $\widetilde{\theta}_t$ by (17)
7:    **end if**
8:    Compute $\widetilde{\beta}_{t-1}$ by (19)
9:    Select X_t by (20)
10:   Receive the reward r_t
11:   Update $V_t = \gamma V_{t-1} + X_tX_t^\top + (1-\gamma)\lambda I_d$
12:   Compute $\widehat{\theta}_{t+1}$ according to (11)
13: **end for**

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 $1 - 1/T$,

$$\text{D-REG}_T \leq \frac{8L^2k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\frac{1}{(1-\gamma)^{3/2}}P_T$$
$$+ C_T^{\text{GLB}}\sqrt{T\log\frac{1}{\gamma} + \log\left(1 + \frac{L^2}{\lambda d(1-\gamma)}\right)}.$$

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

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

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 $\widetilde{\mathcal{O}}(k_\mu^2 c_\mu^{-3/4}d^{3/4}P_T^{1/4}T^{3/4})$. $\qquad\square$

APPENDIX C
ANALYSIS OF SCB-WEIGHTUCB

In this section, we first present SCB-WeightUCB algorithm in Algorithm 6, Then, in Appendix C-A we present the proof of the estimation error upper bound of our SCB-WeightUCB algorithm (Lemma 3). Finally, in Appendix C-B, we provide the proof of dynamic regret upper bound (Theorem 3).

A. Proof of Lemma 3

Proof. Based on the estimator equation (11), we know that

$$g_t(\widehat{\theta}_t) = \lambda c_\mu\widehat{\theta}_t + \sum_{s=1}^{t-1}w_{t-1,s}\mu(X_s^\top\widehat{\theta}_t)X_s$$
$$= \sum_{s=1}^{t-1}w_{t-1,s}r_sX_s, \tag{48}$$

and then by the mean value theorem, we know that

$$g_t(\theta_1) - g_t(\theta_2) = G_t(\theta_1, \theta_2)(\theta_1 - \theta_2), \tag{49}$$

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

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

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

$$H_t(\theta) \triangleq \lambda c_\mu I_d + \sum_{s=1}^{t-1} w_{t-1,s}\mu'(X_s^\top\theta)X_s X_s^\top. \tag{50}$$

By Assumption 2, the mean value theorem (44) on g_t, the projection (17) and Lemma 28, we have

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

then based on the model assumption (16), the function g_t (13) and the $g_t(\widehat{\theta}_t)$ (48), we have,

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

then, by Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
&\left|\mu(\mathbf{x}^\top\widetilde{\theta}_t) - \mu(\mathbf{x}^\top\theta_t)\right| \\
&\le 2\sqrt{1+2S}\frac{k_\mu}{\sqrt{c_\mu}}\|\mathbf{x}\|_{V_{t-1}^{-1}}\left(E_t + F_t\right),
\end{aligned} \tag{51}
$$

where

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

This two terms can be bounded separately.

Lemma 12. *For any $t \in [T]$, we have*

$$
\left\|\sum_{s=1}^{t-1} w_{t-1,s}(\mu(X_s^\top\theta_t) - \mu(X_s^\top\theta_s))X_s\right\|_{H_t^{-1}(\theta_t)}
$$
$$
\le L\frac{k_\mu}{\sqrt{c_\mu}}\sqrt{d}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^p w_{t-1,s}}\|\theta_p - \theta_{p+1}\|_2. \tag{52}
$$

Lemma 13. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have for all $t \in [T]$,*

$$
\left\|\sum_{s=1}^{t-1} w_{t-1,s}\eta_s X_s - \lambda c_\mu\theta_t\right\|_{H_t^{-1}(\theta_t)}
$$
$$
\le \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}}\log\frac{1}{\delta} + \frac{2m}{\sqrt{\lambda c_\mu}}d\log(2) + \sqrt{\lambda c_\mu}S \tag{53}
$$
$$
+ \frac{dm}{\sqrt{\lambda c_\mu}}\log\left(1 + \frac{L^2 k_\mu \sum_{s=1}^{t-1} w_{t-1,s}}{\lambda c_\mu d}\right),
$$

Based on the inequality (51), Lemma 10 and Lemma 11, 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 $t \in [T]$,

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

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

Proof of Lemma 12. Since $\forall \theta, H_t(\theta) \succeq c_\mu V_{t-1}$, we have

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

Then use Lemma 10 and we complete the proof. □

Proof of Lemma 13. We define $\widetilde{\eta}_s \triangleq \frac{\sqrt{w_{t-1,s}}\eta_s}{|m|}$, $\widetilde{X}_s \triangleq \sqrt{w_{t-1,s}}X_s$ and notice that $\forall t \in [T], s \in [t-1], |w_{t-1,s}| \leq 1$, then $\widetilde{\eta}_s$ is bounded by 1 with variance $\widetilde{\sigma}_s$, then we have

$$F_t = \left\| \sum_{s=1}^{t-1} w_{t-1,s}\eta_s X_s - \lambda c_\mu \theta_t \right\|_{H_t^{-1}(\theta_t)}$$

$$\leq \left\| \sum_{s=1}^{t-1} w_{t-1,s}\eta_s X_s \right\|_{H_t^{-1}(\theta_t)} + \sqrt{\lambda c_\mu}S$$

$$= \left\| \sum_{s=1}^{t-1} \widetilde{\eta}_s \widetilde{X}_s \right\|_{\widetilde{H}_t^{-1}(\theta_t)} + \sqrt{\lambda c_\mu}S$$

where $\widetilde{H}_t(\theta) \triangleq \frac{\lambda c_\mu}{m^2}I_d + \sum_{s=1}^{t-1} \frac{\mu'(X_s^\top\theta)}{m^2}\widetilde{X}_s\widetilde{X}_s^\top$. By the model assumption (16), we know that $\widetilde{\sigma}_t^2 = \frac{\mathbb{E}[\widetilde{\eta}_t^2|\mathcal{F}_{t-1}]}{m^2} \leq \frac{\mathbb{E}[\eta_t^2|\mathcal{F}_t]}{m^2} = \frac{\text{Var}[r_t|\mathcal{F}_t]}{m^2} = \frac{\mu'(X_t\theta_t)}{m^2}$, then from the Self-normalized concentration inequality for self-concordant bandits [33, Theorem 1], restated in Theorem 8, we can get the bound for the first term $\left\| \sum_{s=1}^{t-1} \widetilde{\eta}_s \widetilde{X}_s \right\|_{\widetilde{H}_t^{-1}(\theta_t)}$ as follows,

$$\left\| \sum_{s=1}^{t-1} w_{t-1,s}\widetilde{\eta}_s X_s \right\|_{H_t^{-1}(\theta_t)}$$

$$\leq \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}}\log\left(\frac{\det(H_t)^{1/2}}{\delta(\lambda c_\mu)^{d/2}} \right) + \frac{2m}{\sqrt{\lambda c_\mu}}d\log(2),$$

then we have

$$\left\| \sum_{s=1}^{t-1} \gamma^{-s}\eta_s X_s \right\|_{\widetilde{H}_t^{-1}(\theta_t)} \leq \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}}\log\frac{1}{\delta}$$

$$+ \frac{dm}{\sqrt{\lambda c_\mu}}\log\left(1 + \frac{L^2 k_\mu \sum_{s=1}^{t-1} w_{t-1,s}}{\lambda c_\mu d} \right) + \frac{2m}{\sqrt{\lambda c_\mu}}d\log(2).$$

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

B. 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 $1 - \delta$,

$$\forall t \in [T], \mu(X_t^{*\top}\theta_t)$$

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

$$\left. + L^2\frac{k_\mu}{\sqrt{\lambda c_\mu}}\sqrt{d}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}\|\theta_p - \theta_{p+1}\|_2} \right),$$

$$\forall t \in [T], \mu(X_t^\top\theta_t)$$

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

$$\left. + L^2\frac{k_\mu}{\sqrt{\lambda c_\mu}}\sqrt{d}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}\|\theta_p - \theta_{p+1}\|_2} \right).$$

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

$$\mu(X_t^{*\top}\theta_t) - \mu(X_t^\top\theta_t) \leq \mu(X_t^{*\top}\widetilde{\theta}_t) - \mu(X_t^\top\widetilde{\theta}_t)$$

$$+ 2\sqrt{1+2S}\left(\frac{2L^2 k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}}\|\theta_p - \theta_{p+1}\|_2 \right.$$

$$\left. + \frac{k_\mu}{\sqrt{c_\mu}}\left(\widetilde{\beta}_{t-1}\|X_t^*\|_{V_{t-1}^{-1}} + \widetilde{\beta}_{t-1}\|X_t\|_{V_{t-1}^{-1}} \right) \right)$$

$$\leq \frac{4\sqrt{1+2S}L^2 k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}}\|\theta_p - \theta_{p+1}\|_2$$

$$+ \frac{4\sqrt{1+2S}k_\mu}{\sqrt{c_\mu}}\widetilde{\beta}_{t-1}\|X_t\|_{V_{t-1}^{-1}},$$

where the last step comes from the arm selection criterion (20) 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,

$$\text{D-REG}_T = \sum_{t=1}^{T} \mu(X_t^{*\top}\theta_t) - \mu(X_t^\top\theta_t)$$

$$\leq \underbrace{\frac{4\sqrt{1+2S}L^2 k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\sum_{t=1}^{T}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}}\|\theta_p - \theta_{p+1}\|_2}_{\text{bias part}}$$

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

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

Now we derive the upper bound for these two parts.

Bias Part. Similar to the proof of inequality (35), we have

$$\frac{4\sqrt{1+2S}L^2 k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\sum_{t=1}^{T}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}}\|\theta_p - \theta_{p+1}\|_2$$

$$\leq \frac{8\sqrt{1+2S}L^2 k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\frac{1}{(1-\gamma)^{3/2}}P_T.$$

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

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

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

For term $\sqrt{\sum_{t=1}^{T}\|X_t\|_{V_{t-1}^{-1}}^2}$, we can use the Lemma 9 to bound it, Let $C_T^{\text{SCB}} \triangleq \frac{4\sqrt{1+2S}k_\mu}{\sqrt{c_\mu}}\widetilde{\beta}_T\sqrt{2\max\{1, L^2/\lambda\}dT}$ then

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

$$\leq C_T^{\text{SCB}}\sqrt{T\log\frac{1}{\gamma} + \log\left(1 + \frac{L^2}{\lambda d(1-\gamma)}\right)}.$$

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 $1 - 1/T$,

$$\text{D-Reg}_T \leq \frac{8\sqrt{1+2S}L^2 k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\frac{1}{(1-\gamma)^{3/2}}P_T$$

$$+ C_T^{\text{SCB}}\sqrt{T\log\frac{1}{\gamma} + \log\left(1 + \frac{L^2}{\lambda d(1-\gamma)}\right)}.$$

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

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

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. $\qquad\square$

APPENDIX D
PIECEWISE-STATIONARY SCB

In this section, we study SCB under piecewise-stationary environment and our work is a direct improvement over [7]. 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.

A. SCB-PW-WeightUCB Algorithm

Inspired by [34], we make a direct improvement over [7]. Just like [7], 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 [7] 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.

Algorithm 7 SCB-PW-WeightUCB

Require: 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 (11)
6: **end for**

Estimator. At iteration t, we adopt the same maximum likelihood estimator (11) with $w_{t,s} = \gamma^{t-s}$ as in the drifting case.

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

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

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

Lemma 14. *For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have $\forall t \in \mathcal{T}(D), \theta_t \in \mathcal{C}_t(\delta)$.*

$$\mathcal{C}_t(\delta) = \Bigg\{\theta \in \Theta \,\Big|\, \|g_t(\theta) - g_t(\widehat{\theta}_t)\|_{H_t^{-1}(\theta)}$$

$$\leq \frac{2L^2 S k_\mu}{\sqrt{\lambda c_\mu}}\frac{\gamma^D}{1-\gamma} + \frac{Lm}{\sqrt{\lambda c_\mu}}\frac{\gamma^D}{1-\gamma} + \breve{\beta}_t\Bigg\},$$

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

The proof of Lemma 14 is presented in Appendix D-B.

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,

$$(X_t, \widetilde{\theta}_t) = \arg\max_{\mathbf{x}\in\mathcal{X}, \theta\in\mathcal{C}_t(\delta)}\mu(\mathbf{x}^\top\theta). \tag{54}$$

The main difference between parameter-based and bonus-based selection criteria is discussed in Section 3.2 of [34]. The overall algorithm is summarized in Algorithm 7.

B. Proof of Lemma 14

Proof. Based on the model assumption (16), the function g_t (13) and the $g_t(\widehat{\theta}_t)$ (48), we have,

$$g_t(\theta_t) - g_t(\widehat{\theta}_t)$$

$$= \lambda c_\mu \theta_t + \sum_{s=1}^{t-1}\gamma^{t-s-1}\mu(X_s^\top\theta_t)X_s - \sum_{s=1}^{t-1}\gamma^{t-s-1}r_s X_s$$

$$= \lambda c_\mu \theta_t + \sum_{s=1}^{t-1} \gamma^{t-s-1} \mu(X_s^\top \theta_t) X_s$$

$$- \sum_{s=1}^{t-1} \gamma^{t-s-1} (\mu(X_s^\top \theta_s) + \eta_s) X_s$$

$$= \sum_{s=1}^{t-1} \gamma^{t-s-1} (\mu(X_s^\top \theta_t) - \mu(X_s^\top \theta_s)) X_s + \lambda c_\mu \theta_t$$

$$- \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s.$$

Then,

$$\| g_t(\theta_t) - g_t(\widehat{\theta}_t) \|_{H_t^{-1}(\theta_t)}$$

$$= \left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} (\mu(X_s^\top \theta_t) - \mu(X_s^\top \theta_s)) X_s \right.$$

$$\left. + \lambda c_\mu \theta_t - \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s \right\|_{H_t^{-1}(\theta_t)}$$

$$\leq \left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} (\mu(X_s^\top \theta_t) - \mu(X_s^\top \theta_s)) X_s \right\|_{H_t^{-1}(\theta_t)}$$

$$+ \left\| \lambda c_\mu \theta_t - \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s \right\|_{H_t^{-1}(\theta_t)}$$

$$\leq \underbrace{\left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} (\mu(X_s^\top \theta_t) - \mu(X_s^\top \theta_s)) X_s \right\|_{H_t^{-1}(\theta_t)}}_{\text{TERM (A)}}$$

$$+ \underbrace{\left\| \sum_{s=1}^{t-D-1} \gamma^{t-s-1} \eta_s X_s \right\|_{H_t^{-1}(\theta_t)}}_{\text{TERM (B)}}$$

$$+ \underbrace{\left\| \sum_{s=t-D}^{t-1} \gamma^{t-s-1} \eta_s X_s - \lambda c_\mu \theta_t \right\|_{H_t^{-1}(\theta_t)}}_{\text{TERM (C)}}.$$

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

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

$$= \left\| \sum_{s=1}^{t-D-1} \gamma^{t-s-1} (\mu(X_s^\top \theta_t) - \mu(X_s^\top \theta_s)) X_s \right\|_{H_t^{-1}(\theta_t)}$$

$$\leq \left\| \sum_{s=1}^{t-D-1} \gamma^{t-s-1} k_\mu X_s^\top (\theta_t - \theta_s) X_s \right\|_{H_t^{-1}(\theta_t)}$$

$$\leq \sum_{s=1}^{t-D-1} \gamma^{t-s-1} k_\mu \| X_s \|_2 \| (\theta_t - \theta_s) \|_2 \| X_s \|_{H_t^{-1}(\theta_t)}$$

$$\leq \frac{2 L^2 S k_\mu}{\sqrt{\lambda c_\mu}} \frac{\gamma^D}{1-\gamma}.$$

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). We define $\widetilde{\eta}_s \triangleq \frac{\sqrt{\gamma^{t-s-1}} \eta_s}{|m|}$, $\widetilde{X}_s \triangleq \sqrt{\gamma^{t-s-1}} X_s$ and notice that $\forall t \in [T], s \in [t-1], \lfloor \gamma^{t-s-1} \rfloor \leq 1$, then $\widetilde{\eta}_s$ is bounded by 1 with variance $\widetilde{\sigma}_s$. Let $\widetilde{H}_t(\theta) \triangleq \frac{\lambda c_\mu}{m^2} I_d + \sum_{s=1}^{t-1} \frac{\mu'(X_s^\top \theta)}{m^2} \widetilde{X}_s \widetilde{X}_s^\top$ and $\widetilde{H}_{t-D:t}(\theta) = \lambda c_\mu I_d + \sum_{s=t-D}^{t-1} \frac{\mu'(X_s^\top \theta)}{m^2} \widetilde{X}_s \widetilde{X}_s^\top$,

$$\left\| \sum_{s=t-D}^{t-1} \gamma^{t-s-1} \eta_s X_s - \lambda c_\mu \theta_t \right\|_{H_t^{-1}(\theta_t)}$$

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

$$\leq \left\| \sum_{s=t-D}^{t-1} \widetilde{\eta}_s \widetilde{X}_s \right\|_{\widetilde{H}_{t-D:t}^{-1}(\theta_t)} + \sqrt{\lambda c_\mu} S,$$

where $\widetilde{H}_t(\theta) \succeq \widetilde{H}_{t-D:t}(\theta)$. Next, we need to bound the term $\| \sum_{s=t-D}^{t-1} \widetilde{\eta}_s \widetilde{X}_s \|_{\widetilde{H}_{t-D:t}^{-1}(\theta_t)}$ using self-normalization bound [33, Theorem 1], restated in Theorem 8, similar to the proof of Lemma 13, we have

$$\left\| \sum_{s=t-D}^{t-1} \gamma^{-s} \eta_s X_s \right\|_{\widetilde{H}_{t-D:t}^{-1}(\theta_t)}$$

$$\leq \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}} \log \left(\frac{\det \left(\widetilde{H}_{t-D:t} \right)^{1/2}}{\delta (\lambda c_\mu)^{d/2}} \right) + \frac{2m}{\sqrt{\lambda c_\mu}} d \log(2)$$

$$\leq \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}} \log \frac{1}{\delta} + \frac{dm}{\sqrt{\lambda c_\mu}} \log \left(1 + \frac{L^2 k_\mu (1 - \gamma^{2D})}{\lambda c_\mu d (1 - \gamma)} \right)$$

$$+ \frac{2m}{\sqrt{\lambda c_\mu}} d \log(2).$$

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

$$\| g_t(\theta_t) - g_t(\widehat{\theta}_t) \|_{H_t^{-1}(\theta_t)}$$

$$\leq \frac{2 L^2 S k_\mu}{\sqrt{\lambda c_\mu}} \frac{\gamma^D}{1-\gamma} + \frac{Lm}{\sqrt{\lambda c_\mu}} \frac{\gamma^D}{1-\gamma} + \breve{\beta}_t,$$

which completes the proof. $\qquad \square$

C. Proof of Theorem 4

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

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

For $t \in \mathcal{T}(D)$, by selection criterion (54),

$$R_t = \mu(X_t^{*\top}\theta_t) - \mu(X_t^\top\theta_t)$$
$$\leq \mu(X_t^\top\widetilde{\theta}_t) - \mu(X_t^\top\widehat{\theta}_t) + \mu(X_t^\top\widehat{\theta}_t) - \mu(X_t^\top\theta_t)$$
$$\leq \alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t)\left|X_t^\top\left(\widetilde{\theta}_t - \widehat{\theta}_t\right)\right| + \alpha(X_t,\theta_t,\widehat{\theta}_t)\left|X_t^\top\left(\theta_t - \widehat{\theta}_t\right)\right|$$
$$\leq \sqrt{1+2S}$$
$$\cdot \Bigg(\alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t) \|X_t\|_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)} \left\|g_t(\widetilde{\theta}_t) - g_t(\widehat{\theta}_t)\right\|_{H_t^{-1}(\widetilde{\theta}_t)}$$
$$+ \alpha(X_t,\theta_t,\widehat{\theta}_t) \|X_t\|_{G_t^{-1}(\theta_t,\widehat{\theta}_t)} \left\|g_t(\theta_t) - g_t(\widehat{\theta}_t)\right\|_{H_t^{-1}(\theta_t)} \Bigg).$$

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 $1-\delta$,

$$\sum_{t\in\mathcal{T}(D)} R_t \leq \sum_{t\in\mathcal{T}(D)} \sqrt{1+2S}\big(\alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t) \|X_t\|_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}\rho_t$$
$$+ \alpha(X_t,\theta_t,\widehat{\theta}_t) \|X_t\|_{G_t^{-1}(\theta_t,\widehat{\theta}_t)}\rho_t\big)$$
$$\leq \sqrt{1+2S}\rho_T\Bigg(\sum_{t\in\mathcal{T}(D)} \alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t) \|X_t\|_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}$$
$$+ \sum_{t\in\mathcal{T}(D)} \alpha(X_t,\theta_t,\widehat{\theta}_t) \|X_t\|_{G_t^{-1}(\theta_t,\widehat{\theta}_t)} \Bigg).$$

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

Based on the definition of g_t (13), we have

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

Then based on the definition of G_t (49), 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

$$\sum_{t\in\mathcal{T}(D)} \alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t) \|X_t\|_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}$$

$$\leq \sqrt{\sum_{t=1}^T \alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t)}\sqrt{\sum_{t=1}^T \alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t)\|X_t\|_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}^2}$$
$$\leq \sqrt{k_\mu T}\sqrt{\sum_{t=1}^T \left\|\widetilde{X}_t\right\|_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}^2}.$$

Then for the term $\sqrt{\sum_{t=1}^T \|\widetilde{X}_t\|_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}^2}$, we can directly use the Lemma 9 to bound it, let $C_T^{\text{PWSCB}} \triangleq \sqrt{2k_\mu \max\{1, L^2 k_\mu/(\lambda c_\mu)\}dT}$ we have

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

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,

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

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

$$\text{D-REG}_T \leq \Gamma_T D$$
$$+ 2\sqrt{1+2S}\rho_T C_T^{\text{PWSCB}}\sqrt{T\log\frac{1}{\gamma} + \log\left(1 + \frac{L^2 k_\mu}{\lambda c_\mu d(1-\gamma)}\right)},$$

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

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

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-degenerate case of $\Gamma_T > d/\sqrt{T}$, We set the discounted

factor optimally as $1 - \gamma = (\Gamma_T/(dT))^{2/3}$ and attain an $\widetilde{\mathcal{O}}(d^{2/3}\Gamma_T^{1/3}T^{2/3})$ regret, which completes the proof. $\square$

APPENDIX E
ANALYSIS OF WEIGHTUCRL

A. Proof of Lemma 4

Proof. Fix $h \in [H]$, based on the reward model assumption (22) and the estimator (23), the estimation error of reward estimation can be decomposed as

$$\widehat{\theta}_h^k - \theta_h^k$$
$$= \left(\Lambda_h^{k-1}\right)^{-1}\left(\sum_{j=1}^{k-1} w_{k-1,j} r_h^j(s_h^j, a_h^j)\phi(s_h^j, a_h^j)\right) - \theta_h^k$$
$$= \left(\Lambda_h^{k-1}\right)^{-1}\left(\sum_{j=1}^{k-1} w_{k-1,j}\phi(s_h^j, a_h^j)^\top \theta_h^j \phi(s_h^j, a_h^j)\right) - \theta_h^k$$
$$= \left(\Lambda_h^{k-1}\right)^{-1}\left(\sum_{j=1}^{k-1} w_{k-1,j}\phi(s_h^j, a_h^j)\phi(s_h^j, a_h^j)^\top \theta_h^j\right)$$
$$\quad - \left(\Lambda_h^{k-1}\right)^{-1}\left(\lambda_\theta I_d + \sum_{j=1}^{k-1} w_{k-1,j}\phi(s_h^j, a_h^j)\phi(s_h^j, a_h^j)^\top\right)\theta_h^k$$
$$= \underbrace{\left(\Lambda_h^{k-1}\right)^{-1}\left(\sum_{j=1}^{k-1} w_{k-1,j}\phi(s_h^j, a_h^j)\phi(s_h^j, a_h^j)^\top \left(\theta_h^j - \theta_h^k\right)\right)}_{\text{bias part}}$$
$$\quad \underbrace{- \left(\Lambda_h^{k-1}\right)^{-1}\lambda_\theta \theta_h^k}_{\text{variance part}}.$$

Then, by the Cauchy-Schwarz inequality, we know that for any $s \in \mathcal{S}, a \in \mathcal{A}$,
$$\left|\phi(s,a)^\top\left(\widehat{\theta}_h^k - \theta_h^k\right)\right| = \|\phi(s,a)\|_{(\Lambda_h^{k-1})^{-1}}\left(A_h^k + B_h^k\right),$$
$$(55)$$

where
$$A_h^k = \left\|\sum_{j=1}^{k-1} w_{k-1,j}\phi(s_h^j, a_h^j)\phi(s_h^j, a_h^j)^\top\left(\theta_h^j - \theta_h^k\right)\right\|_{(\Lambda_h^{k-1})^{-1}},$$
$$B_h^k = \left\|\lambda_\theta \theta_h^k\right\|_{(\Lambda_h^{k-1})^{-1}}.$$

The above two terms can be bounded separately,

Term A_h^k. The first step is to extract the variations of the parameter θ_h^k as follows,

$$\left\|\sum_{j=1}^{k-1} w_{k-1,j}\phi(s_h^j, a_h^j)\phi(s_h^j, a_h^j)^\top\left(\theta_h^j - \theta_h^k\right)\right\|_{(\Lambda_h^{k-1})^{-1}}$$
$$= \left\|\sum_{j=1}^{k-1} w_{k-1,j}\phi(s_h^j, a_h^j)\phi(s_h^j, a_h^j)^\top \sum_{p=j}^{k-1}\left(\theta_h^p - \theta_h^{p+1}\right)\right\|_{(\Lambda_h^{k-1})^{-1}}$$
$$= \left\|\sum_{p=1}^{k-1}\sum_{j=1}^{p} w_{k-1,j}\phi(s_h^j, a_h^j)\phi(s_h^j, a_h^j)^\top\left(\theta_h^p - \theta_h^{p+1}\right)\right\|_{(\Lambda_h^{k-1})^{-1}}$$

$$\le \sum_{p=1}^{k-1}\left\|\sum_{j=1}^{p} w_{k-1,j}\phi(s_h^j, a_h^j)\left\|\phi(s_h^j, a_h^j)\right\|_2\left\|\theta_h^p - \theta_h^{p+1}\right\|_2\right\|_{(\Lambda_h^{k-1})^{-1}}$$
$$\le L_\phi\sum_{p=1}^{k-1}\sum_{j=1}^{p} w_{k-1,j}\left\|\phi(s_h^j, a_h^j)\right\|_{(\Lambda_h^{k-1})^{-1}}\left\|\theta_h^p - \theta_h^{p+1}\right\|_2,$$

and term $\sum_{j=1}^{p} w_{k-1,j}\left\|\phi(s_h^j, a_h^j)\right\|_{(\Lambda_h^{k-1})^{-1}}$ can be able to further derive an expression about weight $w_{k-1,j}$ as follows,

$$\sum_{j=1}^{p} w_{k-1,j}\left\|\phi(s_h^j, a_h^j)\right\|_{(\Lambda_h^{k-1})^{-1}}$$
$$\le \sqrt{\sum_{j=1}^{p} w_{k-1,j}}\sqrt{\sum_{j=1}^{p} w_{k-1,j}\left\|\phi(s_h^j, a_h^j)\right\|^2_{(\Lambda_h^{k-1})^{-1}}}$$
$$\le \sqrt{d}\sqrt{\sum_{j=1}^{p} w_{k-1,j}}.$$

In above, the first step holds by the Cauchy-Schwarz inequality. Besides, the last step follows the Lemma 25 by letting $X_j = \sqrt{w_{k-1,j}}\phi(s_h^j, a_h^j)$ and $U_{k-1} = \Lambda_h^{k-1}$, which means for Term 1 we have

$$\left\|\sum_{j=1}^{k-1} w_{k-1,j}\phi(s_h^j, a_h^j)\phi(s_h^j, a_h^j)^\top\left(\theta_h^j - \theta_h^k\right)\right\|_{(\Lambda_h^{k-1})^{-1}}$$
$$\le L_\phi\sqrt{d}\sum_{p=1}^{k-1}\sqrt{\sum_{j=1}^{p} w_{k-1,j}}\left\|\theta_h^p - \theta_h^{p+1}\right\|_2, \quad (56)$$

Term B_h^k.
$$\left\|\lambda_\theta \theta_h^k\right\|_{(\Lambda_h^{k-1})^{-1}} \le \frac{\lambda_\theta}{\sqrt{\lambda_{\min}(\Lambda_h^{k-1})}}\left\|\theta_h^k\right\|_2 \le \sqrt{\lambda_\theta}S_\theta. \quad (57)$$

Plug Eq (56) and Eq (57) into Eq (55) and we have
$$\left\|\widehat{\theta}_h^k - \theta_h^k\right\|_{\Lambda_h^{k-1}}$$
$$\le L_\phi\sqrt{d}\sum_{p=1}^{k-1}\sqrt{\sum_{j=1}^{p} w_{k-1,j}}\left\|\theta_h^p - \theta_h^{p+1}\right\|_2 + \sqrt{\lambda_\theta}S_\theta,$$

further we have
$$\left|\phi(s,a)^\top\left(\widehat{\theta}_h^k - \theta_h^k\right)\right|$$
$$\le \|\phi(s,a)\|_{(\Lambda_h^{k-1})^{-1}}$$
$$\quad \cdot \left(L_\phi\sqrt{d}\sum_{p=1}^{k-1}\sqrt{\sum_{j=1}^{p} w_{k-1,j}}\left\|\theta_h^p - \theta_h^{p+1}\right\|_2 + \sqrt{\lambda_\theta}S_\theta\right)$$
$$\le L_\phi^2\sqrt{\frac{d}{\lambda_\theta}}\sum_{p=1}^{k-1}\sqrt{\sum_{j=1}^{p} w_{k-1,j}}\left\|\theta_h^p - \theta_h^{p+1}\right\|_2$$
$$\quad\quad + \beta_\theta\|\phi(s,a)\|_{(\Lambda_h^{k-1})^{-1}},$$

where $\|\phi(s,a)\|_{(\Lambda_h^{k-1})^{-1}} \le \|\phi(s,a)\|_2/\sqrt{\lambda_\theta}$ and $\beta_\theta \triangleq \sqrt{\lambda_\theta}S_\theta$, which completes the proof. $\square$

B. Proof of Lemma 5

Proof. Fix $h \in [H]$, based on the model assumption (22) and estimator (25), we have

$$
\widehat{\mathbf{w}}_h^k = \left(\Sigma_h^{k-1}\right)^{-1} \left(\sum_{j=1}^{k-1} \alpha_{k-1,j} V_{h+1}^j(s_{h+1}^j) \psi_{h+1}^j\left(s_h^j, a_h^j\right)\right)
$$

$$
= \left(\Sigma_h^{k-1}\right)^{-1} \left(\sum_{j=1}^{k-1} \alpha_{k-1,j} \left(\psi_{h+1}^j\left(s_h^j, a_h^j\right)^\top \mathbf{w}_h^j + \eta_{h+1}^j\right)\right.
$$
$$
\left. \cdot \psi_{h+1}^j\left(s_h^j, a_h^j\right)\right), \qquad (58)
$$

where we define the noise as $\eta_{h+1}^j \triangleq V_{h+1}^j\left(s_{h+1}^j\right) - \left[\mathbb{P}_h^j V_{h+1}^j\right]\left(s_h^j, a_h^j\right)$, and we have

$$
\mathbf{w}_h^k = \left(\Sigma_h^{k-1}\right)^{-1} \left(\lambda_{\mathbf{w}} I_d\right.
$$
$$
\left. + \sum_{j=1}^{k-1} \alpha_{k-1,j} \psi_{h+1}^j\left(s_h^j, a_h^j\right) \psi_{h+1}^j\left(s_h^j, a_h^j\right)^\top\right) \mathbf{w}_h^k, \qquad (59)
$$

combine Eq (58) and Eq (59) and we have the estimation error

$$
\widehat{\mathbf{w}}_h^k - \mathbf{w}_h^k
$$
$$
= \underbrace{\left(\Sigma_h^{k-1}\right)^{-1} \left(\sum_{j=1}^{k-1} \alpha_{k-1,j} \psi_{h+1}^j\left(s_h^j, a_h^j\right) \psi_{h+1}^j\left(s_h^j, a_h^j\right)^\top \right.}_{\text{bias part}}
$$
$$
\left. \cdot \left(\mathbf{w}_h^j - \mathbf{w}_h^k\right)\right)
$$
$$
+ \underbrace{\left(\Sigma_h^{k-1}\right)^{-1} \left(\sum_{j=1}^{k-1} \alpha_{k-1,j} \eta_{h+1}^j \psi_{h+1}^j\left(s_h^j, a_h^j\right) - \lambda_{\mathbf{w}} \mathbf{w}_h^k\right)}_{\text{variance part}}.
$$

Then, by the Cauchy-Schwarz inequality, we know that for any $s \in \mathcal{S}, a \in \mathcal{A}$,

$$
\left|\psi_{h+1}^k\left(s, a\right)^\top \left(\widehat{\mathbf{w}}_h^k - \mathbf{w}_h^k\right)\right|
$$
$$
\leq \left\|\psi_{h+1}^k\left(s, a\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}} \left(C_h^k + D_h^k\right), \qquad (60)
$$

where

$$
C_h^k = \left\|\sum_{j=1}^{k-1} \alpha_{k-1,j} \right.
$$
$$
\left. \cdot \psi_{h+1}^j\left(s_h^j, a_h^j\right) \psi_{h+1}^j\left(s_h^j, a_h^j\right)^\top \left(\mathbf{w}_h^j - \mathbf{w}_h^k\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}
$$

$$
D_h^k = \left\|\sum_{j=1}^{k-1} \alpha_{k-1,j} \eta_{h+1}^j \psi_{h+1}^j\left(s_h^j, a_h^j\right) - \lambda_{\mathbf{w}} \mathbf{w}_h^k\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}.
$$

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

Lemma 15. *For any $k \in [K]$, we have*

$$
C_h^k \leq H L_\psi \sqrt{d} \sum_{p=1}^{k-1} \sqrt{\sum_{j=1}^p \alpha_{k-1,j}} \left\|\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right\|_2.
$$

Lemma 16. *If $\forall k, j \in [K], \alpha_{k-1,j} \leq 1$, for any $\delta \in (0,1)$, with probability at least $1 - \delta$, the following holds for all $k \in [K]$,*

$$
D_h^k \leq H \sqrt{\frac{1}{2} \log \frac{1}{\delta} + \frac{d}{4} \log \left(1 + \frac{H^2 L_\psi^2 \sum_{j=1}^{k-1} \alpha_{k-1,j}}{\lambda_{\mathbf{w}} d}\right)}
$$
$$
+ \sqrt{\lambda_{\mathbf{w}}} S_{\mathbf{w}}.
$$

Based on the inequality (60), Lemma 15, Lemma 16, and $\left\|\psi_{h+1}^k\left(s, a\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}} \leq \left\|\psi_{h+1}^k\left(s, a\right)\right\|_2 / \sqrt{\lambda_{\mathbf{w}}} \leq H L_\psi / \sqrt{\lambda_{\mathbf{w}}}$, with probability at least $1 - \delta$, the following holds for all $k \in [K]$,

$$
\left|\psi_{h+1}^k\left(s, a\right)^\top \left(\widehat{\mathbf{w}}_h^k - \mathbf{w}_h^k\right)\right|
$$
$$
\leq \Gamma_{h,\mathbf{w}}^{k-1} + \beta_{\mathbf{w}}^{k-1} \left\|\psi_{h+1}^k\left(s, a\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}},
$$

where

$$
\Gamma_{h,\mathbf{w}}^{k-1} \triangleq H^2 L_\psi^2 \sqrt{\frac{d}{\lambda_{\mathbf{w}}}} \sum_{p=1}^{k-1} \sqrt{\sum_{j=1}^p \alpha_{k-1,j}} \left\|\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right\|_2
$$

$$
\beta_{\mathbf{w}}^{k-1} \triangleq H \sqrt{\frac{1}{2} \log \frac{1}{\delta} + \frac{d}{4} \log \left(1 + \frac{H^2 L_\psi^2 \sum_{j=1}^{k-1} \alpha_{k-1,j}}{\lambda_{\mathbf{w}} d}\right)}
$$
$$
+ \sqrt{\lambda_{\mathbf{w}}} S_{\mathbf{w}},
$$

which completes the proof. $\square$

Proof of Lemma 15. The first step is to extract the variations of the parameter $\mathbf{w}_h^k$ as follows,

$$
\left\|\sum_{j=1}^{k-1} \alpha_{k-1,j} \psi_{h+1}^j\left(s_h^j, a_h^j\right)\right.
$$
$$
\left. \cdot \psi_{h+1}^j\left(s_h^j, a_h^j\right)^\top \left(\mathbf{w}_h^j - \mathbf{w}_h^k\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}
$$
$$
= \left\|\sum_{j=1}^{k-1} \alpha_{k-1,j} \psi_{h+1}^j\left(s_h^j, a_h^j\right)\right.
$$
$$
\left. \cdot \psi_{h+1}^j\left(s_h^j, a_h^j\right)^\top \sum_{p=j}^{k-1} \left(\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}
$$
$$
= \left\|\sum_{p=1}^{k-1} \sum_{j=1}^p \alpha_{k-1,j} \psi_{h+1}^j\left(s_h^j, a_h^j\right)\right.
$$
$$
\left. \cdot \psi_{h+1}^j\left(s_h^j, a_h^j\right)^\top \left(\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}
$$
$$
\leq \sum_{p=1}^{k-1} \left\|\sum_{j=1}^p \alpha_{k-1,j} \psi_{h+1}^j\left(s_h^j, a_h^j\right)\right.
$$
$$
\left. \cdot \psi_{h+1}^j\left(s_h^j, a_h^j\right)^\top \left(\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}
$$

$$\leq HL_\psi \sum_{p=1}^{k-1} \sum_{j=1}^{p} \alpha_{k-1,j} \left\| \psi_{h+1}^j \left(s_h^j, a_h^j \right) \right\|_{(\Sigma_h^{k-1})^{-1}}$$
$$\cdot \left\| \mathbf{w}_h^p - \mathbf{w}_h^{p+1} \right\|_2 ,$$

and term $\sum_{j=1}^{p} \alpha_{k-1,j} \left\| \psi_{h+1}^j \left(s_h^j, a_h^j \right) \right\|_{(\Sigma_h^{k-1})^{-1}}$ can be able to further derive an expression about weight $\alpha_{k-1,j}$ as follows,

$$\sum_{j=1}^{p} \alpha_{k-1,j} \left\| \psi_{h+1}^j \left(s_h^j, a_h^j \right) \right\|_{(\Sigma_h^{k-1})^{-1}}$$
$$\leq \sqrt{\sum_{j=1}^{p} \alpha_{k-1,j}} \sqrt{\sum_{j=1}^{p} \alpha_{k-1,j} \left\| \psi_{h+1}^j \left(s_h^j, a_h^j \right) \right\|_{(\Sigma_h^{k-1})^{-1}}^2}$$
$$\leq \sqrt{d} \sqrt{\sum_{j=1}^{p} \alpha_{k-1,j}}.$$

In above, the second last step holds by the Cauchy-Schwarz inequality. Besides, the last step follows the Lemma 25 by letting $X_j = \sqrt{\alpha_{k-1,j}} \psi_{h+1}^j \left(s_h^j, a_h^j \right)$ and $U_{k-1} = \Sigma_h^{k-1}$. Hence we complete the proof. $\qquad \square$

Proof of Lemma 16.

$$\left\| \sum_{j=1}^{k-1} \alpha_{k-1,j} \eta_{h+1}^j \psi_{h+1}^j \left(s_h^j, a_h^j \right) - \lambda_\mathbf{w} \mathbf{w}_h^k \right\|_{(\Sigma_h^{k-1})^{-1}}$$
$$\leq \left\| \sum_{j=1}^{k-1} \alpha_{k-1,j} \eta_{h+1}^j \psi_{h+1}^j \left(s_h^j, a_h^j \right) \right\|_{(\Sigma_h^{k-1})^{-1}}$$
$$+ \left\| \lambda_\mathbf{w} \mathbf{w}_h^k \right\|_{(\Sigma_h^{k-1})^{-1}}$$
$$\leq \left\| \sum_{j=1}^{k-1} \alpha_{k-1,j} \eta_{h+1}^j \psi_{h+1}^j \left(s_h^j, a_h^j \right) \right\|_{(\Sigma_h^{k-1})^{-1}} + \sqrt{\lambda_\mathbf{w}} S_\mathbf{w}.$$

Let $\widetilde{\eta}_{h+1}^j \triangleq \sqrt{\alpha_{k-1,j}} \eta_{h+1}^j$ and $X_j \triangleq \sqrt{\alpha_{k-1,j}} \psi_{h+1}^j \left(s_h^j, a_h^j \right)$, then we have notice that since the reward $r \in [0,1]$, and $\alpha_{k-1,j} \leq 1$, the noise $\widetilde{\eta}_{h+1}^j$ is bounded by:

$$\widetilde{\eta}_{h+1}^j = \sqrt{\alpha_{k-1,j}} \left(V_{h+1}^j \left(s_{h+1}^j \right) - \left[\mathbb{P}_h V_{h+1}^j \right] \left(s_h^j, a_h^j \right) \right) \leq H,$$

based on Lemma 23, we find that the noise $\widetilde{\eta}_{h+1}^j$ is $\frac{H}{2}$-sub-Gaussian. Then, by Theorem 7, we have with probability at least $1 - \delta$, the following holds for all $k \in [K]$.

$$\left\| \sum_{j=1}^{k-1} \widetilde{\eta}_{h+1}^j X_j \right\|_{(\Sigma_h^{k-1})^{-1}}$$
$$\leq \sqrt{\frac{H^2}{2} \log \left(\frac{\det(\Sigma_h^{k-1})^{\frac{1}{2}} \det(\Sigma_{0,h})^{-\frac{1}{2}}}{\delta} \right)}$$

where

$$\det(\Sigma_h^{k-1}) \leq \left(\frac{\mathrm{trace}(\Sigma_h^{k-1})}{d} \right)^d$$
$$= \left(\frac{d\lambda_\mathbf{w} + \sum_{j=1}^{k-1} \left\| \alpha_{k-1,j} \psi_{h+1}^j \left(s_h^j, a_h^j \right) \right\|_2^2}{d} \right)^d$$
$$= \left(\frac{d\lambda_\mathbf{w} + H^2 L_\psi^2 \sum_{j=1}^{k-1} \alpha_{k-1,j}}{d} \right)^d$$
$$\det(\Sigma_{0,h}) \leq \lambda_\mathbf{w}^d,$$

so we have

$$\left\| \sum_{j=1}^{k-1} \alpha_{k-1,j} \eta_{h+1}^j \psi_{h+1}^j \left(s_h^j, a_h^j \right) \right\|_{(\Sigma_h^{k-1})^{-1}}$$
$$\leq H \sqrt{\frac{1}{2} \log \frac{1}{\delta} + \frac{d}{4} \log \left(1 + \frac{H^2 L_\psi^2 \sum_{j=1}^{k-1} \alpha_{k-1,j}}{\lambda_\mathbf{w} d} \right)}.$$

which completes the proof. $\qquad \square$

C. Proof of Theorem 5

Proof. To prove the theorem, we first introduce the following lemma

Lemma 17. *We define the model prediction error as*

$$E_h^k(s,a) = r_h^k(s,a) + \mathbb{P}_h^k V_{h+1}^k(s,a) - Q_h^k(s,a), \qquad (61)$$

then with probability at least $1 - 2\delta$, the following holds for all $k \in [K]$, $h \in [H]$ and $\forall s \in \mathcal{S}, a \in \mathcal{A}$,

$$-2\beta_\theta \left\| \phi(s,a) \right\|_{(\Lambda_h^{k-1})^{-1}} - 2\beta_\mathbf{w}^{k-1} \left\| \psi_{h+1}^k(s,a) \right\|_{(\Sigma_h^{k-1})^{-1}}$$
$$- \Gamma_{h,\theta}^{k-1} - \Gamma_{h,\mathbf{w}}^{k-1} \leq E_h^k(s,a) \leq \Gamma_{h,\theta}^{k-1} + \Gamma_{h,\mathbf{w}}^{k-1}.$$

We can further connect the dynamic regret to the model prediction error, by the following Lemma.

Lemma 18. *For the policies $\{\pi_h^k\}_{h \in [H], k \in [K]}$ with $a_h^k = \arg\max_{a \in \mathcal{A}} Q_h^k(s_h^k, a)$, and the optimal policies and $\delta \in (0,1)$, we have the following decomposition holds with probability at least $1 - 2\delta$,*

$$\text{D-REG}_T \leq \sum_{k=1}^{K} \sum_{h=1}^{H} \left(\mathbb{E}_{\pi_{*,h}^k} \left[E_h^k(s_h^k, a_h^k) \right] - E_h^k(s_h^k, a_h^k) \right)$$
$$+ 4H \sqrt{2T \log(1/\delta)}.$$

Based on Lemma 18 and notice that $\forall s \in \mathcal{S}, a \in \mathcal{A},$ $\left| E_h^k(s,a) \right| \le 2H$, we have

$$\mathbb{E}_{\pi_{*,h}^k}\left[E_h^k(s_h^k, a_h^k)\right] - E_h^k(s_h^k, a_h^k)$$

$$\le \min\left\{4H, 2\Gamma_{h,\theta}^{k-1} + 2\Gamma_{h,\mathbf{w}}^{k-1} + 2\beta_\theta \left\|\phi(s_h^k, a_h^k)\right\|_{(\Lambda_h^{k-1})^{-1}} \right.$$
$$\left. + 2\beta_{\mathbf{w}}^{k-1}\left\|\psi_{h+1}^k(s_h^k, a_h^k)\right\|_{(\Sigma_h^{k-1})^{-1}}\right\}$$

$$\le 2\Gamma_{h,\theta}^{k-1} + 2\Gamma_{h,\mathbf{w}}^{k-1} + \min\left\{4H, 2\beta_\theta\left\|\phi(s_h^k, a_h^k)\right\|_{(\Lambda_h^{k-1})^{-1}}\right\}$$
$$+ \min\left\{4H, 2\beta_{\mathbf{w}}^{k-1}\left\|\psi_{h+1}^k(s_h^k, a_h^k)\right\|_{(\Sigma_h^{k-1})^{-1}}\right\}$$

$$\le 2\Gamma_{h,\theta}^{k-1} + 2\Gamma_{h,\mathbf{w}}^{k-1} + 4H\beta_\theta \min\left\{1, \left\|\phi(s_h^k, a_h^k)\right\|_{(\Lambda_h^{k-1})^{-1}}\right\}$$
$$+ 4\beta_{\mathbf{w}}^{k-1}\min\left\{1, \left\|\psi_{h+1}^k(s_h^k, a_h^k)\right\|_{(\Sigma_h^{k-1})^{-1}}\right\},$$

the last inequality comes from that $\beta_\theta > 1, \beta_{\mathbf{w}}^{k-1} \ge H$. so we have with probability at least $1 - 4\delta$,

$$\text{D-REG}_T \le \underbrace{4H\beta_\theta \sum_{k=1}^K \sum_{h=1}^H \min\left\{1, \left\|\phi(s_h^k, a_h^k)\right\|_{(\Lambda_h^{k-1})^{-1}}\right\}}_{\text{variance part1}}$$

$$+ \underbrace{4\sum_{k=1}^K \sum_{h=1}^H \beta_{\mathbf{w}}^{k-1}\min\left\{1, \left\|\psi_{h+1}^k(s_h^k, a_h^k)\right\|_{(\Sigma_h^{k-1})^{-1}}\right\}}_{\text{variance part2}}$$

$$+ \underbrace{2\sum_{k=1}^K\sum_{h=1}^H \Gamma_{h,\theta}^{k-1} + 2\sum_{k=1}^K\sum_{h=1}^H \Gamma_{h,\mathbf{w}}^{k-1}}_{\text{bias part}} + 4H\sqrt{2T\log(1/\delta)}.$$

Bias. Now we set $w_{k,j} = \gamma^{k-j}, \gamma \in (0,1)$,

$$2\sum_{k=1}^K\sum_{h=1}^H \Gamma_{h,\theta}^{k-1}$$

$$= 2L_\phi^2 \sqrt{\frac{d}{\lambda_\theta}} \sum_{k=1}^K\sum_{h=1}^H\sum_{p=1}^{k-1} \sqrt{\sum_{j=1}^p w_{k-1,j}} \left\|\theta_h^p - \theta_h^{p+1}\right\|_2$$

$$= 2L_\phi^2 \sqrt{\frac{d}{\lambda_\theta}} \sum_{p=1}^{K-1}\sum_{h=1}^H\sum_{k=p+1}^K \sqrt{\sum_{j=1}^p w_{k-1,j}} \left\|\theta_h^p - \theta_h^{p+1}\right\|_2$$

$$= 2L_\phi^2 \sqrt{\frac{d}{\lambda_\theta}} \sum_{p=1}^{K-1}\sum_{h=1}^H\sum_{k=p+1}^K \gamma^{\frac{k-1}{2}}\sqrt{\sum_{j=1}^p \gamma^{-j}} \left\|\theta_h^p - \theta_h^{p+1}\right\|_2$$

$$= 2L_\phi^2 \sqrt{\frac{d}{\lambda_\theta}} \sum_{p=1}^{K-1}\sum_{h=1}^H \frac{\gamma^{\frac{p}{2}} - \gamma^{\frac{K}{2}}}{1 - \gamma^{\frac{1}{2}}}\sqrt{\frac{\gamma^{-p} - 1}{1 - \gamma}} \left\|\theta_h^p - \theta_h^{p+1}\right\|_2$$

$$\le 4L_\phi^2 \sqrt{\frac{d}{\lambda_\theta}} \frac{1}{(1-\gamma)^{3/2}} \sum_{p=1}^{K-1}\sum_{h=1}^H \left\|\theta_h^p - \theta_h^{p+1}\right\|_2, \quad (62)$$

then we set $\alpha_{k,j} = \gamma^{k-j}, \gamma \in (0,1)$ and have

$$2\sum_{k=1}^K\sum_{h=1}^H \Gamma_{h,\mathbf{w}}^{k-1}$$

$$= 2H^2 L_\psi^2 \sqrt{\frac{d}{\lambda_{\mathbf{w}}}} \sum_{k=1}^K\sum_{h=1}^H\sum_{p=1}^{k-1} \sqrt{\sum_{j=1}^p \alpha_{k-1,j}} \left\|\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right\|_2$$

$$= 2H^2 L_\psi^2 \sqrt{\frac{d}{\lambda_{\mathbf{w}}}} \sum_{p=1}^{K-1}\sum_{h=1}^H\sum_{k=p+1}^K \sqrt{\sum_{j=1}^p \alpha_{k-1,j}} \left\|\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right\|_2$$

$$= 2H^2 L_\psi^2 \sqrt{\frac{d}{\lambda_{\mathbf{w}}}} \sum_{p=1}^{K-1}\sum_{h=1}^H\sum_{k=p+1}^K \gamma^{\frac{k-1}{2}}\sqrt{\sum_{j=1}^p \gamma^{-j}} \left\|\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right\|_2$$

$$= 2H^2 L_\psi^2 \sqrt{\frac{d}{\lambda_{\mathbf{w}}}} \sum_{p=1}^{K-1}\sum_{h=1}^H \frac{\gamma^{\frac{p}{2}} - \gamma^{\frac{K}{2}}}{1 - \gamma^{\frac{1}{2}}}\sqrt{\frac{\gamma^{-p} - 1}{1 - \gamma}} \left\|\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right\|_2$$

$$\le 4H^2 L_\psi^2 \sqrt{\frac{d}{\lambda_{\mathbf{w}}}} \frac{1}{(1-\gamma)^{3/2}} \sum_{p=1}^{K-1}\sum_{h=1}^H \left\|\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right\|_2, \quad (63)$$

Variance. For variance part 1, we have

$$4H\beta_\theta \sum_{k=1}^K\sum_{h=1}^H \min\left\{1, \left\|\phi(s_h^k, a_h^k)\right\|_{(\Lambda_h^{k-1})^{-1}}\right\}$$

$$\le 4H^2\beta_\theta\sqrt{K}\sqrt{\sum_{k=1}^K \min\left\{1, \left\|\phi(s_h^k, a_h^k)\right\|^2_{(\Lambda_h^{k-1})^{-1}}\right\}}.$$

Based on the Lemma 9 (Potential Lemma), and let $X_k = \phi(s_h^k, a_h^k), U_k = \Lambda_h^{k-1}$, we know that $\forall h \in [H]$, we have

$$\sum_{k=1}^K \min\left\{1, \left\|\phi(s_h^k, a_h^k)\right\|^2_{(\Lambda_h^{k-1})^{-1}}\right\}$$

$$\le 2d\left(K\log\frac{1}{\gamma} + \log\left(1 + \frac{L_\phi^2}{\lambda_\theta d(1-\gamma)}\right)\right),$$

so we have

$$4H\beta_\theta \sum_{k=1}^K\sum_{h=1}^H \min\left\{1, \left\|\phi(s_h^k, a_h^k)\right\|_{(\Lambda_h^{k-1})^{-1}}\right\}$$

$$\le 4H^2\beta_\theta\sqrt{K}\sqrt{2d\left(K\log\frac{1}{\gamma} + \log\left(1 + \frac{L_\phi^2}{\lambda_\theta d(1-\gamma)}\right)\right)}.$$

For variance part 2, we have

$$4\sum_{k=1}^K\sum_{h=1}^H \beta_{\mathbf{w}}^{k-1}\min\left\{1, \left\|\psi_{h+1}^k(s_h^k, a_h^k)\right\|_{(\Sigma_h^{k-1})^{-1}}\right\}$$

$$\le 4\beta_{\mathbf{w}}^K\sum_{k=1}^K\sum_{h=1}^H \min\left\{1, \left\|\psi_{h+1}^k(s_h^k, a_h^k)\right\|_{(\Sigma_h^{k-1})^{-1}}\right\} \quad (64)$$

$$\le 4H\beta_{\mathbf{w}}^K\sqrt{K}\sqrt{\sum_{k=1}^K \min\left\{1, \left\|\psi_{h+1}^k(s_h^k, a_h^k)\right\|^2_{(\Sigma_h^{k-1})^{-1}}\right\}}$$

Based on potential lemma, we know that $\forall h \in [H]$, we have

$$\sum_{k=1}^K \min\left\{1, \left\|\psi_{h+1}^k(s_h^k, a_h^k)\right\|^2_{(\Sigma_h^{k-1})^{-1}}\right\}$$

$$\le 2d\left(K\log\frac{1}{\gamma} + \log\left(1 + \frac{H^2 L_\psi^2}{\lambda_{\mathbf{w}} d(1-\gamma)}\right)\right),$$

so we have

$$4 \sum_{k=1}^{K} \sum_{h=1}^{H} \beta_{\mathbf{w}}^{k-1} \min \left\{ 1, \left\| \psi_{h+1}^k \left(s_h^k, a_h^k \right) \right\|_{\left(\Sigma_h^{k-1} \right)^{-1}} \right\}$$

$$\leq 4H \beta_{\mathbf{w}}^K \sqrt{K} \sqrt{2d \left(K \log \frac{1}{\gamma} + \log \left(1 + \frac{H^2 L_\psi^2}{\lambda_{\mathbf{w}} d(1-\gamma)} \right) \right)}.$$

Since there is a term $HK\sqrt{\log(1/\gamma)}$ in the regret bound, we cannot let γ close to 0, so we set $\gamma \geq 1/K$ and have $\log(1/\gamma) \leq C(1-\gamma)$, where $C = \log K/(1 - 1/K)$. We set $\lambda_\theta = d$, and $\lambda_{\mathbf{w}} = H^2 d$. Combining the upper bounds of the bias and variance parts and with confidence level $\delta = 1/(4T)$, by union bound we have the following dynamic regret bound with probability at least $1 - 1/T$,

$$\text{D-REG}_T$$
$$\leq \mathcal{O}\Big(\frac{1}{(1-\gamma)^{3/2}} P_T^\theta + H \frac{1}{(1-\gamma)^{3/2}} P_T^{\mathbf{w}} + HdHK\sqrt{1-\gamma}$$
$$+ H^{3/2} d\sqrt{HK} \Big)$$
$$\leq \mathcal{O}\left(Hd \left(\frac{1}{(1-\gamma)^{3/2}} \Delta + HK\sqrt{1-\gamma} \right) + H^{3/2} d\sqrt{HK} \right)$$

Furthermore, by setting the discounted factor optimally as $\gamma = 1 - \max\left\{ 1/K, \sqrt{\Delta/T} \right\}$, we have

$$\text{D-REG}_T \leq \begin{cases} \widetilde{\mathcal{O}}\left(Hd\Delta^{1/4} T^{3/4} \right) & \text{when } \Delta \geq H/K, \\ \widetilde{\mathcal{O}}\left(dH^{3/2}\sqrt{T} \right) & \text{when } \Delta < H/K. \end{cases} \quad (65)$$

$\square$

Proof of Lemma 17. We first consider the upper bound of E_h^k, based on the definition of Q_h^k (26) and model assumption (22) and Eq. (24), we have $\forall a \in \mathcal{A}, s \in \mathcal{S}$,

$$r_h^k(s,a) + \left[\mathbb{P}_h^k V_{h+1}^k \right](s,a) - Q_h^k(s,a)$$
$$= r_h^k(s,a) + \left[\mathbb{P}_h^k V_{h+1}^k \right](s,a) - \phi(s,a)^\top \widehat{\theta}_h^k - \psi_{h+1}^k(s,a)^\top \widehat{\mathbf{w}}_h^k$$
$$- \beta_\theta \|\phi(s,a)\|_{\left(\Lambda_h^{k-1} \right)^{-1}} - \beta_{\mathbf{w}}^{k-1} \|\psi_{h+1}^k(s,a)\|_{\left(\Sigma_h^{k-1} \right)^{-1}}$$
$$= \phi(s,a)^\top \left(\theta_h^k - \widehat{\theta}_h^k \right) + \psi_{h+1}^k(s,a)^\top \left(\mathbf{w}_h^k - \widehat{\mathbf{w}}_h^k \right)$$
$$- \beta_\theta \|\phi(s,a)\|_{\left(\Lambda_h^{k-1} \right)^{-1}} - \beta_{\mathbf{w}}^{k-1} \|\psi_{h+1}^k(s,a)\|_{\left(\Sigma_h^{k-1} \right)^{-1}}$$
$$\leq \Gamma_{h,\theta}^{k-1} + \Gamma_{h,\mathbf{w}}^{k-1},$$

where the last inequality comes from Lemma 4 and Lemma 5. Similarly, we can get the lower bound of E_h^k, $\forall a \in \mathcal{A}, s \in \mathcal{S}$,

$$Q_h^k(s,a) - r_h^k(s,a) - \left[\mathbb{P}_h^k V_{h+1}^k \right](s,a)$$
$$= \phi(s,a)^\top \widehat{\theta}_h^k + \beta_\theta \|\phi(s,a)\|_{\left(\Lambda_h^{k-1} \right)^{-1}} + \psi_{h+1}^k(s,a)^\top \widehat{\mathbf{w}}_h^k$$
$$+ \beta_{\mathbf{w}}^{k-1} \|\psi_{h+1}^k(s,a)\|_{\left(\Sigma_h^{k-1} \right)^{-1}} - r_h^k(s,a) - \left[\mathbb{P}_h^k V_{h+1}^k \right](s,a)$$
$$= \phi(s,a)^\top \left(\widehat{\theta}_h^k - \theta_h^k \right) + \psi_{h+1}^k(s,a)^\top \left(\widehat{\mathbf{w}}_h^k - \mathbf{w}_h^k \right)$$
$$+ \beta_\theta \|\phi(s,a)\|_{\left(\Lambda_h^{k-1} \right)^{-1}} + \beta_{\mathbf{w}}^{k-1} \|\psi_{h+1}^k(s,a)\|_{\left(\Sigma_h^{k-1} \right)^{-1}}$$
$$\leq \Gamma_{h,\theta}^{k-1} + \Gamma_{h,\mathbf{w}}^{k-1} + 2\beta_\theta \|\phi(s,a)\|_{\left(\Lambda_h^{k-1} \right)^{-1}}$$
$$+ 2\beta_{\mathbf{w}}^{k-1} \|\psi_{h+1}^k(s,a)\|_{\left(\Sigma_h^{k-1} \right)^{-1}},$$

thus completes the proof. $\square$

Proof of Lemma 18. We first decompose the one step dynamic regret,

$$V_1^{k,\pi_*^k}(s_1^k) - V_1^{k,\pi^k}(s_1^k)$$
$$= \underbrace{V_1^{k,\pi_*^k}(s_1^k) - V_1^k(s_1^k)}_{\text{TERM (1)}} + \underbrace{V_1^k(s_1^k) - V_1^{k,\pi^k}(s_1^k)}_{\text{TERM (2)}}.$$

TERM (1). We first have $\forall s \in \mathcal{S}$,

$$V_h^{k,\pi_*^k}(s) - V_h^k(s)$$
$$= \mathbb{E}_{a \sim \pi_{*,h}^k(s)} \left[Q_h^{k,\pi_*^k}(s,a) \right] - \mathbb{E}_{a \sim \pi_h^k(s)} \left[Q_h^k(s,a) \right]$$
$$= \mathbb{E}_{a \sim \pi_{*,h}^k(s)} \left[Q_h^{k,\pi_*^k}(s,a) \right] - \mathbb{E}_{a \sim \pi_{*,h}^k(s)} \left[Q_h^k(s,a) \right]$$
$$+ \mathbb{E}_{a \sim \pi_{*,h}^k(s)} \left[Q_h^k(s,a) \right] - \mathbb{E}_{a \sim \pi_h^k(s)} \left[Q_h^k(s,a) \right]$$
$$\leq \mathbb{E}_{a \sim \pi_{*,h}^k(s)} \left[Q_h^{k,\pi_*^k}(s,a) - Q_h^k(s,a) \right],$$

where the last inequality comes from $\pi_h^k(s) = \arg\max_{a \in \mathcal{A}} Q_h^k(s,a)$. Then we have

$$Q_h^{k,\pi_*^k}(s,a) - Q_h^k(s,a)$$
$$= r_h^k(s,a) + \left[\mathbb{P}_h^k V_{h+1}^{k,\pi_*^k} \right](s,a) - r_h^k(s,a) - \left[\mathbb{P}_h^k V_{h+1}^k \right](s,a)$$
$$+ r_h^k(s,a) + \left[\mathbb{P}_h^k V_{h+1}^k \right](s,a) - Q_h^k(s,a)$$
$$= \left[\mathbb{P}_h^k \left(V_{h+1}^{k,\pi_*^k} - V_{h+1}^k \right) \right](s,a) + E_h^k(s,a),$$

where the last equality comes from the definition of model prediction error. For notational simplicity, we define the operators $\mathbb{J}_h^k f(s) = \left\langle f(x,\cdot), \pi_{*,h}^k(\cdot \mid s) \right\rangle$, then we have

$$V_h^{k,\pi_*^k}(s) - V_h^k(s)$$
$$\leq \mathbb{J}_h^k \mathbb{P}_h^k \left(V_{h+1}^{k,\pi_*^k} - V_{h+1}^k \right)(s,a) + \mathbb{J}_h^k E_h^k(s,a),$$

recursively expanding the above inequality and we have

$$V_1^{k,\pi_*^k}(s_1^k) - V_1^k(s_1^k)$$
$$\leq \left(\prod_{h=1}^{H} \mathbb{J}_h^k \mathbb{P}_h^k \right) \left(V_{H+1}^{k,\pi_*^k}(s_{H+1}^k) - V_{H+1}^k(s_{H+1}) \right)$$
$$+ \sum_{h=1}^{H} \left(\prod_{j=1}^{H} \mathbb{J}_j^k \mathbb{P}_j^k \right) \mathbb{J}_h^k E_h^k(s_h^k, a_h^k)$$
$$\leq \sum_{h=1}^{H} \mathbb{E}_{\pi_{*,h}^k} \left[E_h^k(s_h^k, a_h^k) \right],$$

where the last inequality comes from that $\forall \pi, V_{H+1}^{k,\pi}(\cdot) = 0, V_{H+1}^k(\cdot) = 0$. Then we have

$$\sum_{k=1}^{K} V_1^{k,\pi_*^k}(s_h^k) - V_1^k(s_h^k)$$
$$\leq \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi_{*,h}^k} \left[E_h^k(s_h^k, a_h^k) \right]. \quad (66)$$

TERM (2). Based on Eq (61), we have $\forall s \in \mathcal{S}, a \in \mathcal{A}$,

$$
\begin{aligned}
E_h^k(s,a) &= r_h^k(s,a) + \left[\mathbb{P}_h^k V_{h+1}^k\right](s,a) - Q_h^k(s,a) \\
&= r_h^k(s,a) + \left[\mathbb{P}_h^k V_{h+1}^k\right](s,a) - Q_h^{k,\pi^k}(s,a) \\
&\quad + Q_h^{k,\pi^k}(s,a) - Q_h^k(s,a) \\
&= \left(\left[\mathbb{P}_h^k V_{h+1}^k\right](s,a) - \left[\mathbb{P}_h^k V_{h+1}^{k,\pi^k}\right](s,a)\right) \\
&\quad + Q_h^{k,\pi^k}(s,a) - Q_h^k(s,a).
\end{aligned}
$$

By applying this equality, we further have

$$
\begin{aligned}
&V_h^k\left(s_h^k\right) - V_h^{k,\pi^k}\left(s_h^k\right) \\
&= \mathbb{E}_{a\sim\pi_h^k(s_h^k)}\left[Q_h^k(s_h^k,a)\right] - \mathbb{E}_{a\sim\pi_h^k(s_h^k)}\left[Q_h^{k,\pi^k}(s_h^k,a)\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + E_h^k - E_h^k \\
&= \mathbb{E}_{a\sim\pi_h^k(s_h^k)}\left[Q_h^k(s_h^k,a)\right] - \mathbb{E}_{a\sim\pi_h^k(s_h^k)}\left[Q_h^{k,\pi^k}(s_h^k,a)\right] \\
&\quad + \left(\left[\mathbb{P}_h^k V_{h+1}^k\right](s_h^k,a_h^k) - \left[\mathbb{P}_h^k V_{h+1}^{k,\pi^k}\right](s_h^k,a_h^k)\right) \\
&\quad + Q_h^{k,\pi^k}\left(s_h^k,a_h^k\right) - Q_h^k\left(s_h^k,a_h^k\right) - E_h^k\left(s_h^k,a_h^k\right).
\end{aligned}
$$

we define

$$
\begin{aligned}
\mathcal{M}_{h,V}^k &= \left(\left[\mathbb{P}_h^k V_{h+1}^k\right]\left(s_h^k,a_h^k\right) - \left[\mathbb{P}_h^k V_{h+1}^{k,\pi^k}\right]\left(s_h^k,a_h^k\right)\right) \\
&\quad - \left(V_{h+1}^k\left(s_{h+1}^k\right) - V_{h+1}^{k,\pi^k}\left(s_{h+1}^k\right)\right) \\
\mathcal{M}_{h,Q}^k &= \mathbb{E}_{a\sim\pi_h^k(s)}\left[Q_h^k(s_h^k,a) - Q_h^{k,\pi^k}(s_h^k,a)\right] \\
&\quad - \left(Q_h^k\left(s_h^k,a_h^k\right) - Q_h^{k,\pi^k}\left(s_h^k,a_h^k\right)\right),
\end{aligned}
$$

then we have

$$
\begin{aligned}
&V_h^k\left(s_h^k\right) - V_h^{\pi^k}\left(s_h^k\right) - \left(V_{h+1}^k\left(s_{h+1}^k\right) - V_{h+1}^{\pi^k}\left(s_{h+1}^k\right)\right) \\
&= \mathcal{M}_{h,V}^k + \mathcal{M}_{h,Q}^k - E_h^k\left(s_h^k,a_h^k\right).
\end{aligned}
$$

Summing up for $k \in [K]$ and $h \in [H]$, since $V_{H+1}^k = 0$, $V_{H+1}^{k,\pi} = 0$, we have

$$
\begin{aligned}
&\sum_{k=1}^K V_1^k\left(s_1^k\right) - V_1^{\pi^k}\left(s_1^k\right) \\
&\leq \sum_{k=1}^K \sum_{h=1}^H \left(-E_h^k\left(s_h^k,a_h^k\right)\right) + \sum_{k=1}^K \sum_{h=1}^H \mathcal{M}_{h,V}^k + \sum_{k=1}^K \sum_{h=1}^H \mathcal{M}_{h,Q}^k.
\end{aligned}
$$

Since $\mathcal{M}_{h,V}^k$ and $\mathcal{M}_{h,Q}^k$ are martingale difference which bounded by $2H$, then based on Lemma 24, we have the following holds each with probability at least $1-\delta$,

$$
\begin{aligned}
\sum_{k=1}^K \sum_{h=1}^H \mathcal{M}_{h,V}^k &\leq 2H\sqrt{2T\log(1/\delta)}, \\
\sum_{k=1}^K \sum_{h=1}^H \mathcal{M}_{h,Q}^k &\leq 2H\sqrt{2T\log(1/\delta)},
\end{aligned}
$$

where $T = KH$. Then the following holds with probability at least $1-2\delta$,

$$
\begin{aligned}
&\sum_{k=1}^K \left(V_1^k\left(s_1^k\right) - V_1^{\pi^k}\left(s_1^k\right)\right) \\
&\leq \sum_{k=1}^K \sum_{h=1}^H \left(-E_h^k\left(s_h^k,a_h^k\right)\right) + 4H\sqrt{2T\log(1/\delta)}.
\end{aligned}
\tag{67}
$$

Combining Eq (66) and Eq (67) and we have

$$
\begin{aligned}
&V_1^{k,\pi_*^k}\left(s_1^k\right) - V_1^{k,\pi^k}\left(s_1^k\right) \\
&= \sum_{k=1}^K \sum_{h=1}^H \left(\mathbb{E}_{\pi_{*,h}^k}\left[E_h^k(s_h^k,a_h^k)\right] - E_h^k\left(s_h^k,a_h^k\right)\right) \\
&\qquad\qquad\qquad\qquad\qquad + 4H\sqrt{2T\log(1/\delta)}.
\end{aligned}
$$

$\square$

APPENDIX F
ANALYSIS OF MNL-WEIGHTUCRL

A. Proof of Lemma 6

Proof. Based on Lemma 2 of [17], we know that

$$
\begin{aligned}
&\left|\left[\widetilde{\mathbb{P}}_h^k V\right](s,a) - \left[\mathbb{P}_h^k V\right](s,a)\right| \\
&= \left|\sum_{s'\in\mathcal{S}_h^k} p_h^k(\psi(s'\mid s,a)^\top \widetilde{\mathbf{w}}_h^k)V(s')\right. \\
&\qquad\qquad\qquad \left. - \sum_{s'\in\mathcal{S}_h^k} p_h^k(\psi(s'\mid s,a)^\top \mathbf{w}_h^k)V(s')\right| \\
&\leq H\max_{s'\in\mathcal{S}_h^k}\left|\psi(s'\mid s,a)^\top\left(\widetilde{\mathbf{w}}_h^k - \mathbf{w}_h^k\right)\right|.
\end{aligned}
$$

We first construct the following Lemma.

Lemma 19. *For any* $\mathbf{x} \in \mathcal{X}$, *and* $\delta \in (0,1)$, $\forall k,j \in [K], \alpha_{k,j} \leq 1$, *with probability at least* $1-\delta$, *the following holds for all* $k \in [K], h \in [H]$

$$
\begin{aligned}
&\left|\psi(s'\mid s,a)^\top\left(\widetilde{\mathbf{w}}_h^k - \mathbf{w}_h^k\right)\right| \\
&\leq \frac{1}{\kappa}\left(\Gamma_{\mathbf{w}}^{k-1} + \bar{\beta}_{\mathbf{w}}^{k-1}\|\psi(s'\mid s,a)\|_{(\bar{\Sigma}_h^{k-1})^{-1}}\right).
\end{aligned}
$$

$\Gamma_{\mathbf{w}}^{k-1} \triangleq L_\psi^2\sqrt{\frac{d}{\lambda_{\mathbf{w}}}}\sum_{p=1}^{k-1}\sqrt{\sum_{j=1}^p \alpha_{k-1,j}}\left\|\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right\|_2$, *and* $\bar{\beta}_{\mathbf{w}}^k$ *is the radius of confidence region set by*

$$
\bar{\beta}_{\mathbf{w}}^k \triangleq \sqrt{\frac{1}{2}\log\frac{1}{\delta} + \frac{d}{4}\log\left(1 + \frac{U L_\psi^2 \sum_{j=1}^k \alpha_{k,j}}{\lambda_{\mathbf{w}} d}\right)} + \sqrt{\lambda_{\mathbf{w}}}\kappa S_{\mathbf{w}}.
$$

Then we have

$$
\begin{aligned}
&\left|\left[\widetilde{\mathbb{P}}_h^k V\right](s,a) - \left[\mathbb{P}_h^k V\right](s,a)\right| \\
&\leq \frac{H}{\kappa}\left(\Gamma_{\mathbf{w}}^{k-1} + \bar{\beta}_{\mathbf{w}}^{k-1}\max_{s'\in\mathcal{S}_h^k}\|\psi(s'\mid s,a)\|_{(\bar{\Sigma}_h^{k-1})^{-1}}\right),
\end{aligned}
$$

where $\Gamma_{\mathbf{w}}^{k-1} \triangleq L_\psi^2\sqrt{\frac{d}{\lambda_{\mathbf{w}}}}\sum_{p=1}^{k-1}\sqrt{\sum_{j=1}^p \alpha_{k-1,j}}\left\|\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right\|_2$, $\bar{\beta}_{\mathbf{w}}^k$ is the radius of confidence region set by

$$
\bar{\beta}_{\mathbf{w}}^k \triangleq \sqrt{\frac{1}{2}\log\frac{1}{\delta} + \frac{d}{4}\log\left(1 + \frac{U L_\psi^2 \sum_{j=1}^k \alpha_{k,j}}{\lambda_{\mathbf{w}} d}\right)} + \sqrt{\lambda_{\mathbf{w}}}\kappa S_{\mathbf{w}}.
$$

$\square$

Proof of Lemma 19. Fix $h \in [H]$, based on the estimator (28), we have

$$g_h^k(\widehat{\mathbf{w}}_h^k) = \lambda_{\mathbf{w}} \kappa \widehat{\mathbf{w}}_h^k + \sum_{j=1}^{k-1} \alpha_{k-1,j} \sum_{s' \in \mathcal{S}_h^j} p_h^j(\bar{\psi}_h^j(s')^\top \widehat{\mathbf{w}}_h^k) \bar{\psi}_h^j(s')$$

$$= \sum_{j=1}^{k-1} \alpha_{k-1,j} \sum_{s' \in \mathcal{S}_h^j} y_h^j(s') \bar{\psi}_h^j(s')$$

$$= \sum_{j=1}^{k-1} \alpha_{k-1,j} \sum_{s' \in \mathcal{S}_h^j} \left(p_h^j(\bar{\psi}_h^j(s')^\top \mathbf{w}_h^j) + \eta_h^j(s') \right) \bar{\psi}_h^j(s'),$$

where we define $\eta_h^j(s') \triangleq y_h^j(s') - p_h^j(\bar{\psi}_h^j(s')^\top \mathbf{w}_h^j)$. Then by the mean value theorem, we know that

$$g_h^k(\mathbf{w}_1) - g_h^k(\mathbf{w}_2) = G_h^k(\mathbf{w}_1 - \mathbf{w}_2), \qquad (68)$$

where $G_h^k(\mathbf{w}_1 - \mathbf{w}_2) \triangleq \int_0^1 \nabla g_h^k (s\mathbf{w}_2 + (1-s)\mathbf{w}_1) \, ds \in \mathbb{R}^{d \times d}$. Notice that for any $\mathbf{w} \in \mathcal{W}$, the gradient of g_h^k is

$$\nabla g_h^k(\mathbf{w}) = \lambda_{\mathbf{w}} \kappa I_d$$
$$+ \sum_{j=1}^{k-1} \alpha_{k-1,j} \left(\sum_{s' \in \mathcal{S}_h^j} p_h^j(\bar{\psi}_h^j(s')^\top \mathbf{w}) \bar{\psi}_h^j(s') \bar{\psi}_h^j(s')^\top \right.$$
$$\left. - \sum_{s' \in \mathcal{S}_h^j} \sum_{s'' \in \mathcal{S}_h^j} p_h^j(\bar{\psi}_h^j(s')^\top \mathbf{w}) p_h^j(\bar{\psi}_h^j(s'')^\top \mathbf{w}) \bar{\psi}_h^j(s') \bar{\psi}_h^j(s'')^\top \right)$$

Based on Lemma 5 of [17], we know that

$$\nabla g_h^k(\mathbf{w})$$
$$\succeq \lambda_{\mathbf{w}} \kappa I_d + \kappa \sum_{j=1}^{k-1} \alpha_{k-1,j} \sum_{s' \in \mathcal{S}_h^j} \bar{\psi}_h^j(s') \bar{\psi}_h^j(s')^\top = \kappa \bar{\Sigma}_h^{k-1},$$

which clearly implies $\forall \mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}, G_h^k(\mathbf{w}_1, \mathbf{w}_2) \succeq \kappa \bar{\Sigma}_h^{k-1}$. By Assumption 6, the mean value theorem (68) on g_h^k and the projection (29), we have

$$\left| \psi(s' \mid s, a)^\top \left(\widetilde{\mathbf{w}}_h^k - \mathbf{w}_h^k \right) \right|$$
$$= \left| \psi(s' \mid s, a)^\top \left(G_h^k(\widetilde{\mathbf{w}}_h^k, \mathbf{w}_h^k) \right)^{-1} \left(g_h^k(\widetilde{\mathbf{w}}_h^k) - g_h^k(\mathbf{w}_h^k) \right) \right|$$
$$\leq \left\| \psi(s' \mid s, a) \right\|_{\left(G_h^k(\widetilde{\mathbf{w}}_h^k, \mathbf{w}_h^k) \right)^{-1}}$$
$$\quad \cdot \left\| g_h^k(\widetilde{\mathbf{w}}_h^k) - g_h^k(\mathbf{w}_h^k) \right\|_{\left(G_h^k(\widetilde{\mathbf{w}}_h^k, \mathbf{w}_h^k) \right)^{-1}}$$
$$\leq \frac{1}{\kappa} \left\| \psi(s' \mid s, a) \right\|_{(\bar{\Sigma}_h^{k-1})^{-1}} \left\| g_h^k(\widetilde{\mathbf{w}}_h^k) - g_h^k(\mathbf{w}_h^k) \right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}$$
$$\leq \frac{1}{\kappa} \left\| \psi(s' \mid s, a) \right\|_{(\bar{\Sigma}_h^{k-1})^{-1}} \left\| g_h^k(\widehat{\mathbf{w}}_h^k) - g_h^k(\mathbf{w}_h^k) \right\|_{(\bar{\Sigma}_h^{k-1})^{-1}},$$

we further have

$$g_h^k(\widehat{\mathbf{w}}_h^k) - g_h^k(\mathbf{w}_h^k)$$
$$= \sum_{j=1}^{k-1} \alpha_{k-1,j} \sum_{s' \in \mathcal{S}_h^j} \left(p_h^j(\bar{\psi}_h^j(s')^\top \mathbf{w}_h^j) + \eta_h^j(s') \right) \bar{\psi}_h^j(s')$$
$$\quad - \lambda_{\mathbf{w}} \kappa \mathbf{w}_h^k - \sum_{j=1}^{k-1} \alpha_{k-1,j} \sum_{s' \in \mathcal{S}_h^j} p_h^j(\bar{\psi}_h^j(s')^\top \mathbf{w}_h^k) \bar{\psi}_h^j(s')$$

$$= \underbrace{\sum_{j=1}^{k-1} \alpha_{k-1,j} \sum_{s' \in \mathcal{S}_h^j} \left(p_h^j(\bar{\psi}_h^j(s')^\top \mathbf{w}_h^j) - p_h^j(\bar{\psi}_h^j(s')^\top \mathbf{w}_h^k) \right) \bar{\psi}_h^j(s')}_{\text{bias part}}$$

$$\underbrace{- \lambda_{\mathbf{w}} \kappa \mathbf{w}_h^k + \sum_{j=1}^{k-1} \alpha_{k-1,j} \sum_{s' \in \mathcal{S}_h^j} \eta_h^j(s') \bar{\psi}_h^j(s')}_{\text{variance part}}.$$

Then, by the Cauchy-Schwarz inequality, we know that for any $s \in \mathcal{S}, a \in \mathcal{A}$,

$$\left| \psi(s' \mid s, a)^\top \left(\widehat{\mathbf{w}}_h^k - \mathbf{w}_h^k \right) \right|$$
$$\leq \frac{1}{\kappa} \left\| \psi(s' \mid s, a) \right\|_{(\bar{\Sigma}_h^{k-1})^{-1}} \left(E_h^k + F_h^k \right), \qquad (69)$$

where

$$E_h^k = \left\| \sum_{j=1}^{k-1} \alpha_{k-1,j} \sum_{s' \in \mathcal{S}_h^j} \left(p_h^j(\bar{\psi}_h^j(s')^\top \mathbf{w}_h^j) \right. \right.$$
$$\left. \left. - p_h^j(\bar{\psi}_h^j(s')^\top \mathbf{w}_h^k) \right) \bar{\psi}_h^j(s') \right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}$$

$$F_h^k = \left\| - \lambda_{\mathbf{w}} \kappa \mathbf{w}_h^k + \sum_{j=1}^{k-1} \alpha_{k-1,j} \sum_{s' \in \mathcal{S}_h^j} \eta_h^j(s') \bar{\psi}_h^j(s') \right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}.$$

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

Lemma 20. *For any $k \in [K]$, we have*

$$E_h^k \leq L_\psi \sqrt{d} \sum_{p=1}^{k-1} \sqrt{\sum_{j=1}^{p} \alpha_{k-1,j}} \left\| \mathbf{w}_h^p - \mathbf{w}_h^{p+1} \right\|_2.$$

Lemma 21. *If $\forall k \in [K], \forall j \in [k-1], \alpha_{k-1,j} \leq 1$, for any $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds for all $k \in [K]$,*

$$F_h^k \leq \sqrt{\frac{1}{2} \log \frac{1}{\delta} + \frac{d}{4} \log \left(1 + \frac{U L_\psi^2 \sum_{j=1}^{k-1} \alpha_{k-1,j}}{\lambda_{\mathbf{w}} d} \right)}$$
$$+ \sqrt{\lambda_{\mathbf{w}}} \kappa S_{\mathbf{w}}.$$

Based on the inequality (69), Lemma 20, Lemma 21, and $\left\| \psi(s' \mid s, a) \right\|_{(\bar{\Sigma}_h^{k-1})^{-1}} \leq \left\| \psi(s' \mid s, a) \right\|_2 / \sqrt{\lambda_{\mathbf{w}}} \leq L_\psi / \sqrt{\lambda_{\mathbf{w}}}$, with probability at least $1-\delta$, the following holds for all $k \in [K]$,

$$\left| \psi(s' \mid s, a)^\top \left(\widetilde{\mathbf{w}}_h^k - \mathbf{w}_h^k \right) \right|$$
$$\leq \frac{1}{\kappa} \left(\Gamma_{\mathbf{w}}^{k-1} + \bar{\beta}_{\mathbf{w}}^{k-1} \left\| \psi(s' \mid s, a) \right\|_{(\bar{\Sigma}_h^{k-1})^{-1}} \right),$$

where

$$\Gamma_{\mathbf{w}}^{k-1} \triangleq L_\psi^2 \sqrt{\frac{d}{\lambda_{\mathbf{w}}}} \sum_{p=1}^{k-1} \sqrt{\sum_{j=1}^{p} \alpha_{k-1,j}} \left\| \mathbf{w}_h^p - \mathbf{w}_h^{p+1} \right\|_2$$

$$\bar{\beta}_{\mathbf{w}}^{k-1} \triangleq \sqrt{\frac{1}{2} \log \frac{1}{\delta} + \frac{d}{4} \log \left(1 + \frac{U L_\psi^2 \sum_{j=1}^{k-1} \alpha_{k-1,j}}{\lambda_{\mathbf{w}} d} \right)}$$
$$+ \sqrt{\lambda_{\mathbf{w}}} \kappa S_{\mathbf{w}},$$

which completes the proof. $\hspace{2em}\square$

Proof of Lemma 20. The first step is to extract the variations of the parameter $\mathbf{w}_h^k$ as follows,

$$\left\|\sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\left(p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^j)\right.\right.$$
$$\left.\left.-p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^k)\right)\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}$$
$$=\left\|\sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\sum_{p=j}^{k-1}\left(p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^p)\right.\right.$$
$$\left.\left.-p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^{p+1})\right)\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}$$
$$=\left\|\sum_{p=1}^{k-1}\sum_{j=1}^{p}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\left(p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^p)\right.\right.$$
$$\left.\left.-p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^{p+1})\right)\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}.$$

The gradient of $p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w})$ is given by

$$\nabla p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w})=p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w})\bar{\psi}_h^j(s')^\top$$
$$-p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w})\sum_{s''\in\mathcal{S}_h^j}p_h^j(\bar{\psi}_h^j(s'')^\top\mathbf{w})\bar{\psi}_h^j(s'')^\top,$$

by the mean value theorem, there exist $\bar{\mathbf{w}}_h^p=\nu\mathbf{w}_h^p+(1-\nu)\mathbf{w}_h^{p+1}$ for some $\nu\in[0,1]$, such that

$$p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^p)-p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^{p+1})$$
$$=p_h^j(\bar{\psi}_h^j(s')^\top\bar{\mathbf{w}}_h^p)\left(\bar{\psi}_h^j(s')^\top-\sum_{s''\in\mathcal{S}_h^j}p_h^j(\bar{\psi}_h^j(s'')^\top\bar{\mathbf{w}}_h^p)\bar{\psi}_h^j(s'')^\top\right)$$
$$\cdot\left(\mathbf{w}_h^p-\mathbf{w}_h^{p+1}\right)$$
$$\leq L_\psi p_h^j(\bar{\psi}_h^j(s')^\top\bar{\mathbf{w}}_h^p)\left(1-\sum_{s''\in\mathcal{S}_h^j}p_h^j(\bar{\psi}_h^j(s'')^\top\bar{\mathbf{w}}_h^p)\right)$$
$$\cdot\left\|\mathbf{w}_h^p-\mathbf{w}_h^{p+1}\right\|_2$$
$$\leq L_\psi\nabla p_h^j(\bar{\psi}_h^j(s')^\top\bar{\mathbf{w}}_h^p)\left\|\mathbf{w}_h^p-\mathbf{w}_h^{p+1}\right\|_2,$$

where we define $\nabla p_h^j(\bar{\psi}_h^j(s')^\top\bar{\mathbf{w}}_h^p)=p_h^j(\bar{\psi}_h^j(s')^\top\bar{\mathbf{w}}_h^p)-p_h^j(\bar{\psi}_h^j(s')^\top\bar{\mathbf{w}}_h^p)\sum_{s''\in\mathcal{S}_h^j}p_h^j(\bar{\psi}_h^j(s'')^\top\bar{\mathbf{w}}_h^p)$. Then we have

$$\left\|\sum_{p=1}^{k-1}\sum_{j=1}^{p}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\left(p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^p)\right.\right.$$
$$\left.\left.-p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^{p+1})\right)\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}$$
$$\leq L_\psi\sum_{p=1}^{k-1}\left\|\sum_{j=1}^{p}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\nabla p_h^j(\bar{\psi}_h^j(s')^\top\bar{\mathbf{w}}_h^p)\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}$$
$$\cdot\left\|\mathbf{w}_h^p-\mathbf{w}_h^{p+1}\right\|_2$$
$$\leq L_\psi\sum_{p=1}^{k-1}\sum_{j=1}^{p}\alpha_{k-1,j}\left|\sum_{s'\in\mathcal{S}_h^j}\nabla p_h^j(\bar{\psi}_h^j(s')^\top\bar{\mathbf{w}}_h^p)\right|$$
$$\cdot\left\|\max_{s'\in\mathcal{S}_h^j}\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}\left\|\mathbf{w}_h^p-\mathbf{w}_h^{p+1}\right\|_2$$
$$\leq L_\psi\sum_{p=1}^{k-1}\sum_{j=1}^{p}\alpha_{k-1,j}\left\|\max_{s'\in\mathcal{S}_h^j}\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}\left\|\mathbf{w}_h^p-\mathbf{w}_h^{p+1}\right\|_2,$$

where the last inequality comes from the fact that $\left|\sum_{s'\in\mathcal{S}_h^j}\nabla p_h^j(\bar{\psi}_h^j(s')^\top\bar{\mathbf{w}}_h^p)\right|\leq 1$. Further, for the term $\sum_{j=1}^{p}\alpha_{k-1,j}\left\|\max_{s'\in\mathcal{S}_h^j}\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}$ can be able to further derive an expression about weight $\alpha_{k-1,j}$ as follows,

$$\sum_{j=1}^{p}\alpha_{k-1,j}\left\|\max_{s'\in\mathcal{S}_h^j}\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}$$
$$\leq\sqrt{\sum_{j=1}^{p}\alpha_{k-1,j}}\sqrt{\sum_{j=1}^{p}\alpha_{k-1,j}\left\|\max_{s'\in\mathcal{S}_h^j}\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}^2}$$
$$\leq\sqrt{\sum_{j=1}^{p}\alpha_{k-1,j}}\sqrt{\sum_{j=1}^{p}\sum_{s'\in\mathcal{S}_h^j}\alpha_{k-1,j}\left\|\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}^2}$$
$$\leq\sqrt{d}\sqrt{\sum_{j=1}^{p}\alpha_{k-1,j}}.$$

In above, the second last step holds by the Cauchy-Schwarz inequality. Besides, the last step follows that

$$\sum_{j=1}^{p}\sum_{s'\in\mathcal{S}_h^j}\alpha_{k-1,j}\left\|\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}^2$$
$$\sum_{j=1}^{p}\sum_{s'\in\mathcal{S}_h^j}\text{Tr}(\alpha_{k-1,j}\bar{\psi}_h^j(s')^\top(\bar{\Sigma}_h^{k-1})^{-1}\bar{\psi}_h^j(s'))$$
$$=\text{Tr}\left((\bar{\Sigma}_h^{k-1})^{-1}\sum_{j=1}^{p}\sum_{s'\in\mathcal{S}_h^j}\alpha_{k-1,j}\bar{\psi}_h^j(s')\bar{\psi}_h^j(s')^\top\right)$$
$$\leq\text{Tr}\left((\bar{\Sigma}_h^{k-1})^{-1}\sum_{j=1}^{k-1}\sum_{s'\in\mathcal{S}_h^j}\alpha_{k-1,j}\bar{\psi}_h^j(s')\bar{\psi}_h^j(s')^\top\right)$$
$$+\text{Tr}\left(U_{t-1}^{-1}\lambda\sum_{i=1}^{d}\mathbf{e}_i\mathbf{e}_i^\top\right)$$
$$=\text{Tr}(I_d)=d.$$

Hence we complete the proof. $\hspace{2em}\square$

Proof of Lemma 21.

$$\left\|-\lambda_\mathbf{w}\kappa\mathbf{w}_h^k+\sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\eta_h^j(s')\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}$$

$$\leq \left\|\sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\eta_h^j(s')\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}} + \left\|\lambda_{\mathbf{w}}\kappa\mathbf{w}_h^k\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}$$

$$\leq \left\|\sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\eta_h^j(s')\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}} + \sqrt{\lambda_{\mathbf{w}}}\kappa S_{\mathbf{w}}.$$

we define $\widetilde{\eta}_j = \sqrt{\alpha_{k-1,j}}\eta_j$ and $X_j = \sqrt{\alpha_{k-1,j}}\bar{\psi}_h^j(s')$, then we have notice that since the reward $r\in[0,1]$, and $\alpha_{k-1,j}\leq 1$, the noise $\widetilde{\eta}_j$ is bounded by:

$$\widetilde{\eta}_j = \sqrt{\alpha_{k-1,j}}\left(y_h^j(s') - p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^j)\right)\leq 1,$$

based on Lemma 23, we find that the noise $\widetilde{\eta}_j$ is $\frac{1}{2}$-sub-Gaussian. Then, by Theorem 7, we have with probability at least $1-\delta$, the following holds for all $k\in[K]$.

$$\left\|\sum_{j=1}^{k-1}\widetilde{\eta}_j X_j\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}} \leq \sqrt{\frac{1}{2}\log\left(\frac{\det(\bar{\Sigma}_h^{k-1})^{\frac{1}{2}}\det(\bar{\Sigma}_h^0)^{-\frac{1}{2}}}{\delta}\right)}$$

where

$$\det(\bar{\Sigma}_h^{k-1}) \leq \left(\frac{\mathrm{trace}(\bar{\Sigma}_h^{k-1})}{d}\right)^d$$

$$= \left(\frac{d\lambda_{\mathbf{w}} + \sum_{j=1}^{k-1}\sum_{s'\in\mathcal{S}_h^j}\alpha_{k-1,j}\left\|\bar{\psi}_h^j(s')\right\|_2^2}{d}\right)^d$$

$$= \left(\frac{d\lambda_{\mathbf{w}} + UL_\psi^2\sum_{j=1}^{k-1}\alpha_{k-1,j}}{d}\right)^d$$

$$\det(\bar{\Sigma}_h^0) \leq (\lambda_{\mathbf{w}})^d,$$

so we have

$$\left\|\sum_{j=1}^{k-1}\alpha_{k-1,j}\eta_j\bar{\psi}_{h+1}^j\left(s_h^j,a_h^j\right)\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}$$

$$\leq \sqrt{\frac{1}{2}\log\frac{1}{\delta} + \frac{d}{4}\log\left(1 + \frac{UL_\psi^2\sum_{j=1}^{k-1}\alpha_{k-1,j}}{\lambda_{\mathbf{w}}d}\right)}.$$

which completes the proof. $\qquad\square$

B. Proof of Theorem 6

Proof. To prove the theorem, we first introduce the following lemma

Lemma 22. *We define the model prediction error as*

$$E_h^k(s,a) = r_h^k(s,a) + \mathbb{P}_h^k\bar{V}_{h+1}^k(s,a) - \bar{Q}_h^k(s,a), \qquad (70)$$

then with probability at least $1-2\delta$, the following holds for all $k\in[K]$, $h\in[H]$ and $\forall s\in\mathcal{S}, a\in\mathcal{A}$,

$$-\Gamma_{h,\theta}^{k-1} - \frac{H}{\kappa}\Gamma_{h,\mathbf{w}}^{k-1} - 2\beta_\theta\|\phi(s,a)\|_{(\Lambda_h^{k-1})^{-1}}$$

$$-2\frac{H}{\kappa}\bar{\beta}_{\mathbf{w}}^{k-1}\max_{s'\in\mathcal{S}_h^k}\|\psi(s'\mid s,a)\|_{(\bar{\Sigma}_h^{k-1})^{-1}} \leq E_h^k(s,a)$$

$$\leq \Gamma_{h,\theta}^{k-1} + \frac{H}{\kappa}\Gamma_{h,\mathbf{w}}^{k-1}.$$

And notice that $\forall s\in\mathcal{S}, a\in\mathcal{A}, \left|E_h^k(s,a)\right|\leq 2H$, we have

$$\mathbb{E}_{\pi_{*,h}^k}\left[E_h^k(s_h^k, a_h^k)\right] - E_h^k(s_h^k, a_h^k)$$

$$\leq \min\Bigg\{4H, 2\Gamma_{h,\theta}^{k-1} + 2\frac{H}{\kappa}\Gamma_{h,\mathbf{w}}^{k-1} + 2\beta_\theta\left\|\phi(s_h^k, a_h^k)\right\|_{(\Lambda_h^{k-1})^{-1}}$$

$$+ 2\frac{H}{\kappa}\bar{\beta}_{\mathbf{w}}^{k-1}\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s_h^k, a_h^k)\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}\Bigg\}$$

$$\leq 2\Gamma_{h,\theta}^{k-1} + 2\frac{H}{\kappa}\Gamma_{h,\mathbf{w}}^{k-1} + \min\left\{4H, 2\beta_\theta\left\|\phi(s_h^k, a_h^k)\right\|_{(\Lambda_h^{k-1})^{-1}}\right\}$$

$$+ \min\left\{4H, 2\frac{H}{\kappa}\bar{\beta}_{\mathbf{w}}^{k-1}\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s_h^k, a_h^k)\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}\right\}$$

$$\leq 2\Gamma_{h,\theta}^{k-1} + 2\frac{H}{\kappa}\Gamma_{h,\mathbf{w}}^{k-1} + 4H\beta_\theta\min\left\{1, \left\|\phi(s_h^k, a_h^k)\right\|_{(\Lambda_h^{k-1})^{-1}}\right\}$$

$$+ 4\frac{H}{\kappa}\bar{\beta}_{\mathbf{w}}^{k-1}\min\left\{1, \max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s_h^k, a_h^k)\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}\right\},$$

By Lemma 18, we can further connect the dynamic regret to the model prediction error, we have with probability at least $1-4\delta$,

$$\text{D-REG}_T \leq \underbrace{4H\beta_\theta\sum_{k=1}^K\sum_{h=1}^H\min\left\{1, \left\|\phi(s_h^k, a_h^k)\right\|_{(\Lambda_h^{k-1})^{-1}}\right\}}_{\text{variance part1}}$$

$$+ \underbrace{4\frac{H}{\kappa}\sum_{k=1}^K\sum_{h=1}^H\bar{\beta}_{\mathbf{w}}^{k-1}\min\left\{1, \max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s_h^k, a_h^k)\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}\right\}}_{\text{variance part2}}$$

$$+ \underbrace{2\sum_{k=1}^K\sum_{h=1}^H\Gamma_{h,\theta}^{k-1} + 2\frac{H}{\kappa}\sum_{k=1}^K\sum_{h=1}^H\Gamma_{h,\mathbf{w}}^{k-1}}_{\text{bias part}} + 4H\sqrt{2T\log(1/\delta)}.$$

Bias. Now we set $w_{k,j} = \gamma^{k-j}, \gamma\in(0,1)$, same as Eq (62), we have

$$2\sum_{k=1}^K\sum_{h=1}^H\Gamma_{h,\theta}^{k-1} \leq 4L_\phi^2\sqrt{\frac{d}{\lambda_\theta}}\frac{1}{(1-\gamma)^{3/2}}\sum_{p=1}^{K-1}\sum_{h=1}^H\left\|\theta_h^p - \theta_h^{p+1}\right\|_2,$$

then we set $\alpha_{k,j} = \gamma^{k-j}, \gamma\in(0,1)$, similar to Eq (63), we have

$$2\frac{H}{\kappa}\sum_{k=1}^K\sum_{h=1}^H\Gamma_{h,\mathbf{w}}^{k-1}$$

$$\leq 4\frac{H}{\kappa}L_\psi^2\sqrt{\frac{d}{\lambda_{\mathbf{w}}}}\frac{1}{(1-\gamma)^{3/2}}\sum_{p=1}^{K-1}\sum_{h=1}^H\left\|\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right\|_2,$$

Variance. Same as Eq (64), we have

$$4H\beta_\theta\sum_{k=1}^K\sum_{h=1}^H\min\left\{1, \left\|\phi(s_h^k, a_h^k)\right\|_{(\Lambda_h^{k-1})^{-1}}\right\}$$

$$\leq 4H\beta_\theta\sqrt{KH}\sqrt{H2d\left(K\log\frac{1}{\gamma} + \log\left(1 + \frac{L_\phi^2}{\lambda_\theta d(1-\gamma)}\right)\right)}.$$

For the second term,

$$4\frac{H}{\kappa}\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\beta}_{\mathbf{w}}^{k-1}\min\left\{1,\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s_h^k,a_h^k)\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}\right\}$$

$$\leq 4\frac{H}{\kappa}\bar{\beta}_{\mathbf{w}}^{K}\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{1,\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s_h^k,a_h^k)\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}\right\}$$

$$\leq 4\frac{H^2}{\kappa}\bar{\beta}_{\mathbf{w}}^{K}\sqrt{K}\sqrt{\sum_{k=1}^{K}\min\left\{1,\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s_h^k,a_h^k)\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}^2\right\}}$$

$$\leq 4\frac{H^2}{\kappa}\bar{\beta}_{\mathbf{w}}^{K}\sqrt{K}\sqrt{\sum_{k=1}^{K}\min\left\{1,\sum_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s_h^k,a_h^k)\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}^2\right\}}.$$

Based on the Lemma 9 (Potential Lemma), we know that $\forall h\in[H]$, we have

$$\sum_{k=1}^{K}\min\left\{1,\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s_h^k,a_h^k)\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}^2\right\}$$
$$\leq 2d\left(K\log\frac{1}{\gamma}+\log\left(1+\frac{UL_\psi^2}{\lambda_{\mathbf{w}}d(1-\gamma)}\right)\right),$$

so we have

$$4\frac{H}{\kappa}\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\beta}_{\mathbf{w}}^{k-1}\min\left\{1,\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s_h^k,a_h^k)\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}\right\}$$
$$\leq 4\frac{H^2}{\kappa}\bar{\beta}_{\mathbf{w}}^{K}\sqrt{K}\sqrt{2d\left(K\log\frac{1}{\gamma}+\log\left(1+\frac{UL_\psi^2}{\lambda_{\mathbf{w}}d(1-\gamma)}\right)\right)}.$$

Since there is a term $HK\sqrt{\log(1/\gamma)}$ in the regret bound, we cannot let γ close to 0, so we set $\gamma\geq 1/K$ and have $\log(1/\gamma)\leq C(1-\gamma)$, where $C=\log K/(1-1/K)$. We set $\lambda_\theta=d$, and $\lambda_{\mathbf{w}}=d$. Combining the upper bounds of the bias and variance parts and with confidence level $\delta=1/(4T)$, by union bound we have the following dynamic regret bound with probability at least $1-1/T$,

$$\text{D-REG}_T\leq\mathcal{O}\left(\frac{1}{(1-\gamma)^{3/2}}P_T^\theta+\frac{H}{\kappa}\frac{1}{(1-\gamma)^{3/2}}P_T^{\mathbf{w}}\right.$$
$$\left.+dH^2K\sqrt{1-\gamma}+\frac{dH^2K}{\kappa}\sqrt{1-\gamma}+H^{3/2}d\sqrt{HK}\right)$$
$$\leq\mathcal{O}\left(\frac{Hd}{\kappa}\left(\frac{1}{(1-\gamma)^{3/2}}\Delta+HK\sqrt{1-\gamma}\right)+H^{3/2}d\sqrt{HK}\right).$$

Furthermore, by setting the discounted factor optimally as $\gamma=1-\max\left\{1/K,\sqrt{\Delta/T}\right\}$, we have

$$\text{D-REG}_T\leq\begin{cases}\widetilde{\mathcal{O}}\left(\kappa^{-1}Hd\Delta^{1/4}T^{3/4}\right) & \text{when }\Delta\geq H/K,\\ \widetilde{\mathcal{O}}\left(\kappa^{-1}dH^{3/2}\sqrt{T}\right) & \text{when }\Delta<H/K.\end{cases}$$

□

Proof of Lemma 22. We first consider the upper bound of E_h^k, based on the definition of $\bar{Q}_h^k$ (26) and model assumption (22) and Eq. (24), we have $\forall a\in\mathcal{A}, s\in\mathcal{S}$,

$$r_h^k(s,a)+\left[\mathbb{P}_h^k\bar{V}_{h+1}^k\right](s,a)-\bar{Q}_h^k(s,a)$$
$$=r_h^k(s,a)+\left[\mathbb{P}_h^k\bar{V}_{h+1}^k\right](s,a)-\phi(s,a)^\top\widehat{\theta}_h^k$$
$$\quad-\beta_\theta\left\|\phi(s,a)\right\|_{(\Lambda_h^{k-1})^{-1}}-[\widetilde{\mathbb{P}}_h^k\bar{V}_{h+1}^k](s,a)$$
$$\quad-\frac{H}{\kappa}\bar{\beta}_{\mathbf{w}}^{k-1}\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s,a)\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}$$
$$=\phi(s,a)^\top\left(\theta_h^k-\widehat{\theta}_h^k\right)-\beta_\theta\left\|\phi(s,a)\right\|_{(\Lambda_h^{k-1})^{-1}}$$
$$\quad+\left(\left[\mathbb{P}_h^k\bar{V}_{h+1}^k\right](s,a)-\left[\widetilde{\mathbb{P}}_h^k\bar{V}_{h+1}^k\right](s,a)\right)$$
$$\quad-\frac{H}{\kappa}\bar{\beta}_{\mathbf{w}}^{k-1}\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s,a)\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}$$
$$\leq\Gamma_{h,\theta}^{k-1}+\frac{H}{\kappa}\Gamma_{h,\mathbf{w}}^{k-1},$$

where the last inequality comes from Lemma 4 and Lemma 6. Similarly, we can get the lower bound of E_h^k, $\forall a\in\mathcal{A}, s\in\mathcal{S}$,

$$\bar{Q}_h^k(s,a)-r_h^k(s,a)-\left[\mathbb{P}_h^k\bar{V}_{h+1}^k\right](s,a)$$
$$=\phi(s,a)^\top\widehat{\theta}_h^k+\beta_\theta\left\|\phi(s,a)\right\|_{(\Lambda_h^{k-1})^{-1}}+[\widetilde{\mathbb{P}}_h^k\bar{V}_{h+1}^k](s,a)$$
$$\quad+\frac{H}{\kappa}\bar{\beta}_{\mathbf{w}}^{k-1}\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s,a)\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}$$
$$\quad-r_h^k(s,a)-\left[\mathbb{P}_h^k\bar{V}_{h+1}^k\right](s,a)$$
$$=\phi(s,a)^\top\left(\widehat{\theta}_h^k-\theta_h^k\right)+\left(\left[\widetilde{\mathbb{P}}_h^k\bar{V}_{h+1}^k\right](s,a)-\left[\mathbb{P}_h^k\bar{V}_{h+1}^k\right](s,a)\right)$$
$$\quad+\beta_\theta\left\|\phi(s,a)\right\|_{(\Lambda_h^{k-1})^{-1}}$$
$$\quad+\frac{H}{\kappa}\bar{\beta}_{\mathbf{w}}^{k-1}\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s,a)\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}$$
$$\leq\Gamma_{h,\theta}^{k-1}+\frac{H}{\kappa}\Gamma_{h,\mathbf{w}}^{k-1}+2\beta_\theta\left\|\phi(s,a)\right\|_{(\Lambda_h^{k-1})^{-1}}$$
$$\quad+2\frac{H}{\kappa}\bar{\beta}_{\mathbf{w}}^{k-1}\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s,a)\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}},$$

thus completes the proof. □

APPENDIX G
TECHNICAL LEMMAS

In this section, we provide several useful lemmas, mainly about concentrations, and some derivatives of self-concordant property.

A. Concentration inequalities

Lemma 23 (Hoeffding's Lemma). *Let Z be a real random variable such that $Z\in[a,b]$ almost surely. Then*

$$\forall\lambda\in\mathbb{R}\quad\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right]\leq\exp\left(\frac{\lambda^2(b-a)^2}{8}\right),$$

or variable $(Z-\mathbb{E}[Z])$ is $\frac{(b-a)}{2}$-sub-Gaussian.

Lemma 24 (Azuma-Hoeffding inequality). *Let $M>0$ be a constant. Let $\{x_i\}_{i=1}^n$ be a martingale difference sequence with respect to a filtration $\{\mathcal{G}_i\}_i$ ($\mathbb{E}\left[x_i\mid\mathcal{G}_i\right]=0$ a.s. and x_i is $\mathcal{G}_{i+1}$-measurable) such that for all $i\in[n],|x_i|\leq M$ holds*

almost surely. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$, we have

$$\sum_{i=1}^{n} x_i \leq M\sqrt{2n\log(1/\delta)}.$$

Theorem 7 (Self-normalized concentration inequality for linear bandits [32, Theorem 1]). *Let $\{F_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\eta_t\}_{t=0}^{\infty}$ be a real-valued stochastic process such that η_t is F_t-measurable and η_t is conditionally R-sub-Gaussian for some $R \geq 0$ i.e., $\forall \lambda \in \mathbb{R}$, $\mathbb{E}\left[e^{\lambda\eta_t} \mid F_{t-1}\right] \leq \exp(\frac{\lambda^2 R^2}{2})$. Let $\{X_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that X_t is F_{t-1}-measurable. Assume that V is a $d \times d$ positive definite matrix. For any $t \geq 0$, define*

$$V_t = V_0 + \sum_{s=1}^{t} X_s X_s^{\top}, \qquad S_t = \sum_{s=1}^{t} \eta_s X_s.$$

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

$$\|S_t\|_{V_t^{-1}} \leq \sqrt{2R^2 \log\left(\frac{\det(V_t)^{\frac{1}{2}}\det(V_0)^{-\frac{1}{2}}}{\delta}\right)}.$$

Theorem 8 (Self-normalized concentration inequality for self-concordant bandits [33, Theorem 1]). *Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\eta_t\}_{t=0}^{\infty}$ be a martingale difference sequence such that η_t is $\mathcal{F}_t$ measurable. 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 $\|X_t\|_2 \leq 1$. Furthermore, assume that conditionally on $\mathcal{F}_t$ we have $|\eta_t| \leq 1$ a.s., and denote $\sigma_t^2 = \mathbb{E}\left[\eta_t^2 \mid \mathcal{F}_{t-1}\right]$. For any $t \geq 0$, define*

$$H_t = \sum_{s=1}^{t} \sigma_s^2 X_s X_s^{\top} + \lambda I_d, \qquad S_t = \sum_{s=1}^{t} \eta_s X_s,$$

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

$$\|S_t\|_{H_t^{-1}} \leq \frac{\sqrt{\lambda}}{2} + \frac{2}{\sqrt{\lambda}}\log\left(\frac{\det(H_t)^{1/2}}{\delta\lambda^{d/2}}\right) + \frac{2}{\sqrt{\lambda}}d\log(2).$$

Lemma 25. *Suppose $U_0 = \lambda I_d$, $U_t = U_{t-1} + A_t A_t^{\top}$, and $A_t \in \mathbb{R}^d$, then*

$$\forall p \in [t-1], \quad \sum_{s=1}^{p} \|A_s\|_{U_{t-1}^{-1}}^2 \leq d. \tag{71}$$

Proof of Lemma 25.

$$\sum_{s=1}^{p} \|A_s\|_{U_{t-1}^{-1}}^2 = \sum_{s=1}^{p} \mathrm{Tr}(A_s^{\top} U_{t-1}^{-1} A_s) = \mathrm{Tr}\left(U_{t-1}^{-1}\sum_{s=1}^{p} A_s A_s^{\top}\right)$$

$$\leq \mathrm{Tr}\left(U_{t-1}^{-1}\sum_{s=1}^{p} A_s A_s^{\top}\right) + \mathrm{Tr}\left(U_{t-1}^{-1}\sum_{s=p+1}^{t-1} A_s A_s^{\top}\right)$$

$$+ \mathrm{Tr}\left(U_{t-1}^{-1}\lambda\sum_{i=1}^{d} \mathbf{e}_i \mathbf{e}_i^{\top}\right) = \mathrm{Tr}(I_d) = d.$$

$\square$

Lemma 26 (Determinant inequality). *We let the $V_t = \sum_{s=1}^{t} w_{t,s}X_s X_s^{\top} + \lambda I_d$, $V_0 = \lambda I_d$. Assume $\|\mathbf{x}\|_2 \leq L$ and we have,*

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

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

$$\mathrm{Tr}(V_t) = \mathrm{Tr}(\lambda I_d) + \sum_{s=1}^{t} w_{t,s}\mathrm{Tr}\left(X_s X_s^{\top}\right)$$
$$= \lambda d + \sum_{s=1}^{t} w_{t,s}\|X_s\|_2^2 \leq \lambda d + L^2\sum_{s=1}^{t} w_{t,s}. \tag{72}$$

Base on the definition of determinant and the upper bound of $\mathrm{Tr}(V_t)$ (72), we can get the upper bound for $\det(V_t)$,

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

$\square$

B. 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 28.

Lemma 27 (Lemma 9 of [33]). *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 [33] which can be easily proven.

Lemma 28. *With G_t defined in (49) and H_t defined in (50), the following inequalities hold*

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

Lemma 29 (Lemma 7 of [36]). *Denote by L_i the absolute value of cumulative rewards for episode i, i.e., $L_i \triangleq \left|\sum_{t=(i-1)\Delta+1}^{i\Delta} r_t(X_t)\right|$, then*

$$\Pr\left[\forall i \in [\lceil T/\Delta\rceil], L_i \leq LS\Delta + 2R\sqrt{\Delta\ln\frac{T}{\sqrt{\Delta}}}\right] \geq 1 - \frac{2}{T}.$$

APPENDIX H
BANDITS OVER BANDITS

A. BOB Algorithm

We divide the T rounds into equal-length episodes of size Δ, such that $\Delta = \lceil d\sqrt{T} \rceil$. In each episode, we run LB-WeightUCB with different discounted factors γ. Specifically, the γ comes from the candidate set $\mathcal{W}$,

$$\mathcal{W} = \left\{ \gamma_i = 1 - d^{-\frac{1}{2}} 2^{1-i} \mid i \in [N] \right\}, \quad (73)$$

where $N = \lceil \log_2(T/\sqrt{d}) \rceil + 1$ is the number of candidate values, and recall that S is the upper bound on the norm of the underlying regression parameters.

To adaptively select the optimal γ, we model the selection procedure as an adversarial bandit problem. In this formulation, each step corresponds to one episode of length Δ, such that has total $\lceil T/\Delta \rceil$ rounds. At each step, a bandit algorithm selects a γ from the candidate set $\mathcal{W}$, observes the cumulative reward from the LB-WeightUCB with corresponding discounted factor, and uses it as feedback to update its selection strategy.

Let $\gamma_{\min}$ ($\gamma_{\max}$) be the minimal (maximal) discounted factor in the candidate set $\mathcal{W}$, then it is evident to verify that

$$\gamma_{\min} = 1 - \frac{1}{\sqrt{d}}, \quad 1 > \gamma_{\max} \geq 1 - \frac{1}{T}.$$

B. Regret Analysis

Theorem 9. *LB-WeightUCB without the knowledge of path-length P_T, together with Bandits-over-Bandits mechanism satisfies with probability at least $1 - 3/T$,*

$$\text{D-REG}_T = \sum_{t=1}^{T} \max_{\mathbf{x} \in \mathcal{X}} \mathbf{x}^\top \theta_t - \sum_{t=1}^{T} X_t^\top \theta_t = \widetilde{\mathcal{O}} \left(d^{3/4} P_T^{1/4} T^{3/4} \right).$$

Proof of Theorem 9. We begin by decomposing the dynamic regret. Let $X_t^* \triangleq \arg\max_{\mathbf{x} \in \mathcal{X}} \mathbf{x}^\top \theta_t$ and we have

$$\text{D-REG}_T = \sum_{t=1}^{T} \left(X_t^{*\top} \theta_t - X_t^\top \theta_t \right)$$

$$= \underbrace{\sum_{t=1}^{T} X_t^{*\top} \theta_t - \sum_{i=1}^{\lceil T/\Delta \rceil} \sum_{t=(i-1)\Delta+1}^{i\Delta} X_t^{\widetilde{\gamma}^{*\top}} \theta_t}_{\text{base-regret}}$$

$$+ \underbrace{\sum_{i=1}^{\lceil T/\Delta \rceil} \sum_{t=(i-1)\Delta+1}^{i\Delta} \left(X_t^{\widetilde{\gamma}^{*\top}} \theta_t - X_t^{\gamma_i \top} \theta_t \right)}_{\text{meta-regret}},$$

where $\widetilde{\gamma}^*$ is the best discounted factor in the candidate set to approximate the optimal discounted factor $\gamma^* = 1 - \max\{1/T, \sqrt{P_T/(dT)}\}$.

Base-regret. Based on (37), and the union bound, we have with probability at least $1 - 2N\delta$,

$$\text{base-regret} = \sum_{t=1}^{T} X_t^{*\top} \theta_t - \sum_{i=1}^{\lceil T/\Delta \rceil} \sum_{t=(i-1)\Delta+1}^{i\Delta} X_t^{\widetilde{\gamma}^{*\top}} \theta_t$$

$$\leq \sum_{i=1}^{\lceil T/\Delta \rceil} \widetilde{\mathcal{O}} \left(\frac{1}{(1-\widetilde{\gamma}^*)^{3/2}} P_i + d(1-\widetilde{\gamma}^*)^{1/2} \Delta \right)$$

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

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

$$\leq \widetilde{\mathcal{O}}(d^{3/4} P_T^{1/4} T^{3/4}). \quad (74)$$

We know that $\gamma^* = 1 - \max\{1/T, \sqrt{P_T/(dT)}\}$ is the optimal discounted factor. Since $\gamma^* \in [\gamma_{\min}, \gamma_{\max}]$, the candidate set $\mathcal{W}$ covers γ^*. Furthermore, due to the geometric spacing of $\mathcal{W}$, there exists some $\widetilde{\gamma}^* \in \mathcal{W}$ such that

$$1 - \widetilde{\gamma}^* \leq 1 - \gamma^* \leq 2(1 - \widetilde{\gamma}^*).$$

Meta-regret. The analysis of meta-regret follows the proof for window-based algorithm [52, Proposition 1].

$$\text{meta-regret}$$

$$= \sum_{i=1}^{\lceil T/\Delta \rceil} \sum_{t=(i-1)\Delta+1}^{i\Delta} \left(X_t^{\widetilde{\gamma}^{*\top}} \theta_t - X_t^{\gamma_i \top} \theta_t \right)$$

$$= \sum_{i=1}^{\lceil T/\Delta \rceil} \sum_{t=(i-1)\Delta+1}^{i\Delta} \left(X_t^{\widetilde{\gamma}^{*\top}} \theta_t + \eta_t - X_t^{\gamma_i \top} \theta_t - \eta_t \right)$$

$$= \sum_{i=1}^{\lceil T/\Delta \rceil} \sum_{t=(i-1)\Delta+1}^{i\Delta} \left(r_t^{\widetilde{\gamma}^*} - r_t^{\gamma_i} \right)$$

$$= \sum_{i=1}^{\lceil T/\Delta \rceil} \left(L_i^{\widetilde{\gamma}^*} - L_i^{\gamma_i} \right),$$

Based on Lemma 29, we know that with probability at least $1 - \frac{2}{T}$, we have $\forall i \in [[T/\Delta]], L_i \leq LS\Delta + 2R\sqrt{\Delta \ln \frac{T}{\sqrt{\Delta}}}$ we define $L_{\max} \triangleq LS\Delta + 2R\sqrt{\Delta \ln \frac{T}{\sqrt{\Delta}}}$. We choose Exp3.IX [53] as the meta algorithm. Then we have with probability at least $1 - \delta$,

$$\sum_{i=1}^{\lceil T/\Delta \rceil} \left(L_i^{\widetilde{\gamma}^*} - L_i^{\gamma_i} \right) = \widetilde{\mathcal{O}} \left(L_{\max} \sqrt{\frac{T}{\Delta} N} \right)$$

$$= \widetilde{\mathcal{O}} \left(\sqrt{\Delta T N} \right) = d^{1/2} T^{3/4}. \quad (75)$$

Combining the upper bounds of base-regret (74) and meta-regret (75), by the union bound, and let $\delta = \frac{1}{(2N+1)T}$, we have with probability at least $1 - 3/T$,

$$\text{D-REG}_T = \widetilde{\mathcal{O}} \left(d^{3/4} P_T^{1/4} T^{3/4} + d^{1/2} T^{3/4} \right).$$

Thus we complete the proof. $\square$

References

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

[2] Y. Russac, C. Vernade, and O. Cappé, "Weighted linear bandits for non-stationary environments," in *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019, pp. 12 040–12 049.

[3] P. Zhao, L. Zhang, Y. Jiang, and Z.-H. Zhou, "A simple approach for non-stationary linear bandits," in *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020, pp. 746–755.

[4] Y. Russac, O. Cappé, and A. Garivier, "Algorithms for non-stationary generalized linear bandits," *ArXiv preprint*, vol. arXiv:2003.10113, 2020.

[5] B. Kim and A. Tewari, "Randomized exploration for non-stationary stochastic linear bandits," in *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020, pp. 71–80.

[6] L. Faury, Y. Russac, M. Abeille, and C. Calauzènes, "Regret bounds for generalized linear bandits under parameter drift," *ArXiv preprint*, vol. arXiv:2103.05750, 2021.

[7] Y. Russac, L. Faury, O. Cappé, and A. Garivier, "Self-concordant analysis of generalized linear bandits with forgetting," in *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021, pp. 658–666.

[8] C.-Y. Wei and H. Luo, "Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach," in *Proceedings of the 34th Conference on Learning Theory (COLT)*, 2021, pp. 4300–4354.

[9] Y. Deng, X. Zhou, B. Kim, A. Tewari, A. Gupta, and N. Shroff, "Weighted Gaussian process bandits for non-stationary environments," in *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022, pp. 6909–6932.

[10] Y. Liu, B. Van Roy, and K. Xu, "Nonstationary bandit learning via predictive sampling," in *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023, pp. 6215–6244.

[11] J. Wang, P. Zhao, and Z.-H. Zhou, "Revisiting weighted strategy for non-stationary parametric bandits," in *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023, vol. 206, 2023, pp. 7913–7942.

[12] S. Tomkins, P. Liao, P. V. Klasnja, and S. A. Murphy, "Intelligentpooling: practical Thompson sampling for mhealth," *Machine Learning*, vol. 110, no. 9, pp. 2685–2727, 2021.

[13] W. Huleihel, S. Pal, and O. Shayevitz, "Learning user preferences in non-stationary environments," in *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021, pp. 1432–1440.

[14] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan, "Provably efficient reinforcement learning with linear function approximation," in *Proceedings of 33rd Conference on Learning Theory (COLT)*, 2020, pp. 2137–2143.

[15] A. Touati and P. Vincent, "Efficient learning in non-stationary linear markov decision processes," *ArXiv preprint*, vol. arXiv:2010.12870, 2020.

[16] T. Hwang and M. Oh, "Model-based reinforcement learning with multinomial logistic function approximation," in *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*, 2023, pp. 7971–7979.

[17] L.-F. Li, Y.-J. Zhang, P. Zhao, and Z.-H. Zhou, "Provably efficient reinforcement learning with multinomial logit function approximation," in *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024, pp. 58 539–58 573.

[18] H. Zhong, Z. Yang, Z. Wang, and C. Szepesvári, "Optimistic policy optimization is provably efficient in non-stationary mdps," *ArXiv preprint*, vol. arXiv: 2110.08984, 2022.

[19] L.-F. Li, P. Zhao, and Z.-H. Zhou, "Near-optimal dynamic regret for adversarial linear mixture mdps," in *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024, pp. 55 858–55 883.

[20] O. Besbes, Y. Gur, and A. Zeevi, "Stochastic multi-armed-bandit problem with non-stationary rewards," *Advances in Neural Information Processing Systems 27 (NeurIPS)*, 2014.

[21] P. Zhao, Y. Zhang, L. Zhang, and Z.-H. Zhou, "Adaptivity and non-stationarity: Problem-dependent dynamic regret for online convex optimization," *Journal of Machine Learning Research*, vol. 25, pp. 98:1–98:52, 2024.

[22] Z.-H. Zhou, "Open-environment machine learning," *National Science Review*, vol. 9, no. 8, 2022.

[23] K. Crammer, Y. Mansour, E. Even-Dar, and J. W. Vaughan, "Regret minimization with concept drift," in *Proceedings of the 23rd Conference on Learning Theory (COLT)*, 2010, pp. 168–180.

[24] C. Chiang, C. Lee, and C. Lu, "Beating bandits in gradually evolving worlds," in *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, 2013, pp. 210–227.

[25] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, pp. 44:1–44:37, 2014.

[26] A. Ghosh, A. Sankararaman, K. Ramchandran, T. Javidi, and A. Mazumdar, "Competing bandits in non-stationary matching markets," *IEEE Transactions on Information Theory*, vol. 70, no. 4, pp. 2831–2850, 2024.

[27] S. Li, S. Zhao, Z. Cao, S. Huang, and S. Chen, "Robust domain adaptation with noisy and shifted label distribution," *Frontiers of Computer Science*, vol. 19, no. 3, p. 193310, 2025.

[28] C. Anagnostopoulos, D. K. Tasoulis, N. M. Adams, N. G. Pavlidis, and D. J. Hand, "Online linear and quadratic discriminant analysis with adaptive forgetting for streaming classification," *Statistical Analysis and Data Mining*, vol. 5, no. 2, pp. 139–166, 2012.

[29] P. Zhao, X. Wang, S. Xie, L. Guo, and Z.-H. Zhou, "Distribution-free one-pass learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, pp. 951–963, 2021.

[30] L. Guo, L. Ljung, and P. Priouret, "Performance analysis of the forgetting factor RLS algorithm," *International Journal of Adaptive Control and Signal Processing*, vol. 7, no. 6, pp. 525–537, 1993.

[31] Y. Chu and C. M. Mak, "A variable forgetting factor diffusion recursive least squares algorithm for distributed estimation," *Signal Processing*, vol. 140, pp. 219–225, 2017.

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

[33] L. Faury, M. Abeille, C. Calauzènes, and O. Fercoq, "Improved optimistic algorithms for logistic bandits," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020, pp. 3052–3060.

[34] M. Abeille, L. Faury, and C. Calauzènes, "Instance-wise minimax-optimal algorithms for logistic bandits," in *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021, pp. 3691–3699.

[35] P. Zhao and L. Zhang, "Non-stationary linear bandits revisited," *ArXiv preprint*, vol. arXiv:2103.05324, 2021.

[36] P. Zhao, L. Zhang, Y. Jiang, and Z.-H. Zhou, "A simple approach for non-stationary linear bandits," *ArXiv preprint*, vol. arXiv:2103.05324, 2021.

[37] J. Suk and S. Kpotufe, "Tracking most significant arm switches in bandits," in *Proceedings of the 35th Conference on Learning Theory (COLT)*, 2022, pp. 2160–2182.

[38] Y. Abbasi-Yadkori, A. György, and N. Lazić, "A new look at dynamic regret for non-stationary stochastic bandits," *Journal of Machine Learning Research*, vol. 24, no. 288, pp. 1–37, 2023.

[39] G. Clerici, P. Laforgue, and N. Cesa-Bianchi, "Linear bandits with memory: from rotting to rising," *ArXiv preprint*, vol. arXiv:2302.08345, 2023.

[40] W. C. Cheung, D. Simchi-Levi, and R. Zhu, "Hedging the drift: Learning to optimize under nonstationarity," *Management Science*, vol. 68, no. 3, pp. 1696–1713, 2022.

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

[42] Y.-J. Zhang, S.-A. Xu, P. Zhao, and M. Sugiyama, "Generalized linear bandits: Almost optimal regret with one-pass update," in *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025, p. to appear.

[43] Z. Jia, L. Yang, C. Szepesvári, and M. Wang, "Model-based reinforcement learning with value-targeted regression," in *Proceedings of the 2nd Annual Conference on Learning for Dynamics and Control (L4DC)*, 2020, pp. 666–686.

[44] A. Ayoub, Z. Jia, C. Szepesvári, M. Wang, and L. Yang, "Model-based reinforcement learning with value-targeted regression," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020, pp. 463–474.

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

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

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

[48] S. Dong, T. Ma, and B. V. Roy, "On the performance of thompson sampling on logistic bandits," in *Proceedings of the 32nd Conference on Learning Theory (COLT)*, 2019, pp. 1158–1160.

[49] C. Tekin and M. Liu, "Online learning of rested and restless bandits," *IEEE Transactions on Information Theory*, vol. 58, no. 8, pp. 5588–5611, 2012.

[50] P. N. Karthik, V. Y. F. Tan, A. Mukherjee, and A. Tajer, "Optimal best arm identification with fixed confidence in restless bandits," *IEEE Transactions on Information Theory*, vol. 70, no. 10, pp. 7349–7384, 2024.

[51] D. Baby and Y. Wang, "Online forecasting of total-variation-bounded sequences," in *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019, pp. 11 069–11 079.

[52] W. C. Cheung, D. Simchi-Levi, and R. Zhu, "Hedging the drift: Learning to optimize under non-stationarity," *ArXiv preprint*, vol. arXiv:1903.01461, 2019.

[53] G. Neu, "Explore no more: Improved high-probability regret bounds for non-stochastic bandits," in *Advances in Neural Information Processing Systems 28 (NIPS)*, 2015, pp. 3168–3176.

Jing Wang received the BSc degree from the Nanjing University of Aeronautics and Astronautics in 2021. Currently, he is working toward the PhD degree with the School of Artificial Intelligence, Nanjing University. His research interest is mainly on machine learning and data mining.

Peng Zhao (Member, IEEE) received the BSc degree from Tongji University in 2016 and the PhD degree from Nanjing University in 2021. He is currently an assistant professor with the School of Artificial Intelligence, Nanjing University. His research interests lie in the foundations of machine learning, with a focus on online learning, bandits, and optimization. He has published over 60 papers in total across top-tier journals, such as *Journal of Machine Learning Research* and IEEE/ACM Transactions, and premier conferences, including ICML, NeurIPS, and COLT. He regularly serves as an area chair for ICML and NeurIPS, and is an action editor for *Machine Learning* (Springer).

Zhi-Hua Zhou (Fellow, IEEE) received the BSc, MSc, and PhD degrees (Hons.) in computer science from Nanjing University, Nanjing, China, in 1996, 1998, and 2000, respectively. He joined Nanjing University as an assistant professor, in 2001, where he is currently a professor and the vice president. He is also the founding director of the LAMDA Group. He has authored the books Ensemble Methods: Foundations and Algorithms, Evolutionary Learning: Advances in Theories and Algorithms, Machine Learning, and has published more than 200 papers in top-tier international journals or conference proceedings. He holds more than 30 patents. His research interests are mainly in artificial intelligence, machine learning, and data mining. He is a fellow of ACM, AAAI, AAAS, etc. He received various awards/honors, including the National Natural Science Award of China, the IEEE Computer Society Edward J. McCluskey Technical Achievement Award, and the CCF-ACM Artificial Intelligence Award. He founded Asian Conference on Machine Learning (ACML). He is the president of IJCAI Trustee, a series editor of Lecture Notes in Artificial Intelligence (Springer), an advisory board member of AI Magazine, the editor-in-chief of *Frontiers of Computer Science*, and the associate editor-in-chief of *Science China Information Sciences*.