NONSTATIONARY GENERALIZED LINEAR BANDITS WITH DISCOUNTED ONLINE MIRROR DESCENT

Anonymous authors
Paper under double-blind review

ABSTRACT

We study nonstationary generalized linear bandits (GLBs), where the expected reward is modeled through a nonlinear link function with an unknown *time-varying* parameter. This framework encompasses a broad class of reward models, including linear, Bernoulli, and binomial rewards. Existing approaches are predominantly based on maximum-likelihood estimation (MLE), using sliding-window, restart, or discounting mechanisms to handle nonstationarity. Although these methods achieve statistically efficient regret guarantees, they generally require revisiting past observations at every round, which leads to computation and memory costs that grow with time; moreover, several of them rely on a non-convex projection step. In this paper, we propose `DOMD-GLB`, a new algorithm for nonstationary GLBs that utilizes *discounted online mirror descent (DOMD)* for parameter estimation, thereby incurring only $\mathcal{O}(1)$ computation and memory costs per round. We prove dynamic regret bounds of order $\tilde{\mathcal{O}}\big(c_\mu^{-1/2} d^{3/4} P_T^{1/4} T^{3/4}\big)$ in drifting environments and $\tilde{\mathcal{O}}\big(c_\mu^{-1/3} d^{2/3} \Gamma_T^{1/3} T^{2/3}\big)$ in piecewise-stationary environments, where d denotes the feature dimension, T the time horizon, P_T the path length, Γ_T the number of change points, and c_μ a curvature parameter associated with the link function, while substantially improving computational efficiency over prior work. To the best of our knowledge, this is the first algorithm for nonstationary GLBs with per-round computation and memory costs independent of time.

1 INTRODUCTION

Nonstationary parametric bandits (Cheung et al., 2019; Russac et al., 2019; Zhao et al., 2020; Russac et al., 2020; Faury et al., 2021; Russac et al., 2021; Wei & Luo, 2021; Deng et al., 2022; Liu et al., 2023; Wang et al., 2023; Hong et al., 2023; Jia et al., 2023; Wang et al., 2024; Li & Li, 2025) model sequential decision-making problems in which the reward distribution is governed by an unknown *time-varying* parameter. This framework has attracted substantial attention in recent years, due to its relevance to a wide range of online applications in evolving environments, including recommendation systems (Tomkins et al., 2008; Huleihel et al., 2021), dynamic pricing (Zhao et al., 2023), and related settings with changing user preferences or market conditions.

In this paper, we study nonstationary generalized linear bandits (GLBs) (Russac et al., 2020; Faury et al., 2021; Russac et al., 2021; Wei & Luo, 2021; Wang et al., 2023) with *time-varying contextual action sets*. At each round t, the learner observes a feasible action set $\mathcal{X}_t \subseteq \mathbb{R}^d$, which may vary arbitrarily over time, selects an action $x_t \in \mathcal{X}_t$, and receives a reward satisfying $\mathbb{E}[r_t \mid x_t] = \mu(x_t^\top \theta_t^\star)$. Here, $\mu : \mathbb{R} \to \mathbb{R}$ is a known nonlinear link function determined by the underlying generalized linear model, and $\theta_t^\star \in \mathbb{R}^d$ is an unknown *time-varying* parameter. Thus, both the available contextual actions $\mathcal{X}_t$ and the underlying reward parameter $\theta_t^\star$ may change over time. To quantify parameter nonstationarity, we consider two standard measures: (i) the path length $P_T := \sum_{t=1}^{T-1} \|\theta_{t+1}^\star - \theta_t^\star\|_2$ in drifting environments, and (ii) the number of change points $\Gamma_T := \sum_{t=1}^{T-1} \mathbb{1}\{\theta_{t+1}^\star \neq \theta_t^\star\}$

To handle nonstationarity, existing GLB methods rely predominantly on MLE-based approaches, including restart (Zhao et al., 2020; Wei & Luo, 2021) and weighted (or discounted) (Russac et al., 2020; Faury et al., 2021; Russac et al., 2021; Wang et al., 2023) strategies (see Table 1). Notably, the GLB guarantee of `MASTER` (Wei & Luo, 2021) is established for a *fixed* action set and does not directly extend to the *time-varying action* sets considered in our setting. Beyond this difference

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of dynamic regret bounds for nonstationary GLM bandits. Here, d is the feature dimension, $1/c_\mu$ measures nonlinearity, and P_T and Γ_T denote the path length and number of change points. For computation and memory, we report only the dependence on t. "SC" denotes self-concordance, and "No poly-time guar." means that no polynomial-time guarantee is available. "Varying" allows time-varying feasible action sets; "Fixed" assumes a fixed action set; and "Fixed*" indicates that the stated result is for a fixed action set but its analysis appears to extend directly to varying action sets. ‡The original `RestartGLB` analysis contains a technical flaw. Subsequent work (Zhao & Zhang, 2021) corrects the underlying linear-bandit analysis and obtains $\widetilde{O}(d^{7/8}P_T^{1/4}T^{3/4})$, but does not provide a corrected GLB guarantee.

	Paper	Algorithm Type	Assumption	Regret	Comput. per Round	Memory
Drifting	Zhao et al. (2020) (`RestartGLB`)	Restart + MLE	Sub-Gaussian Noise + Fixed*	$\tilde{\mathcal{O}}\left(c_\mu^{-1}d^{2/3}P_T^{1/3}T^{2/3}\right)^{\ddagger}$	$\mathcal{O}(T^{2/3})$	$\mathcal{O}(T^{2/3})$
	Faury et al. (2021) (`BVD-GLM-UCB`)	Discount + MLE	Bounded Reward + Varying	$\tilde{\mathcal{O}}\left(c_\mu^{-1}d^{9/10}P_T^{1/5}T^{4/5}\right)$	No poly-time guar.	$\mathcal{O}(t)$
	Wei & Luo (2021) (`MASTER + GLM-UCB`)	Restart + MLE	Bounded Reward + Fixed	$\tilde{\mathcal{O}}\left(c_\mu^{-1}dP_T^{1/3}T^{2/3}\right)$	No poly-time guar.	$\mathcal{O}(t\log t)$
	Wang et al. (2023) (`GLB-WeightUCB`)	Discount + MLE	Sub-Gaussian Noise + Fixed*	$\tilde{\mathcal{O}}\left(c_\mu^{-3/4}d^{3/4}P_T^{1/4}T^{3/4}\right)$	No poly-time guar.	$\mathcal{O}(t)$
	Wang et al. (2023) (`SCB-WeightUCB`)	Discount + MLE	Bounded Reward + SC + Fixed*	$\tilde{\mathcal{O}}\left(c_\mu^{-1/2}d^{3/4}P_T^{1/4}T^{3/4}\right)$	No poly-time guar.	$\mathcal{O}(t)$
	This Work (`DOMD-GLB`, Theorem 2)	Discount + OMD	Bounded Reward + Varying	$\tilde{\mathcal{O}}\left(c_\mu^{-1/2}d^{3/4}P_T^{1/4}T^{3/4}\right)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$
Piecewise Stationary	Russac et al. (2020) (`D-GLUCB`)	Discount + MLE	Bounded Reward + Varying	$\tilde{\mathcal{O}}\left(c_\mu^{-1}d^{2/3}\Gamma_T^{1/3}T^{2/3}\right)$	No poly-time guar.	$\mathcal{O}(t)$
	Wei & Luo (2021) (`MASTER + GLM-UCB`)	Restart + MLE	Bounded Reward + Fixed	$\tilde{\mathcal{O}}(c_\mu^{-1}d\sqrt{\Gamma_T T})$	No poly-time guar.	$\mathcal{O}(t\log t)$
	Russac et al. (2021) (`SC-D-GLUCB`)	Discount + MLE	Bounded Reward + SC + Varying	$\tilde{\mathcal{O}}\left(c_\mu^{-1/3}d^{2/3}\Gamma_T^{1/3}T^{2/3}\right)$	$\mathcal{O}(t)$	$\mathcal{O}(t)$
	Wang et al. (2023) (`SCB-PW-WeightUCB`)	Discount + MLE	Bounded Reward + SC + Fixed*	$\tilde{\mathcal{O}}\left(d^{2/3}\Gamma_T^{1/3}T^{2/3}\right)$	No poly-time guar.	$\mathcal{O}(t)$
	This Work (`DOMD-GLB`, Theorem 3)	Discount + OMD	Bounded Reward + Varying	$\tilde{\mathcal{O}}\left(c_\mu^{-1/3}d^{2/3}\Gamma_T^{1/3}T^{2/3}\right)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$

in problem scope, existing MLE-based methods also incur substantial computational and memory costs, since their estimators must be repeatedly recomputed from retained observations. Moreover, several prior works (Faury et al., 2021; Russac et al., 2020; Wang et al., 2023) rely on a *non-convex projection step* of the form $\widetilde{\theta}_t \in \arg\min_\theta \left\| \sum_s \left(\mu(x_s^\top \theta) - \mu(x_s^\top \widehat{\theta}_t)\right) x_s \right\|_{V_t^{-1}}$, where $V_t \in \mathbb{R}^{d\times d}$ is positive definite. This projection is generally difficult to compute in polynomial time.

On the other hand, recent work on *stationary* GLBs (Zhang et al., 2025) showed that online mirror descent (OMD) can estimate the model parameter in a *one-pass* manner, without storing past data. This yields constant per-round computation and memory costs while retaining confidence guarantees comparable to those of statistically efficient MLE-based methods, such as Lee et al. (2024). This suggests OMD as a natural starting point for developing computationally efficient algorithms for nonstationary GLBs. This leads to the following question:

Can we design a nonstationary GLB algorithm with constant computation and memory costs?

In this paper, we answer this question affirmatively by proposing `DOMD-GLB`, a discounted OMD (DOMD)-based algorithm that achieves both statistical and computational efficiency.

Technical difficulties. The stationary OMD analysis of Zhang et al. (2025) does not directly extend to the discounted setting, and several key components of the analysis must be redesigned. The first challenge lies in the *mirror-descent geometry*. In stationary GLBs (Zhang et al., 2025), the additively accumulated Hessian naturally induces the geometry underlying the OMD update. In the nonstationary setting, however, this geometry must incorporate forgetting: stale curvature information should be progressively downweighted rather than accumulated uniformly. Consequently, the stationary geometry is no longer compatible with the discounted recursion, and it is not clear *a priori* what the appropriate discounted analogue of the stationary Bregman divergence should be. Constructing this geometry is therefore a central difficulty in extending OMD to nonstationary GLBs.

The second challenge lies in the *concentration analysis*. The confidence analysis of Zhang et al. (2025) is based on a *mix-loss* argument (Vovk, 2001), which constructs a suitable exponentiated stochastic process and shows that it forms a supermartingale, thereby allowing the use of *Ville's inequality* (Ville, 1939) to derive an anytime concentration bound. In the discounted setting, however,

the corresponding discounted process is generally not a supermartingale, and this argument therefore breaks down. As a result, a different concentration analysis is required (see Remark B.1).

The third challenge is to control the *loss drift* induced by nonstationarity. In the DOMD confidence analysis, the reference parameter is no longer fixed over time, since the current parameter may differ from the parameters underlying past observations. This mismatch creates an additional *loss-drift term* that is absent in the stationary setting. The main difficulty is therefore to isolate and control this term within the DOMD confidence argument. Unlike prior nonstationary MLE-based analyses (e.g., Wang et al. 2023), where the effect of parameter drift can be isolated as an explicit additive *parameter-bias term*, DOMD requires controlling the drift in *loss space*, through terms of the form $\ell_s(\theta_t^\star) - \ell_s(\theta_s^\star)$. Thus, controlling loss drift requires a separate analysis tailored to the DOMD estimator.

Contributions. To address these difficulties, we develop a proper discounted mirror-descent geometry (Subsection 4.1), a new mix-loss-based analysis tailored to DOMD (Lemma B.2), and a separate loss-drift bound in terms of the nonstationarity measure (Lemma B.4). Together, these yield a high-probability confidence bound for the DOMD estimator (Theorem 1), which in turn leads to dynamic regret guarantees in both drifting (Theorem 2) and piecewise-stationary environments (Theorem 3). Our contributions are summarized as follows:

- We propose a DOMD method for parameter estimation in nonstationary generalized linear bandits (GLBs). The resulting update requires only $\mathcal{O}(d^3)$ computation and $\mathcal{O}(d^2)$ memory per round, both independent of the time index t (Subsection 4.1).
- We establish a new confidence bound for DOMD in nonstationary GLBs by extending the *mix-loss* method (Vovk, 2001) to the discounted recursion and by explicitly controlling the drift terms that arise from nonstationarity (Theorem 1).
- In the drifting setting, we prove the dynamic regret bound $\tilde{\mathcal{O}}\big(c_\mu^{-1/2} d^{3/4} P_T^{1/4} T^{3/4}\big)$, matching the best-known regret order among discounted MLE-based methods (Wang et al., 2023) while using a substantially more efficient update (Theorem 2).
- In the piecewise-stationary setting, we obtain the regret bound $\tilde{\mathcal{O}}\big(c_\mu^{-1/3} d^{2/3} \Gamma_T^{1/3} T^{2/3}\big)$, showing that the DOMD framework also provides a strong statistical guarantee when nonstationarity is measured by the number of parameter changes (Theorem 3).

2 Related Works

Nonstationary parametric bandits. Nonstationary parametric bandits have been studied extensively across linear and generalized linear models (Cheung et al., 2019; Russac et al., 2019; Zhao et al., 2020; Russac et al., 2020; Faury et al., 2021; Russac et al., 2021; Wei & Luo, 2021; Deng et al., 2022; Liu et al., 2023; Hong et al., 2023; Jia et al., 2023; Wang et al., 2023; 2024; Li & Li, 2025). In the nonstationary GLB setting most relevant to our work, existing methods rely on MLE-based estimators, adapting them to nonstationarity through windowing, discounting, or restarting (Russac et al., 2020; Faury et al., 2021; Russac et al., 2021; Wang et al., 2023). Among restart-based approaches, `MASTER` (Wei & Luo, 2021) provides an adaptive black-box framework that detects nonstationarity and restarts the base algorithm, achieving the optimal dynamic regret $\tilde{\mathcal{O}}(\min\{\sqrt{\Gamma_T T},\, P_T^{1/3} T^{2/3}\})$. Although these methods enjoy strong statistical guarantees, they typically require repeated re-optimization over past data, leading to computation and memory costs that grow with time, the window size, or the restart period. Moreover, some also rely on a non-convex projection step that is generally not computable in polynomial time.

Stationary generalized linear bandits. The stationary GLB literature has long faced a trade-off between statistical sharpness and computational efficiency. The pioneering optimism-based algorithm of Filippi et al. (2010) requires repeated maximum likelihood estimation (MLE) over the full history, and its regret scales with the global slope parameter $1/c_\mu$, which can be exponentially large in the parameter-norm bound. Subsequent work has pursued two complementary directions: statistical sharpness and computational efficiency. On the statistical side, recent MLE-based analyses use self-concordance and local curvature to replace this global dependence with sharper instance-dependent quantities, yielding nearly optimal regret bounds in terms of local parameters such as $\kappa^\star$ (Lee et al., 2024; Liu et al., 2024; Sawarni et al., 2024). On the computational side, efficient methods such as `GLOC` (Jun et al., 2017) use online updates and achieve constant per-round cost, but their regret

bounds still depend linearly on $1/c_\mu$. More recently, Zhang et al. (2025) showed that these two goals can be achieved *simultaneously*, obtaining nearly optimal regret while maintaining fully online updates with constant time and memory per round.

3 PRELIMINARIES

Notations. For $M \succeq 0$ and $x \in \mathbb{R}^d$, let $\|x\|_M := \sqrt{x^\top M x}$ and $\|x\|_2 := \sqrt{x^\top x}$. We write $[T] := \{1, \dots, T\}$, use $A \succeq B$ to mean that $A - B$ is positive semidefinite, and denote the first and second derivatives of $f : \mathbb{R} \to \mathbb{R}$ by f' and f'', respectively.

3.1 NONSTATIONARY GENERALIZED LINEAR BANDIT

We consider a T-round contextual bandit problem. At each round $t \in [T]$, the learner observes a feasible action set $\mathcal{X}_t \subset \mathbb{R}^d$, chooses an action $x_t \in \mathcal{X}_t$, and then receives a stochastic reward $r_t \in \mathbb{R}$. The environment is governed by an unknown time-varying parameter sequence $\{\theta_t^\star\}_{t=1}^\infty$. Conditioned on the chosen action x_t, the reward follows a canonical exponential-family distribution whose natural parameter is given by the linear predictor $x_t^\top \theta_t^\star$, i.e., $\Pr[r_t | x_t, \theta_t^\star] = \exp\Big(\frac{r_t x_t^\top \theta_t^\star - m(x_t^\top \theta_t^\star)}{g(\tau)} + h(r_t, \tau)\Big)$, where $m : \mathbb{R} \to \mathbb{R}$ is convex and three times continuously differentiable, $g(\tau) > 0$ is a known dispersion term that controls the variability of the distribution, and $h(\cdot, \tau)$ is the base measure. Let $\mu(z) := m'(z)$. Then the conditional mean satisfies $\mathbb{E}[r_t | x_t, \theta_t^\star] = \mu(x_t^\top \theta_t^\star)$ (Proposition 3.1 of Wainwright & Jordan 2008). We assume that the parameter path $\{\theta_t^\star\}_{t \geq 1}$ is fixed independently of the reward noise process, following the standard exogenous-drift formulation in the nonstationary bandit literature (Cheung et al., 2019; Russac et al., 2019; 2021; Zhao et al., 2020; Wang et al., 2023).

We make the following standard assumptions used in GLBs (Filippi et al., 2010):

Assumption 1 (Boundedness)**.** *For all $t \geqslant 1$, we assume $\theta_t^\star \in \Theta := \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leqslant S\}$, $\|x\|_2 \leqslant 1$ for all $x \in \mathcal{X}_t$, and $r_t \in [0, R]$ almost surely for some known $R \in \mathbb{R}$.*

Assumption 2 (Bounded link function)**.** *The link function μ is twice differentiable over its feasible domain. Moreover, there exist constants $c_\mu, k_\mu > 0$ such that $c_\mu \leqslant \mu'(z) \leqslant k_\mu$ for all $z \in [-S, S]$. Consequently, the function m is strictly convex and μ is strictly increasing.*

Proposition 1 (Self-concordance, Lemma 2.2 of Sawarni et al. 2024)**.** *For any GLM supported on $[0, R]$ almost surely, the link function $\mu(\cdot)$ satisfies $|\mu''(z)| \leqslant R \cdot \mu'(z)$ for all $z \in \mathbb{R}$.*

Therefore, under the bounded rewards in Assumption 1, self-concordance is automatically satisfied. As a result, in previous works such as (Russac et al., 2021; Wang et al., 2023), the self-concordance assumption is in fact redundant, since bounded rewards are already imposed. More importantly, this induced self-concordant property is essential for obtaining regret bounds with improved dependence on $1/c_\mu$, which can be exponentially large in the parameter-norm bound, i.e., $1/c_\mu = \mathcal{O}(e^S)$.

In this paper, we consider two standard forms of nonstationarity.

Setting 1: Drifting environment. The environment exhibits gradual drift over time, and nonstationarity is measured by the path length $P_T := \sum_{t=1}^{T-1} \|\theta_{t+1}^\star - \theta_t^\star\|_2$, which quantifies the cumulative variation of the underlying parameter sequence.

Setting 2: Piecewise-stationary environment. The environment is piecewise stationary over time, and nonstationarity is measured by the number of parameter changes $\Gamma_T := \sum_{t=1}^{T-1} \mathbb{1}\{\theta_{t+1}^\star \neq \theta_t^\star\}$, which counts how many times the underlying parameter changes over the horizon.

Our goal is to minimize the cumulative *dynamic regret* $\mathbf{Reg}_T := \sum_{t=1}^T \big[\mu((x_t^\star)^\top \theta_t^\star) - \mu(x_t^\top \theta_t^\star)\big]$, where $x_t^\star \in \arg\max_{x \in \mathcal{X}_t} \mu(x^\top \theta_t^\star) = \arg\max_{x \in \mathcal{X}_t} x^\top \theta_t^\star$ denotes an optimal action at round t.

4 ALGORITHM AND MAIN RESULTS

In this section, we first introduce a DOMD method (Subsection 4.1), then present the OMD-based nonstationary GLB algorithm `DOMD-GLB` (Subsection 4.2), and finally establish dynamic regret guarantees for both drifting (Subsection 4.3) and piecewise-stationary (Subsection 4.4) settings.

4.1 Discounted Online Mirror Descent

We estimate the current parameter via *discounted online mirror descent* (DOMD). For each round t, we define the loss function as the negative log-likelihood $\ell_t(\theta) := \left(m(x_t^\top\theta) - r_t x_t^\top\theta\right)/g(\tau)$. Given the current iterate θ_t, we use the local quadratic surrogate $\widetilde{\ell}_t(\theta) := \ell_t(\theta_t) + \langle\nabla\ell_t(\theta_t),\, \theta-\theta_t\rangle + \frac{1}{2}\|\theta-\theta_t\|^2_{\nabla^2\ell_t(\theta_t)}$. Let $H_t = \lambda I_d + \sum_{s=1}^{t-1}\gamma^{t-1-s}\nabla^2\ell_s(\theta_{s+1})$, and $A_t := \gamma H_t + (1-\gamma)\lambda I_d$, where $\gamma\in(0,1)$ is the discount factor. At each round t, we update the parameter via:

$$\theta_{t+1} \in \operatorname*{arg\,min}_{\theta\in\Theta}\left\{\widetilde{\ell}_t(\theta) + \frac{1}{2\eta}\|\theta-\theta_t\|^2_{A_t}\right\}, \qquad \text{where } \eta := 1 + RS, \tag{1}$$

and subsequently update the curvature matrix via $H_{t+1} = A_t + \nabla^2\ell_t(\theta_{t+1}) = A_t + \frac{\mu'(x_t^\top\theta_{t+1})}{g(\tau)}x_t x_t^\top$.

Role of the surrogate loss. The surrogate loss $\widetilde{\ell}_t(\theta)$ plays a central role in balancing statistical efficiency and computational efficiency. Unlike a purely first-order linearization, $\widetilde{\ell}_t(\theta)$ retains the local curvature of the current loss through $\nabla^2\ell_t(\theta_t)$, which is important for capturing the nonlinear geometry of the GLM. At the same time, it converts the round-t update into a quadratic optimization problem. Indeed, after removing terms independent of θ, equation 1 is equivalent to

$$\theta_{t+1} \in \operatorname*{arg\,min}_{\theta\in\Theta}\left\{\langle\nabla\ell_t(\theta_t),\, \theta-\theta_t\rangle + \frac{1}{2}\|\theta-\theta_t\|^2_{\nabla^2\ell_t(\theta_t)+\eta^{-1}A_t}\right\}.$$

Thus, the algorithm preserves second-order information from the current sample while avoiding the need to solve a global weighted MLE problem over the entire historical dataset.

Discounted curvature versus stationary curvature. To highlight the role of discounting, consider the curvature matrix that would arise in a stationary environment, $H_t^{\mathrm{stat}} := \lambda I_d + \sum_{s=1}^{t-1}\nabla^2\ell_s(\theta_{s+1})$ (e.g., Zhang et al. 2025). In the stationary case, such an undiscounted accumulation is natural, since all past observations are generated under the same underlying parameter and should contribute equally to the local geometry. In contrast, in a nonstationary environment, old observations are collected under outdated parameters and should therefore have progressively less influence on the current update. For this reason, we use the discounted curvature matrix $H_t = \lambda I_d + \sum_{s=1}^{t-1}\gamma^{t-1-s}\nabla^2\ell_s(\theta_{s+1})$, which exponentially downweights stale curvature information. This discounting allows the estimator to adapt to drift while still retaining enough historical curvature to stabilize estimation.

Why A_t instead of H_t? A further distinction, specific to the discounted setting, is that the update rule in equation 1 uses A_t, not H_t, in the proximal geometry. This is important. The matrix H_t is the *post-round-*$(t-1)$ discounted curvature matrix, whereas the actual geometry needed for the round-t update is the *pre-update* matrix $A_t = \gamma H_t + (1-\gamma)\lambda I_d = \lambda I_d + \sum_{s=1}^{t-1}\gamma^{t-s}\nabla^2\ell_s(\theta_{s+1})$. Hence, A_t is exactly the one-step-aged version of H_t: every past curvature contribution receives one additional factor of γ before the new sample is incorporated. This is the correct geometry at round t, since historical information from round s should enter the round-t update with weight γ^{t-s}, not γ^{t-1-s}.

Remark 1 (Constant computation and memory costs)**.** *The DOMD update has a recursive form: it depends only on the current sample, the current estimate θ_t, and the maintained curvature matrix H_t. Thus, it can be implemented without storing or revisiting past observations, requiring only $\mathcal{O}(d^2)$ memory and $\mathcal{O}(d^3)$ computation per round, both independent of t. By contrast, window-based methods (Russac et al., 2020) must store and process a window of past observations, so their costs scale with the window size $\tau = \mathcal{O}(T^{2/3})$. Weighted MLE-based methods (Russac et al., 2021; Wang et al., 2023) similarly revisit past weighted observations, leading to costs that grow with t.*

Theorem 1 (Discounted-OMD confidence bound)**.** *Let $\eta = 1+RS$, $\alpha = \frac{3\eta}{2}$ and $\iota = \sqrt{\frac{2\eta k_\mu}{g(\tau)}}$. Suppose $\lambda \geqslant \max\left\{\frac{6\eta R k_\mu S}{g(\tau)}, \frac{32\alpha d R^2}{7}\right\}$. For each $t\geqslant 2$, define the confidence set*

$$\mathcal{C}_t(\delta) := \left\{\theta\in\mathbb{R}^d \mid \|\theta-\theta_t\|_{H_t} \leqslant \beta_t(\delta) + \iota\mathcal{V}_t\right\},$$

where $\mathcal{V}_t := \sum_{p=1}^{t-1}\gamma^{(t-1-p)/2}\sqrt{\frac{1-\gamma^p}{1-\gamma}}\|\theta^\star_{p+1}-\theta^\star_p\|_2$, and

$$\beta_t^2(\delta) := 4\lambda S^2 + 2\eta\left(1+\frac{R^2}{g(\tau)k_\mu}\right)\log\left(\frac{\pi^2 t^2}{3\delta}\right) + 2\eta\left(3\eta+\frac{1}{2}\right)d\log\left(1+\frac{k_\mu(1-\gamma^{t-1})}{\lambda d\, g(\tau)(1-\gamma)}\right).$$

Then, under Assumptions 1 and 2, we have $\Pr\left[\forall t\geqslant 2,\ \theta_t^\star\in\mathcal{C}_t(\delta)\right]\geqslant 1-\delta$.

Discussion of Theorem 1. Theorem 1 shows that one can construct an ellipsoidal confidence set $\mathcal{C}_t(\delta)$ around the DOMD iterate θ_t that is both *statistically efficient*, in the sense that the radius has no explicit $1/c_\mu$ dependence, and *computationally efficient*, with constant per-round computation and storage costs. In contrast to non-discounted OMD (Zhang et al., 2025), which treats all past observations equally and is therefore better suited to stationary environments, our confidence bound explicitly incorporates forgetting through the discounted matrix H_t and the drift term $\mathcal{V}_t$. As a result, stale data are progressively downweighted, so the confidence set tracks the current parameter more accurately in nonstationary environments. The proof is deferred to Appendix B.

Proof sketch of Theorem 1. We provide a proof sketch and highlight the technical novelties.

1) Reduction to discounted inverse regret. We first upper bound the estimation error of θ_t as:

$$\|\theta_t - \theta_t^\star\|_{H_t}^2 \leqslant 4\lambda S^2 + 2\eta \underbrace{\sum_{s=1}^{t-1} \gamma^{t-1-s}\big(\ell_s(\theta_t^\star) - \ell_s(\theta_{s+1})\big)}_{\text{discounted inverse-regret}} - \frac{2}{3}\sum_{s=1}^{t-1} \gamma^{t-1-s}\|\theta_{s+1} - \theta_s\|_{A_s}^2. \quad (2)$$

Note that the last term on the right-hand side is measured in the A_s-norm because the DOMD update uses the pre-update geometry A_s, whereas the left-hand side is measured in the post-update matrix H_t. This geometric mismatch is specific to the discounted setting and is absent in the stationary OMD analysis (Zhang & Sugiyama, 2024; Lee & Oh, 2024; 2025b; Zhang et al., 2025).

2) Decomposition of discounted inverse regret. To analyze the discounted inverse-regret, we introduce the mix loss $m_s(\Pi_s) := -\log \mathbb{E}_{\theta\sim\Pi_s}\big[e^{-\ell_s(\theta)}\big]$, where $\Pi_s := \mathcal{N}(\theta_s, \alpha A_s^{-1})$. Then, we have

$$\begin{aligned}
&\sum_{s=1}^{t-1}\gamma^{t-1-s}\big(\ell_s(\theta_t^\star) - \ell_s(\theta_{s+1})\big) \\
&= \underbrace{\sum_{s=1}^{t-1}\gamma^{t-1-s}\big(\ell_s(\theta_s^\star) - m_s(\Pi_s)\big)}_{\text{concentration term}} + \sum_{s=1}^{t-1}\gamma^{t-1-s}\big(m_s(\Pi_s) - \ell_s(\theta_{s+1})\big) + \underbrace{\sum_{s=1}^{t-1}\gamma^{t-1-s}\big(\ell_s(\theta_t^\star) - \ell_s(\theta_s^\star)\big)}_{\text{loss-drift term}}.
\end{aligned}$$

Unlike the stationary setting of Zhang et al. (2025), our decomposition contains an additional *loss-drift* term due to the time-varying comparator $\theta_t^\star$. The middle term is controlled by Lemma B.3, which yields a discounted log-determinant term and a movement term; with $\alpha = 3\eta/2$, the latter is canceled by the negative last term in Equation equation 2. Therefore, the main remaining difficulties are to control the concentration term and the loss-drift term.

3) Fixed-time exponential bound for the concentration term. For the concentration term, the stationary OMD analysis of Zhang et al. (2025) based on a supermartingale and Ville's inequality (Ville, 1939) no longer applies. Indeed, letting $Z_s := \ell_s(\theta_s^\star) - m_s(\Pi_s)$ and $S_t := \sum_{s=1}^{t-1}\gamma^{t-1-s}Z_s$, the discounted recursion becomes $S_{t+1} = \gamma S_t + Z_t$, so the process e^{S_t} is generally not a supermartingale. We therefore first derive a *fixed-time* exponential bound, and then obtain a uniform-in-t statement by applying a union bound. Specifically, fixing the time t and writing $w_{s,t} := \gamma^{t-1-s} \in (0, 1]$, the likelihood-ratio identity gives $\mathbb{E}[e^{Z_s} \mid \widetilde{\mathcal{F}}_s] = 1$, where $\widetilde{\mathcal{F}}_s$ denotes a suitable filtration; its precise definition is given in the appendix. Since $x \mapsto x^{w_{s,t}}$ is concave on $\mathbb{R}_+$, conditional Jensen yields $\mathbb{E}[e^{w_{s,t}Z_s} \mid \widetilde{\mathcal{F}}_s] \leqslant \big(\mathbb{E}[e^{Z_s} \mid \widetilde{\mathcal{F}}_s]\big)^{w_{s,t}} = 1$. Iterating over s gives $\mathbb{E}[e^{\sum_{s=1}^{t-1} w_{s,t}Z_s}] \leqslant 1$, and therefore Markov's inequality implies that the concentration term is at most $\log(1/\delta)$ with probability at least $1-\delta$. Thus, unlike the stationary case, concentration is obtained through a fixed-time exponential argument tailored to discounting. Applying this fixed-time bound with $\delta_t = 6\delta/(\pi^2 t^2)$, one can then take a union bound over t to obtain the desired guarantee uniformly over t (see Lemma B.2).

4) Control of the loss-drift term. The main difficulty is that the drift of the comparator is measured through past losses ℓ_s, rather than directly through parameter differences. This gives rise to an additional loss-space error term that does not appear in the stationary OMD analysis (Zhang et al., 2025). It is also structurally different from the explicit parameter-bias terms used in nonstationary MLE-based analyses (Wang et al., 2023). We control this term directly in loss space by writing

$$\ell_s(\theta_t^\star) - \ell_s(\theta_s^\star) = \underbrace{\frac{m(x_s^\top\theta_t^\star) - m(x_s^\top\theta_s^\star) - \mu(x_s^\top\theta_s^\star)x_s^\top(\theta_t^\star - \theta_s^\star)}{g(\tau)}}_{\text{deterministic term}} - \underbrace{\frac{\eta_s\, x_s^\top(\theta_t^\star - \theta_s^\star)}{g(\tau)}}_{\text{stochastic term}}.$$

Algorithm 1 `DOMD-GLB`: **D**iscounted **O**nline **M**irror **D**escent for **G**eneralized **L**inear **B**andit

Require: Discount factor $\gamma \in (0,1)$, regularization $\lambda > 0$, confidence level δ.
1: Initialize $\theta_1 \in \Theta$ and $H_1 \leftarrow \lambda I_d$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Compute the pre-update matrix $A_t \leftarrow \gamma H_t + (1-\gamma)\lambda I_d$.
4: Select $x_t \in \mathcal{X}_t$ according to equation 3 and observe the reward r_t.
5: Update the online estimator θ_{t+1} according to equation 1.
6: Compute the post-update matrix $H_{t+1} \leftarrow A_t + \nabla^2 \ell_t(\theta_{t+1})$.
7: **end for**

The first term on the right-hand side is deterministic. By the smoothness of the GLM loss under Assumption 2, it is bounded by $\frac{k_\mu}{2g(\tau)} \sum_{s=1}^{t-1} \gamma^{t-1-s} \big(x_s^\top(\theta_t^\star - \theta_s^\star)\big)^2 \leqslant \frac{k_\mu}{2g(\tau)} \mathcal{V}_t^2$. Furthermore, the second term is stochastic. Since the comparator path is *oblivious*, i.e., the entire sequence $\{\theta_u^\star\}_{u \geqslant 1}$ is independent of the realized rewards, the coefficients $\gamma^{t-1-s} x_s^\top(\theta_t^\star - \theta_s^\star)$ are predictable. Therefore, a discounted Hoeffding–Chernoff argument yields a deviation term of order $\sqrt{\mathcal{V}_t^2 \log(1/\delta)}$. By Young's inequality, this term is split into a quadratic variation term, which is absorbed into the deterministic variation bound, and a logarithmic confidence penalty. Combining the deterministic and stochastic parts, we obtain $\sum_{s=1}^{t-1} \gamma^{t-1-s} \big(\ell_s(\theta_t^\star) - \ell_s(\theta_s^\star)\big) \leqslant \frac{k_\mu}{g(\tau)} \mathcal{V}_t^2 + \frac{R^2}{4g(\tau)k_\mu} \log \frac{1}{\delta}$. Therefore, the loss-drift term is controlled by the discounted variation measure $\mathcal{V}_t$ (Lemma B.4).

4.2 Action Selection

We select actions using an implementable optimistic score that does not require prior knowledge of the variation budget $\mathcal{V}_t$. An exact optimism-in-the-face-of-uncertainty (OFU) rule (Abbasi-Yadkori et al., 2011) over the full confidence set $\mathcal{C}_t(\delta)$ would include the drift-dependent part of the confidence radius, and hence would not be directly implementable without knowing $\mathcal{V}_t$. We therefore use the following computable UCB-type rule:

$$x_t \in \arg\max_{x \in \mathcal{X}_t} \left\{ x^\top \theta_t + \beta_t(\delta) \|x\|_{H_t^{-1}} \right\}. \tag{3}$$

The full procedure is summarized in Algorithm 1. Note that the same algorithm handles both drifting and piecewise-stationary environments without prior knowledge of the nonstationarity type.

4.3 Dynamic Regret Bound under Drifting Setting

Theorem 2. *Let $T \geqslant 2$, $\lambda = \max\left\{\frac{6\eta R k_\mu S}{g(\tau)}, \frac{32\alpha d R^2}{7}, \frac{c_\mu}{g(\tau)}\right\}$, and $\alpha = \frac{3\eta}{2}$. Suppose Assumptions 1 and 2 hold. Then, for every $\gamma \in [1/T, 1)$, with probability at least $1-\delta$, `DOMD-GLB` achieves*

$$\mathbf{Reg}_T = \tilde{\mathcal{O}}\left(\frac{k_\mu}{\sqrt{c_\mu}}\, d\sqrt{T} + \frac{k_\mu}{\sqrt{c_\mu}}\, dT\sqrt{1-\gamma} + \frac{k_\mu^{3/2}}{\sqrt{c_\mu}}\, \frac{P_T}{(1-\gamma)^{3/2}} \right).$$

Moreover, under the tuned choice $\gamma = 1 - \min\left\{1 - \frac{1}{T},\ \max\left\{\frac{1}{T},\ \sqrt{\frac{\sqrt{k_\mu}\, P_T}{dT}}\right\}\right\}$, we obtain

$$\mathbf{Reg}_T = \begin{cases} \tilde{\mathcal{O}}\left(\frac{k_\mu}{\sqrt{c_\mu}}\, d\sqrt{T}\right), & \text{if } 0 \leqslant P_T < \dfrac{d}{\sqrt{k_\mu}\, T}, \\ \tilde{\mathcal{O}}\left(\frac{k_\mu^{9/8}}{\sqrt{c_\mu}}\, d^{3/4} P_T^{1/4} T^{3/4}\right), & \text{if } \dfrac{d}{\sqrt{k_\mu}\, T} \leqslant P_T \leqslant \dfrac{dT}{\sqrt{k_\mu}} \left(1 - \dfrac{1}{T}\right)^2, \\ \tilde{\mathcal{O}}\left(\frac{k_\mu^{3/2}}{\sqrt{c_\mu}}\, P_T\right), & \text{if } P_T > \dfrac{dT}{\sqrt{k_\mu}} \left(1 - \dfrac{1}{T}\right)^2. \end{cases}$$

Discussion of Theorem 2. In the intermediate-P_T regime, Theorem 2 yields the regret bound $\tilde{\mathcal{O}}\big(c_\mu^{-1/2} d^{3/4} P_T^{1/4} T^{3/4}\big)$, matching the best-known regret order of the discounted MLE-based method `SCB-WeightUCB` (Wang et al., 2023), while replacing the discounted MLE update with DOMD.

This leads to a substantial computational advantage: our per-round computation and memory costs are constant, both independent of t, whereas weighted MLE-based methods (Faury et al., 2021; Wang et al., 2023) must revisit past weighted observations, so their costs grow with t. Moreover, those methods rely on a non-convex projection step, which is generally not guaranteed to be computable in polynomial time. We also note that `MASTER` (Wei & Luo, 2021) can achieve the sharper regret bound $\tilde{\mathcal{O}}\big(c_\mu^{-1} d P_T^{1/3} T^{2/3}\big)$, but for a different problem setting: it assumes a fixed action set, whereas we allow the feasible action set $\mathcal{X}_t$ to vary over time. Hence, the two guarantees are not directly comparable, and the existing `MASTER` result does not apply to our setting. In addition, the `MASTER` meta-algorithm itself incurs $\mathcal{O}(\log t)$ computation and memory overhead per round, and thus its computational and storage costs still depend on t, unlike our constant-cost update.

Among methods applicable, or directly extendable, to time-varying action sets, our result achieves the best currently supported regret order while providing substantially improved computational and memory efficiency. The seemingly sharper `RestartGLB` bound of Zhao et al. (2020) is based on an analysis later shown to contain a technical flaw; the corrected linear-bandit analysis (Zhao & Zhang, 2021) instead yields a $T^{3/4}$-type, rather than $T^{2/3}$-type, regret bound, while no corrected GLB guarantee has been established. The proof is deferred to Appendix C.

4.4 Dynamic Regret Bound under Piecewise-Stationary Setting

Theorem 3. *In the piecewise-stationary setting, under the same parameter choices and assumptions as in Theorem 2, for every $\gamma \in [1/T, 1)$, with probability at least $1-\delta$, `DOMD-GLB` satisfies*

$$\mathbf{Reg}_T = \tilde{\mathcal{O}}\left(\frac{k_\mu}{\sqrt{c_\mu}}\, d\sqrt{T} + \frac{k_\mu}{\sqrt{c_\mu}}\, dT\sqrt{1-\gamma} + \frac{\Gamma_T}{1-\gamma} \right).$$

Moreover, for the tuned choice $\gamma = 1 - \min\left\{1 - \frac{1}{T},\ \max\left\{\frac{1}{T},\ \left(\frac{\Gamma_T \sqrt{c_\mu}}{k_\mu d T}\right)^{2/3}\right\}\right\}$, we obtain

$$\mathbf{Reg}_T = \begin{cases} \tilde{\mathcal{O}}\left(\frac{k_\mu}{\sqrt{c_\mu}}\, d\sqrt{T}\right), & \text{if } 0 \leqslant \Gamma_T < \dfrac{k_\mu d}{\sqrt{c_\mu T}}, \\ \tilde{\mathcal{O}}\left(\frac{k_\mu^{2/3}}{c_\mu^{1/3}}\, d^{2/3} \Gamma_T^{1/3} T^{2/3}\right), & \text{if } \dfrac{k_\mu d}{\sqrt{c_\mu T}} \leqslant \Gamma_T \leqslant \dfrac{k_\mu dT}{\sqrt{c_\mu}} \left(1 - \dfrac{1}{T}\right)^{3/2}, \\ \tilde{\mathcal{O}}(\Gamma_T), & \text{if } \Gamma_T > \dfrac{k_\mu dT}{\sqrt{c_\mu}} \left(1 - \dfrac{1}{T}\right)^{3/2}. \end{cases}$$

Discussion of Theorem 3. In the intermediate-Γ_T regime, Theorem 3 gives the regret bound $\tilde{\mathcal{O}}\big(c_\mu^{-1/3} d^{2/3} \Gamma_T^{1/3} T^{2/3}\big)$. Compared with the best-known discounted MLE-based regret bound $\tilde{\mathcal{O}}\big(d^{2/3} \Gamma_T^{1/3} T^{2/3}\big)$, achieved by `SCB-PW-WeightUCB` (Wang et al., 2023), our result incurs an additional dependence on c_μ. However, that method relies on a non-convex projection step and thus does not generally admit a polynomial-time guarantee. Moreover, since it is MLE-based, its memory cost grows linearly with t. From this perspective, the additional $1/c_\mu$ dependence in our bound can be viewed as the cost of achieving substantially better computation and memory efficiency. We believe that further improving the dependence on $1/c_\mu$ while retaining an online estimation approach is nontrivial, and we leave this direction for future work. The proof is deferred to Appendix D.

5 Numerical Experiments

In this section, we empirically evaluate the proposed algorithm. At each round t, the learner is presented with N arms, whose feature vectors are drawn i.i.d. from a standard Gaussian distribution and normalized to unit ℓ_2-norm. After selecting an arm x_t, the learner observes a reward $r_t \sim \mathrm{Bernoulli}(\sigma(x_t^\top \theta_t))$, where $\sigma(z) = (1 + e^{-z})^{-1}$ and θ_t is the time-varying parameter. We consider two types of non-stationarity. In the *drifting* setting, θ_t evolves continuously along a circle of radius S and completes one full rotation, so the total angular change is 2π and the resulting variation budget satisfies $P_T \leqslant 2S\pi$. In the *piecewise-stationary* setting, θ_t remains constant within each segment and changes once at $t = T/2$ through a sign flip $\theta \leftarrow -\theta$, so $\Gamma_T = 1$. Throughout, we set $T = 5000$, $d = 5$, $N = 30$, and $S \in \{1.0, 3.0\}$. We report the mean cumulative dynamic regret and the mean per-round computation time over 20 random seeds, along with one standard deviation for each.

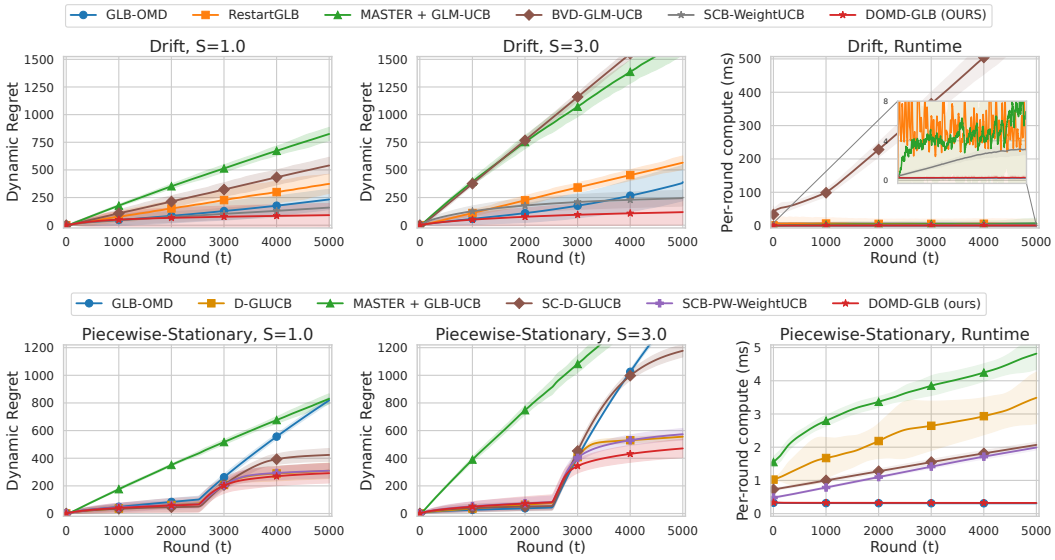


Figure 1: Cumulative dynamic regret and runtime for varying S in two nonstationary settings.

Baselines. For the drifting experiments, we compare `DOMD-GLB` with the following baselines: (a) `GLB-OMD` (Zhang et al., 2025), a stationary OMD-based algorithm; (b) `RestartGLB` (Zhao et al., 2020), a restart-based algorithm; (c) `MASTER + GLM-UCB` (Wei & Luo, 2021), an adaptive restart-based algorithm; (d) `BVD-GLM-UCB` (Faury et al., 2021); and (e) `SCB-WeightUCB` (Wang et al., 2023). For the piecewise-stationary experiments, we compare `DOMD-GLB` with: (a) `GLB-OMD`; (b) `D-GLUCB` (Russac et al., 2020); (c) `MASTER + GLM-UCB`; (d) `SC-D-GLUCB` (Russac et al., 2021); and (e) `SCB-PW-WeightUCB` (Wang et al., 2023). We use the theoretically prescribed values for λ, γ and the restart period H. For all methods, we scale the theoretical confidence radii by $1/5$, since the original bonuses are too conservative in practice. For `MASTER`, we set the block-scale parameter to $n = 12$. For algorithms that require non-convex projection steps, and therefore do not admit polynomial-time computational guarantees, such as `BVD-GLM-UCB`, `SCB-WeightUCB`, `D-GLUCB`, and `SCB-PW-WeightUCB`, we use projected gradient descent as an approximation.

Results. Figure 1 shows that `DOMD-GLB` consistently outperforms all other baselines across all settings, while incurring only a constant per-round computational cost. `MASTER` performs particularly poorly, with its cumulative regret remaining nearly linear over the considered horizon. Importantly, its GLB guarantee is established only for a fixed action set, whereas our experiments use time-varying feasible action sets. Thus, `MASTER` is operated outside the setting covered by its existing theoretical guarantee, which may partly explain its poor empirical performance. This behavior is also consistent with the practical limitations of `MASTER` reported by Gerogiannis et al. (2025). Moreover, in the piecewise-stationary setting, although `SCB-PW-WeightUCB` achieves a tighter regret bound than ours by a factor of $c_\mu^{-1/3}$, its empirical performance is worse than ours. This is because its UCB bonus is more conservative, scaling with $c_\mu^{-1/2}$ (Equation (17) in Wang et al. 2023), whereas our UCB bonus has no such harmful dependence (Equation equation 3).

6 Conclusion

In this paper, to the best of our knowledge, we introduce the first algorithm for nonstationary GLBs built on a novel discounted online mirror descent (DOMD) method. Using this DOMD method, `DOMD-GLB` achieves $\tilde{\mathcal{O}}(c_\mu^{-1/2} d^{3/4} P_T^{1/4} T^{3/4})$ and $\tilde{\mathcal{O}}(c_\mu^{-1/3} d^{2/3} \Gamma_T^{1/3} T^{2/3})$ dynamic regret bounds in the drifting and piecewise-stationary settings, respectively, while requiring only constant per-round computation and memory costs. We believe this opens a new and promising research direction toward nonstationary parametric bandit algorithms that are both computationally and statistically efficient.

AI USE STATEMENT

In this work, we used generative AI tools for auxiliary tasks including rapid checking of mathematical derivations and proofs and assistance with writing experimental code. We did not use generative AI tools to develop the core research ideas or the main technical arguments and proofs, which were developed by the authors, and the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to assist with writing, editing, and improving the presentation of the paper. We have reviewed all AI-assisted work. In particular, AI-assisted proof checks were independently verified against our own derivations, and AI-assisted experimental code was reviewed and tested for correctness before use. All AI-assisted text was also revised by the authors for technical accuracy and consistency with the intended claims. We take responsibility for the final content of this work, including all text, claims, proofs, code, and experimental results produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure that both the theoretical and empirical results in this work are reproducible. All assumptions, algorithmic details, and parameter choices required for our theoretical results are stated explicitly in the main paper, and complete proofs of the main theorems and supporting lemmas are provided in the appendix. The proposed algorithm is described in pseudocode together with precise update rules and confidence parameters. For the empirical results, we provide the experimental setup, baseline implementations, hyperparameter choices, and evaluation protocol needed to reproduce the reported comparisons. Source code for reproducing all experiments is included in the supplementary material.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1079–1087. PMLR, 2019.

Wooseong Cho, Taehyun Hwang, Joongkyu Lee, and Min-hwan Oh. Randomized exploration for reinforcement learning with multinomial logistic function approximation. *Advances in Neural Information Processing Systems*, 37:76643–76720, 2024.

Hyunjun Choi, Joongkyu Lee, and Min-hwan Oh. True impact of cascade length in contextual cascading bandits. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yuntian Deng, Xingyu Zhou, Baekjin Kim, Ambuj Tewari, Abhishek Gupta, and Ness Shroff. Weighted gaussian process bandits for non-stationary environments. In *International Conference on Artificial Intelligence and Statistics*, pp. 6909–6932. PMLR, 2022.

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

Louis Faury, Marc Abeille, Kwang-Sung Jun, and Clément Calauzènes. Jointly efficient and optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 546–580. PMLR, 2022.

Sarah Filippi, Olivier Cappé, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, pp. 586–594, Red Hook, NY, USA, 2010. Curran Associates Inc.

Argyrios Gerogiannis, Yu-Han Huang, and Venugopal Veeravalli. Is prior-free black-box non-stationary reinforcement learning feasible? In *International Conference on Artificial Intelligence and Statistics*, pp. 2692–2700. PMLR, 2025.

Kihyuk Hong, Yuhang Li, and Ambuj Tewari. An optimization-based algorithm for non-stationary kernel bandits without prior knowledge. In *International Conference on Artificial Intelligence and Statistics*, pp. 3048–3085. PMLR, 2023.

Wasim Huleihel, Soumyabrata Pal, and Ofer Shayevitz. Learning user preferences in non-stationary environments. In *International Conference on Artificial Intelligence and Statistics*, pp. 1432–1440. PMLR, 2021.

Su Jia, Qian Xie, Nathan Kallus, and Peter I Frazier. Smooth non-stationary bandits. In *International Conference on Machine Learning*, pp. 14930–14944. PMLR, 2023.

Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Joongkyu Lee and Min-hwan Oh. Nearly minimax optimal regret for multinomial logistic bandit. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Joongkyu Lee and Min-hwan Oh. Combinatorial reinforcement learning with preference feedback. In *Forty-second International Conference on Machine Learning*, 2025a.

Joongkyu Lee and Min-hwan Oh. Improved online confidence bounds for multinomial logistic bandits. In *Forty-second International Conference on Machine Learning*, 2025b.

Joongkyu Lee, Seouh-won Yi, and Min-hwan Oh. Preference-based reinforcement learning beyond pairwise comparisons: Benefits of multiple options. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. A unified confidence sequence for generalized linear models, with applications to bandits. *Advances in Neural Information Processing Systems*, 37:124640–124685, 2024.

Long-Fei Li, Yu-Yang Qian, Peng Zhao, and Zhi-Hua Zhou. Provably efficient online rlhf with one-pass reward modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Shaoang Li and Jian Li. Constrained feedback learning for non-stationary multi-armed bandits. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Shuai Liu, Alex Ayoub, Flore Sentenac, Xiaoqi Tan, and Csaba Szepesvári. Almost free: Self-concordance in natural exponential families and an application to bandits. *Advances in Neural Information Processing Systems*, 37:70828–70874, 2024.

Yueyang Liu, Benjamin Van Roy, and Kuang Xu. Nonstationary bandit learning via predictive sampling. In *International Conference on Artificial Intelligence and Statistics*, pp. 6215–6244. PMLR, 2023.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Yoan Russac, Louis Faury, Olivier Cappé, and Aurélien Garivier. Self-concordant analysis of generalized linear bandits with forgetting. In *International Conference on Artificial Intelligence and Statistics*, pp. 658–666. PMLR, 2021.

Ayush Sawarni, Nirjhar Das, Siddharth Barman, and Gaurav Sinha. Generalized linear bandits with limited adaptivity. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=FTPDBQuT4G.

Sabina Tomkins, Peng Liao, S Yeung, P Klasnja, and S Murphy. Intelligentpooling: Practical thompson sampling for mhealth; 2020. *URL https://arxiv. org/abs*, 2008.

Jean Ville. *Etude critique de la notion de collectif*, volume 3. Gauthier-Villars Paris, 1939.

Volodya Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

Jing Wang, Peng Zhao, and Zhi-Hua Zhou. Revisiting weighted strategy for non-stationary parametric bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 7913–7942. PMLR, 2023.

Jing Wang, Yu-Jie Zhang, Peng Zhao, and Zhi-Hua Zhou. Heavy-tailed linear bandits: Huber regression with one-pass update. In *International Conference on Machine Learning*, pp. 65595–65617. PMLR, 2025.

Zhiyong Wang, Jize Xie, Yi Chen, John Lui, and Dongruo Zhou. Variance-dependent regret bounds for non-stationary linear bandits. *arXiv preprint arXiv:2403.10732*, 2024.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Conference on learning theory*, pp. 4300–4354. PMLR, 2021.

Yu-Jie Zhang and Masashi Sugiyama. Online (multinomial) logistic bandit: Improved regret and constant computation cost. *Advances in Neural Information Processing Systems*, 36, 2024.

Yu-Jie Zhang, Sheng-An Xu, Peng Zhao, and Masashi Sugiyama. Generalized linear bandits: Almost optimal regret with one-pass update. *arXiv preprint arXiv:2507.11847*, 2025.

Peng Zhao and Lijun Zhang. Non-stationary linear bandits revisited. *arXiv preprint arXiv:2103.05324*, 19, 2021.

Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. A simple approach for non-stationary linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 746–755. PMLR, 2020.

Zifeng Zhao, Feiyu Jiang, Yi Yu, and Xi Chen. High-dimensional dynamic pricing under non-stationarity: Learning and earning with change-point detection. *arXiv preprint arXiv:2303.07570*, 2023.

Appendix

Table of Contents

A Further Related Work **13**

B Proof of Theorem 1 **13**

B.1 Main Proof of Theorem 1 13

B.2 Proofs of Lemmas for Theorem 1 15

B.3 Auxiliary Lemmas for Theorem 1 22

C Proof of Theorem 2 **28**

C.1 Main Proof of Theorem 2 28

C.2 Proofs of Lemmas for Theorem 2 31

D Proof of Theorem 3 **33**

D.1 Main Proof of Theorem 3 33

D.2 Proofs of Lemmas for Theorem 3 37

A Further Related Work

In this section, we discuss additional related work that complements Section 2.

OMD in bandits and reinforcement learning. Online mirror descent (OMD) has recently emerged as a useful tool for designing jointly efficient bandit algorithms, replacing repeated batch MLE by one-pass updates whose per-round cost remains independent of time. In logistic bandits, Faury et al. (2022); Zhang & Sugiyama (2024) developed computationally efficient UCB-type algorithms based on OMD. This approach has since been extended to broader settings, including multinomial logit (MNL) bandits (Lee & Oh, 2024; 2025b) and GLBs (Zhang et al., 2025). Related OMD-style ideas have also been explored in other online decision-making problems, including heavy-tailed linear bandits (Wang et al., 2025), cascading bandits (Choi et al., 2025), reinforcement learning (Cho et al., 2024; Lee & Oh, 2025a), and RLHF/PbRL (Li et al., 2025; Lee et al., 2025). However, to the best of our knowledge, the use of OMD in nonstationary settings remains largely unexplored.

B Proof of Theorem 1

In this section, we present the proof of Theorem 1. We first give the main proof (Subsection B.1), then provide the proofs of the key lemmas used therein (Subsection B.2), and finally present several auxiliary technical lemmas (Subsection B.3).

B.1 Main Proof of Theorem 1

Proof of Theorem 1. Compared with the stationary analysis of Zhang et al. (2025), our nonstationary setting requires three essential changes: (i) the potential and Gaussian perturbation must be defined using the discounted curvature A_t rather than the stationary matrix $H_t^{\mathrm{stat}} := \lambda I_d + \sum_{s=1}^{t-1} \nabla^2 \ell_s(\theta_{s+1})$; (ii) the discounted exponential process is no longer additive, so Ville's inequality and the resulting anytime concentration argument no longer apply directly (see Remark B.1); and (iii) the drifting comparator sequence $\{\theta_t^\star\}_{t\geqslant 1}$ creates an additional drift term that must be controlled separately.

We begin the proof by stating the one-step discounted inequality lemma.

Lemma B.1 (One-step discounted OMD inequality)**.** *Let $\eta = 1 + RS$ and $\lambda \geqslant \frac{6\eta R k_\mu S}{g(\tau)}$. Then, under Assumptions 1 and 2, for every $\theta \in \Theta$ and every $t \geqslant 1$, we have*

$$\|\theta_{t+1} - \theta\|_{H_{t+1}}^2 \leqslant \|\theta_t - \theta\|_{A_t}^2 + 2\eta\big(\ell_t(\theta) - \ell_t(\theta_{t+1})\big) - \frac{2}{3}\|\theta_{t+1} - \theta_t\|_{A_t}^2.$$

Consequently, for every $\theta \in \Theta$ and every $t \geqslant 2$,

$$\|\theta_t - \theta\|_{H_t}^2 \leqslant 4\lambda S^2 + 2\eta \sum_{s=1}^{t-1} \gamma^{t-1-s}\big(\ell_s(\theta) - \ell_s(\theta_{s+1})\big) - \frac{2}{3}\sum_{s=1}^{t-1} \gamma^{t-1-s}\|\theta_{s+1} - \theta_s\|_{A_s}^2.$$

The proof is provided in Appendix B.2.1.

By applying Lemma B.1 with $\theta = \theta_t^\star$, we get

$$\|\theta_t - \theta_t^\star\|_{H_t}^2 \leqslant 4\lambda S^2 + 2\eta \sum_{s=1}^{t-1} \gamma^{t-1-s}\big(\ell_s(\theta_t^\star) - \ell_s(\theta_{s+1})\big) - \frac{2}{3}\sum_{s=1}^{t-1} \gamma^{t-1-s}\|\theta_{s+1} - \theta_s\|_{A_s}^2. \quad \text{(B.1)}$$

For each round t, let $\Pi_t := \mathcal{N}(\theta_t, \alpha A_t^{-1})$, where $\alpha = 3\eta/2$, denote the Gaussian distribution centered at the current iterate θ_t, with covariance matrix αA_t^{-1}. Associated with this Gaussian perturbation, we define the mix loss as follows:

$$m_t(\Pi_t) := -\log \mathbb{E}_{\theta\sim\Pi_t}\Big[e^{-\ell_t(\theta)}\Big].$$

Now, we decompose $\sum_{s=1}^{t-1} \gamma^{t-1-s}\big(\ell_s(\theta_t^\star) - \ell_s(\theta_{s+1})\big)$ in equation B.1 as follows:

$$\begin{aligned}
&\sum_{s=1}^{t-1}\gamma^{t-1-s}\big(\ell_s(\theta_t^\star) - \ell_s(\theta_{s+1})\big) \\
&= \sum_{s=1}^{t-1}\gamma^{t-1-s}\big(\ell_s(\theta_s^\star) - m_s(\Pi_s)\big) + \sum_{s=1}^{t-1}\gamma^{t-1-s}\big(m_s(\Pi_s) - \ell_s(\theta_{s+1})\big) + \sum_{s=1}^{t-1}\gamma^{t-1-s}\big(\ell_s(\theta_t^\star) - \ell_s(\theta_s^\star)\big).
\end{aligned} \quad \text{(B.2)}$$

We next present three lemmas that bound the above three terms.

Lemma B.2 (Fixed-time discounted mix-loss concentration)**.** *Define $\Pi_t := \mathcal{N}(\theta_t, \alpha A_t^{-1})$ and $m_t(\Pi_t) := -\log \mathbb{E}_{\theta\sim\Pi_t}\big[e^{-\ell_t(\theta)}\big]$. Then, for any fixed $t \geqslant 2$, with probability at least $1 - \delta$, we have*

$$\sum_{s=1}^{t-1} \gamma^{t-1-s}\big(\ell_s(\theta_s^\star) - m_s(\Pi_s)\big) \leqslant \log\frac{1}{\delta}.$$

The proof is provided in Appendix B.2.2.

Remark B.1 (Why the discounted case only yields a fixed-time bound)**.** *Our exponential argument differs from Lemma 5 of Zhang et al. (2025). In the undiscounted setting of Zhang et al. (2025), the cumulative process is additive in time, which allows one to construct a nonnegative supermartingale and then apply Ville's inequality (Ville, 1939) to obtain an anytime bound. By contrast, in our discounted setting, for a fixed terminal time t we consider $\sum_{s=1}^{t-1} \gamma^{t-1-s} Z_s$, where $Z_s := \ell_s(\theta_s^\star) - m_s(\Pi_s)$. For each fixed t, one can show that $\mathbb{E}[e^{\sum_{s=1}^{t-1} \gamma^{t-1-s} Z_s}] \leqslant 1$, and hence a fixed-time high-probability bound follows by Markov's inequality. However, the process $(e^{\sum_{s=1}^{t-1} \gamma^{t-1-s} Z_s})_{t\geqslant 2}$ is generally not a supermartingale, so Ville's inequality cannot be applied. To see this, define $S_t := \sum_{s=1}^{t-1} \gamma^{t-1-s} Z_s$. Then, the obstruction is the discounted recursion $S_{t+1} = \gamma S_t + Z_t$, which gives $e^{S_{t+1}} = e^{\gamma S_t} e^{Z_t}$. Since $\mathbb{E}[e^{Z_t}|\widetilde{\mathcal{F}}_t] = 1$, we get $\mathbb{E}[e^{S_{t+1}}|\widetilde{\mathcal{F}}_t] = e^{\gamma S_t}$. A supermartingale argument would require $\mathbb{E}[e^{S_{t+1}}|\widetilde{\mathcal{F}}_t] \leqslant e^{S_t}$, but this is not guaranteed because $\gamma < 1$ and S_t need not be nonnegative. Therefore, unlike Lemma 5 of Zhang et al. (2025), our discounted analysis yields concentration only for each fixed t, rather than an anytime bound uniform over all t.*

Lemma B.3 (Discounted mix-loss gap bound)**.** *Define* $\Pi_t := \mathcal{N}(\theta_t, \alpha A_t^{-1})$ *and* $m_t(\Pi_t) := -\log \mathbb{E}_{\theta\sim\Pi_t}\big[e^{-\ell_t(\theta)}\big]$. *Let* $\lambda \geqslant 32\alpha d R^2/7$. *Then, for every* $t \geqslant 2$*, under Assumption 1, we have*

$$\sum_{s=1}^{t-1} \gamma^{t-1-s}\big(m_s(\Pi_s) - \ell_s(\theta_{s+1})\big) \leqslant \left(3\eta + \frac{1}{2}\right) d \log\left(1 + \frac{k_\mu(1-\gamma^{t-1})}{\lambda d\, g(\tau)(1-\gamma)}\right) + \frac{1}{3\eta}\sum_{s=1}^{t-1} \gamma^{t-1-s}\|\theta_{s+1} - \theta_s\|_{A_s}^2.$$

The proof is provided in Appendix B.2.3.

Lemma B.4 (Drift-loss bound)**.** *Under Assumptions 1 and 2, for* $t \geqslant 2$*, with probability at least* $1-\delta$*, we have*

$$\sum_{s=1}^{t-1} \gamma^{t-1-s}\big(\ell_s(\theta_t^\star) - \ell_s(\theta_s^\star)\big) \leqslant \frac{k_\mu}{g(\tau)}\,\mathcal{V}_t^2 + \frac{R^2}{4\,g(\tau)\,k_\mu}\log\frac{1}{\delta},$$

where $\mathcal{V}_t = \sum_{p=1}^{t-1}\gamma^{(t-1-p)/2}\sqrt{\frac{1-\gamma^p}{1-\gamma}}\|\theta_{p+1}^\star - \theta_p^\star\|_2$.

The proof is provided in Appendix B.2.4.

Consequently, on the right-hand side of equation B.2, the first term is bounded by Lemma B.2 with probability $1-\delta/2$; the second term is bounded deterministically by Lemma B.3; the third term is bounded by Lemma B.4 with probability $1-\delta/2$. Substituting these three bounds into equation B.1 and using $2\eta\cdot\frac{1}{2\alpha} = 2\eta\cdot\frac{1}{3\eta} = \frac{2}{3}$, we obtain, on an event of probability at least $1-\delta$,

$$\|\theta_t - \theta_t^\star\|_{H_t}^2 \leqslant 4\lambda S^2 + 2\eta\log\frac{2}{\delta} + 2\eta\left(3\eta + \frac{1}{2}\right) d\log\left(1 + \frac{k_\mu(1-\gamma^{t-1})}{\lambda d\, g(\tau)(1-\gamma)}\right) + \frac{2\eta R^2}{g(\tau)k_\mu}\log\frac{2}{\delta} + \frac{2\eta k_\mu}{g(\tau)}\mathcal{V}_t^2, \qquad \forall t \geqslant 2.$$

To obtain an anytime statement, for each $t \geqslant 2$, set $\delta_t := \frac{6\delta}{\pi^2 t^2}$. Applying the above fixed-time bound with δ replaced by δ_t, we obtain that, for each fixed $t \geqslant 2$, the desired inequality holds with probability at least $1-\delta_t$. Moreover,

$$\sum_{t=2}^{\infty}\delta_t = \frac{6\delta}{\pi^2}\sum_{t=2}^{\infty}\frac{1}{t^2} \leqslant \frac{6\delta}{\pi^2}\sum_{t=1}^{\infty}\frac{1}{t^2} = \delta.$$

Therefore, by a union bound over all $t \geqslant 2$, with probability at least $1-\delta$, the stated inequality holds simultaneously for all $t \geqslant 2$. This completes the proof of Theorem 1. □

B.2 Proofs of Lemmas for Theorem 1

B.2.1 Proof of Lemma B.1

Proof of Lemma B.1. Fix $t \geqslant 1$ and $\theta \in \Theta$. Define

$$f_t(z) := \frac{m(z) - r_t z}{g(\tau)}, \qquad z \in \mathbb{R}.$$

Then, for every $\theta \in \Theta$,

$$\ell_t(\theta) = f_t(x_t^\top\theta), \qquad f_t'(z) = \frac{\mu(z) - r_t}{g(\tau)}, \qquad f_t''(z) = \frac{\mu'(z)}{g(\tau)}, \qquad f_t'''(z) = \frac{\mu''(z)}{g(\tau)}.$$

Hence, by Assumption 2, we have

$$\frac{c_\mu}{g(\tau)} \leqslant f_t''(z) \leqslant \frac{k_\mu}{g(\tau)}, \qquad |f_t'''(z)| \leqslant \frac{R k_\mu}{g(\tau)}.$$

We first lower bound $\ell_t(\theta) - \ell_t(\theta_{t+1})$. By the integral form of the Bregman divergence,

$$\begin{aligned}\ell_t(\theta) - \ell_t(\theta_{t+1}) &= f_t(x_t^\top \theta) - f_t(x_t^\top \theta_{t+1}) \\ &= f_t'(x_t^\top \theta_{t+1})\, x_t^\top(\theta - \theta_{t+1}) \\ &\quad + \big(x_t^\top(\theta - \theta_{t+1})\big)^2 \int_0^1 (1-s)\, f_t''\big(x_t^\top \theta_{t+1} + s\, x_t^\top(\theta - \theta_{t+1})\big)\, \mathrm{d}s.\end{aligned}$$

Using $f_t'' = \mu'/g(\tau)$ and applying Lemma B.5 with

$$v = x_t^\top \theta_{t+1}, \qquad z_1 = x_t^\top \theta, \qquad z_2 = x_t^\top \theta_{t+1},$$

we obtain

$$\int_0^1 (1-s)\, f_t''\big(x_t^\top \theta_{t+1} + s\, x_t^\top(\theta - \theta_{t+1})\big)\, \mathrm{d}s \geqslant \frac{\mu'(x_t^\top \theta_{t+1})}{g(\tau)\big(2 + R|x_t^\top(\theta - \theta_{t+1})|\big)}.$$

Therefore,

$$\ell_t(\theta) - \ell_t(\theta_{t+1}) \geqslant f_t'(x_t^\top \theta_{t+1})\, x_t^\top(\theta - \theta_{t+1}) + \frac{\mu'(x_t^\top \theta_{t+1})}{g(\tau)\big(2 + R|x_t^\top(\theta - \theta_{t+1})|\big)}\big(x_t^\top(\theta - \theta_{t+1})\big)^2.$$

Since $\theta, \theta_{t+1} \in \Theta$ and $\|x_t\|_2 \leqslant 1$,

$$|x_t^\top(\theta - \theta_{t+1})| \leqslant \|\theta - \theta_{t+1}\|_2 \leqslant 2S.$$

Hence, by the definition of $\eta = 1 + RS$, we get

$$\begin{aligned}\ell_t(\theta) - \ell_t(\theta_{t+1}) &\geqslant f_t'(x_t^\top \theta_{t+1})\, x_t^\top(\theta - \theta_{t+1}) + \frac{1}{2\eta}\frac{\mu'(x_t^\top \theta_{t+1})}{g(\tau)}\big(x_t^\top(\theta - \theta_{t+1})\big)^2 \\ &= f_t'(x_t^\top \theta_{t+1})\, x_t^\top(\theta - \theta_{t+1}) + \frac{1}{2\eta}\|\theta - \theta_{t+1}\|^2_{\nabla^2 \ell_t(\theta_{t+1})}. \qquad \text{(B.3)}\end{aligned}$$

Next, expand $f_t'(x_t^\top \theta_{t+1})$ around $x_t^\top \theta_t$. By Taylor's theorem, there exists ξ_t between $x_t^\top \theta_t$ and $x_t^\top \theta_{t+1}$ such that

$$f_t'(x_t^\top \theta_{t+1}) = f_t'(x_t^\top \theta_t) + f_t''(x_t^\top \theta_t)\, x_t^\top(\theta_{t+1} - \theta_t) + \frac{1}{2} f_t'''(\xi_t)\big(x_t^\top(\theta_{t+1} - \theta_t)\big)^2.$$

Therefore,

$$\begin{aligned}&f_t'(x_t^\top \theta_{t+1})\, x_t^\top(\theta - \theta_{t+1}) \\ &\geqslant \Big(f_t'(x_t^\top \theta_t) + f_t''(x_t^\top \theta_t)\, x_t^\top(\theta_{t+1} - \theta_t)\Big) x_t^\top(\theta - \theta_{t+1}) - \frac{R k_\mu}{2g(\tau)}\big(x_t^\top(\theta_{t+1} - \theta_t)\big)^2 |x_t^\top(\theta - \theta_{t+1})| \\ &\geqslant \Big(f_t'(x_t^\top \theta_t) + f_t''(x_t^\top \theta_t)\, x_t^\top(\theta_{t+1} - \theta_t)\Big) x_t^\top(\theta - \theta_{t+1}) - \frac{R k_\mu S}{g(\tau)}\big(x_t^\top(\theta_{t+1} - \theta_t)\big)^2, \qquad \text{(B.4)}\end{aligned}$$

where the last inequality holds because $|x_t^\top(\theta - \theta_{t+1})| \leqslant 2S$ (Assumption 1).

Moreover, by the definitions of f_t and ℓ_t, we have

$$f_t'(x_t^\top \theta_t)\, x_t^\top(\theta - \theta_{t+1}) = \langle \nabla \ell_t(\theta_t),\, \theta - \theta_{t+1} \rangle, \qquad \text{(B.5)}$$

and

$$f_t''(x_t^\top \theta_t)\, x_t^\top(\theta_{t+1} - \theta_t)\, x_t^\top(\theta - \theta_{t+1}) = \Big\langle \nabla^2 \ell_t(\theta_t)(\theta_{t+1} - \theta_t),\, \theta - \theta_{t+1} \Big\rangle. \qquad \text{(B.6)}$$

By combining equation B.3, equation B.4, equation B.5, and equation B.6, we obtain

$$\begin{aligned}\ell_t(\theta) - \ell_t(\theta_{t+1}) &\geqslant \Big\langle \nabla \ell_t(\theta_t) + \nabla^2 \ell_t(\theta_t)(\theta_{t+1} - \theta_t),\, \theta - \theta_{t+1} \Big\rangle \\ &\quad + \frac{1}{2\eta}\|\theta - \theta_{t+1}\|^2_{\nabla^2 \ell_t(\theta_{t+1})} - \frac{R k_\mu S}{g(\tau)}\big(x_t^\top(\theta_{t+1} - \theta_t)\big)^2 \\ &\geqslant \langle \nabla \widetilde{\ell}_t(\theta_{t+1}),\, \theta - \theta_{t+1} \rangle + \frac{1}{2\eta}\|\theta - \theta_{t+1}\|^2_{\nabla^2 \ell_t(\theta_{t+1})} - \frac{R k_\mu S}{g(\tau)}\big(x_t^\top(\theta_{t+1} - \theta_t)\big)^2. \\ &\qquad \big(\nabla \widetilde{\ell}_t(\theta_{t+1}) = \nabla \ell_t(\theta_t) + \nabla^2 \ell_t(\theta_t)(\theta_{t+1} - \theta_t)\big)\end{aligned}$$

Multiplying both sides by 2η, we get

$$2\eta\big(\ell_t(\theta)-\ell_t(\theta_{t+1})\big) \geqslant 2\eta\langle\nabla\widetilde{\ell}_t(\theta_{t+1}),\,\theta-\theta_{t+1}\rangle+\|\theta-\theta_{t+1}\|^2_{\nabla^2\ell_t(\theta_{t+1})} - \frac{2\eta R k_\mu S}{g(\tau)}\big(x_t^\top(\theta_{t+1}-\theta_t)\big)^2.$$

Because $\|x_t\|_2 \leqslant 1$, $A_t \succeq \lambda I_d$, and $\lambda \geqslant \frac{6\eta R k_\mu S}{g(\tau)}$, we have

$$\frac{2\eta R k_\mu S}{g(\tau)}\big(x_t^\top(\theta_{t+1}-\theta_t)\big)^2 \leqslant \frac{2\eta R k_\mu S}{g(\tau)}\|\theta_{t+1}-\theta_t\|_2^2 \leqslant \frac{2\eta R k_\mu S}{g(\tau)}\frac{1}{\lambda}\|\theta_{t+1}-\theta_t\|_{A_t}^2 \leqslant \frac{1}{3}\|\theta_{t+1}-\theta_t\|_{A_t}^2.$$

Hence, we obtain

$$2\eta\big(\ell_t(\theta)-\ell_t(\theta_{t+1})\big) \geqslant 2\eta\langle\nabla\widetilde{\ell}_t(\theta_{t+1}),\,\theta-\theta_{t+1}\rangle+\|\theta-\theta_{t+1}\|^2_{\nabla^2\ell_t(\theta_{t+1})}-\frac{1}{3}\|\theta_{t+1}-\theta_t\|_{A_t}^2. \tag{B.7}$$

Now, for any $\theta\in\Theta$, we define $F_t(\theta) := \widetilde{\ell}_t(\theta)+\frac{1}{2\eta}\|\theta-\theta_t\|_{A_t}^2$. Since F_t is convex and differentiable on the convex set Θ, and $\theta_{t+1}\in\arg\min_{\theta\in\Theta}F_t(\theta)$, we have

$$\big\langle\nabla F_t(\theta_{t+1}),\,\theta-\theta_{t+1}\big\rangle \geqslant 0.$$

That is,

$$\eta\langle\nabla\widetilde{\ell}_t(\theta_{t+1}),\,\theta-\theta_{t+1}\rangle \geqslant -\big\langle A_t(\theta_{t+1}-\theta_t),\,\theta-\theta_{t+1}\big\rangle. \tag{B.8}$$

Substituting equation B.8 into equation B.7 yields

$$2\eta\big(\ell_t(\theta)-\ell_t(\theta_{t+1})\big) \geqslant -2\big\langle A_t(\theta_{t+1}-\theta_t),\,\theta-\theta_{t+1}\big\rangle+\|\theta-\theta_{t+1}\|^2_{\nabla^2\ell_t(\theta_{t+1})}-\frac{1}{3}\|\theta_{t+1}-\theta_t\|_{A_t}^2. \tag{B.9}$$

Finally, since $\theta-\theta_t=(\theta-\theta_{t+1})+(\theta_{t+1}-\theta_t)$, the three-point identity gives

$$-2\big\langle A_t(\theta_{t+1}-\theta_t),\,\theta-\theta_{t+1}\big\rangle = \|\theta-\theta_{t+1}\|_{A_t}^2-\|\theta-\theta_t\|_{A_t}^2+\|\theta_{t+1}-\theta_t\|_{A_t}^2.$$

Plugging this into equation B.9, we have

$$2\eta\big(\ell_t(\theta)-\ell_t(\theta_{t+1})\big) \geqslant \|\theta-\theta_{t+1}\|_{A_t}^2-\|\theta-\theta_t\|_{A_t}^2+\frac{2}{3}\|\theta_{t+1}-\theta_t\|_{A_t}^2+\|\theta-\theta_{t+1}\|^2_{\nabla^2\ell_t(\theta_{t+1})}.$$

Since $H_{t+1}=A_t+\nabla^2\ell_t(\theta_{t+1})$, we conclude that

$$\|\theta_{t+1}-\theta\|_{H_{t+1}}^2 \leqslant \|\theta_t-\theta\|_{A_t}^2+2\eta\big(\ell_t(\theta)-\ell_t(\theta_{t+1})\big)-\frac{2}{3}\|\theta_{t+1}-\theta_t\|_{A_t}^2.$$

This proves the one-step inequality, i.e., the first statement.

To unroll this bound, using $A_t=\gamma H_t+(1-\gamma)\lambda I_d$, we get

$$\|\theta_t-\theta\|_{A_t}^2 \leqslant \gamma\|\theta_t-\theta\|_{H_t}^2+(1-\gamma)\lambda\|\theta_t-\theta\|_2^2 \leqslant \gamma\|\theta_t-\theta\|_{H_t}^2+4(1-\gamma)\lambda S^2,$$

where the last step uses $\theta_t,\theta\in\Theta$. Therefore,

$$\|\theta_{t+1}-\theta\|_{H_{t+1}}^2 \leqslant \gamma\|\theta_t-\theta\|_{H_t}^2+4(1-\gamma)\lambda S^2+2\eta\big(\ell_t(\theta)-\ell_t(\theta_{t+1})\big)-\frac{2}{3}\|\theta_{t+1}-\theta_t\|_{A_t}^2.$$

Iterating this inequality from $s=1$ to $t-1$, we obtain

$$\|\theta_t-\theta\|_{H_t}^2 \leqslant \gamma^{t-1}\|\theta_1-\theta\|_{H_1}^2+4(1-\gamma)\lambda S^2\sum_{s=1}^{t-1}\gamma^{t-1-s} + 2\eta\sum_{s=1}^{t-1}\gamma^{t-1-s}\big(\ell_s(\theta)-\ell_s(\theta_{s+1})\big)-\frac{2}{3}\sum_{s=1}^{t-1}\gamma^{t-1-s}\|\theta_{s+1}-\theta_s\|_{A_s}^2. \tag{B.10}$$

Since $H_1=\lambda I_d$ and $\theta_1,\theta\in\Theta$,

$$\|\theta_1-\theta\|_{H_1}^2=\lambda\|\theta_1-\theta\|_2^2\leqslant 4\lambda S^2.$$

Therefore,

$$\gamma^{t-1}\|\theta_1-\theta\|_{H_1}^2+4(1-\gamma)\lambda S^2\sum_{s=1}^{t-1}\gamma^{t-1-s}\leqslant 4\gamma^{t-1}\lambda S^2+4(1-\gamma)\lambda S^2\cdot\frac{1-\gamma^{t-1}}{1-\gamma}\leqslant 4\lambda S^2.$$

Substituting this into equation B.10 gives

$$\|\theta_t-\theta\|_{H_t}^2\leqslant 4\lambda S^2+2\eta\sum_{s=1}^{t-1}\gamma^{t-1-s}\big(\ell_s(\theta)-\ell_s(\theta_{s+1})\big)-\frac{2}{3}\sum_{s=1}^{t-1}\gamma^{t-1-s}\|\theta_{s+1}-\theta_s\|_{A_s}^2.$$

This proves the second statement. □

B.2.2 Proof of Lemma B.2

Proof of Lemma B.2. Fix $t\geqslant 2$. For simplicity, let $w_{s,t}:=\gamma^{t-1-s}\in(0,1]$ and $Z_s:=\ell_s(\theta_s^\star)-m_s(\Pi_s)$. Recall that $\ell_s(\theta)=\frac{m(x_s^\top\theta)-r_sx_s^\top\theta}{g(\tau)}=-\log\Pr\left[r_s\mid x_s,\theta\right]+h(r_s,\tau)$. Hence, we have

$$m_s(\Pi_s)=-\log\int e^{-\ell_s(\theta)}\,\mathrm{d}\Pi_s(\theta)=h(r_s,\tau)-\log\int p(r_s\mid x_s^\top\theta)\,\mathrm{d}\Pi_s(\theta).$$

Therefore, we can write Z_s as:

$$Z_s=\log\frac{\int p(r_s\mid x_s^\top\theta)\,\mathrm{d}\Pi_s(\theta)}{p(r_s\mid x_s^\top\theta_s^\star)}.$$

Define the filtration before observing r_t by $\mathcal{F}_t:=\sigma(x_1,r_1,\ldots,x_{t-1},r_{t-1},x_t)$, and the enlarged filtration $\widetilde{\mathcal{F}}_t:=\mathcal{F}_t\vee\sigma(\theta_s^\star:s\geqslant 1)$. Then, since Π_s is $\widetilde{\mathcal{F}}_s$-measurable and $r_s\mid\widetilde{\mathcal{F}}_s\sim p(\cdot\mid x_s^\top\theta_s^\star)$, we obtain

$$\begin{aligned}\mathbb{E}[e^{Z_s}\mid\widetilde{\mathcal{F}}_s]&=\mathbb{E}\left[\frac{\int p(r_s\mid x_s^\top\theta)\,\mathrm{d}\Pi_s(\theta)}{p(r_s\mid x_s^\top\theta_s^\star)}\;\middle|\;\widetilde{\mathcal{F}}_s\right]\\&=\int\frac{\int p(r\mid x_s^\top\theta)\,\mathrm{d}\Pi_s(\theta)}{p(r\mid x_s^\top\theta_s^\star)}\,p(r\mid x_s^\top\theta_s^\star)\,\mathrm{d}r\\&=\int\int p(r\mid x_s^\top\theta)\,\mathrm{d}\Pi_s(\theta)\,\mathrm{d}r\\&=\int 1\,\mathrm{d}\Pi_s(\theta)=1.\end{aligned}$$

Because $0<w_{s,t}\leqslant 1$, the map $x\mapsto x^{w_{s,t}}$ is concave on $\mathbb{R}_+$. Therefore, conditional Jensen implies

$$\mathbb{E}[e^{w_{s,t}Z_s}\mid\widetilde{\mathcal{F}}_s]=\mathbb{E}[(e^{Z_s})^{w_{s,t}}\mid\widetilde{\mathcal{F}}_s]\leqslant\big(\mathbb{E}[e^{Z_s}\mid\widetilde{\mathcal{F}}_s]\big)^{w_{s,t}}=1.\tag{B.11}$$

We now iterate conditional expectations. By equation B.11, we get

$$\mathbb{E}\Big[e^{\sum_{s=1}^{t-2}w_{s,t}Z_s}\,\mathbb{E}[e^{w_{t-1,t}Z_{t-1}}\mid\widetilde{\mathcal{F}}_{t-1}]\Big]\leqslant\mathbb{E}\Big[e^{\sum_{s=1}^{t-2}w_{s,t}Z_s}\Big].$$

Applying the same argument recursively to $t-2,t-3,\ldots,1$ and using equation B.11 at each step, we get

$$\mathbb{E}[e^{\sum_{s=1}^{t-1}w_{s,t}Z_s}]\leqslant 1.$$

Hence, by Markov's inequality, we obtain

$$\Pr\left(\sum_{s=1}^{t-1}w_{s,t}Z_s\geqslant\log\frac{1}{\delta}\right)=\Pr\left(e^{\sum_{s=1}^{t-1}w_{s,t}Z_s}\geqslant\frac{1}{\delta}\right)\leqslant\delta\mathbb{E}[e^{\sum_{s=1}^{t-1}w_{s,t}Z_s}]\leqslant\delta.$$

Therefore, with probability at least $1-\delta$, we have

$$\sum_{s=1}^{t-1}w_{s,t}Z_s\leqslant\log\frac{1}{\delta}.$$

Recalling $w_{s,t}=\gamma^{t-1-s}$ and $Z_s=\ell_s(\theta_s^\star)-m_s(\Pi_s)$ proves the lemma. □

B.2.3 PROOF OF LEMMA B.3

Proof of Lemma B.3. Fix s. Define $Q_s := \mathcal{N}(\theta_{s+1}, \alpha H_{s+1}^{-1})$, where $\alpha = 3\eta/2$. Then, by the variational formula for the mix loss (Lemma B.8), we have

$$m_s(\Pi_s) \leqslant \mathbb{E}_{\theta\sim Q_s}[\ell_s(\theta)] + D_{\mathrm{KL}}(Q_s\|\Pi_s).$$

Recall the definition of the Bregman divergence

$$D_{\ell_s}(\theta, \theta_{s+1}) := \ell_s(\theta) - \ell_s(\theta_{s+1}) - \langle \nabla\ell_s(\theta_{s+1}),\, \theta - \theta_{s+1}\rangle.$$

Taking expectation under Q_s and using that Q_s is centered at θ_{s+1}, so that $\mathbb{E}_{Q_s}[\theta - \theta_{s+1}] = 0$, gives

$$\mathbb{E}_{Q_s}[\ell_s(\theta)] - \ell_s(\theta_{s+1}) = \mathbb{E}_{Q_s}[D_{\ell_s}(\theta, \theta_{s+1})].$$

Hence, we get

$$m_s(\Pi_s) - \ell_s(\theta_{s+1}) \leqslant \underbrace{\mathbb{E}_{Q_s}\big[D_{\ell_s}(\theta, \theta_{s+1})\big]}_{=:I_{1,s}} + \underbrace{D_{\mathrm{KL}}(Q_s\|\Pi_s)}_{=:I_{2,s}}. \tag{B.12}$$

To bound $I_{1,s}$, we apply Lemma B.7, which yields

$$I_{1,s} \leqslant 2\alpha\Big(\log\det(H_{s+1}) - \log\det(A_s)\Big). \tag{B.13}$$

We next turn to the term $I_{2,s}$. Since $\Pi_s = \mathcal{N}(\theta_s, \alpha A_s^{-1})$ and $Q_s = \mathcal{N}(\theta_{s+1}, \alpha H_{s+1}^{-1})$, the Gaussian KL formula gives

$$I_{2,s} = \frac{1}{2}\left[\log\frac{\det(H_{s+1})}{\det(A_s)} + \mathrm{Tr}(A_s H_{s+1}^{-1}) - d + \frac{1}{\alpha}\|\theta_{s+1} - \theta_s\|_{A_s}^2\right].$$

Since $H_{s+1} \succeq A_s$, we have $A_s^{1/2} H_{s+1}^{-1} A_s^{1/2} \preceq I_d$, so every eigenvalue of $A_s H_{s+1}^{-1}$ lies in $(0, 1]$. Therefore, we get

$$\mathrm{Tr}(A_s H_{s+1}^{-1}) - d \leqslant 0.$$

This gives

$$\begin{aligned} I_{2,s} &\leqslant \frac{1}{2}\left[\log\frac{\det(H_{s+1})}{\det(A_s)} + \frac{1}{\alpha}\|\theta_{s+1} - \theta_s\|_{A_s}^2\right] \\ &= \frac{1}{2}\Big(\log\det(H_{s+1}) - \log\det(A_s)\Big) + \frac{1}{2\alpha}\|\theta_{s+1} - \theta_s\|_{A_s}^2 \end{aligned} \tag{B.14}$$

Plugging equation B.13 and equation B.14 into equation B.12, we obtain

$$m_s(\Pi_s) - \ell_s(\theta_{s+1}) \leqslant \left(2\alpha + \frac{1}{2}\right)\Big(\log\det(H_{s+1}) - \log\det(A_s)\Big) + \frac{1}{2\alpha}\|\theta_{s+1} - \theta_s\|_{A_s}^2. \tag{B.15}$$

For simplicity, let $L_s := \log\det(H_s)$. Since $A_s = \gamma H_s + (1-\gamma)\lambda I_d$, concavity of $\log\det(\cdot)$ implies

$$\log\det(A_s) \geqslant \gamma L_s + (1-\gamma) d\log\lambda.$$

Therefore

$$\log\det(H_{s+1}) - \log\det(A_s) = L_{s+1} - \log\det(A_s) \leqslant L_{s+1} - \gamma L_s - (1-\gamma)d\log\lambda. \tag{B.16}$$

Multiplying both sides of equation B.16 by γ^{t-1-s} and summing over $s = 1, \ldots, t-1$ yields

$$\begin{aligned} \sum_{s=1}^{t-1}\gamma^{t-1-s}\Big(\log\det(H_{s+1}) - \log\det(A_s)\Big) \leqslant \sum_{s=1}^{t-1}\gamma^{t-1-s}L_{s+1} - \sum_{s=1}^{t-1}\gamma^{t-s}L_s \\ - (1-\gamma)d\log\lambda\sum_{s=1}^{t-1}\gamma^{t-1-s}. \end{aligned}$$

The first two sums can be telescoped as follows:

$$\sum_{s=1}^{t-1}\gamma^{t-1-s}L_{s+1} - \sum_{s=1}^{t-1}\gamma^{t-s}L_s = L_t - \gamma^{t-1}L_1 = L_t - \gamma^{t-1}d\log\lambda,$$

where the last equality holds since $L_1 = d\log\lambda$. Moreover, evaluating the geometric sum gives

$$(1-\gamma)\sum_{s=1}^{t-1}\gamma^{t-1-s} = 1-\gamma^{t-1}.$$

Hence, we get

$$\begin{aligned}\sum_{s=1}^{t-1}\gamma^{t-1-s}\Big(\log\det(H_{s+1}) - \log\det(A_s)\Big) &\leqslant L_t - \gamma^{t-1}d\log\lambda - (1-\gamma^{t-1})d\log\lambda \\ &= L_t - d\log\lambda. \end{aligned} \tag{B.17}$$

On the other hand, by the definition of $H_t = \lambda I_d + \sum_{s=1}^{t-1}\gamma^{t-1-s}\frac{\mu'(x_s^\top\theta_{s+1})}{g(\tau)}x_s x_s^\top$, we have

$$\begin{aligned}\mathrm{Tr}(H_t) &= \lambda d + \sum_{s=1}^{t-1}\gamma^{t-1-s}\frac{\mu'(x_s^\top\theta_{s+1})}{g(\tau)}\|x_s\|_2^2 \\ &\leqslant \lambda d + \frac{k_\mu}{g(\tau)}\sum_{s=1}^{t-1}\gamma^{t-1-s} && (\|x\|_2 \leqslant 1, \text{ Assumption 1}) \\ &= \lambda d + \frac{k_\mu(1-\gamma^{t-1})}{g(\tau)(1-\gamma)}. && (\textstyle\sum_{s=1}^{t-1}\gamma^{t-1-s} = (1-\gamma^{t-1})/(1-\gamma))\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}L_t - d\log\lambda = \log\det(H_t) - d\log\lambda &\leqslant \log\left(\frac{\mathrm{Tr}(H_t)}{d}\right)^d - d\log\lambda && \text{(AM-GM inequality)} \\ &\leqslant d\log\left(1 + \frac{k_\mu(1-\gamma^{t-1})}{\lambda d\, g(\tau)(1-\gamma)}\right). \end{aligned} \tag{B.18}$$

Finally, summing equation B.15 with weights γ^{t-1-s} yields

$$\begin{aligned}&\sum_{s=1}^{t-1}\gamma^{t-1-s}\big(m_s(\Pi_s) - \ell_s(\theta_{s+1})\big) \\ &\quad\leqslant \left(2\alpha + \frac{1}{2}\right)\sum_{s=1}^{t-1}\gamma^{t-1-s}\Big(\log\det(H_{s+1}) - \log\det(A_s)\Big) + \frac{1}{2\alpha}\sum_{s=1}^{t-1}\gamma^{t-1-s}\|\theta_{s+1}-\theta_s\|_{A_s}^2 \\ &\quad\leqslant \left(2\alpha + \frac{1}{2}\right)\big(L_t - d\log\lambda\big) + \frac{1}{2\alpha}\sum_{s=1}^{t-1}\gamma^{t-1-s}\|\theta_{s+1}-\theta_s\|_{A_s}^2 && \text{(Eqn. equation B.17)} \\ &\quad\leqslant \left(2\alpha + \frac{1}{2}\right)d\log\left(1 + \frac{k_\mu(1-\gamma^{t-1})}{\lambda d\, g(\tau)(1-\gamma)}\right) + \frac{1}{2\alpha}\sum_{s=1}^{t-1}\gamma^{t-1-s}\|\theta_{s+1}-\theta_s\|_{A_s}^2. \\ &&& \text{(Eqn. equation B.18)}\end{aligned}$$

Finally, substituting $\alpha = \frac{3\eta}{2}$ completes the proof. □

B.2.4 Proof of Lemma B.4

Proof of Lemma B.4. Fix $t \geqslant 2$. Using the definition of ℓ_s, we obtain

$$\begin{aligned}\ell_s(\theta_t^\star) - \ell_s(\theta_s^\star) &= \frac{m(x_s^\top\theta_t^\star) - r_s\, x_s^\top\theta_t^\star - m(x_s^\top\theta_s^\star) + r_s\, x_s^\top\theta_s^\star}{g(\tau)} \\ &= \frac{m(x_s^\top\theta_t^\star) - m(x_s^\top\theta_s^\star) - r_s\, x_s^\top(\theta_t^\star - \theta_s^\star)}{g(\tau)} \\ &= \frac{m(x_s^\top\theta_t^\star) - m(x_s^\top\theta_s^\star) - \mu(x_s^\top\theta_s^\star)\, x_s^\top(\theta_t^\star - \theta_s^\star)}{g(\tau)} - \frac{\eta_s\, x_s^\top(\theta_t^\star - \theta_s^\star)}{g(\tau)}. \\ &&& (r_s = \mu(x_s^\top\theta_s^\star) + \eta_s)\end{aligned}$$

Therefore, we get

$$\sum_{s=1}^{t-1} \gamma^{t-1-s}\big(\ell_s(\theta_t^\star) - \ell_s(\theta_s^\star)\big) = B_t^{\mathrm{det}} - M_t,$$

where the deterministic term is defined by

$$B_t^{\mathrm{det}} := \frac{1}{g(\tau)} \sum_{s=1}^{t-1} \gamma^{t-1-s}\Big(m(x_s^\top \theta_t^\star) - m(x_s^\top \theta_s^\star) - \mu(x_s^\top \theta_s^\star)\, x_s^\top(\theta_t^\star - \theta_s^\star)\Big),$$

and the stochastic term is

$$M_t := \frac{1}{g(\tau)} \sum_{s=1}^{t-1} \gamma^{t-1-s} \eta_s\, x_s^\top(\theta_t^\star - \theta_s^\star).$$

i) Deterministic part. By Assumption 1, both $x_s^\top \theta_t^\star$ and $x_s^\top \theta_s^\star$ belong to $[-S, S]$. Since $m' = \mu$, $m'' = \mu'$, and Assumption 2 gives $\mu'(z) \leqslant k_\mu$ for all $z \in [-S, S]$, Taylor's theorem implies that

$$m(x_s^\top \theta_t^\star) - m(x_s^\top \theta_s^\star) - \mu(x_s^\top \theta_s^\star)\, x_s^\top(\theta_t^\star - \theta_s^\star) \leqslant \frac{k_\mu}{2}\big(x_s^\top(\theta_t^\star - \theta_s^\star)\big)^2.$$

Consequently, we can bound B_t^{det} as follows:

$$\begin{aligned} B_t^{\mathrm{det}} &\leqslant \frac{k_\mu}{2g(\tau)} \sum_{s=1}^{t-1} \gamma^{t-1-s}\big(x_s^\top(\theta_t^\star - \theta_s^\star)\big)^2 \\ &\leqslant \frac{k_\mu}{2g(\tau)}\, \mathcal{V}_t^2. \qquad \text{(Lemma B.9)} \end{aligned}$$

ii) Stochastic part. For fixed t, let

$$c_{s,t} := \gamma^{t-1-s}\, x_s^\top(\theta_t^\star - \theta_s^\star), \qquad s = 1, \dots, t-1.$$

Define the filtration before observing r_t by $\mathcal{F}_t := \sigma(x_1, r_1, \dots, x_{t-1}, r_{t-1}, x_t)$, and the enlarged filtration $\widetilde{\mathcal{F}}_t := \mathcal{F}_t \vee \sigma(\theta_u^\star : u \geqslant 1)$. Since the path $\{\theta_u^\star\}_{u=1}^T$ is *oblivious*, i.e., the entire sequence of underlying parameters does not depend on the realized rewards or on the learner's internal randomness, and x_s is $\widetilde{\mathcal{F}}_s$-measurable, each $c_{s,t}$ is $\widetilde{\mathcal{F}}_s$-measurable.

Moreover, because $r_s \in [0, R]$ almost surely (Assumption 1) and $\mu(x_s^\top \theta_s^\star) = \mathbb{E}[r_s | \widetilde{\mathcal{F}}_s] \in [0, R]$, we have $\eta_s \in [-\mu(x_s^\top \theta_s^\star),\ R - \mu(x_s^\top \theta_s^\star)]$ almost surely. Thus, conditioned on $\widetilde{\mathcal{F}}_s$, the centered random variable η_s lies in an interval of length R. Hence, by Hoeffding's lemma, for every $\bar{\lambda} > 0$, we get

$$\mathbb{E}\Big[\exp(-\bar{\lambda} c_{s,t} \eta_s) \,\Big|\, \widetilde{\mathcal{F}}_s\Big] \leqslant \exp\left(\frac{\bar{\lambda}^2 c_{s,t}^2 R^2}{8}\right).$$

Let $Y_t := \sum_{s=1}^{t-1} c_{s,t} \eta_s$. Iterating conditional expectations from $s = t-1$ down to $s = 1$ yields

$$\mathbb{E}\Big[e^{-\bar{\lambda} Y_t}\Big] \leqslant \exp\left(\frac{\bar{\lambda}^2 R^2}{8} \sum_{s=1}^{t-1} c_{s,t}^2\right).$$

Therefore, for every $u > 0$ and every $\bar{\lambda} > 0$, Chernoff's bound gives

$$\begin{aligned} \mathbb{P}(-M_t \geqslant u) &= \mathbb{P}\left(-\sum_{s=1}^{t-1} c_{s,t}\eta_s \geqslant g(\tau)u\right) \\ &\leqslant \exp\big(-\bar{\lambda} g(\tau) u\big)\, \mathbb{E}\Big[e^{-\bar{\lambda} Y_t}\Big] \qquad \text{(Chernoff's bound)} \\ &\leqslant \exp\left(-\bar{\lambda} g(\tau) u + \frac{\bar{\lambda}^2 R^2}{8} \sum_{s=1}^{t-1} c_{s,t}^2\right). \end{aligned}$$

Optimizing over $\bar{\lambda} > 0$ gives $\bar{\lambda}^\star = \frac{4g(\tau)u}{R^2 \sum_{s=1}^{t-1} c_{s,t}^2}$, and thus

$$\mathbb{P}(-M_t \geqslant u) \leqslant \exp\left(-\frac{2g(\tau)^2 u^2}{R^2 \sum_{s=1}^{t-1} c_{s,t}^2}\right).$$

Hence, with probability at least $1-\delta$, we have

$$\begin{aligned}
-M_t &\leqslant \frac{R}{g(\tau)}\sqrt{\frac{1}{2}\log\frac{1}{\delta}\sum_{s=1}^{t-1} c_{s,t}^2} \\
&= \frac{R}{g(\tau)}\sqrt{\frac{1}{2}\log\frac{1}{\delta}\sum_{s=1}^{t-1} \gamma^{2(t-1-s)}\big(x_s^\top(\theta_t^\star - \theta_s^\star)\big)^2} \\
&\leqslant \frac{R}{g(\tau)}\sqrt{\frac{1}{2}\log\frac{1}{\delta}}\,\mathcal{V}_t. \qquad \text{(Lemma B.9)}
\end{aligned}$$

Applying Young's inequality $ab \leqslant \frac{\varepsilon}{2}a^2 + \frac{1}{2\varepsilon}b^2$ with $a = \mathcal{V}_t$, $b = \frac{R}{g(\tau)}\sqrt{\frac{1}{2}\log\frac{1}{\delta}}$, and $\varepsilon = \frac{k_\mu}{g(\tau)}$, we obtain

$$-M_t \leqslant \frac{k_\mu}{2g(\tau)}\mathcal{V}_t^2 + \frac{R^2}{4g(\tau)k_\mu}\log\frac{1}{\delta}.$$

Combining the bounds on $B_t^{\det}$ and $-M_t$, we conclude that, with probability at least $1-\delta$,

$$\sum_{s=1}^{t-1}\gamma^{t-1-s}\big(\ell_s(\theta_t^\star) - \ell_s(\theta_s^\star)\big) \leqslant \frac{k_\mu}{g(\tau)}\,\mathcal{V}_t^2 + \frac{R^2}{4g(\tau)k_\mu}\log\frac{1}{\delta}.$$

This completes the proof. □

B.3 Auxiliary Lemmas for Theorem 1

Lemma B.5 (Self-concordance control)**.** *Under Assumption 1, for every $v, z_1, z_2 \in \mathbb{R}$,*

$$\int_0^1 (1-s)\mu'\big(v + s(z_1 - z_2)\big)\,\mathrm{d}s \geqslant \frac{\mu'(v)}{2 + R|z_1 - z_2|}, \tag{B.19}$$

and

$$\int_0^1 (1-s)\mu'\big(v + s(z_1 - z_2)\big)\,\mathrm{d}s \leqslant \exp\left(\frac{R^2(z_1 - z_2)^2}{4}\right)\mu'(v). \tag{B.20}$$

Proof of Lemma B.5. Under the bounded rewards assumption (Assumption 1), Proposition 1 implies that $|\mu''(z)| \leqslant R\,\mu'(z)$ for all $z \in \mathbb{R}$. Equivalently,

$$\left|\frac{d}{du}\log\mu'(u)\right| = \left|\frac{\mu''(u)}{\mu'(u)}\right| \leqslant R.$$

Hence, for any $s \in [0,1]$,

$$\begin{aligned}
\log\mu'\big(v + s(z_1 - z_2)\big) - \log\mu'(v) &= \int_0^1 \frac{d}{dw}\log\mu'\big(v + ws(z_1 - z_2)\big)\,\mathrm{d}w \\
&= \int_0^1 \frac{\mu''\big(v + ws(z_1 - z_2)\big)}{\mu'\big(v + ws(z_1 - z_2)\big)}\,s(z_1 - z_2)\,\mathrm{d}w.
\end{aligned}$$

Therefore,

$$\big|\log\mu'\big(v + s(z_1 - z_2)\big) - \log\mu'(v)\big| \leqslant \int_0^1 R\,s|z_1 - z_2|\,\mathrm{d}w = Rs|z_1 - z_2|,$$

which implies

$$e^{-Rs|z_1 - z_2|}\mu'(v) \leqslant \mu'\big(v + s(z_1 - z_2)\big) \leqslant e^{Rs|z_1 - z_2|}\mu'(v).$$

For the lower bound, let $u := R|z_1 - z_2|$. Then

$$\int_0^1 (1-s)\mu'\big(v + s(z_1 - z_2)\big)\,\mathrm{d}s \geqslant \mu'(v)\int_0^1 (1-s)e^{-us}\,\mathrm{d}s$$
$$= \mu'(v)\frac{u - 1 + e^{-u}}{u^2} \qquad (u > 0),$$

where the last equality follows from a direct calculation.

Hence it suffices to show

$$\frac{u - 1 + e^{-u}}{u^2} \geqslant \frac{1}{2+u} \qquad \forall u > 0.$$

Define $q(u) := (2+u)(u - 1 + e^{-u}) - u^2$. Then $q(0) = 0$, and

$$q'(u) = 1 - (1+u)e^{-u} \geqslant 0 \qquad \forall u > 0,$$

because $e^u > 1 + u$. Therefore $q(u) > 0$ for all $u > 0$, which proves equation B.19.

For the upper bound, using $\mu'\big(v + s(z_1 - z_2)\big) \leqslant e^{us}\mu'(v)$, we obtain

$$\int_0^1 (1-s)\mu'\big(v + s(z_1 - z_2)\big)\,\mathrm{d}s \leqslant \mu'(v)\int_0^1 (1-s)e^{us}\,\mathrm{d}s.$$

Let $B \sim \mathrm{Beta}(1,2)$, whose density is $2(1-s)\mathbf{1}\{0 \leqslant s \leqslant 1\}$. Then

$$\int_0^1 (1-s)e^{us}\,\mathrm{d}s = \frac{1}{2}\mathbb{E}[e^{uB}].$$

Since $B \in [0,1]$, Hoeffding's lemma yields

$$\mathbb{E}[e^{u(B-\mathbb{E}B)}] \leqslant e^{u^2/8}. \tag{B.21}$$

Using $\mathbb{E}B = 1/3$, we get

$$\int_0^1 (1-s)e^{us}\,\mathrm{d}s = \frac{1}{2}\mathbb{E}[e^{uB}] = \frac{1}{2}e^{u\mathbb{E}B}\mathbb{E}[e^{u(B-\mathbb{E}B)}] = \frac{1}{2}e^{u/3}\mathbb{E}[e^{u(B-\mathbb{E}B)}]$$
$$\leqslant \frac{1}{2}\exp\left(\frac{u}{3} + \frac{u^2}{8}\right). \qquad \text{(Eqn. equation B.21)}$$

Moreover, $\frac{u^2}{8} + \frac{2}{9} - \frac{u}{3} = \frac{(3u-4)^2}{72} \geqslant 0$, so $\frac{u}{3} \leqslant \frac{u^2}{8} + \frac{2}{9}$. Hence

$$\int_0^1 (1-s)e^{us}\,\mathrm{d}s \leqslant \frac{1}{2}\exp\left(\frac{u}{3} + \frac{u^2}{8}\right) \leqslant \frac{1}{2}e^{2/9}e^{u^2/4} \leqslant e^{u^2/4}, \qquad (e^{2/9}/2 < 1)$$

which proves equation B.20. □

Lemma B.6 (Gaussian exponential moment)**.** *For any $\alpha, c > 0$, let $Z \sim \mathcal{N}(0, \alpha H^{-1})$, where $H \succeq \lambda I_d$. If $\lambda \geqslant \frac{64\alpha dc}{7}$, then, we have*

$$\mathbb{E}\Big[e^{c\|Z\|_2^2}\Big] \leqslant \frac{4}{3}.$$

Proof of Lemma B.6. Let $\xi \sim \mathcal{N}(0, I_d)$ and write $Z = \sqrt{\alpha}\, H^{-1/2}\xi$. Since $H^{-1} \preceq \lambda^{-1} I_d$, we have pointwise

$$\|Z\|_2^2 = \alpha\, \xi^\top H^{-1}\xi \leqslant \frac{\alpha}{\lambda}\|\xi\|_2^2.$$

Therefore

$$\mathbb{E}\Big[e^{c\|Z\|_2^2}\Big] \leqslant \mathbb{E}\Big[e^{\frac{\alpha c}{\lambda}\|\xi\|_2^2}\Big]. \tag{B.22}$$

Now write $\xi = (\xi_1, \dots, \xi_d)$ with $\xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Then, since $\|\xi\|_2^2 = \sum_{i=1}^d \xi_i^2$, by defining $h := \alpha c/\lambda$, we get

$$\mathbb{E}\Big[e^{h\|\xi\|_2^2}\Big] = \mathbb{E}\left[\exp\left(h\sum_{i=1}^d \xi_i^2\right)\right] = \prod_{i=1}^d \mathbb{E}\Big[e^{h\xi_i^2}\Big].$$

For each i, we have

$$\begin{aligned}\mathbb{E}\Big[e^{h\xi_i^2}\Big] &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \exp\left(hx^2 - \frac{x^2}{2}\right) dx \\ &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \exp\left(-\frac{1-2h}{2}x^2\right) dx \\ &= \frac{1}{\sqrt{1-2h}}. \qquad \left(\tfrac{\alpha c}{\lambda} \leqslant \tfrac{7}{64d} < \tfrac{1}{2}\right)\end{aligned}$$

Hence, we get

$$\mathbb{E}\Big[e^{h\|\xi\|_2^2}\Big] = \prod_{i=1}^d \mathbb{E}\Big[e^{h\xi_i^2}\Big] = (1-2h)^{-d/2}.$$

Plugging this into equation B.22 and substituting $h = \alpha c/\lambda$, we obtain

$$\mathbb{E}\Big[e^{c\|Z\|_2^2}\Big] \leqslant \left(1 - \frac{2\alpha c}{\lambda}\right)^{-d/2}.$$

Finally, our assumption $\lambda \geqslant \frac{64\alpha dc}{7}$ implies $\frac{2\alpha c}{\lambda} \leqslant \frac{7}{32d} < 1$. Since the map $x \mapsto (1-x)^{-d/2}$ is increasing on $(0,1)$, it follows that

$$\left(1 - \frac{2\alpha c}{\lambda}\right)^{-d/2} \leqslant \left(1 - \frac{7}{32d}\right)^{-d/2}.$$

Now use the elementary bound $-\log(1-x) \leqslant \frac{x}{1-x}$ for $x \in (0,1)$. With $x = \frac{7}{32d}$, we obtain

$$\begin{aligned}\left(1 - \frac{7}{32d}\right)^{-d/2} &= \exp\left(-\frac{d}{2}\log\left(1 - \frac{7}{32d}\right)\right) \leqslant \exp\left(\frac{d}{2}\cdot\frac{7/(32d)}{1-7/(32d)}\right) \\ &= \exp\left(\frac{7/64}{1-7/(32d)}\right).\end{aligned}$$

Finally, since $d \geqslant 1$, we get $1 - \frac{7}{32d} \geqslant 1 - \frac{7}{32} = \frac{25}{32}$, and therefore

$$\frac{7/64}{1-7/(32d)} \leqslant \frac{7/64}{25/32} = \frac{7}{50}.$$

Hence, we have

$$\left(1 - \frac{2\alpha c}{\lambda}\right)^{-d/2} \leqslant \exp\left(\frac{7/64}{1-7/(32d)}\right) \leqslant e^{7/50} < \frac{4}{3},$$

which proves the lemma. □

Lemma B.7. *Suppose $\|x_s\|_2 \leqslant 1$ and $\lambda \geqslant 32\alpha dR^2/7$. For each $s \geqslant 1$, let $Q_s := \mathcal{N}(\theta_{s+1}, \alpha H_{s+1}^{-1})$ be the Gaussian distribution centered at θ_{s+1} with covariance matrix αH_{s+1}^{-1}. Also define the Bregman divergence of ℓ_s at θ_{s+1} by $D_{\ell_s}(\theta, \theta_{s+1}) := \ell_s(\theta) - \ell_s(\theta_{s+1}) - \langle \nabla\ell_s(\theta_{s+1}),\, \theta - \theta_{s+1}\rangle$. Then, we have*

$$\mathbb{E}_{\theta\sim Q_s}[D_{\ell_s}(\theta, \theta_{s+1})] \leqslant 2\alpha\Big(\log\det(H_{s+1}) - \log\det(A_s)\Big).$$

Proof of Lemma B.7. For convenience, define $\omega_s := \frac{\mu'(x_s^\top\theta_{s+1})}{g(\tau)}$, so that

$$\nabla^2\ell_s(\theta_{s+1}) = \frac{\mu'(x_s^\top\theta_{s+1})}{g(\tau)}x_s x_s^\top = \omega_s x_s x_s^\top,$$

By the integral form of the Bregman divergence, we can write $D_{\ell_s}(\theta, \theta_{s+1})$ as follows:

$$D_{\ell_s}(\theta, \theta_{s+1}) = \frac{\big(x_s^\top(\theta - \theta_{s+1})\big)^2}{g(\tau)} \int_0^1 (1-\nu)\mu'\big(x_s^\top\theta_{s+1} + \nu x_s^\top(\theta - \theta_{s+1})\big)\,\mathrm{d}\nu,$$

Then, by applying Lemma B.5, we obtain

$$D_{\ell_s}(\theta, \theta_{s+1}) \leqslant \exp\left(\frac{R^2\big(x_s^\top(\theta - \theta_{s+1})\big)^2}{4}\right) \frac{\mu'(x_s^\top\theta_{s+1})}{g(\tau)} \big(x_s^\top(\theta - \theta_{s+1})\big)^2.$$

Since

$$\|\theta - \theta_{s+1}\|^2_{\nabla^2\ell_s(\theta_{s+1})} = (\theta - \theta_{s+1})^\top \nabla^2\ell_s(\theta_{s+1})(\theta - \theta_{s+1}) = \frac{\mu'(x_s^\top\theta_{s+1})}{g(\tau)} \big(x_s^\top(\theta - \theta_{s+1})\big)^2,$$

it follows that

$$\begin{aligned} D_{\ell_s}(\theta, \theta_{s+1}) &\leqslant \exp\left(\frac{R^2\big(x_s^\top(\theta - \theta_{s+1})\big)^2}{4}\right) \|\theta - \theta_{s+1}\|^2_{\nabla^2\ell_s(\theta_{s+1})} \\ &\leqslant \exp\left(\frac{R^2\|\theta - \theta_{s+1}\|_2^2}{4}\right) \|\theta - \theta_{s+1}\|^2_{\nabla^2\ell_s(\theta_{s+1})}. \end{aligned} \tag{B.23}$$

where the last inequality holds because $\|x_s\|_2 \leqslant 1$. Now let $W_{s+1} := \theta - \theta_{s+1} \sim \mathcal{N}(0, \alpha H_{s+1}^{-1})$. By taking expectations on both sides of equation B.23 and applying Cauchy–Schwarz, we obtain

$$\mathbb{E}_{Q_s}\big[D_{\ell_s}(\theta, \theta_{s+1})\big] \leqslant \sqrt{\mathbb{E}\big[e^{R^2\|W_{s+1}\|_2^2/2}\big]\ \mathbb{E}\Big[\|W_{s+1}\|^4_{\nabla^2\ell_s(\theta_{s+1})}\Big]}. \tag{B.24}$$

Because $H_{s+1} \succeq \lambda I_d$ and $\lambda \geqslant 32\alpha d R^2/7 = 64\alpha dc/7$, Lemma B.6 with $c = R^2/2$ yields

$$\mathbb{E}\Big[e^{R^2\|W_{s+1}\|_2^2/2}\Big] \leqslant \frac{4}{3}. \tag{B.25}$$

It remains to bound the fourth moment. Let $\Sigma_s := \alpha H_{s+1}^{-1}$ and $M_s := \Sigma_s^{1/2}\nabla^2\ell_s(\theta_{s+1})\Sigma_s^{1/2}$. Since $W_{s+1} \sim \mathcal{N}(0, \Sigma_s)$, we may write

$$W_{s+1} = \Sigma_s^{1/2}\xi, \qquad \xi \sim \mathcal{N}(0, I_d).$$

Therefore,

$$W_{s+1}^\top \nabla^2\ell_s(\theta_{s+1}) W_{s+1} = \xi^\top M_s \xi.$$

Now diagonalize M_s as

$$M_s = U\Lambda U^\top, \qquad \Lambda = \mathrm{diag}(\rho_1, \ldots, \rho_d),$$

where U is an orthogonal matrix and $\rho_1, \ldots, \rho_d \geqslant 0$ are the eigenvalues of M_s. Since $U^\top\xi \sim \mathcal{N}(0, I_d)$, letting $\zeta := U^\top\xi$, we obtain

$$W_{s+1}^\top \nabla^2\ell_s(\theta_{s+1}) W_{s+1} = \xi^\top M_s \xi = \zeta^\top\Lambda\zeta = \sum_{i=1}^d \rho_i\zeta_i^2,$$

where $\zeta_1, \ldots, \zeta_d$ are i.i.d. standard Gaussian random variables. Hence

$$\|W_{s+1}\|^4_{\nabla^2\ell_s(\theta_{s+1})} = (\xi^\top M_s\xi)^2 = \left(\sum_{i=1}^d \rho_i\zeta_i^2\right)^2 = \sum_{i=1}^d \rho_i^2\zeta_i^4 + \sum_{i\neq j} \rho_i\rho_j\zeta_i^2\zeta_j^2.$$

Taking expectations and using $\mathbb{E}[\zeta_i^4] = 3$ and $\mathbb{E}[\zeta_i^2\zeta_j^2] = 1$ for $i \neq j$, we get

$$\begin{aligned} \mathbb{E}\Big[\|W_{s+1}\|^4_{\nabla^2\ell_s(\theta_{s+1})}\Big] &= \mathbb{E}\left[(\xi^\top M_s\xi)^2\right] = 3\sum_{i=1}^d \rho_i^2 + \sum_{i\neq j}\rho_i\rho_j \\ &= 2\sum_{i=1}^d \rho_i^2 + \left(\sum_{i=1}^d \rho_i\right)^2 \\ &= 2\,\mathrm{Tr}(M_s^2) + \mathrm{Tr}(M_s)^2. \end{aligned}$$

Recall that $M_s = \Sigma_s^{1/2} \nabla^2 \ell_s(\theta_{s+1}) \Sigma_s^{1/2}$. Since $\nabla^2 \ell_s(\theta_{s+1}) = \frac{\mu'(x_s^\top \theta_{s+1})}{g(\tau)} x_s x_s^\top$, the matrix M_s is positive semidefinite and has rank at most one. Hence, if $\lambda_s(M_s)$ denotes its unique possibly nonzero eigenvalue, then

$$\mathrm{Tr}(M_s) = \lambda_s(M_s), \qquad \mathrm{Tr}(M_s^2) = \lambda_s(M_s)^2,$$

and therefore

$$\mathrm{Tr}(M_s)^2 = \mathrm{Tr}(M_s^2).$$

Therefore, we obtain

$$\begin{aligned}
\mathbb{E}\Big[\|W_{s+1}\|^4_{\nabla^2 \ell_s(\theta_{s+1})}\Big] &= 2\,\mathrm{Tr}(M_s^2) + \mathrm{Tr}(M_s)^2 \\
&= 3\,\mathrm{Tr}(M_s)^2 && (M_s \text{ is rank-one}) \\
&= 3\,\mathrm{Tr}\big(\Sigma_s \nabla^2 \ell_s(\theta_{s+1})\big)^2 && (\text{cyclicity of trace}) \\
&= 3\alpha^2\,\mathrm{Tr}\big(H_{s+1}^{-1} \nabla^2 \ell_s(\theta_{s+1})\big)^2 . && \text{(B.26)}
\end{aligned}$$

Substituting equation B.25 and equation B.26) into equation B.24, we obtain

$$\mathbb{E}_{Q_s}\big[D_{\ell_s}(\theta, \theta_{s+1})\big] \leqslant 2\alpha\ \mathrm{Tr}(H_{s+1}^{-1} \nabla^2 \ell_s(\theta_{s+1})). \tag{B.27}$$

Now write $H_{s+1} = A_s + \omega_s x_s x_s^\top$ and $u_s := x_s^\top A_s^{-1} x_s$. By the Sherman–Morrison formula,

$$\begin{aligned}
\mathrm{Tr}(H_{s+1}^{-1} \nabla^2 \ell_s(\theta_{s+1})) &= \omega_s x_s^\top H_{s+1}^{-1} x_s \\
&= \omega_s \left(x_s^\top A_s^{-1} x_s - \frac{\omega_s\, x_s^\top A_s^{-1} x_s\, x_s^\top A_s^{-1} x_s}{1 + \omega_s x_s^\top A_s^{-1} x_s} \right) \\
&= \omega_s \left(u_s - \frac{\omega_s u_s^2}{1 + \omega_s u_s} \right) = \frac{\omega_s u_s}{1 + \omega_s u_s}. && \text{(B.28)}
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
\log\det(H_{s+1}) - \log\det(A_s) &= \log\det(A_s + \omega_s x_s x_s^\top) - \log\det(A_s) \\
&= \log\Big(\det(A_s)\big(1 + \omega_s x_s^\top A_s^{-1} x_s\big)\Big) - \log\det(A_s) \\
&= \log(1 + \omega_s u_s). && \text{(B.29)}
\end{aligned}$$

Since $\log(1+y) \geqslant y/(1+y)$ for every $y > -1$, it follows that

$$\begin{aligned}
\mathrm{Tr}(H_{s+1}^{-1} \nabla^2 \ell_s(\theta_{s+1})) &= \frac{\omega_s u_s}{1 + \omega_s u_s} && (\text{Eqn. equation B.28}) \\
&\leqslant \log(1 + \omega_s u_s) && (\log(1+y) \geqslant y/(1+y)) \\
&= \log\det(H_{s+1}) - \log\det(A_s). && (\text{Eqn. equation B.29})
\end{aligned}$$

Combining this with equation B.27 completes the proof. □

Lemma B.8 (Variational formula for mix loss, Lemma 12 of Zhang et al. 2025)**.** *Let P be a probability distribution defined over $\mathbb{R}^d$ and Δ be the set of all measurable distributions. For any loss function $\ell : \mathbb{R}^d \to \mathbb{R}$, we have*

$$-\frac{1}{\alpha} \log\Big(\mathbb{E}_{\theta\sim P}\Big[e^{-\alpha \ell(\theta)}\Big]\Big) = \mathbb{E}_{\theta\sim P_\star}[\ell(\theta)] + \frac{1}{\alpha}\mathrm{KL}(P_\star \| P),$$

where $P_\star = \arg\min_{P' \in \Delta}\ \mathbb{E}_{\theta\sim P'}[\ell(\theta)] + \frac{1}{\alpha}\mathrm{KL}(P' \| P)$ is the optimal solution. Furthermore, for any distribution Q defined over $\mathbb{R}^d$, we have

$$-\frac{1}{\alpha} \log\Big(\mathbb{E}_{\theta\sim P}\Big[e^{-\alpha \ell(\theta)}\Big]\Big) = \mathbb{E}_{\theta\sim Q}[\ell(\theta)] + \frac{1}{\alpha}\mathrm{KL}(Q \| P) - \frac{1}{\alpha}\mathrm{KL}(Q \| P_\star).$$

Lemma B.9 (Kernel compression)**.** *For $t \geqslant 2$, define $\mathcal{V}_t := \sum_{p=1}^{t-1} \gamma^{(t-1-p)/2} \sqrt{\frac{1-\gamma^p}{1-\gamma}} \|\theta^\star_{p+1} - \theta^\star_p\|_2$. Assume that $\|x_s\|_2 \leqslant 1$ for all $s \in [T]$. Then, we have*

$$\sum_{s=1}^{t-1} \gamma^{t-1-s} \left(x_s^\top(\theta_t^\star - \theta_s^\star)\right)^2 \leqslant \mathcal{V}_t^2 \qquad \textit{and} \qquad \sum_{s=1}^{t-1} \gamma^{2(t-1-s)} \left(x_s^\top(\theta_t^\star - \theta_s^\star)\right)^2 \leqslant \mathcal{V}_t^2.$$

Proof of Lemma B.9. Fix $t \geqslant 2$. For simplicity, define $\zeta_{s,t} := x_s^\top(\theta_t^\star - \theta_s^\star) = x_s^\top \sum_{p=s}^{t-1} \big(\theta_{p+1}^\star - \theta_p^\star\big)$. Then, for each $s \in \{1, \dots, t-1\}$, we have

$$\zeta_{s,t} = x_s^\top \sum_{p=s}^{t-1} (\theta_{p+1}^\star - \theta_p^\star) = \sum_{p=1}^{t-1} \mathbb{1}\{s \leqslant p\}\, x_s^\top (\theta_{p+1}^\star - \theta_p^\star).$$

Define $u_t := \big(\gamma^{(t-1-s)/2}\zeta_{s,t}\big)_{s=1}^{t-1} \in \mathbb{R}^{t-1}$, and, for each $p \in \{1, \dots, t-1\}$,

$$v_{p,t} := \Big(\mathbb{1}\{s \leqslant p\}\, \gamma^{(t-1-s)/2}\, x_s^\top (\theta_{p+1}^\star - \theta_p^\star)\Big)_{s=1}^{t-1} \in \mathbb{R}^{t-1}.$$

Then, by the above decomposition of $\zeta_{s,t}$, we have

$$u_t = \sum_{p=1}^{t-1} v_{p,t}.$$

We now bound $\|v_{p,t}\|_2$. For each p, we get

$$\begin{aligned}
\|v_{p,t}\|_2^2 &= \sum_{s=1}^{t-1} \mathbb{1}\{s \leqslant p\}\, \gamma^{t-1-s} \big(x_s^\top (\theta_{p+1}^\star - \theta_p^\star)\big)^2 \\
&= \sum_{s=1}^{p} \gamma^{t-1-s} \big(x_s^\top (\theta_{p+1}^\star - \theta_p^\star)\big)^2 \\
&\leqslant \sum_{s=1}^{p} \gamma^{t-1-s}\, \|x_s\|_2^2\, \|\theta_{p+1}^\star - \theta_p^\star\|_2^2 \\
&\leqslant \|\theta_{p+1}^\star - \theta_p^\star\|_2^2 \sum_{s=1}^{p} \gamma^{t-1-s}. && (\|x_s\|_2^2 \leqslant 1)
\end{aligned}$$

Since

$$\sum_{s=1}^{p} \gamma^{t-1-s} = \gamma^{t-1-p} \sum_{r=0}^{p-1} \gamma^r = \gamma^{t-1-p} \frac{1-\gamma^p}{1-\gamma},$$

it follows that

$$\|v_{p,t}\|_2 \leqslant \gamma^{(t-1-p)/2} \sqrt{\frac{1-\gamma^p}{1-\gamma}}\, \|\theta_{p+1}^\star - \theta_p^\star\|_2.$$

Next, by the triangle inequality in $\mathbb{R}^{t-1}$, we obtain

$$\|u_t\|_2 = \left\| \sum_{p=1}^{t-1} v_{p,t} \right\|_2 \leqslant \sum_{p=1}^{t-1} \|v_{p,t}\|_2 \leqslant \sum_{p=1}^{t-1} \gamma^{(t-1-p)/2} \sqrt{\frac{1-\gamma^p}{1-\gamma}}\, \|\theta_{p+1}^\star - \theta_p^\star\|_2 = \mathcal{V}_t.$$

Therefore, by the definition of u_t, we get

$$\|u_t\|_2^2 = \sum_{s=1}^{t-1} \gamma^{t-1-s} \zeta_{s,t}^2 \leqslant \mathcal{V}_t^2,$$

which proves the first inequality.

It remains to prove the second inequality. Since $\gamma \in (0,1)$, for every $s \leqslant t-1$,

$$\gamma^{2(t-1-s)} \leqslant \gamma^{t-1-s}.$$

Therefore,

$$\sum_{s=1}^{t-1} \gamma^{2(t-1-s)} \zeta_{s,t}^2 \leqslant \sum_{s=1}^{t-1} \gamma^{t-1-s} \zeta_{s,t}^2 \leqslant \mathcal{V}_t^2.$$

This completes the proof. □

C Proof of Theorem 2

In this section, we provide the proof of Theorem 2.

C.1 Main Proof of Theorem 2

We begin by stating useful lemmas that will be used throughout the proof. First, define the high-probability event

$$\mathcal{E} := \bigcap_{t \geqslant 2} \{\theta_t^\star \in \mathcal{C}_t(\delta)\}. \tag{C.1}$$

By Theorem 1, this event satisfies

$$\Pr(\mathcal{E}) \geqslant 1 - \delta.$$

On $\mathcal{E}$, the confidence bound in parameter space directly translates into a corresponding prediction bound for the mean reward, which will be used repeatedly in the regret analysis.

Lemma C.1 (Prediction bound under the discounted-OMD confidence event)**.** *Under the same parameter choices as in Theorem 1, and under Assumptions 1 and 2, on the event $\mathcal{E}$, the following holds for every $t \geqslant 2$ and every feasible action $x \in \mathcal{X}_t$:*

$$\left|x^\top(\theta_t - \theta_t^\star)\right| \leqslant \beta_t(\delta)\|x\|_{H_t^{-1}} + \frac{\iota}{\sqrt{\lambda}}\mathcal{V}_t.$$

The proof is provided in Appendix C.2.1.

Lemma C.2 is a discounted version of the standard elliptical potential lemma, closely related to Lemma 8 of Faury et al. (2021); the main difference is that their analysis is based on $V_t = \lambda I_d + \sum_{s=1}^{t-1} \gamma^{t-1-s} x_s x_s^\top$, whereas ours uses a Hessian-based matrix $H_t = \lambda I_d + \sum_{s=1}^{t-1} \gamma^{t-1-s} \nabla^2 \ell_s(\theta_{s+1})$.

Lemma C.2 (Discounted elliptical potential)**.** *Assume $\lambda \geqslant c_\mu / g(\tau)$. Then for every $T \geqslant 1$, we have*

$$\sum_{t=1}^{T} \|x_t\|^2_{H_t^{-1}} \leqslant \frac{2g(\tau)}{c_\mu}\left[d\log\left(1 + \frac{k_\mu}{\lambda d\, g(\tau)(1-\gamma)}\right) + Td\log\frac{1}{\gamma}\right].$$

The proof is provided in Appendix C.2.2.

Lemma C.3 provides a bound on the cumulative drift kernel $\mathcal{V}_t$, similarly to prior work such as Wang et al. (2023). The geometric discounting suppresses the influence of past parameter changes, so the total contribution is controlled by the path length P_T.

Lemma C.3 (Summing the drift kernel)**.** *Let $\mathcal{V}_t := \sum_{p=1}^{t-1} \gamma^{(t-1-p)/2}\sqrt{\frac{1-\gamma^p}{1-\gamma}}\|\theta_{p+1}^\star - \theta_p^\star\|_2$. Then, for every horizon $T \geqslant 2$, we have*

$$\sum_{t=2}^{T} \mathcal{V}_t \leqslant \frac{1+\sqrt{\gamma}}{(1-\gamma)^{3/2}}\, P_T \leqslant \frac{2}{(1-\gamma)^{3/2}}\, P_T.$$

The proof is provided in Appendix C.2.3.

Now, we are ready to provide the proof of Theorem 2.

Proof of Theorem 2. Throughout the proof, we condition on the event $\mathcal{E} := \bigcap_{t\geqslant 2}\{\theta_t^\star \in \mathcal{C}_t(\delta)\}$, defined in equation C.1, which holds with probability at least $1-\delta$ by Theorem 1.

Let $R_t := \mu((x_t^\star)^\top\theta_t^\star) - \mu(x_t^\top\theta_t^\star)$ be the instantaneous pseudo-regret. Then, by Lemma C.1, with probability at least $1-\delta$, for every $x \in \mathcal{X}_t$, we have

$$\begin{aligned}
R_t &= \mu((x_t^\star)^\top\theta_t^\star) - \mu(x_t^\top\theta_t^\star) \\
&\leqslant k_\mu\left((x_t^\star - x_t)^\top\theta_t^\star\right) && (k_\mu\text{-Lipschitz continuity of } \mu) \\
&\leqslant k_\mu\left((x_t^\star)^\top\theta_t + \beta_t(\delta)\|x_t^\star\|_{H_t^{-1}} + \frac{\iota}{\sqrt{\lambda}}\mathcal{V}_t - \left(x_t^\top\theta_t - \beta_t(\delta)\|x_t\|_{H_t^{-1}} - \frac{\iota}{\sqrt{\lambda}}\mathcal{V}_t\right)\right) \\
& && \text{(Lemma C.1)} \\
&\leqslant 2k_\mu\beta_t(\delta)\|x_t\|_{H_t^{-1}} + 2\frac{k_\mu\iota}{\sqrt{\lambda}}\mathcal{V}_t. && \text{(action-selection rule in Eqn. equation 3)}
\end{aligned}$$

Summing over $t = 1, \dots, T$ yields

$$
\begin{aligned}
\mathbf{Reg}_T &\leqslant R_1 + 2k_\mu \sum_{t=2}^{T} \beta_t(\delta) \|x_t\|_{H_t^{-1}} + 2\frac{k_\mu \iota}{\sqrt{\lambda}} \sum_{t=2}^{T} \mathcal{V}_t \\
&\leqslant R_1 + 2k_\mu \beta_T(\delta) \sum_{t=2}^{T} \|x_t\|_{H_t^{-1}} + 2\frac{k_\mu \iota}{\sqrt{\lambda}} \sum_{t=2}^{T} \mathcal{V}_t && (\beta_t(\delta) \text{ is nondecreasing}) \\
&\leqslant R_1 + 2k_\mu \beta_T(\delta) \sqrt{T \sum_{t=2}^{T} \|x_t\|^2_{H_t^{-1}}} + 2\frac{k_\mu \iota}{\sqrt{\lambda}} \sum_{t=2}^{T} \mathcal{V}_t && (\text{Cauchy-Schwarz inequality}) \\
&\leqslant R_1 + 2k_\mu \beta_T(\delta) \sqrt{T \frac{2g(\tau)}{c_\mu} \left(d \log\left(1 + \frac{k_\mu}{\lambda d\, g(\tau)(1-\gamma)}\right) + Td \log \frac{1}{\gamma} \right)} + 2\frac{k_\mu \iota}{\sqrt{\lambda}} \sum_{t=2}^{T} \mathcal{V}_t \\
& && (\text{Lemma C.2}) \\
&\leqslant R_1 + 2k_\mu \beta_T(\delta) \sqrt{T \frac{2g(\tau)}{c_\mu} \left[d \log\left(1 + \frac{k_\mu}{\lambda d\, g(\tau)(1-\gamma)}\right) + Td \log \frac{1}{\gamma} \right]} \\
&\quad + \frac{4k_\mu \iota}{\sqrt{\lambda}} \frac{P_T}{(1-\gamma)^{3/2}} && (\text{Lemma C.3}) \\
&\leqslant 2k_\mu S + 2k_\mu \beta_T(\delta) \sqrt{T \frac{2g(\tau)}{c_\mu} \left[d \log\left(1 + \frac{k_\mu}{\lambda d\, g(\tau)(1-\gamma)}\right) + Td \log \frac{1}{\gamma} \right]} \\
&\quad + \frac{4k_\mu \iota}{\sqrt{\lambda}} \frac{P_T}{(1-\gamma)^{3/2}}, && (\text{C.2})
\end{aligned}
$$

where the last inequality holds because $R_1 \leqslant k_\mu(\|x_1^\star\|_2 + \|x_1\|_2)\|\theta_1^\star\|_2 \leqslant 2k_\mu S$.

By Theorem 1, under

$$
\lambda \geqslant \max\left\{ \frac{6\eta R k_\mu S}{g(\tau)}, \frac{32\alpha d R^2}{7}, \frac{c_\mu}{g(\tau)} \right\}, \quad \eta = 1 + RS, \quad \alpha = \frac{3\eta}{2}, \quad \iota = \sqrt{\frac{2\eta k_\mu}{g(\tau)}},
$$

we have

$$
\beta_T(\delta) = \tilde{\mathcal{O}}(\sqrt{d}) \qquad \text{and} \qquad \frac{k_\mu \iota}{\sqrt{\lambda}} = \tilde{\mathcal{O}}\left(\frac{k_\mu^{3/2}}{\sqrt{c_\mu}} \right). \tag{C.3}
$$

Moreover, since $\gamma \in [1/T, 1)$, we have

$$
\log \frac{1}{\gamma} = \begin{cases} \mathcal{O}(1-\gamma), & \text{if } \gamma \in [1/2, 1), \\ \mathcal{O}(\log T) = \tilde{\mathcal{O}}(1), & \text{if } \gamma \in [1/T, 1/2), \end{cases}
$$

and hence, in both cases,

$$
T\sqrt{\log(1/\gamma)} = \tilde{\mathcal{O}}\big(T\sqrt{1-\gamma}\big).
$$

Therefore, we have

$$
\sqrt{\frac{2g(\tau)T}{c_\mu} \left[d \log\left(1 + \frac{k_\mu}{\lambda d\, g(\tau)(1-\gamma)}\right) + Td \log \frac{1}{\gamma} \right]} = \tilde{\mathcal{O}}\left(\sqrt{\frac{Td}{c_\mu}} + T\sqrt{\frac{d(1-\gamma)}{c_\mu}} \right). \tag{C.4}
$$

Plugging equation C.3 and equation C.4 into equation C.2, we obtain

$$
\mathbf{Reg}_T = \tilde{\mathcal{O}}\left(\frac{k_\mu}{\sqrt{c_\mu}}\, d\sqrt{T} + \frac{k_\mu}{\sqrt{c_\mu}}\, dT\sqrt{1-\gamma} + \frac{k_\mu^{3/2}}{\sqrt{c_\mu}} \frac{P_T}{(1-\gamma)^{3/2}} \right).
$$

This proves the first bound.

Now we prove the second bound. For the tuned choice, let

$$
\gamma = 1 - \min\left\{ 1 - \frac{1}{T},\ \max\left\{ \frac{1}{T},\ \sqrt{\frac{\sqrt{k_\mu} P_T}{dT}} \right\} \right\}.
$$

Then $\gamma \in [1/T, 1)$, and therefore the first bound applies.

Case 1. If $\frac{d}{\sqrt{k_\mu}\,T} \leqslant P_T \leqslant \frac{dT}{\sqrt{k_\mu}}\left(1-\frac{1}{T}\right)^2$, then $1-\gamma = \sqrt{\frac{\sqrt{k_\mu}P_T}{dT}}$. This implies that

$$\frac{k_\mu}{\sqrt{c_\mu}}\,dT\sqrt{1-\gamma} = \tilde{\mathcal{O}}\left(\frac{k_\mu^{9/8}}{\sqrt{c_\mu}}\,d^{3/4}P_T^{1/4}T^{3/4}\right),$$

and

$$\frac{k_\mu^{3/2}}{\sqrt{c_\mu}}\frac{P_T}{(1-\gamma)^{3/2}} = \tilde{\mathcal{O}}\left(\frac{k_\mu^{9/8}}{\sqrt{c_\mu}}\,d^{3/4}P_T^{1/4}T^{3/4}\right).$$

Moreover, under $P_T \geqslant \frac{d}{\sqrt{k_\mu}\,T}$, the first term $\frac{k_\mu}{\sqrt{c_\mu}}\,d\sqrt{T}$ is dominated by the same quantity. Therefore, we get

$$\mathbf{Reg}_T = \tilde{\mathcal{O}}\left(\frac{k_\mu^{9/8}}{\sqrt{c_\mu}}\,d^{3/4}P_T^{1/4}T^{3/4}\right).$$

Case 2. If $0 \leqslant P_T < \frac{d}{\sqrt{k_\mu}\,T}$, then $1-\gamma = \frac{1}{T}$. In this case, we have

$$\frac{k_\mu}{\sqrt{c_\mu}}\,dT\sqrt{1-\gamma} = \tilde{\mathcal{O}}\left(\frac{k_\mu}{\sqrt{c_\mu}}\,d\sqrt{T}\right),$$

and

$$\frac{k_\mu^{3/2}}{\sqrt{c_\mu}}\frac{P_T}{(1-\gamma)^{3/2}} = \tilde{\mathcal{O}}\left(\frac{k_\mu^{3/2}}{\sqrt{c_\mu}}\,P_T T^{3/2}\right) < \tilde{\mathcal{O}}\left(\frac{k_\mu}{\sqrt{c_\mu}}\,d\sqrt{T}\right).$$

Hence, we obtain

$$\mathbf{Reg}_T \leqslant \tilde{\mathcal{O}}\left(\frac{k_\mu}{\sqrt{c_\mu}}\,d\sqrt{T}\right).$$

Case 3. If $P_T > \frac{dT}{\sqrt{k_\mu}}\left(1-\frac{1}{T}\right)^2$, then $1-\gamma = 1-\frac{1}{T}$, that is, $\gamma = \frac{1}{T}$. In this case, we have

$$\frac{k_\mu}{\sqrt{c_\mu}}\,dT\sqrt{1-\gamma} = \tilde{\mathcal{O}}\left(\frac{k_\mu}{\sqrt{c_\mu}}\,dT\right),$$

and

$$\frac{k_\mu^{3/2}}{\sqrt{c_\mu}}\frac{P_T}{(1-\gamma)^{3/2}} = \tilde{\mathcal{O}}\left(\frac{k_\mu^{3/2}}{\sqrt{c_\mu}}\,P_T\right).$$

Moreover, the first term $\frac{k_\mu}{\sqrt{c_\mu}}\,d\sqrt{T}$ is dominated by $\frac{k_\mu}{\sqrt{c_\mu}}\,dT$. Since $T \geqslant 2$, we have $\left(1-\frac{1}{T}\right)^2 \geqslant \frac{1}{4}$, and thus

$$P_T > \frac{dT}{\sqrt{k_\mu}}\left(1-\frac{1}{T}\right)^2 \geqslant \frac{dT}{4\sqrt{k_\mu}}.$$

Therefore, the first term can be bounded as

$$\frac{k_\mu}{\sqrt{c_\mu}}\,d\sqrt{T} \leqslant \frac{k_\mu}{\sqrt{c_\mu}}\,dT \leqslant 4\frac{k_\mu^{3/2}}{\sqrt{c_\mu}}\,P_T.$$

Hence, all three terms are dominated by $\tilde{\mathcal{O}}\left(\frac{k_\mu^{3/2}}{\sqrt{c_\mu}}\,P_T\right)$, and we obtain

$$\mathbf{Reg}_T \leqslant \tilde{\mathcal{O}}\left(\frac{k_\mu^{3/2}}{\sqrt{c_\mu}}\,P_T\right).$$

This completes the proof of Theorem 2. □

Remark C.1 (On the clipped choice of γ)**.** *The clipped definition of γ is necessary because our generic regret bound is proved only for admissible $\gamma \in [1/T, 1)$, whereas the unconstrained minimizer of the simplified upper bound need not belong to this interval. Earlier works often stated only the interior regimes and implicitly assumed that the resulting choice of γ was feasible. For example, Faury et al. (2021) effectively restrict to a sublinear-P_T regime in the proof, which allows them to exclude the large-P_T case and thus ensures that the tuned choice $\gamma = 1 - \left(\frac{P_T}{dT}\right)^{2/3}$ remains admissible. Similarly, Wang et al. (2023) first derives generic bounds for admissible $\gamma \in (1/T, 1)$, but then states tuned two-case corollaries without an explicit feasibility check. A similar issue also appears in Russac et al. (2020; 2021) for the piecewise-stationary setting, where the tuning implicitly relies on a sublinear-Γ_T regime so that the proposed choice $\gamma \approx 1 - \left(\frac{\Gamma_T}{dT}\right)^{2/3}$ remains admissible. By contrast, our three-case bound is fully rigorous: the clipping step enforces admissibility explicitly, so the resulting bound applies uniformly over all parameter regimes. If the same feasibility check were made explicit in the earlier analyses, their tuned bounds would likewise take an analogous multi-case form, rather than covering only the interior one- or two-case regimes.*

C.2 Proofs of Lemmas for Theorem 2

C.2.1 Proof of Lemma C.1

Proof of Lemma C.1. Fix any $t \geqslant 2$ and any feasible action $x \in \mathcal{X}_t$. On the event $\mathcal{E}$, we have $\theta_t^\star \in \mathcal{C}_t(\delta)$, and hence, by the definition of $\mathcal{C}_t(\delta)$, we have

$$\|\theta_t^\star - \theta_t\|_{H_t} \leqslant \beta_t(\delta) + \iota \mathcal{V}_t.$$

Therefore, by Cauchy–Schwarz inequality,

$$\begin{aligned} |x^\top(\theta_t^\star - \theta_t)| &\leqslant \|x\|_{H_t^{-1}} \|\theta_t^\star - \theta_t\|_{H_t} \\ &\leqslant \beta_t(\delta)\|x\|_{H_t^{-1}} + \iota \mathcal{V}_t \|x\|_{H_t^{-1}} \\ &\leqslant \beta_t(\delta)\|x\|_{H_t^{-1}} + \frac{\iota}{\sqrt{\lambda}} \mathcal{V}_t, \end{aligned}$$

where the last inequality holds because

$$\|x\|_{H_t^{-1}} \leqslant \frac{\|x\|_2}{\sqrt{\lambda}} \leqslant \frac{1}{\sqrt{\lambda}}. \qquad \text{(Assumption 1)}$$

This completes the proof. □

C.2.2 Proof of Lemma C.2

Proof of Lemma C.2. Since $A_t = \gamma H_t + (1-\gamma)\lambda I_d \preceq \gamma H_t + (1-\gamma) H_t = H_t$, inverse monotonicity on positive definite matrices gives $A_t^{-1} \succeq H_t^{-1}$, which implies $\|x_t\|^2_{A_t^{-1}} \geqslant \|x_t\|^2_{H_t^{-1}}$.

Now, we write

$$\nabla^2 \ell_t(\theta_{t+1}) = \frac{\mu'(x_t^\top \theta_{t+1})}{g(\tau)} x_t x_t^\top =: h_t x_t x_t^\top,$$

where Assumption 2 ensures that $\frac{c_\mu}{g(\tau)} \leqslant h_t \leqslant \frac{k_\mu}{g(\tau)}$.

Then, we have

$$\begin{aligned} \log\det(H_{t+1}) - \log\det(A_t) &= \log\det(A_t + h_t x_t x_t^\top) - \log\det(A_t) \\ &= \log\det\left(I_d + h_t A_t^{-1/2} x_t x_t^\top A_t^{-1/2}\right) \\ &= \log\left(1 + h_t \|x_t\|^2_{A_t^{-1}}\right) \\ &\geqslant \log\left(1 + \frac{c_\mu}{g(\tau)} \|x_t\|^2_{H_t^{-1}}\right). \\ &\qquad (h_t \geqslant c_\mu/g(\tau) \text{ and } \|x_t\|^2_{A_t^{-1}} \geqslant \|x_t\|^2_{H_t^{-1}}) \end{aligned}$$

Since $\|x_t\|^2_{H_t^{-1}} \leqslant 1/\lambda$ and $\lambda \geqslant c_\mu/g(\tau)$, we have $0 \leqslant (c_\mu/g(\tau))\|x_t\|^2_{H_t^{-1}} \leqslant 1$. Using $\log(1+u) \geqslant u/2$ for $u \in [0,1]$, we get

$$\log\det(H_{t+1}) - \log\det(A_t) \geqslant \frac{c_\mu}{2g(\tau)}\|x_t\|^2_{H_t^{-1}},$$

which implies

$$\sum_{t=1}^{T} \|x_t\|^2_{H_t^{-1}} \leqslant \frac{2g(\tau)}{c_\mu} \sum_{t=1}^{T} \big(\log\det(H_{t+1}) - \log\det(A_t)\big). \tag{C.5}$$

Because $A_t = \gamma H_t + (1-\gamma)\lambda I_d \succeq \gamma H_t$, monotonicity of $\log\det$ gives

$$\log\det(A_t) \geqslant \log\det(\gamma H_t) = d\log\gamma + \log\det(H_t).$$

Hence, we get

$$\log\det(H_{t+1}) - \log\det(A_t) \leqslant \log\det(H_{t+1}) - \log\det(H_t) - d\log\gamma.$$

Summing from $t=1$ to T yields

$$\begin{aligned}\sum_{t=1}^{T} \big(\log\det(H_{t+1}) - \log\det(A_t)\big) &\leqslant \log\det(H_{T+1}) - \log\det(H_1) + Td\log\frac{1}{\gamma} \\ &= \log\det(H_{T+1}) - d\log\lambda + Td\log\frac{1}{\gamma}. \qquad (H_1 = \lambda I_d)\end{aligned}$$

Also, we have

$$H_{T+1} = \lambda I_d + \sum_{s=1}^{T} \gamma^{T-s}\frac{\mu'(x_s^\top\theta_{s+1})}{g(\tau)} x_s x_s^\top \preceq \lambda I_d + \frac{k_\mu}{g(\tau)}\sum_{s=1}^{T}\gamma^{T-s} x_s x_s^\top.$$

Using $\|x_s\|_2 \leqslant 1$, we get

$$\mathrm{Tr}(H_{T+1}) \leqslant \lambda d + \frac{k_\mu}{g(\tau)}\sum_{s=1}^{T}\gamma^{T-s} \leqslant \lambda d + \frac{k_\mu}{g(\tau)(1-\gamma)}.$$

By AM–GM, we obtain

$$\det(H_{T+1}) \leqslant \left(\frac{\mathrm{Tr}(H_{T+1})}{d}\right)^d \leqslant \left(\lambda + \frac{k_\mu}{d\,g(\tau)(1-\gamma)}\right)^d.$$

which implies.

$$\begin{aligned}\sum_{t=1}^{T} \big(\log\det(H_{t+1}) - \log\det(A_t)\big) &\leqslant \log\det(H_{T+1}) - d\log\lambda + Td\log\frac{1}{\gamma} \\ &\leqslant d\log\left(1 + \frac{k_\mu}{\lambda d\, g(\tau)(1-\gamma)}\right) + Td\log\frac{1}{\gamma}. \end{aligned} \tag{C.6}$$

Substituting equation C.6 into equation C.5 completes the proof. □

C.2.3 Proof of Lemma C.3

Proof of Lemma C.3. By the definition of $\mathcal{V}_t$, we have

$$\begin{aligned}
\sum_{t=2}^{T} \mathcal{V}_t &\leqslant \sum_{t=2}^{T} \sum_{p=1}^{t-1} \frac{\gamma^{(t-1-p)/2}}{\sqrt{1-\gamma}} \|\theta^\star_{p+1} - \theta^\star_p\|_2 \\
&= \sum_{p=1}^{T-1} \|\theta^\star_{p+1} - \theta^\star_p\|_2 \sum_{t=p+1}^{T} \frac{\gamma^{(t-1-p)/2}}{\sqrt{1-\gamma}} \\
&= \frac{1}{\sqrt{1-\gamma}} \sum_{p=1}^{T-1} \|\theta^\star_{p+1} - \theta^\star_p\|_2 \sum_{k=0}^{T-1-p} \gamma^{k/2} \\
&\leqslant \frac{1}{\sqrt{1-\gamma}} \sum_{p=1}^{T-1} \|\theta^\star_{p+1} - \theta^\star_p\|_2 \cdot \frac{1}{1-\sqrt{\gamma}} \\
&= \frac{1}{\sqrt{1-\gamma}(1-\sqrt{\gamma})} P_T \\
&= \frac{1+\sqrt{\gamma}}{(1-\gamma)^{3/2}} P_T, \qquad (1-\sqrt{\gamma} = \tfrac{1-\gamma}{1+\sqrt{\gamma}})
\end{aligned}$$

which proves the first inequality. Since $1 + \sqrt{\gamma} \leqslant 2$, the second inequality follows. □

D Proof of Theorem 3

In this section, we prove Theorem 3. The overall argument closely follows that of Appendix C; the only additional ingredient is a sharper bound on $\sum_t \mathcal{V}_t$ over rounds that are sufficiently far from the most recent change point.

D.1 Main Proof of Theorem 3

As in the proof of Theorem 2, we condition on the good event

$$\mathcal{E} := \bigcap_{t \geqslant 2} \{\theta^\star_t \in \mathcal{C}_t(\delta)\},$$

which, by Theorem 1, holds with probability at least $\Pr(\mathcal{E}) \geqslant 1 - \delta$. We further use Lemma C.1 and Lemma C.2 from Appendix C.

For each integer $D \geqslant 1$, let $\mathcal{G}_D$ denote the set of rounds $t \in [T]$ such that the environment parameter has not changed during the previous D rounds. More precisely,

$$\mathcal{G}_D := \big\{t \in [T] : \theta^\star_s = \theta^\star_t \text{ for all } s \in \{\max\{t-D, 1\}, \dots, t-1\}\big\}.$$

Equivalently, $t \in \mathcal{G}_D$ if the parameter value at round t coincides with its value at every round in the most recent D-step history prior to t (or from round 1 onward when $t \leqslant D$).

We first quantify how many rounds fail to belong to $\mathcal{G}_D$. Under piecewise stationarity, a round is excluded from $\mathcal{G}_D$ only if a changepoint occurred within its most recent D-step history. Therefore, each changepoint can affect only a length-D window of subsequent rounds, which yields the following bound on the number of such "bad" rounds.

Lemma D.1 (Bad-round count under piecewise stationarity)**.** *For every integer $D \geqslant 1$, we have*

$$|[T] \backslash \mathcal{G}_D| \leqslant \Gamma_T D.$$

The proof is deferred to Appendix D.2.1.

We next control the cumulative drift term over the complementary set of "good" rounds $\mathcal{G}_D$. On these rounds, the parameter has remained unchanged throughout the most recent D rounds, so any contribution to the drift quantity $\mathcal{V}_t$ must come from changes that occurred strictly earlier. Because the discounted kernel downweights older rounds geometrically, those remote changes are suppressed by a factor of order $\gamma^{D/2}$, leading to the following estimate.

Lemma D.2 (Piecewise drift-kernel sum)**.** *Under the piecewise-stationary assumption, for every integer $D \geqslant 1$,*

$$\sum_{t\in\mathcal{G}_D\cap\{2,\dots,T\}} \mathcal{V}_t \leqslant \frac{2\gamma^{D/2}}{(1-\gamma)^{3/2}}\, P_T \leqslant \frac{4S\Gamma_T\,\gamma^{D/2}}{(1-\gamma)^{3/2}}.$$

The proof is deferred to Appendix D.2.2.

Now, we are ready to provide the proof of Theorem 3.

Proof of Theorem 3. We work on the event $\mathcal{E} := \bigcap_{t\geqslant 2}\{\theta_t^\star \in \mathcal{C}_t(\delta)\}$, defined in equation C.1, which satisfies $\Pr(\mathcal{E}) \geqslant 1-\delta$ by Theorem 1.

Let $R_t := \mu((x_t^\star)^\top\theta_t^\star) - \mu(x_t^\top\theta_t^\star)$ and $x_t^\star \in \arg\max_{x\in\mathcal{X}_t}\mu(x^\top\theta_t^\star)$, so that $\mathbf{Reg}_T = \sum_{t=1}^T R_t$. Fix any integer $D \geqslant 1$. We decompose the regret as

$$\mathbf{Reg}_T = R_1 + \sum_{t\in\mathcal{G}_D\cap\{2,\dots,T\}} R_t + \sum_{t\in([T]\setminus\mathcal{G}_D)\cap\{2,\dots,T\}} R_t.$$

For every $t \geqslant 2$, the same argument as in the proof of Theorem 2, using Lemma C.1 and our action-selection rule in equation 3 (i.e., UCB optimality of x_t), yields

$$R_t \leqslant 2k_\mu\beta_t(\delta)\|x_t\|_{H_t^{-1}} + 2\frac{k_\mu\iota}{\sqrt{\lambda}}\mathcal{V}_t.$$

Hence, we have

$$\begin{aligned}
\sum_{t\in\mathcal{G}_D\cap\{2,\dots,T\}} R_t &\leqslant 2k_\mu \sum_{t\in\mathcal{G}_D\cap\{2,\dots,T\}} \beta_t(\delta)\|x_t\|_{H_t^{-1}} + 2\frac{k_\mu\iota}{\sqrt{\lambda}} \sum_{t\in\mathcal{G}_D\cap\{2,\dots,T\}} \mathcal{V}_t \\
&\leqslant 2k_\mu\beta_T(\delta)\sum_{t=2}^T \|x_t\|_{H_t^{-1}} + 2\frac{k_\mu\iota}{\sqrt{\lambda}} \sum_{t\in\mathcal{G}_D\cap\{2,\dots,T\}} \mathcal{V}_t \\
&\leqslant 2k_\mu\beta_T(\delta)\sqrt{T\sum_{t=2}^T \|x_t\|^2_{H_t^{-1}}} + 2\frac{k_\mu\iota}{\sqrt{\lambda}} \sum_{t\in\mathcal{G}_D\cap\{2,\dots,T\}} \mathcal{V}_t \\
&\leqslant 2k_\mu\beta_T(\delta)\sqrt{\frac{2g(\tau)T}{c_\mu}\left[d\log\left(1+\frac{k_\mu}{\lambda d\, g(\tau)(1-\gamma)}\right) + Td\log\frac{1}{\gamma}\right]} \\
&\quad + \frac{8Sk_\mu\iota}{\sqrt{\lambda}}\frac{\Gamma_T\,\gamma^{D/2}}{(1-\gamma)^{3/2}}. \qquad \text{(Lemma C.2 and Lemma D.2)}
\end{aligned}$$

On the other hand, by Assumption 1, every conditional mean reward belongs to $[0, R]$, and hence $0 \leqslant R_t \leqslant R$, for all $t \in [T]$. Therefore, we get

$$\begin{aligned}
R_1 + \sum_{t\in([T]\setminus\mathcal{G}_D)\cap\{2,\dots,T\}} R_t &\leqslant R\big(1 + |[T]\setminus\mathcal{G}_D|\big) \\
&\leqslant R(\Gamma_T D + 1). \qquad \text{(Lemma D.1)}
\end{aligned}$$

Putting the above two bounds together, we obtain that, for every integer $D \geqslant 1$,

$$\begin{aligned}
\mathbf{Reg}_T &\leqslant R(\Gamma_T D + 1) + 2k_\mu\beta_T(\delta)\sqrt{\frac{2g(\tau)T}{c_\mu}\left[d\log\left(1+\frac{k_\mu}{\lambda d\, g(\tau)(1-\gamma)}\right) + Td\log\frac{1}{\gamma}\right]} \\
&\quad + \frac{8Sk_\mu\iota}{\sqrt{\lambda}}\frac{\Gamma_T\,\gamma^{D/2}}{(1-\gamma)^{3/2}}. \qquad \text{(D.1)}
\end{aligned}$$

Now we define $C_{\mathrm{pw}} := \max\left\{e,\ \frac{8Sk_\mu\iota}{\sqrt{\lambda}}\right\}$ and $D_\gamma := \left\lceil \frac{2\log\left(C_{\mathrm{pw}}T/(1-\gamma)\right)}{\log(1/\gamma)} \right\rceil$. Then, we have

$$\frac{D_\gamma}{2}\log\frac{1}{\gamma} \geqslant \log\frac{C_{\mathrm{pw}}T}{1-\gamma}.$$

Equivalently,

$$\gamma^{D_\gamma/2} = \exp\left(-\frac{D_\gamma}{2}\log\frac{1}{\gamma}\right) \leqslant \frac{1-\gamma}{C_{\mathrm{pw}}T} \leqslant \frac{(1-\gamma)\sqrt{\lambda}}{8Sk_\mu\iota T}.$$

Moreover, since $\gamma \in [1/T, 1)$, we have

$$\log\frac{1}{\gamma} = \begin{cases}\Theta(1-\gamma), & \text{if } \gamma \in [1/2, 1),\\ \Omega(1), & \text{if } \gamma \in [1/T, 1/2),\end{cases}$$

and therefore

$$D_\gamma = \left\lceil \frac{2\log\big(C_{\mathrm{pw}}T/(1-\gamma)\big)}{\log(1/\gamma)} \right\rceil = \begin{cases}\tilde{\mathcal{O}}\left(\dfrac{1}{1-\gamma}\right), & \text{if } \gamma \in [1/2, 1),\\ \mathcal{O}(\log T) = \tilde{\mathcal{O}}\left(\dfrac{1}{1-\gamma}\right), & \text{if } \gamma \in [1/T, 1/2).\end{cases}$$

Hence, in both cases,

$$D_\gamma = \tilde{\mathcal{O}}\left(\frac{1}{1-\gamma}\right).$$

Hence, substituting $D = D_\gamma$ into equation D.1 yields

$$\mathbf{Reg}_T \leqslant R(\Gamma_T D_\gamma + 1) + 2k_\mu\beta_T(\delta)\sqrt{\frac{2g(\tau)T}{c_\mu}\left[d\log\left(1 + \frac{k_\mu}{\lambda d\, g(\tau)(1-\gamma)}\right) + Td\log\frac{1}{\gamma}\right]} + \frac{8Sk_\mu\iota}{\sqrt{\lambda}}\frac{\Gamma_T\,\gamma^{D_\gamma/2}}{(1-\gamma)^{3/2}}. \tag{D.2}$$

The middle term is handled exactly as in the proof of Theorem 2:

$$2k_\mu\beta_T(\delta)\sqrt{\frac{2g(\tau)T}{c_\mu}\left[d\log\left(1 + \frac{k_\mu}{\lambda d\, g(\tau)(1-\gamma)}\right) + Td\log\frac{1}{\gamma}\right]} = \tilde{\mathcal{O}}\left(\frac{k_\mu}{\sqrt{c_\mu}}\, d\sqrt{T} + \frac{k_\mu}{\sqrt{c_\mu}}\, dT\sqrt{1-\gamma}\right). \tag{D.3}$$

Also, we have

$$R(\Gamma_T D_\gamma + 1) = \tilde{\mathcal{O}}\left(1 + \frac{\Gamma_T}{1-\gamma}\right), \tag{D.4}$$

and, using $\gamma^{D_\gamma/2} \leqslant \frac{(1-\gamma)\sqrt{\lambda}}{8Sk_\mu\iota T}$, we get

$$\frac{8Sk_\mu\iota}{\sqrt{\lambda}}\frac{\Gamma_T\,\gamma^{D_\gamma/2}}{(1-\gamma)^{3/2}} \leqslant \frac{\Gamma_T}{T\sqrt{1-\gamma}} \leqslant \tilde{\mathcal{O}}\left(\frac{\Gamma_T}{1-\gamma}\right). \tag{D.5}$$

Plugging equation D.3, equation D.4, and equation D.5 into equation D.2, we obtain

$$\mathbf{Reg}_T \leqslant \tilde{\mathcal{O}}\left(\frac{k_\mu}{\sqrt{c_\mu}}\, d\sqrt{T} + \frac{k_\mu}{\sqrt{c_\mu}}\, dT\sqrt{1-\gamma} + \frac{\Gamma_T}{1-\gamma}\right).$$

This proves the first bound.

Now we prove the second bound. For the clipped tuned choice, let

$$\gamma = 1 - \min\left\{1 - \frac{1}{T},\ \max\left\{\frac{1}{T},\ \left(\frac{\Gamma_T\sqrt{c_\mu}}{k_\mu dT}\right)^{2/3}\right\}\right\}.$$

By construction, we have $\gamma \in [1/T, 1)$.

Case 1. If $\frac{k_\mu d}{\sqrt{c_\mu T}} \leqslant \Gamma_T \leqslant \frac{k_\mu dT}{\sqrt{c_\mu}} \left(1 - \frac{1}{T}\right)^{3/2}$, then $\frac{1}{T} \leqslant \left(\frac{\Gamma_T \sqrt{c_\mu}}{k_\mu dT}\right)^{2/3} \leqslant 1 - \frac{1}{T}$, and therefore $1 - \gamma = \left(\frac{\Gamma_T \sqrt{c_\mu}}{k_\mu dT}\right)^{2/3}$. This implies that

$$\frac{k_\mu}{\sqrt{c_\mu}} dT\sqrt{1-\gamma} = \tilde{\mathcal{O}}\left(\frac{k_\mu^{2/3}}{c_\mu^{1/3}} d^{2/3}\Gamma_T^{1/3}T^{2/3}\right),$$

and

$$\frac{\Gamma_T}{1-\gamma} = \tilde{\mathcal{O}}\left(\frac{k_\mu^{2/3}}{c_\mu^{1/3}} d^{2/3}\Gamma_T^{1/3}T^{2/3}\right).$$

Moreover, under $\Gamma_T \geqslant \frac{k_\mu d}{\sqrt{c_\mu T}}$, the first term $\frac{k_\mu}{\sqrt{c_\mu}} d\sqrt{T}$ is dominated by the same quantity. Therefore, we get

$$\mathbf{Reg}_T = \tilde{\mathcal{O}}\left(\frac{k_\mu^{2/3}}{c_\mu^{1/3}} d^{2/3}\Gamma_T^{1/3}T^{2/3}\right).$$

Case 2. If $0 \leqslant \Gamma_T < \frac{k_\mu d}{\sqrt{c_\mu T}}$, then $\left(\frac{\Gamma_T \sqrt{c_\mu}}{k_\mu dT}\right)^{2/3} < \frac{1}{T}$, and hence $1 - \gamma = \frac{1}{T}$. In this case, we have

$$\frac{k_\mu}{\sqrt{c_\mu}} dT\sqrt{1-\gamma} = \tilde{\mathcal{O}}\left(\frac{k_\mu}{\sqrt{c_\mu}} d\sqrt{T}\right),$$

and

$$\frac{\Gamma_T}{1-\gamma} = \Gamma_T T < \frac{k_\mu}{\sqrt{c_\mu}} d\sqrt{T}.$$

Hence, we obtain

$$\mathbf{Reg}_T \leqslant \tilde{\mathcal{O}}\left(\frac{k_\mu}{\sqrt{c_\mu}} d\sqrt{T}\right).$$

Case 3. If $\Gamma_T > \frac{k_\mu dT}{\sqrt{c_\mu}} \left(1 - \frac{1}{T}\right)^{3/2}$, then $\left(\frac{\Gamma_T \sqrt{c_\mu}}{k_\mu dT}\right)^{2/3} > 1 - \frac{1}{T}$, and thus $1 - \gamma = 1 - \frac{1}{T}$, that is $\gamma = \frac{1}{T}$. In this case, we have

$$\frac{k_\mu}{\sqrt{c_\mu}} dT\sqrt{1-\gamma} = \tilde{\mathcal{O}}\left(\frac{k_\mu}{\sqrt{c_\mu}} dT\right),$$

and

$$\frac{\Gamma_T}{1-\gamma} = \tilde{\mathcal{O}}(\Gamma_T).$$

Moreover, the first term $\frac{k_\mu}{\sqrt{c_\mu}} d\sqrt{T}$ is dominated by $\frac{k_\mu}{\sqrt{c_\mu}} dT$. Since $T \geqslant 2$, we have $\left(1 - \frac{1}{T}\right)^{3/2} \geqslant \frac{1}{2\sqrt{2}}$. Therefore,

$$\Gamma_T > \frac{k_\mu dT}{\sqrt{c_\mu}} \left(1 - \frac{1}{T}\right)^{3/2} \geqslant \frac{1}{2\sqrt{2}} \frac{k_\mu}{\sqrt{c_\mu}} dT,$$

which implies

$$\frac{k_\mu}{\sqrt{c_\mu}} dT \leqslant 2\sqrt{2}\,\Gamma_T.$$

Hence, all three terms are dominated by

$$\tilde{\mathcal{O}}(\Gamma_T),$$

and we obtain

$$\mathbf{Reg}_T \leqslant \tilde{\mathcal{O}}(\Gamma_T).$$

This completes the proof of Theorem 3. □

D.2 Proofs of Lemmas for Theorem 3

D.2.1 Proof of Lemma D.1

Proof of Lemma D.1. Let $\nu_1, \dots, \nu_{\Gamma_T}$ be the change points, that is, the rounds at which the parameter value changes. Fix any round $t \in [T] \backslash \mathcal{G}_D$. By the definition of $\mathcal{G}_D$, there exists $s \in \big\{\max\{t - D, 1\}, \dots, t-1\big\}$ such that

$$\theta_s^\star \neq \theta_t^\star.$$

Since the sequence $\{\theta_u^\star\}_{u=1}^T$ is piecewise constant, the fact that $\theta_s^\star \neq \theta_t^\star$ implies that the parameter must change at least once between rounds s and t. Hence, there exists a change point ν_j, for some $j \in [\Gamma_T]$, such that $\nu_j \in \{s+1, \dots, t\}$. Because $s \geqslant (t-D) \vee 1 = \max\{t - D, 1\}$, we have $s + 1 \geqslant \max\{t - D, 1\} + 1$. Therefore, $\nu_j \in \{s+1, \dots, t\} \subseteq \{\max\{t - D, 1\} + 1, \dots, t\}$. That is,

$$\max\{t - D, 1\} + 1 \leqslant \nu_j \leqslant t.$$

From $\nu_j \geqslant \max\{t - D, 1\} + 1$, we obtain $\nu_j \geqslant t - D + 1$, because $\max\{t - D, 1\} \geqslant t - D$. Rearranging this inequality gives

$$t \leqslant \nu_j + D - 1.$$

Combining this with $\nu_j \leqslant t$, we obtain

$$\nu_j \leqslant t \leqslant \nu_j + D - 1.$$

Since also $t \leqslant T$, it follows that

$$t \in \{\nu_j, \nu_j + 1, \dots, \min\{\nu_j + D - 1, T\}\}.$$

Because the choice of $t \in [T] \backslash \mathcal{G}_D$ was arbitrary, we conclude that every bad round belongs to at least one such interval. Therefore, we get

$$[T] \backslash \mathcal{G}_D \subseteq \bigcup_{j=1}^{\Gamma_T} \{\nu_j, \nu_j + 1, \dots, \min\{\nu_j + D - 1, T\}\}.$$

For each $j \in [\Gamma_T]$, the set $\{\nu_j, \nu_j + 1, \dots, \min\{\nu_j + D - 1, T\}\}$ contains at most D integers. Indeed, if $\nu_j + D - 1 \leqslant T$, then its cardinality is exactly $(\nu_j + D - 1) - \nu_j + 1 = D$, whereas if $\nu_j + D - 1 > T$, then truncation at T only makes the set smaller. Hence,

$$\Big|\{\nu_j, \nu_j + 1, \dots, \min\{\nu_j + D - 1, T\}\}\Big| \leqslant D \qquad \text{for all } j \in [\Gamma_T].$$

Taking cardinalities and using the union bound for finite sets, we obtain

$$|[T] \backslash \mathcal{G}_D| \leqslant \sum_{j=1}^{\Gamma_T} \Big|\{\nu_j, \nu_j + 1, \dots, \min\{\nu_j + D - 1, T\}\}\Big| \leqslant \sum_{j=1}^{\Gamma_T} D = \Gamma_T D.$$

This completes the proof. □

D.2.2 Proof of Lemma D.2

Proof of Lemma D.2. Fix any $t \in \mathcal{G}_D \cap \{2, \dots, T\}$. By definition,

$$\begin{aligned}
\mathcal{V}_t &= \sum_{p=1}^{t-1} \gamma^{(t-1-p)/2} \sqrt{\frac{1-\gamma^p}{1-\gamma}} \, \|\theta_{p+1}^\star - \theta_p^\star\|_2 \\
&\leqslant \sum_{p=1}^{t-1} \frac{\gamma^{(t-1-p)/2}}{\sqrt{1-\gamma}} \, \|\theta_{p+1}^\star - \theta_p^\star\|_2 \qquad (1 - \gamma^p \leqslant 1) \\
&\leqslant \sum_{p=1}^{t-D-1} \frac{\gamma^{(t-1-p)/2}}{\sqrt{1-\gamma}} \, \|\theta_{p+1}^\star - \theta_p^\star\|_2,
\end{aligned}$$

where the last inequality holds because, for $t \in \mathcal{G}_D$, the parameter does not change over the most recent D rounds before t. Here, the sum is interpreted as 0 if $t - D - 1 < 1$.

Summing over $t \in \mathcal{G}_D \cap \{2, \dots, T\}$ and then enlarging the index set to all $t = 2, \dots, T$, we get

$$\begin{aligned}
\sum_{t \in \mathcal{G}_D \cap \{2,\dots,T\}} \mathcal{V}_t &\leqslant \sum_{t=2}^{T} \sum_{p=1}^{t-D-1} \frac{\gamma^{(t-1-p)/2}}{\sqrt{1-\gamma}} \|\theta^\star_{p+1} - \theta^\star_p\|_2 \\
&= \frac{1}{\sqrt{1-\gamma}} \sum_{p=1}^{T-1} \|\theta^\star_{p+1} - \theta^\star_p\|_2 \sum_{t=p+D+1}^{T} \gamma^{(t-1-p)/2} \\
&\leqslant \frac{1}{\sqrt{1-\gamma}} \sum_{p=1}^{T-1} \|\theta^\star_{p+1} - \theta^\star_p\|_2 \sum_{k=D}^{\infty} \gamma^{k/2} \\
&= \frac{\gamma^{D/2}}{\sqrt{1-\gamma}(1-\sqrt{\gamma})} \sum_{p=1}^{T-1} \|\theta^\star_{p+1} - \theta^\star_p\|_2 = \frac{\gamma^{D/2}}{\sqrt{1-\gamma}(1-\sqrt{\gamma})} P_T,
\end{aligned}$$

For each fixed p, where $P_T := \sum_{p=1}^{T-1} \|\theta^\star_{p+1} - \theta^\star_p\|_2$. Using

$$1 - \sqrt{\gamma} = \frac{1-\gamma}{1+\sqrt{\gamma}} \qquad \text{and} \qquad 1 + \sqrt{\gamma} \leqslant 2,$$

we obtain

$$\sum_{t \in \mathcal{G}_D \cap \{2,\dots,T\}} \mathcal{V}_t \leqslant \frac{2\gamma^{D/2}}{(1-\gamma)^{3/2}} P_T. \tag{D.6}$$

It remains to bound P_T under piecewise stationarity. Let $\nu_1, \dots, \nu_{\Gamma_T}$ be the change points. If $p \notin \{\nu_1 - 1, \dots, \nu_{\Gamma_T} - 1\}$, then $\theta^\star_{p+1} = \theta^\star_p$, so

$$\|\theta^\star_{p+1} - \theta^\star_p\|_2 = 0.$$

On the other hand, if $p = \nu_j - 1$ for some $j \in [\Gamma_T]$, then by the triangle inequality,

$$\|\theta^\star_{p+1} - \theta^\star_p\|_2 \leqslant \|\theta^\star_{p+1}\|_2 + \|\theta^\star_p\|_2 \leqslant 2S,$$

since $\theta^\star_p, \theta^\star_{p+1} \in \Theta \subseteq \{\theta : \|\theta\|_2 \leqslant S\}$. Because there are exactly Γ_T such indices,

$$P_T = \sum_{p=1}^{T-1} \|\theta^\star_{p+1} - \theta^\star_p\|_2 \leqslant 2S\Gamma_T.$$

Substituting this into equation D.6 yields

$$\sum_{t \in \mathcal{G}_D \cap \{2,\dots,T\}} \mathcal{V}_t \leqslant \frac{4S\Gamma_T \, \gamma^{D/2}}{(1-\gamma)^{3/2}}.$$

This completes the proof. □