Variance-Adaptive Optimal Algorithm for Reinforcement Learning with Multinomial Logit Function Approximation

Wonyoung Kim
Chung-Ang University
wyk7@cau.ac.kr

Garud Iyengar
Columbia University
garud@columbia.edu

Assaf Zeevi
Columbia University
assaf@gsb.columbia.edu

Min-hwan Oh
Seoul National University
minoh@snu.ac.kr

Abstract

Reinforcement learning with multinomial logistic (MNL) function approximation has become an important framework due to its flexibility and broad applicability. While existing studies have established regret guarantees under worst-case analysis, they do not capture how performance depends on the variability of the interaction between the learner and the environment. In this paper, we develop a new theoretical analysis for MNL-based Markov decision processes that yields explicit variance-adaptive regret bounds. Our algorithm is computationally efficient and achieves the instance-wise optimal rate of regret, narrowing the gap between upper and lower bounds. Our numerical experiments validate that our method learns optimal policies more efficiently than conventional approaches.

1 Introduction

Reinforcement learning (RL) studies how an agent can learn to make sequential decisions by interacting with an environment so as to maximize cumulative rewards [Puterman, 2014, Sutton and Barto, 2018]. In the classical *tabular* setting— Markov decision processes (MDPs) with finite state and action spaces, decades of work have developed a rich theoretical understanding: algorithms achieve near-minimax optimal regret and sample-efficiency guarantees that scale with the number of states/actions and the horizon [Kearns and Singh, 2002, Brafman and Tennenholtz, 2002, Azar et al., 2017, Dann et al., 2017, Zhang et al., 2020, Li et al., 2021, Zhang et al., 2024, Lee and Oh, 2025a]. However, tabular methods become infeasible as the state–action space grows. *feature-based* RL is a way to extend tabular RL to large-scale problems, where either the transition model and/or value functions are approximated in a low-dimensional function class. In particular, *linear* function approximation [Ayoub et al., 2020, Ishfaq et al., 2021, Zanette et al., 2020, Jin et al., 2023] represents the transition model as linear combinations of known feature vectors, enabling theoretical tractability that scales with the feature dimension and generalization beyond visited states

However, the linear MDP paradigm can be fundamentally limited for modeling *transition probabilities*. In many environments, the next-state distribution is required to lie on a probability simplex, whereas a generic linear parametrization does not naturally preserve non-negativity and normalization, and may therefore impose an unnatural structure on the dynamics [Hwang and Oh, 2023, Lee and Oh, 2024a]. This motivates parametric models that retain the benefits of low-dimensional feature representations while respecting the geometry of probability distributions. To that end, the *multinomial logistic* (MNL) MDP model has been proposed as a principled alternative [Hwang and Oh, 2023, Cho et al.,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Regret bound comparison of MNL-based RL algorithms. Here, d denotes the feature dimension, H is the episode length, and K is the total number of episodes. The parameter κ represents the problem-dependent lower bound on the variance over the bounded parameter set as defined in (10), while $\sigma_{k,h}^2$ denotes the realized variance along an interaction sample as defined in (11).

	Paper	Regret Bound	Variance-Adaptive	Optimal
Upper Bound	Hwang and Oh [2023]	$\tilde{O}(\kappa^{-1/2} dH^{3/2}\sqrt{K})$	No	No
	Cho et al. [2024]	$\tilde{O}(dH^2\sqrt{K})$	No	No
	Li et al. [2024]	$\tilde{O}(dH^2\sqrt{K})$	No	No
	This work (Corollary 11)	$\tilde{O}(dH^{3/2}\sqrt{K})$	No	**Yes**
	This work (Theorem 6)	$\tilde{O}\left(dH\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\sigma_{k,h}^2}\right)$	**Yes**	**Yes**
Lower Bound	Park et al. [2025]	$\Omega(dH^{3/2}\sqrt{K})$	No	-
	This work (Theorem 8)	$\Omega\left(dH\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\sigma_{k,h}^2}\right)$	**Yes**	-

2024]: by parameterizing transition probabilities through a softmax/logistic structure, MNL-MDPs produce valid distributions by construction and can capture richer, non-linear dependence on features using a compact parameterization.

Despite its promises, the existing theoretical results are still dominated by *worst-case* regret analyses. That is, existing bounds do not reflect the *realized variance* induced by the actual interaction trajectory between an algorithm and a specific MDP instance [Hwang and Oh, 2023, Cho et al., 2024, Li et al., 2024]. From the viewpoint of modern learning theory, this is unsatisfying: refined, data-dependent guarantees—motivated by ideas such as local complexity and variance-sensitive error bounds—often provide a sharper characterization of performance when the observed sample path is "easy" [Bartlett and Mendelson, 2006, Xu and Zeevi, 2025]. Moreover, even at the level of worst-case regret, optimality in MNL-MDPs has not been fully settled. Table 1 summarizes the state of the art and highlights the missing piece: regret bounds that are simultaneously (i) *variance-adaptive* and (ii) *optimal*.

In this work, we establish the first variance-adaptive regret upper and lower bounds for RL with MNL function approximation, significantly bridging the gap between pessimistic worst-case theory and variance-adaptive empirical performance. Our technical contribution is centered on a novel integration of the Online Newton Step (ONS) algorithm with an estimation framework to derive an exact ellipsoid confidence set for the parameter estimator. To achieve this, we introduce three key theoretical innovations: (i) a tighter regret analysis for ONS applied specifically to log-sum-exp functions; (ii) a new Hessian approximation technique that stabilizes curvature estimation in high-dimensional discrete spaces; and (iii) a refined second-order expansion of the confidence bound that captures local variance more accurately than traditional first-order methods. Collectively, these ingredients yield a variance-adaptive regret guarantee that reflects the intrinsic difficulty of the MDP instance, providing a sharper and more robust theoretical foundation for MNL-based reinforcement learning. The main contribution of this paper is as follows:

- **Variance-adaptive optimal regret bound:** We establish the first variance-dependent regret upper bound (Theorem 6) for MNL-MDPs. Unlike previous minimax analyses, our bound is *adaptive to the variance on the realized sample path*, capturing the intrinsic difficulty of the MDP instance by adaptively scaling with the local variance and information geometry of the observed data sequence.
- **First optimal regret for for MNL-MDPs:** We show that our established bounds are provably optimal. To complement the upper bound, we prove a lower bound (Theorem 8) that matches the variance-dependent regret upper bound up to polylogarithmic factors. We also provide the minimax upper bound that matches the lower bound for the first time in MNL-MDPs.
- **Efficient Hessian estimation and confidence ellipsoids:** On the technical side, we propose a new Hessian estimation technique that significantly improves upon existing exponential error dependencies (Lemma 3 and Lemma 27). Furthermore, we develop a specialized

estimation scheme that transforms online curvature updates into a *self-normalized confidence ellipsoid* (Theorem 2), providing a rigorous theoretical foundation for parameter estimation, which can be of independent interest beyond MNL-MDPs.

- **Variance-adaptive algorithm for MNL-MDP:** Our proposed algorithm uses *second-order approximated upper confidence bound exploration strategy* to capture the realized variance and achieve variance-adaptive regret bound. This algorithm achieves the first regret guarantees in this setting that are simultaneously variance-adaptive and optimal up to logarithmic factors, effectively bridging the gap between pessimistic regret bound and empirical efficiency. Our numerical experiments show that the proposed algorithm is computationally efficient with superior performance.

2 Related Works

Table 1 summarizes the regret comparison for RL algorithms with MNL function approximation. Hwang and Oh [2023] first established a provably efficient algorithm for this setting. Subsequent works by Cho et al. [2024] and Li et al. [2024] provided κ-free regret bounds by employing randomized exploration and online mirror descent (OMD), respectively, where κ is a problem-dependent constant that serves as a lower bound on the minimum eigenvalue of the Hessian across the feasible parameter space (Definition 2). However, these results are not instance-dependent and fail to adapt to the realized variance along the sample path. In contrast, our work is the first to establish regret bounds for MNL-MDPs that are both variance-adaptive and strictly tighter than existing theoretical guarantees.

While the MNL bandit model has been extensively studied within the context of assortment optimization and discrete choice modeling [Agrawal et al., 2017, 2019, Oh and Iyengar, 2021, Chen et al., 2024, Wang et al., 2018, Perivier and Goyal, 2022, Lee et al., 2024], it remains distinct from the MNL-MDP framework. Notably, neither is a special case of the other. In the MNL bandit setting, the agent selects an assortment of items where the resulting rewards (e.g., a customer's selection) follow a multinomial distribution. Conversely, in an MNL-MDP, the agent selects a single action that determines state transition probabilities via an MNL function.

Our instance-dependent constants are closely related to the problem-dependent quantities studied in logistic and multinomial logistic (MNL) bandits. Abeille et al. [2021] and Faury et al. [2022] established instance-dependent lower bounds and proposed algorithms that exploit self-concordance to move the dependence on κ into lower-order (warm-up) terms. More recently, Lee and Oh [2025b] and Boudart et al. [2026] extended these techniques to multinomial choices, removing the dependence on the parameter norm and achieving variance-dependent, minimax-optimal regret bounds. While Boudart et al. [2026] considers the variance along the optimal trajectory, our regret bound is both instance-dependent and algorithm-dependent, explicitly capturing the empirical variance along the selected action trajectory. Crucially, neither variance quantity dominates the other in general—their relative magnitude depends entirely on the algorithm's online exploratory choices—rendering the two bounds strictly incomparable.

3 Multinomial Logit Markov Decision Process

We consider episodic Markov decision processes (MDPs) denoted by $\mathcal{M}(\mathcal{S}, \mathcal{A}, H, \mathbb{P} = \{\mathbb{P}_h : h \in [H]\}, r)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, H is the length of the episode, K is the total number of episodes, $\{\mathbb{P}_h : h \in [H]\}$ is the collection of state transition probability distributions for each $h \in [H]$, and r is a reward function. In each episode $k \in [K]$, the environment chooses an initial state $s_{k,1} \in \mathcal{S}$ and the state evolves for H steps. At each step $h \in [H]$ in every episode k, the decision maker observes the state $s_{k,h} \in \mathcal{S}$, selects an action $a_{k,h} \in \mathcal{A}$, and receives an immediate reward $r(s_{k,h}, a_{k,h}) \in [0, 1]$, and the next state is generated from the transition probability distribution $\mathbb{P}_h(\cdot|s_{k,h}, a_{k,h})$. This process repeats until the end of the episode. A policy $\pi : \mathcal{S} \times [H] \to \mathcal{A}$ is a (potentially stochastic) function that determines which action the decision maker chooses in state s and step $h \in [H]$, i.e., $a \sim \pi(s, h)$. The value $V_h^\pi(s)$ of policy π in state $s \in \mathcal{S}$ and step $h \in [H]$ is defined as the expected sum of rewards under the policy π until the end of the episode when starting

from $s_h = s$, i.e.,

$$V_h^\pi(s) = \mathbb{E}_\pi\big[\sum_{h'=h}^{H} r(s_{h'}, a_{h'})\big| s_h = s\big],$$

where $\mathbb{E}_\pi$ is the expectation over the trajectory generated by the policy-environment interaction measure. The action-value function $Q_h(s,a;\pi)$ of the policy π is defined as

$$Q_h^\pi(s,a) := \mathbb{E}_\pi\Big[\sum_{h'=h}^{H} r\big(s_{h'}, a_{h'}\big)\Big| s_h = s, a_h = a\Big].$$

We define the optimal policy $\pi^\star := (\pi_1^\star, \ldots, \pi_H^\star)$ to be a policy that achieves the highest possible value at each step $h \in [H]$ and for every state $s \in \mathcal{S}$. We denote the value of the optimal policy $V_h^\star(s) := V_h^{\pi^\star}(s)$ and the action-value function of the optimal policy $Q_h^\star(s,a) := Q_h^{\pi^\star}(s,a)$, respectively. Thus, $V_h^{\pi^\star}(s) = \max_{a\in\mathcal{A}} Q_h^{\pi^\star}(s,a)$. For simplicity, we define $\mathbb{P}_h V_{h+1}(s,a) := \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)}[V_{h+1}(s')]$. Recall that the action-value functions satisfy the Bellman recursion:

$$Q_h^\pi(s,a) = r(s,a) + \mathbb{P}_h V_{h+1}^\pi(s,a), \quad Q_h^\star(s,a) = r(s,a) + \mathbb{P}_h V_{h+1}^\star(s,a).$$

where $V_{H+1}^\pi(s) = V_{H+1}^\star(s) = 0$. Let $\Pi_K = \{\pi_k : k \in [K]\}$ denote a collection of policies, where π_k is the policy of the decision maker for the $k \in [K]$-th episode. Then the cumulative regret of Π_K for MDP $\mathcal{M} := \mathcal{M}(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$ is defined as

$$\mathrm{R}_{\Pi_K}(\mathcal{M}) := \sum_{k=1}^{K} V_1^\star(s_{k,1}) - V_1^{\pi_k}(s_{k,1}).$$

The goal of the decision maker is to choose Π_K that minimizes regret. Note that this is equivalent to choosing Π_K that maximizes cumulative rewards over K episodes. We assume that the transition kernel of the MDP $\mathcal{M}(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$ is given by an MNL model.

Definition 1 (Multinomial Logit MDP)**.** Let $\mathbf{e}_s$ denote the s-th Euclidean basis. The state transition function when an action $a_h \in \mathcal{A}$ is taken at the state s_h is given by

$$\mathbb{P}_h(s_{h+1}|s_h, a_h, \boldsymbol{\theta}_h^\star) = \frac{\exp\big(\mathbf{e}_{s_{h+1}}^\top \boldsymbol{\Phi}(s_h, a_h)\boldsymbol{\theta}_h^\star\big)}{\sum_{\tilde{s}\in\mathcal{S}_h(s_h,a_h)} \exp\big(\mathbf{e}_{\tilde{s}}^\top \boldsymbol{\Phi}(s_h, a_h)\boldsymbol{\theta}_h^\star\big)}, \tag{1}$$

where $\mathcal{S}_h(s,a)$ is the set of states that are reachable from $s \in \mathcal{S}$ in step $h \in [H]$ using action $a \in \mathcal{A}$. The feature map $\boldsymbol{\Phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{|\mathcal{S}|\times d}$ is known, and the transition parameter $\boldsymbol{\Theta}^\star = [\boldsymbol{\theta}_h^\star]_{h\in[H]}$ is unknown. The feature map $\boldsymbol{\Phi}$ and $\boldsymbol{\theta}^\star$ are bounded: $\|\boldsymbol{\Phi}(s,a)^\top \mathbf{e}_{s'}\|_2 \le B_{\boldsymbol{\Phi}}$ for all $s', s \in \mathcal{S}$, and $a \in \mathcal{A}$, and $\max_{h\in[H]} \|\boldsymbol{\theta}_h^\star\|_2 \le B_{\boldsymbol{\theta}}$.

Note that each step h has a different parameter $\boldsymbol{\theta}_h^\star$, which makes the problem harder to learn than the homogeneous transition probability studied in Hwang and Oh [2023]. While the size of the state space $|\mathcal{S}|$ may be exponentially large, the cardinality of the reachable set $\mathcal{S}_h(s,a)$ may be small. Let $U := \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \big|\{s' \in \mathcal{S} : P(s' \mid s, a) > 0\}\big|$ denote the maximum size of the reachable set. As noted by Hwang and Oh [2023], in many practical environments the reachable set is localized, so U is much smaller than the size of the full state space, $U \ll |\mathcal{S}|$. In spatially localized MDPs such as grid worlds and robot navigation, an agent on an $N \times N$ grid ($|\mathcal{S}| = N^2$) that moves in one of four cardinal directions can reach only the adjacent cells, so $U \le 4$ however large N is. In the RiverSwim environment, the agent can only move to a neighboring state or stay in its current state, so $U \le 3$ for any chain length. In assortment recommendation, a customer chooses among the K items displayed from a catalog of $|\mathcal{S}|$ items. Display constraints limit transitions to this candidate set, so $U = K \ll |\mathcal{S}|$.

MNL-MDPs model sequential decision problems in which a single action leads to one of several reachable next states, and the relative transition probabilities among those states follow a MNL structure. For example, in recommendation or online engagement systems, one displayed policy/action may lead a user to several possible future states (e.g., click, continue browsing, switch category, or exit), and the long-term objective depends on how these state transitions unfold over time.

Algorithm 1 Online-to-Confidence-Ellipsoid Estimator Update (OCEE)

Input: online estimator $\boldsymbol{\theta}$, information matrix $\mathbf{H}$, learning rate η, transition sample (s, a, s'), parameter bound $B_{\boldsymbol{\theta}}$, moment vector $\boldsymbol{\gamma}$.

1: Compute the gradient and update the information matrix:

$$\mathbf{g} = \boldsymbol{\Phi}(s,a)^\top \big(-\mathbf{e}_{s'} + \nabla L_{\mathcal{S}(s,a)}(\boldsymbol{\Phi}(s,a)\boldsymbol{\theta}) \big), \quad \mathbf{H} \leftarrow \mathbf{H} + \mathbf{g}\mathbf{g}^\top$$

2: **Newton step**: $\tilde{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta} - \eta \mathbf{H}^{-1}\mathbf{g}$

3: **Projection step**: $\boldsymbol{\theta} \leftarrow \operatorname{argmin}_{\|\boldsymbol{\theta}\|_2 \le B_{\boldsymbol{\theta}}} \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_{\mathbf{H}}^2$

4: Update the moment vector: $\boldsymbol{\gamma} \leftarrow \boldsymbol{\gamma} + \mathbf{g}\mathbf{g}^\top \boldsymbol{\theta}$

5: Compute the estimator: $\widehat{\boldsymbol{\theta}} = \mathbf{H}^{-1}\boldsymbol{\gamma}$

Output: Estimator $\widehat{\boldsymbol{\theta}}$, updated online estimator $\boldsymbol{\theta}$, updated Hessian $\mathbf{H}$, updated moment vector $\boldsymbol{\gamma}$.

4 Multinomial Logit Parameter Estimation Procedure

In this section, we present our proposed online estimation method for the MNL-MDP. We first propose an online algorithm for maximizing the likelihood of a collection of multinomial random variables (Section 4.1). Next, we show how to convert a regret bound of the negative log-likelihood into a confidence region for the parameter $\boldsymbol{\Theta}^\star$ that is compatible with the UCB-style algorithms (Section 4.2).

4.1 Tight Regret Bound for Online Newton Step

Let $L_{\mathcal{S}_h(s,a)}(\boldsymbol{\eta}) := \log \sum_{\bar{s} \in \mathcal{S}_h(s,a)} \exp(\mathbf{e}_{\bar{s}}^\top \boldsymbol{\eta})$ denote the log-sum function. For step $h \in [H]$ in episode $k \in [K]$, the negative log-likelihood is

$$\begin{aligned} \ell_{k,h}(\boldsymbol{\theta}) &:= -\log \mathbb{P}_h(s_{k,h+1} \mid s_{k,h}, a_{k,h}, \boldsymbol{\theta}) \\ &= -\mathbf{e}_{s_{k,h+1}}^\top \boldsymbol{\Phi}(s_{k,h}, a_{k,h})\boldsymbol{\theta} + L_{\mathcal{S}_h(s_{k,h},a_{k,h})}\big(\boldsymbol{\Phi}(s_{k,h}, a_{k,h})\boldsymbol{\theta}\big). \end{aligned} \tag{2}$$

For each $h \in [H]$, let $\boldsymbol{\theta}_{0,h} := \mathbf{0}$ and given $\boldsymbol{\theta}_{k-1,h}$ we define

$$\boldsymbol{\theta}_{k,h} := \boldsymbol{\theta}_{k-1,h} - \eta \widehat{\mathbf{H}}_{k,h}^{-1} \nabla \ell_{k,h}(\boldsymbol{\theta}_{k-1,h}) \tag{3}$$

the Online Newton Step (ONS) estimate of $\boldsymbol{\theta}_h^\star$, where

$$\widehat{\mathbf{H}}_{k,h} := \sum_{\nu=1}^{k} \nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) \nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top + \epsilon \mathbf{I}_d \tag{4}$$

denote the *observed information matrix* over the k transitions from step h to $h+1$ with transition samples $\{(s_{\nu,h}, a_{\nu,h}, s_{\nu,h+1})\}_{\nu=1}^{k}$ and initial parameter $\epsilon > 0$. Our analysis utilizes the Fisher Information Identity:

$$\mathbb{E}[\nabla \ell_{\nu,h}(\theta_h^\star) \nabla \ell_{\nu,h}(\theta_h^\star)^\top \mid \mathcal{F}_{\nu,h}] = \nabla^2 \ell_{\nu,h}(\theta_h^\star),$$

where $\mathcal{F}_{\nu,h}$ is the sigma-algebra generated by $\cup_{h'=1}^{H} \cup_{\nu'=1}^{\nu-1} \{s_{\nu',h'}, a_{\nu',h'}\}$ and $\cup_{h'=1}^{h} \{s_{\nu,h'}, a_{\nu,h'}\}$. This identity allows us to relate the second-order curvature directly to the variance of the gradients at the true parameter $\theta_h^\star$.

If the gradient $\nabla \ell_{\nu,h}(\theta_h^\star)$ is approximated by the gradient at the current iterate $\theta_{\nu,h}$, the resulting error is quadratic in the difference $\|\nabla \ell_{\nu,h}(\theta_{\nu-1,h}) - \nabla \ell_{\nu,h}(\theta_h^\star)\|$.

Lemma 1 (Quadratic Error Approximation for the True Hessian)**.** *Let* $\mathbf{g}_{\nu,h} := \nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})$. *Then,*

$$\mathbb{E}\big[\mathbf{g}_{\nu,h}\mathbf{g}_{\nu,h}^\top \,\big|\, \mathcal{F}_{\nu,h}\big] - \nabla^2 \ell_{\nu,h}(\boldsymbol{\theta}_h^\star) = \big(\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) - \nabla \ell_{\nu,h}(\boldsymbol{\theta}_h^\star)\big)\big(\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) - \nabla \ell_{\nu,h}(\boldsymbol{\theta}_h^\star)\big)^\top. \tag{5}$$

The cumulative rank-one update of $\mathbf{g}_{\nu,h}\mathbf{g}_{\nu,h}^\top$ effectively captures the curvature in the direction of the gradient. As these gradients accumulate over iterations, the resulting matrix provides a robust approximation of the second-order structure without relying on global Hessian stability constants.

4.2 Online to Confidence Set Conversion

While the ONS estimate $\boldsymbol{\theta}_{k,h}$ method is computationally efficient and gives a reasonable estimator, the error bound of the estimator is not guaranteed, i.e., the confidence ellipsoid is not established. Instead of using the online estimate $\boldsymbol{\theta}_{\nu,h}$ we incorporate all online estimates to the proposed estimator that achieves an error bound normalized by the Hessian matrix. Given a regularization parameter $\epsilon > 0$, we define our estimator $\widehat{\boldsymbol{\theta}}_{k,h}$ for $\boldsymbol{\theta}^\star_h$ in episode k as the solution of the optimization problem:

$$\widehat{\boldsymbol{\theta}}_{k,h} := \operatorname*{argmin}_{\boldsymbol{\theta}_h \in \mathbb{R}^d} \left\{ \sum_{\nu=1}^{k} \left(\mathbf{g}_{\nu,h}^\top (\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h) \right)^2 + \epsilon \|\boldsymbol{\theta}_h\|^2 \right\} \Rightarrow \widehat{\boldsymbol{\theta}}_{k,h} = \widehat{\mathbf{H}}_{k,h}^{-1} \Big(\sum_{\nu=1}^{k} \mathbf{g}_{\nu,h} \mathbf{g}_{\nu,h}^\top \boldsymbol{\theta}_{\nu-1,h} \Big). \quad (6)$$

This estimator can be updated in a constant time with rank-1 update. The following theorem presents the confidence ellipsoid of the proposed estimator $\widehat{\boldsymbol{\theta}}_{k,h}$.

Theorem 2 (Self-Normalized Bound for the Estimator)**.** *Define* $C_{\boldsymbol{\Phi},\boldsymbol{\theta}} := (e-1)(6 + 8B_{\boldsymbol{\Phi}} B_{\boldsymbol{\theta}} + 2B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2)$. *Set the learning rate* $\eta_{\boldsymbol{\Phi},\boldsymbol{\theta}} := (e-1)(3 + 4B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2)$. *For the estimator* $\widehat{\boldsymbol{\theta}}_{k,h}$ *defined in* (6)*, with probability at least* $1 - 3\delta$, $\left\| \widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}^\star_h \right\|_{\widehat{\mathbf{H}}_{k,h}} \le \beta_k$, *for all* $k \in [K]$ *where* $\beta_k = \sqrt{\epsilon} B_{\boldsymbol{\theta}} + \sqrt{\epsilon B_{\boldsymbol{\theta}}^2 + 4\gamma_k}$ *and* $\gamma_k = \tilde{O}(d B_{\boldsymbol{\Phi}}^4 B_{\boldsymbol{\theta}}^4)$.

The proof of Theorem 2 is in Appendix C.6. The rate of radius β_k is $O(B_{\boldsymbol{\theta}}^2 B_{\boldsymbol{\Phi}}^2 \sqrt{d \log dk})$. Our self-normalized bound involves the observed information matrix $\widehat{\mathbf{H}}_{k,h}$ instead of the Hessian estimate $\tilde{\mathbf{H}}_{k,h} := \sum_{\nu=1}^{k} \nabla^2 \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) + \epsilon \mathbf{I}_d$. Zhang and Sugiyama [2024] proposed an online mirror descent algorithm that establishes a similar self-normalized bound. Their self-normalized bound is normalized by the Hessian estimate $\tilde{\mathbf{H}}_{k,h}$, which involves the second order derivative.

Following the standard self-concordant analysis (see e.g., Sun and Tran-Dinh [2019]), there exists a constant $C > 0$ such that

$$\nabla^2 \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) \succeq \exp\left(-C \left\| \boldsymbol{\Phi}(s_{\nu,h}, a_{\nu,h})(\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}^\star_h) \right\|_2 \right) \nabla^2 \ell_{\nu,h}(\boldsymbol{\theta}^\star_h). \quad (7)$$

Therefore, it suffers from the exponential difference $\exp B_{\boldsymbol{\Phi}} B_{\boldsymbol{\theta}}$ when obtaining the self-normalized bound that involves the true Hessian, $\mathbf{H}^\star_{k,h} := \sum_{\nu=1}^{k} \nabla^2 \ell_{\nu,h}(\boldsymbol{\theta}^\star) + B_{\boldsymbol{\Phi}}^2 \mathbf{I}_d$. In contrast, our proposed Hessian estimate $\widehat{\mathbf{H}}_{k,h} = \sum_{\nu=1}^{k} \nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) \nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top + \epsilon \mathbf{I}_d$ yields the self-normalized bound with $\mathbf{H}^\star_{k,h}$. The following lemma shows that we can approximate $\mathbf{H}^\star_{k,h}$ with our proposed information matrix $\widehat{\mathbf{H}}_{k,h}$ and without using the second order derivatives of the log-likelihood.

Lemma 3 (Information Matrix Dominates the True Hessian)**.** *For any step* $h \in [H]$*, if the regularization parameter is set to the required value* $\epsilon = B_{\boldsymbol{\Phi}}^2 + 4B_{\boldsymbol{\Phi}}^2 \log(d/\delta)$*, then with probability at least* $1 - \delta$*, we have* $\widehat{\mathbf{H}}_{k,h} \succeq (1 - e^{-1}) \mathbf{H}^\star_{k,h}$ *for all* $k \in [K]$

Lemma 3 shows $\widehat{\mathbf{H}}_{k,h}$ dominates $\mathbf{H}^\star_{k,h}$ and directly implies $\|\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}^\star_h\|_{\mathbf{H}^\star_{k,h}} = O(\sqrt{d \log dk})$. Thus, our estimator has a $\mathbf{H}^\star_{k,h}$-normalized bound without involving the exponential approximation of the second derivative in (7). This paves the way for the improvement over the previous regret upper bounds by the factor of at most $\exp(B_{\boldsymbol{\Phi}} B_{\boldsymbol{\theta}})$

5 Variance-Adaptive MNL-MDP Algorithm

With our proposed estimator $\widehat{\boldsymbol{\theta}}_{k,h}$, let $\mathcal{C}_{k,h} := \left\{ \boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{k,h}\|^2_{\widehat{\mathbf{H}}_{k,h}} \le \beta_k \right\}$, denote the confidence ellipsoid. The optimism in the face of uncertainty (OFU) principle requires the decision maker to compute

$$\operatorname*{argmax}_{\boldsymbol{\theta} \in \mathcal{C}_{k-1,h}} \mathbb{E}\left[\widehat{V}_{k,h+1}(s') \middle| s, a, \boldsymbol{\theta} \right] = \operatorname*{argmax}_{\boldsymbol{\theta} \in \mathcal{C}_{k-1,h}} \left\{ \frac{\sum_{s' \in \mathcal{S}_h(s,a)} \widehat{V}_{k,h+1}(s') \exp\left(\mathbf{e}_{s'}^\top \boldsymbol{\Phi}(s,a) \boldsymbol{\theta} \right)}{\sum_{s' \in \mathcal{S}_h(s,a)} \exp\left(\mathbf{e}_{s'}^\top \boldsymbol{\Phi}(s,a) \boldsymbol{\theta} \right)} \right\}, \quad (8)$$

where $\widehat{V}_{k,h+1}(s) = \max_{a \in \mathcal{A}} \widehat{Q}_{k,h}(s,a)$ is the estimated value function. However, computing the solution for (8) is intractable. Hwang and Oh [2023] use the upper bound:

$$\mathbb{E}[\widehat{V}_{k,h+1}(s')|s,a,\boldsymbol{\theta}] \le \mathbb{E}[\widehat{V}_{k,h+1}(s')|s,a,\widehat{\boldsymbol{\theta}}] + 2H\kappa^{-1/2} \left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star_h \right\|_{\tilde{\mathbf{A}}_{k-1,h}} \max_{s' \in \mathcal{S}_h(s,a)} \|\mathbf{e}_{s'}^\top \boldsymbol{\Phi}(s,a)\|_{\tilde{\mathbf{A}}_{k-1,h}^{-1}}.$$

Algorithm 2 Variance adaptive algorithm for MNL-MDP (`VA-MNL`)

Input: Confidence level $\delta \in (0,1)$, Initial observed information matrix $\mathbf{H}_{0,h} = B_{\boldsymbol{\Phi}}^2(1 + 4\log\frac{d}{\delta})\mathbf{I}_d$, Initial online estimate $\boldsymbol{\theta}_{0,h} = \mathbf{0}$, Learning rate $\eta_{\boldsymbol{\Phi},\boldsymbol{\theta}} := (e-1)(3 + 4B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2)$, Initial moment vector $\boldsymbol{\gamma}_{0,h} = \mathbf{0}$.

1: **for** $k = 1$ to K **do**
2: **for** $h = H$ to 1 **do**
3: Compute $\widehat{Q}_{k,h}(s,a)$ as in (9) for all s and a.
4: **end for**
5: **for** $h = 1$ to H **do**
6: Select action $a_{k,h} = \text{argmax}_{a\in\mathcal{A}}\widehat{Q}_{k,h}(s_{k,h},a)$
7: Observe $s_{k,h+1}$ and update parameters:

$$\widehat{\boldsymbol{\theta}}_{k,h}, \boldsymbol{\theta}_{k,h}, \mathbf{H}_{k,h}, \boldsymbol{\gamma}_{k,h} \leftarrow \texttt{OCEE}(\boldsymbol{\theta}_{k-1,h}, \mathbf{H}_{k-1,h}, \eta_{\boldsymbol{\Phi},\boldsymbol{\theta}}, (s_{k,h}, a_{k,h}, s_{k,h+1}), B_{\boldsymbol{\theta}}, \boldsymbol{\gamma}_{k-1,h})$$

8: **end for**
9: **end for**

where κ is the lower bound of the Hessian defined in (10) and $\tilde{\mathbf{A}}_{k-1,h} := \sum_{\nu=1}^{k-1}\boldsymbol{\Phi}(s_{\nu,h},a_{\nu,h})^\top\boldsymbol{\Phi}(s_{\nu,h},a_{\nu,h}) + \mathbf{I}_d$ is the Gram matrix without the variance term. Note that the $\kappa^{-1/2}$ term propagates to the cumulative regret bound. Cho et al. [2024] and Li et al. [2024] refined this approximation in terms of Hessian $\tilde{\mathbf{H}}_{k,h}$ to remove $\sqrt{\kappa}$ in the leading term of the regret bound. However, they use the first-order Taylor expansion and this overlooks the realized variance of the state-transition and the interaction.

In contrast, we approximate the expected value function with the second-order Taylor expansion. Recall that $L_{\mathcal{S}_h(s,a)}(\boldsymbol{\eta}) := \log\sum_{\bar{s}\in\mathcal{S}_h(s,a)}\exp(\mathbf{e}_{\bar{s}}^\top\boldsymbol{\eta})$ is the log-sum function and let $\mathbf{V}_{k,h}(s,a) := \nabla^2 L_{\mathcal{S}_h(s,a)}(\boldsymbol{\Phi}(s,a)\widehat{\boldsymbol{\theta}}_{k-1,h})$. Let $P_H(x) := \max\{0, \min\{x, H\}\}$ denote the projection function onto $[0,H]$. The proposed second-order upper bound for Q-value is

$$\begin{aligned}\widehat{Q}_{k,h}(s,a) := P_H\Big(r(s,a) &+ \sum_{s'\in\mathcal{S}_{k,h}(s,a)}\widehat{V}_{k,h+1}(s')\,\mathbb{P}(\mathbf{e}_{s'}\mid s,a,\widehat{\boldsymbol{\theta}}_{k-1,h}) \\ &+\beta_{k-1}\Big\|\sum_{s'\in\mathcal{S}_{k,h}(s,a)}\widehat{V}_{k,h+1}(s')\mathbf{e}_{s'}^\top\mathbf{V}_{k,h}(s,a)\boldsymbol{\Phi}(s,a)\Big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}} \\ &+\beta_{k-1}^2\max_{s\in\mathcal{S}_{k,h}(s,a)}\widehat{V}_{k,h+1}(s)\left\|\mathbf{e}_{s'}^\top\boldsymbol{\Phi}(s,a)\right\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}}^2\Big).\end{aligned} \tag{9}$$

In Lemma 4, we establish that the $\widehat{Q}_h$ is an optimistic Q function with high probability.

Lemma 4 (Optimism of the Second Order UCB)**.** *For any $(s,a)\in\mathcal{S}\times\mathcal{A}$, with probability at least $1-3\delta$, $Q_{\star,h}(s,a) \le \widehat{Q}_{k,h}(s,a)$, for all $k\in[K]$ and $h\in[H]$.*

Algorithm 2, termed the *Variance-Adaptive algorithm for MNL-MDP* (`VA-MNL`), combines the online-to-confidence set conversion and a second-order approximate upper confidence bound. At the start of each episode k, the algorithm maintains for every stage h a parameter estimate $\widehat{\boldsymbol{\theta}}_{k-1,h}$ together with the associated observed information matrix $\mathbf{H}_{k-1,h}$, which defines a confidence ellipsoid. The Q-value estimates $\widehat{Q}_{k,h}$ are computed by incorporating both the current estimate and an exploration bonus derived from the confidence set and the second-order Taylor expansion. At each decision point h, the action $a_{k,h}$ is selected as the maximizer of $\widehat{Q}_{k,h}$. After observing the next state $s_{k,h+1}$ and the feature vector $\boldsymbol{\Phi}(s_{k,h},a_{k,h})$, the online estimate $\boldsymbol{\theta}_{k,h}$ and the observed information matrix $\mathbf{H}_{k,h}$ are updated with the outer product of the stochastic gradient as in Algorithm 1 The algorithm updates $\widehat{\boldsymbol{\theta}}_{k,h}$ as in (6) by integrating the previous online estimates $\{\boldsymbol{\theta}_{\nu,h} : \nu\in[k]\}$.

The computational time per step arises from computing the stochastic gradient, updating the inverse of $\mathbf{H}_{k-1,h}$ using the Sherman–Morrison rank-one formula, and projecting the updated parameter back into the feasible set. The gradient computation scales linearly with d, whereas the Hessian inverse update and projection each require $O(d^2)$ operations. In addition, the Q-value computation involves a log-sum-exp evaluation over the set of reachable states $\mathcal{S}_h(s,a)$, which costs $O(|\mathcal{S}_h(s,a)|\,d)$ per

state–action pair. When the number of reachable states is constant, the total cost of processing one decision point is $O(d^2)$.

6 Regret Analysis

We define a problem-dependent constant that significantly affects the regret bound.

Definition 2 (Fisher Information Lower Bound)**.** Let

$$p_{\theta_h}(s' \mid s,a) := \frac{\exp\big(\mathbf{e}_{s'}^\top \boldsymbol{\Phi}(s,a)\theta_h\big)}{\sum_{\bar{s}\in\mathcal{S}_h(s,a)} \exp\big(\mathbf{e}_{\bar{s}}^\top \boldsymbol{\Phi}(s,a)\theta_h\big)}$$

denote the MNL transition probability at step $h \in [H]$ for $s' \in \mathcal{S}_h(s,a)$. We define the Fisher information lower bound as $\kappa > 0$ such that

$$\kappa := \inf_{\substack{h,(s,a):\,|\mathcal{S}_h(s,a)|\geq 2 \\ \|\theta_h\|_2 \leq B_\theta}} \min_{s'\neq\bar{s}} p_{\theta_h}(s' \mid s,a)\, p_{\theta_h}(\bar{s} \mid s,a). \tag{10}$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

Remark 5. The constant κ gives a lower bound on the curvature of the MNL log-likelihood relative to the centered feature Gram matrix. Throughout, let $\phi(s' \mid s,a)^\top := \mathbf{e}_{s'}^\top \boldsymbol{\Phi}(s,a)$ denote the feature row of $s' \in \mathcal{S}_h(s,a)$. Writing $C(p) := \mathrm{diag}(p) - pp^\top$ and letting $\bar{\boldsymbol{\Phi}}(s,a)$ be the matrix of centered feature rows, we have $\nabla_\theta^2 \ell_{h,s,a}(\theta) = \bar{\boldsymbol{\Phi}}(s,a)^\top C(p_\theta)\bar{\boldsymbol{\Phi}}(s,a)$. Using $C(p) = \sum_{i<j} p_i p_j (e_i - e_j)(e_i - e_j)^\top$, we obtain

$$\bar{\Phi}^\top C(p_\theta)\bar{\Phi} \succeq \sum_{s'\neq\bar{s}} p_\theta(s' \mid s,a) p_\theta(\bar{s} \mid s,a) \bar{\phi}(s' \mid s,a)\bar{\phi}(s' \mid s,a)^\top \succeq \kappa\, \bar{\Phi}(s,a)^\top \bar{\Phi}(s,a).$$

Thus the result is a relative curvature bound on the centered feature space. For singleton supports we set $\bar{\Phi}(s,a) = 0$ and $\sigma^2(s,a) = 0$, and no positive-curvature condition is imposed.

Definition 2 is essential in learning the multinomial model [Jézéquel et al., 2021, Oh and Iyengar, 2021, 2019, Amani and Thrampoulidis, 2021, Lee and Oh, 2024b, Zhang and Sugiyama, 2024, Perivier and Goyal, 2022]. The constant κ^{-1} can be exponential in $B_{\boldsymbol{\Phi}}$ and $B_{\boldsymbol{\theta}}$ [Faury et al., 2020]; eliminating the κ^{-1} in the dominating order of the regret bound has been a crucial task in MNL-MDP literature [Cho et al., 2024, Li et al., 2024].

We define the variance over a specific action-state pair

$$\sigma^2(s,a) := \max_{\mathbf{x}:\|\mathbf{x}\|_\infty \leq 1} \mathbf{x}^\top \nabla^2 L_{\mathcal{S}_h(s,a)}(\boldsymbol{\Phi}(s,a)\boldsymbol{\theta}_h^\star)\mathbf{x} \tag{11}$$

as the maximum quadratic value of the Hessian over ℓ_∞-ball. The quantity $\sigma^2(s,a)$ is a reward-independent envelope of the dispersion of the next-state distribution:

$$\sigma^2(s,a) = \sup_{\mathbf{x}:\|\mathbf{x}\|_\infty \leq 1} \mathrm{Var}\big[\mathbf{x}_{s_{h+1}} \mid s_h = s, a_h = a\big].$$

For a binary transition with probability p, $\sigma^2(s,a) = 4p(1-p)$. For a value function $V : \mathcal{S} \to [0,H]$, the variance of the normalized value,

$$\mathrm{Var}[V(s_{h+1})/H \mid s,a] = p(1-p)\frac{(V(1)-V(0))^2}{H^2} = \frac{\sigma^2(s,a)}{4H^2}(V(1)-V(0))^2,$$

which can be much smaller than $\sigma^2(s,a)$. Thus $\sigma^2(s,a)$ is a worst-case transition variance envelope that uniformly controls the reward- and value-level martingale variances used in our analysis.

It is easy to show that the range of $\sigma^2(s,a)$ is in $[4\kappa, 1]$. The maximum over the hypercube is achieved at the endpoints $\{-1,1\}^{\mathcal{S}_h(s,a)}$. Then, with simple algebra, $\sigma^2(s,a) = 1 - 4\min_{\tilde{\mathcal{S}}\subseteq\mathcal{S}_h(s,a)} \big(1/2 - \mathbb{P}(s_{h+1} \in \tilde{\mathcal{S}} \mid s,a)\big)^2$ For given (s,a), if there is a subset $\tilde{\mathcal{S}} \subseteq \mathcal{S}_h(s,a)$ such that $\mathbb{P}(s_{h+1} \in \tilde{\mathcal{S}}) = 1/2$, then $\sigma^2(s,a) = 1$, where the variance reaches the maximum. On the other hand, there are other cases with $\sigma^2(s,a) = O(\kappa)$. For example, in the logistic case $\mathcal{S}_h(s,a) = \{0,1\}$, if $\mathbb{P}(s_{h+1} = 0 | s,a) = \frac{1-\sqrt{1-4\kappa}}{2}$ then $\sigma^2(s,a) = 4\kappa$. In this case, replacing HK with $\sum_{k=1}^K \sum_{h=1}^H \sigma^2(s_{k,h}, a_{k,h})$ in the regret bound improves by a factor of $\kappa^{-1/2}$, or equivalently by $\exp(B_{\boldsymbol{\Phi}} B_{\boldsymbol{\theta}}/2)$ factor.

Theorem 6 (Regret Upper Bound of VA-MNL). *Let $a_{k,h}$ denote the action selected by VA-MNL at step $h \in [H]$ in episode $k \in [K]$ and $s_{k,h}$ denote the realized state. Let $\sigma_{k,h}^2 := \sigma^2(s_{k,h}, a_{k,h})$. If the algorithm Π_K is set to VA-MNL then for any MNL-MDP $\mathcal{M}$, with probability at least $1 - 5\delta$,*

$$R_{\Pi_K}(\mathcal{M}_{\boldsymbol{\Theta}^\star}) = O\bigg(dH \log K \sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H} \sigma_{k,h}^2 \log \frac{KH}{\delta}} + \kappa^{-1} d^{\frac{5}{2}} H^{\frac{5}{2}} \log^{\frac{3}{2}} K\bigg).$$

The explicit expression of the regret bound is presented in Section B.5. Note that the variances $\sigma_{k,h}^2$ are random variables which depend on the action chosen by the algorithm and states evolving accordingly and involves the parameters $\{\boldsymbol{\theta}_h^\star : h \in [H]\}$; thus, our regret bound is variance-adaptive. Because regret measures cumulative performance along the agent's actual trajectory, a tight regret bound should naturally reflect the trajectory-wise variance induced by the algorithm's decisions. Standard purely instance-dependent bounds only depend on properties of the optimal policy or worst-case environment trajectories, failing to capture the variance introduced by the exploration trajectory chosen by the algorithm. In contrast, our bound is both instance- and algorithm-dependent, directly tracking the empirical variance along the algorithm's actual execution path and providing a more precise characterization of regret behavior. This is the first regret upper bound that depends on the variance of the multinomial transition.

When $K = \Omega(\kappa^{-3} d^3 H^2 \log K / \log \frac{KH}{\delta})$ is sufficiently large, $\kappa^{-1} \ll \sum_{k,h} \sigma_{k,h}^2$, and the $\tilde{O}\big(\kappa^{-1} d^{5/2} H^{5/2} \log^{3/2} K\big)$ term is dominated by $\tilde{O}(dH\sqrt{\sum_{k,h} \sigma_{k,h}^2})$. The best known regret bound for MNL-MDP is $\tilde{O}(dH^{3/2}\sqrt{HK} + \kappa^{-1} d^2 H^2)$ [Cho et al., 2024]. Our regret bound has the sum of the variance $\sum_{k=1}^{K}\sum_{h=1}^{H} \sigma_{k,h}^2$ of the multinomial distribution, and improves by at most a factor of $\kappa^{-1/2}$. Further, our upper bound is optimal in that it matches the lower bound in Theorem 8 up to logarithmic factors. Because $\sigma_{k,h}^2 \le 1$, Theorem 6 directly implies $O(dH^{3/2}\sqrt{K} \log KH)$ regret bound (Corollary 11). Further, by developing a novel bound for the sum of variance of the UCB $\widehat{Q}_{k,h}$ (Lemma 7), we establish an improved regret bound by the factor of $\sqrt{H}$ in the main order term.

Deferring the standard computations to Appendix B.5, we provide a novel optimistic law of total variance that enables us to derive the variance-adaptive optimal regret bound.

Lemma 7 (The Optimistic Law of Total Variance). *The sum of variance of the value function $\hat{V}_{k,h+1}$*

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{V}\Big[\widehat{V}_{k,h+1}(s_{k,h+1}) \Big| s_{k,h}, a_{k,h}\Big] \\
&\le 12eH \sum_{k=1}^{K}\sum_{h=1}^{H} \sigma_{k,h}^2 + 1200eH^4\kappa^{-2}\beta_{K-1}^4 \big(d \log \frac{K+1}{d} + 3\log\frac{1}{\delta}\big).
\end{aligned}$$

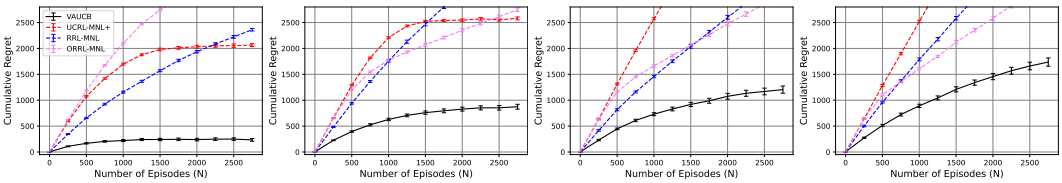
Lemma 7 paves the way to improving the $\sqrt{H}$ term in the regret bound to achieve the optimal regret rate. This is the first variance bound for the *estimated* value function $\widehat{V}_{k,h}$; previous results obtain a bound for the true value function $V_h^\star$.

We provide the variance-dependent lower bound for the MNL-MDP that matches our upper bound with its proof in Appendix B.7.

Theorem 8 (Variance-Adaptive Lower Bound for MNL-MDP). *Assume that $d \ge 2$, $H \ge 4$, and $K \ge \max\{144H(d-1)^2, \kappa^{-2}H^{-1}\}$. Given any parameter $\boldsymbol{\Theta}^\star = (\boldsymbol{\theta}_1^\star, \dots, \boldsymbol{\theta}_H^\star) \in \mathbb{R}^{d\times H}$, there exists a MNL-MDP instance $\mathcal{M}_{\boldsymbol{\Theta}^\star} := \mathcal{M}(\mathcal{S}, \mathcal{A}, H, \boldsymbol{\Theta}^\star, \boldsymbol{\Phi}, r)$ such that with probability at least $1/e$, the regret of any algorithm Π_K,*

$$R_{\Pi_K}(\mathcal{M}_{\boldsymbol{\Theta}^\star}) \ge \frac{2Hd\sqrt{\sum_{h=1}^{H}\sum_{k=1}^{K}\sigma^2(s_{k,h}, a_{k,h})}}{17},$$

where $\{s_{k,h}, a_{k,h} : k \in [K], h \in [H]\}$ denote the state and action sample generated by the interaction between Π_K and $\mathcal{M}_{\boldsymbol{\Theta}^\star}$.



(a) $S = 4, H = 12$ (b) $S = 8, H = 24$ (c) $S = 12, H = 36$ (d) $S = 16, H = 48$

Figure 1: Empirical evaluation of the cumulative regret on the RiverSwim environment across four pairs of state S and episode length H. We compare the proposed `VA-MNL` against `UCRL-MNL+`, `RRL-MNL`, and `ORRL-MNL`. Results are averaged over 20 repeated experiments where the lines represent mean reward and shaded regions indicate standard deviation. The proposed `VA-MNL` yields less cumulative regret and finds the optimal policy faster than other methods.

Theorem 8 implies that no algorithm can achieve better rate than the specified term with high probability greater than $1 - e^{-1}$ as well as in expectation. This is the first variance-adaptive lower bound for the MNL-MDP for any given parameter $\boldsymbol{\Theta}^\star$. The lower bound in Theorem 8 involves the sum of variances $\sum_{h=1}^{H} \sum_{k=1}^{K} \sigma_{k,h}^2$ that depends on the choice of the algorithm Π_K. This captures that the difficulty of the problem arises from the variance of the interaction between the MDP and the algorithm.

Li et al. [2024] proved an $\Omega(dH\sqrt{\kappa K})$ lower bound for the MNL-MDP by reducing the instance into $H/2$ independent logistic bandits and applied the lower bound in Abeille et al. [2021]. Recently, Park et al. [2025] proved an $\Omega(dH^{3/2}\sqrt{K})$ lower bound. However, both bounds represents the worst case over a set of MDPs, which overly pessimistic. In contrast, Theorem 8 provides an instance-dependent lower bound where the parameter $\boldsymbol{\Theta}^\star$ is fixed and the choice of MNL-MDP is restricted.

The novelty in establishing the lower bound is in obtaining a variance-dependent bound for the KL-divergence between the two distinct multinomial distributions for the change-of-measure technique (Lemma 39). To derive the lower bound, we first involve the regret bound maximized over a local neighborhood of a specified instance parametrized by $\boldsymbol{\Theta}^\star$. Based on the fact that the regret $\mathrm{R}_{\Pi_K}(\mathcal{M}_{\boldsymbol{\Theta}})$ is Lipchitz continuous on $\boldsymbol{\Theta}$, we shrink the radius of the neighborhood to zero to obtain a lower bound for the specific instance parametrized by $\boldsymbol{\Theta}^\star$. The details of the proof in Appendix B.7.

7 Experiments

To evaluate the empirical efficiency of the proposed `VA-MNL` algorithm, we conduct a series of experiments on the RiverSwim environment, a standard benchmark specifically designed to validate the exploration–exploitation capabilities of the algorithms for MNL-MDP [Ayoub et al., 2020, Hwang and Oh, 2023, Cho et al., 2024]. The environment isolates the exploration challenge in a simple sequential setting: the agent must choose between a locally attractive action with a small reward and a more difficult direction that is necessary to reach the larger long-term reward. This makes RiverSwim a natural testbed for assessing whether the algorithm performs the better long-horizon decision-making. We defer the descriptions of the RiverSwim environment in Appendix A.

Figure 1 illustrates the cumulative regret across varying problem scales, ranging from $S = 4, H = 12$ to $S = 16, H = 48$. We compare the performance of `VA-MNL` against the state-of-the-art online algorithms: `UCRL-MNL+`, `RRL-MNL`, and `ORRL-MNL` [Cho et al., 2024]. The proposed `VA-MNL` demonstrates significantly superior convergence rates and stability across all scales. Notably, as the state space complexity increases, the performance gap between `VA-MNL` and the baseline algorithms widens. This performance gain is attributed to our novel parameter estimation with the observed information matrix and the second order upper confidence bound that captures the variance to avoid excessive exploration, leading to the rapid stabilization of the regret curves.

Acknowledgments and Disclosure of Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2022-NR071853; RS-2023-00222663; RS-2026-25495485),

by a grant of Korean ARPA-H Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (RS-2024-00512375), and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2021-II211341, Artificial Intelligence Graduate School Program, Chung-Ang University). Garud Iyengar's research was partially supported by ONR grant N000142312374 and NSF grant EFMA-2132142.

References

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

M. Abeille, L. Faury, and C. Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3691–3699. PMLR, 2021.

S. Agrawal, V. Avadhanula, V. Goyal, and A. Zeevi. Thompson sampling for the mnl-bandit. In *Conference on Learning Theory (COLT)*, pages 76–78, 2017.

S. Agrawal, V. Avadhanula, V. Goyal, and A. Zeevi. Mnl-bandit: A dynamic learning approach to assortment selection. *Operations Research*, 67(5):1453–1485, 2019.

S. Amani and C. Thrampoulidis. Ucb-based algorithms for multinomial logistic regression bandits. *Advances in Neural Information Processing Systems*, 34:2913–2924, 2021.

A. Ayoub, Z. Jia, C. Szepesvari, M. Wang, and L. Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

P. L. Bartlett and S. Mendelson. Local rademacher complexities and empirical minimization. *Annals of Statistics*, 34, 2006.

P. Boudart, P. Gaillard, and A. Rudi. Enjoying non-linearity in multinomial logistic bandits: A minimax-optimal algorithm. In *International Conference on Algorithmic Learning Theory*, volume 313 of *Proceedings of Machine Learning Research*. PMLR, 2026.

R. I. Brafman and M. Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

X. Chen, A. Krishnamurthy, and Y. Wang. Robust dynamic assortment optimization in the presence of outlier customers. *Operations Research*, 72(3):999–1015, 2024.

W. Cho, T. Hwang, J. Lee, and M.-h. Oh. Randomized exploration for reinforcement learning with multinomial logistic function approximation. *Advances in Neural Information Processing Systems*, 37:76643–76720, 2024.

C. Dann, T. Lattimore, and E. Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

L. Faury, M. Abeille, C. Calauzènes, and O. Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR, 2020.

L. Faury, M. Abeille, K.-S. Jun, and C. Calauzènes. Jointly efficient and optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*. PMLR, 2022.

E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

T. Hwang and M.-h. Oh. Model-based reinforcement learning with multinomial logistic function approximation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 7971–7979, 2023.

H. Ishfaq, Q. Cui, V. Nguyen, A. Ayoub, Z. Yang, Z. Wang, D. Precup, and L. Yang. Randomized exploration in reinforcement learning with general value function approximation. In *International Conference on Machine Learning*, pages 4607–4616. PMLR, 2021.

R. Jézéquel, P. Gaillard, and A. Rudi. Mixability made efficient: Fast online multiclass logistic regression. *Advances in Neural Information Processing Systems*, 34:23692–23702, 2021.

C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. Provably efficient reinforcement learning with linear function approximation. *Mathematics of Operations Research*, 2023.

K.-S. Jun, A. Bhargava, R. Nowak, and R. Willett. Scalable generalized linear bandits: Online computation and hashing. *Advances in Neural Information Processing Systems*, 30, 2017.

M. Kearns and S. Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49:209–232, 2002.

H. Lee and M.-h. Oh. Minimax optimal reinforcement learning with quasi-optimism. In *The Thirteenth International Conference on Learning Representations*, 2025a.

J. Lee and M.-h. Oh. Demystifying linear mdps and novel dynamics aggregation framework. In *The Twelfth International Conference on Learning Representations*, 2024a.

J. Lee and M.-h. Oh. Nearly minimax optimal regret for multinomial logistic bandit. *Advances in Neural Information Processing Systems*, 37:109003–109065, 2024b.

J. Lee and M.-h. Oh. Improved online confidence bounds for multinomial logistic bandits. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025b.

J. Lee, S.-Y. Yun, and K.-S. Jun. Improved regret bounds of (multinomial) logistic bandits via regret-to-confidence-set conversion. In *International Conference on Artificial Intelligence and Statistics*, pages 4474–4482. PMLR, 2024.

G. Li, L. Shi, Y. Chen, Y. Gu, and Y. Chi. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 17762–17776, 2021.

L.-F. Li, Y.-J. Zhang, P. Zhao, and Z.-H. Zhou. Provably efficient reinforcement learning with multinomial logit function approximation. *Advances in Neural Information Processing Systems*, 37:58539–58573, 2024.

M.-h. Oh and G. Iyengar. Thompson sampling for multinomial logit contextual bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

M.-h. Oh and G. Iyengar. Multinomial logit contextual bandits: Provable optimality and practicality. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 9205–9213, 2021.

J. Park, J. Kwon, and D. Lee. Infinite-horizon reinforcement learning with multinomial logit function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 361–369. PMLR, 2025.

N. Perivier and V. Goyal. Dynamic pricing and assortment under a contextual mnl demand. *Advances in Neural Information Processing Systems*, 35:3461–3474, 2022.

M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

T. Sun and Q. Tran-Dinh. Generalized self-concordant functions: a recipe for newton-type methods. *Mathematical Programming*, 178(1):145–213, 2019.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

Q. Tran-Dinh, Y.-H. Li, and V. Cevher. Composite convex minimization involving self-concordant-like cost functions. In *Modelling, Computation and Optimization in Information Systems and Management Sciences: Proceedings of the 3rd International Conference on Modelling, Computation and Optimization in Information Systems and Management Sciences-MCO 2015-Part I*, pages 155–168. Springer, 2015.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Y. Wang, X. Chen, and Y. Zhou. Near-optimal policies for dynamic multinomial logit assortment selection models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1235–1244, 2018.

Y. Xu and A. Zeevi. Towards optimal problem dependent generalization error bounds in statistical learning theory. *Mathematics of Operations Research*, 50(1):40–67, 2025.

A. Zanette, D. Brandfonbrener, E. Brunskill, M. Pirotta, and A. Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR, 2020.

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

Z. Zhang, Y. Zhou, and X. Ji. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33:15198–15207, 2020.

Z. Zhang, Y. Chen, J. D. Lee, and S. S. Du. Settling the sample complexity of online reinforcement learning. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 5213–5219. PMLR, 2024.

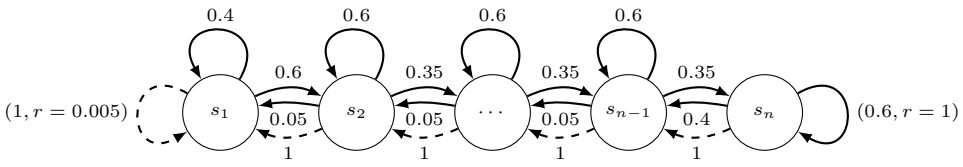


Figure 2: The *RiverSwim* MDP environment with n states. This environment is commonly used to evaluate exploration in RL. Solid arrows represent the swim right action (R); dashed arrows represent the go left action (L).

A Experimental Settings

The RiverSwim environment consists of a linear chain of S states, $\mathcal{S} = \{1, 2, \ldots, S\}$, where the agent begins at the leftmost state. At each state i, the agent can choose between two actions: Left (L) and Right (R). The environment is characterized by a "current" that favors movement toward state 1, creating a challenging exploration problem:

Dynamics: At interior states $i \in \{2, \ldots, S-1\}$, taking action L moves the agent deterministically to $i-1$. Conversely, action R involves a stochastic transition: the agent succeeds in moving to $i+1$ with probability $p_{R\rightarrow} = 0.35$, remains in i with probability $p_{R\circ} = 0.35$, or is pushed back to $i-1$ with probability $p_{R\leftarrow} = 0.30$.

Rewards: A small reward $r_{\text{left}} = 0.005$ is granted for taking L at the leftmost boundary ($i = 1$), providing a "local" trap for sub-optimal policies. A significantly larger reward $r_{\text{right}} = 1.0$ is only obtainable at the rightmost boundary ($i = S$) by taking R, necessitating deep exploration against the current.

Computation Resources: All experiments were conducted on an EC2 c7i.xlarge instance featuring custom 4th Generation Intel Xeon Scalable processors with a dedicated x86_64 architecture to ensure high-performance execution of the numerical simulations.

Hyperparameters: For our proposed algorithm, the hyperparameter δ was tuned over the candidate set $\{0.001, 0.005, 0.01, 0.05, 0.1\}$. For the benchmark algorithms (UCRL-MNL+, RRL-MNL, and ORRL-MNL), we grid-searched the regularization parameter over $\{0.5, 1, 2, 4\}$ and the step size over $\{0.001, 0.01, 0.1, 1\}$.

Variance envelope: Since all the reachable set is size 1 for $\sigma^2(s, L) = 1$ for all state $s \in S$. For the right action $a = R$, $r(s, R) = 1 - 4 \cdot (0.1)^2 = 0.96$. The variance envelop is close to 1 since the distribution of states are nearly equally distributed over the reachable set or concentrated on one state. Even the high variance envelop $\sigma(s, a)$ case our proposed method performs better than the benchmarks.

B Missing Proofs

The appendix is organized around three results: the confidence set for the online estimator (Theorem 2), the variance-adaptive regret upper bound (Theorem 6), and the variance-adaptive lower bound (Theorem 8). Technical tools shared by all three are collected in Appendix C.

Part I: Confidence set for the online estimator.

1. **Moments and curvature of the MNL model.** Proposition 19 identifies the conditional mean and covariance of the next-state indicator with ∇L and $\nabla^2 L$. Proposition 20 shows that the log-sum function is $\sqrt{6}$-self-concordant. Proposition 21, together with Lemmas 17 and 18, lower-bounds the Bregman divergence of the loss by the local curvature at either endpoint, without any dependence on κ.
2. **From online regret to a self-normalized bound.** Lemma 27 bounds the squared gradient projections by the Bregman divergence plus the Fisher-information norm. Combined with Freedman's inequality, Lemma 26 shows that the online regret of the iterates $\{\boldsymbol{\theta}_{\nu-1,h}\}$ controls $\sum_{\nu} \big(\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top (\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}^\star_h)\big)^2$.

3. κ**-free regret of ONS.** Lemma 24 bounds the online regret of the online Newton step by $\widetilde{O}(d)$, using the generalized-projection argument and the elliptical potential lemma (Lemma 22).
4. **Confidence ellipsoid.** Theorem 2 combines Steps 2–3 with the optimality of $\widehat{\boldsymbol{\theta}}_{k,h}$ to obtain $\|\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}^\star_h\|_{\widehat{\mathbf{H}}_{k,h}} \le \beta_k$.
5. **Empirical-to-population Hessian.** Lemma 3 uses the matrix Chernoff bound (Lemma 30) and the covariance inequality (Proposition 40) to show $\widehat{\mathbf{H}}_{k,h} \succeq (1 - e^{-1})\mathbf{H}^\star_{k,h}$. This lets the analysis replace the data-driven matrix $\widehat{\mathbf{H}}_{k,h}$ by the population Fisher information.

Part II: Variance-adaptive regret upper bound.

6. **Kernel expansions.** Lemma 34 gives a second-order expansion of the difference between two MNL transition kernels. The first-order term is a Hessian-weighted inner product, and the remainder is quadratic in $\|\boldsymbol{\Phi}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\|_\infty$. Lemma 35 gives the corresponding first-order (total-variation) bound.
7. **Optimism.** Lemma 4 applies Lemma 34 with the confidence set from Theorem 2 and uses backward induction over h to show $Q^\star_h \le \widehat{Q}_{k,h}$.
8. **Hessian swap.** Lemma 25 replaces the estimated Hessian $\nabla^2 L(\boldsymbol{\Phi}\widehat{\boldsymbol{\theta}}_{k-1,h})$ in the bonus with the true Hessian $\nabla^2 L(\boldsymbol{\Phi}\boldsymbol{\theta}^\star_h)$. It does this by controlling the third derivative of L, and the cost is a lower-order term of order $\kappa^{-1}\beta_{k-1}H \max_s \|\boldsymbol{\Phi}^\top \mathbf{e}_s\|^2_{\mathbf{A}^{-1}_{k-1,h}}$.
9. **Regret decomposition (Theorem 6).** By optimism, the per-episode regret is at most $\widehat{Q}_{n,1} - Q^{\pi_n}_{n,1}$. Unrolling this over h splits the total regret into three pieces:
   (a) *Martingale term* $\sum_{k,h} \Delta_{k,h}$. Its conditional variance is at most $H^2\sigma^2_{k,h}$, so the self-normalized bound (Lemma 31) gives $\widetilde{O}\big(H\sqrt{\sum_{k,h}\sigma^2_{k,h}}\big)$.
   (b) *First-order bonus* $\sum_{k,h} \beta_{k-1}\alpha_{k,h}$. Cauchy–Schwarz splits it into a trace term and a variance term. The trace term is bounded by the matrix potential lemma (Lemma 33). The variance term is the conditional variance of $\widehat{V}_{k,h+1}$ and is bounded by Step 10. Together these give the leading term $\widetilde{O}\big(dH\sqrt{\sum_{k,h}\sigma^2_{k,h}}\big)$.
   (c) *Second-order bonus* $\sum_{k,h} H\beta^2_{k-1}\chi_{k,h}$. Using $\widehat{\mathbf{H}}_{k,h} \succeq \kappa\mathbf{A}_{k,h}$ and the elliptical potential lemma for features (Lemma 23), it is a lower-order term in κ^{-1}.
10. **Total variance of the optimistic values.** Lemma 7 shows that $\sum_{k,h} \mathrm{Var}[\widehat{V}_{k,h+1}(s_{k,h+1}) \mid s_{k,h}, a_{k,h}]$ is at most $O(H)\sum_{k,h}\sigma^2_{k,h}$ plus a burn-in term of order $H^4\kappa^{-2}\beta^4_{K-1}d\log K$. The proof goes as follows:
    - The variance sum is rewritten as the second moment of a telescoping martingale.
    - The Bellman residual of $\widehat{V}$ is decomposed into a reward fluctuation (bounded by $\sigma^2_{k,h}$), a kernel-estimation error (Lemma 35) and the bonus terms (Lemma 25).
    - The conditional expectations are converted to realized sums by the Chernoff bound (Lemma 29).
11. **Problem-dependent rate.** Corollary 11 follows from $\sigma^2_{k,h} \le 1$.

Part III: Variance-adaptive lower bound.

12. **Hard instance and reduction.** We build a chain MNL-MDP in which, at each step, the agent either moves to an absorbing rewarding state or continues along the chain. The success probability depends on how well the action matches a hidden sign vector $\mathbf{u}_h \in \{\pm\sqrt{\Delta}\}^{d-1}$. Lemma 12 lower-bounds the regret by $\frac{H\epsilon}{8(d-1)}$ times the expected number of sign errors, which reduces the problem to $(d-1)\lfloor H/2\rfloor$ binary identification problems.

13. **Variance-weighted information.** Lemma 39 bounds the KL divergence between two interaction measures by $\frac{1}{2}\sum_{k,h}\Delta_{k,h}^2\sigma_{k,h}^2$ plus a cubic remainder, which is controlled by self-concordance and κ. Lemma 15 keeps $\sigma_{k,h}^2$ comparable across nearby parameters.
14. **Averaging argument (Theorem 13).** Pinsker's inequality is applied to each pair of instances that differ in one sign. Averaging over all sign patterns (an Assouad-type argument), choosing $\epsilon \propto (d-1)\sqrt{\Xi}/(KH)$, and letting $\Delta \to 0$ gives an expected lower bound of order $Hd\sqrt{\mathbb{E}\sum_{k,h}\sigma_{k,h}^2}$.
15. **From expectation to high probability.** The anti-concentration inequality for bounded nonnegative variables (Lemma 32, proved via the entropic variational principle, Proposition 41) turns the expected bound into one in terms of the realized $\sum_{k,h}\sigma_{k,h}^2$, holding with probability at least δ.

Technical tools (Appendix C).

- Self-concordance: Lemma 15, Propositions 16–21.
- Potential lemmas: Lemmas 22, 23 and 33.
- Concentration: Freedman (Lemma 28), scalar and matrix Chernoff (Lemmas 29 and 30), the self-normalized bound (Lemma 31) and anti-concentration (Lemma 32).
- Matrix algebra: Lemmas 37 and 38.

B.1 Proof of Lemma 1

Lemma 1 (Quadratic Error Approximation for the True Hessian)**.** *Let* $\mathbf{g}_{\nu,h} := \nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})$. *Then,*

$$\mathbb{E}\big[\mathbf{g}_{\nu,h}\mathbf{g}_{\nu,h}^\top \,\big|\, \mathcal{F}_{\nu,h}\big] - \nabla^2\ell_{\nu,h}(\boldsymbol{\theta}_h^\star) = \big(\nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) - \nabla\ell_{\nu,h}(\boldsymbol{\theta}_h^\star)\big)\big(\nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) - \nabla\ell_{\nu,h}(\boldsymbol{\theta}_h^\star)\big)^\top. \tag{5}$$

Let $\boldsymbol{\xi}_{\nu,h} := \mathbf{e}_{s_{\nu,h+1}} - \mathbb{E}[\mathbf{e}_{s_{\nu,h+1}} \mid \mathcal{F}_{\nu,h}]$ denote the residual. By Proposition 19, $\boldsymbol{\xi}_{\nu,h} = \mathbf{e}_{s_{\nu,h+1}} - \nabla L_{\mathcal{S}_h(s_h,a_h)}\big(\boldsymbol{\Phi}(s_{k,h},a_{k,h})\boldsymbol{\theta}_h^\star\big)$ and $\boldsymbol{\xi}_{\nu,h}^\top\boldsymbol{\Phi}(s_{\nu,h},a_{\nu,h}) = \nabla\ell_{\nu,h}(\boldsymbol{\theta}_h^\star)$. Thus,

$$\nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) = \nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) - \nabla\ell_{\nu,h}(\boldsymbol{\theta}_h^\star) + \boldsymbol{\xi}_{\nu,h}^\top\boldsymbol{\Phi}(s_{\nu,h},a_{\nu,h}).$$

Taking the conditional expectation given $\mathcal{F}_{\nu,h}$, and using $\mathbb{E}[\boldsymbol{\xi}_{\nu,h} \mid \mathcal{F}_{\nu,h}] = \mathbf{0}$ together with $\boldsymbol{\Phi}(s_{\nu,h},a_{\nu,h})^\top\mathbb{E}[\boldsymbol{\xi}_{\nu,h}\boldsymbol{\xi}_{\nu,h}^\top \mid \mathcal{F}_{\nu,h}]\boldsymbol{\Phi}(s_{\nu,h},a_{\nu,h}) = \nabla^2\ell_{\nu,h}(\boldsymbol{\theta}_h^\star)$ (Proposition 19) completes the proof.

B.2 Proof of Theorem 2

Theorem 2 (Self-Normalized Bound for the Estimator)**.** *Define* $C_{\boldsymbol{\Phi},\boldsymbol{\theta}} := (e-1)(6 + 8B_{\boldsymbol{\Phi}}B_{\boldsymbol{\theta}} + 2B_{\boldsymbol{\Phi}}^2B_{\boldsymbol{\theta}}^2)$. *Set the learning rate* $\eta_{\boldsymbol{\Phi},\boldsymbol{\theta}} := (e-1)(3 + 4B_{\boldsymbol{\Phi}}^2B_{\boldsymbol{\theta}}^2)$. *For the estimator* $\widehat{\boldsymbol{\theta}}_{k,h}$ *defined in* (6)*, with probability at least* $1-3\delta$, $\big\|\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{k,h}} \le \beta_k$, *for all* $k \in [K]$ *where* $\beta_k = \sqrt{\epsilon}B_{\boldsymbol{\theta}} + \sqrt{\epsilon B_{\boldsymbol{\theta}}^2 + 4\gamma_k}$ *and* $\gamma_k = \tilde{O}(dB_{\boldsymbol{\Phi}}^4B_{\boldsymbol{\theta}}^4)$.

Proof. Since $(a+b)^2 \le 2a^2 + 2b^2$, for each $\nu \in [k]$,

$$\begin{aligned}
&\|\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h^\star\|_{\widehat{\mathbf{H}}_{\nu,h}-\widehat{\mathbf{H}}_{\nu-1,h}}^2 \\
&= \big(\nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})(\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h^\star)\big)^2 \\
&\le 2\big(\nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})(\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_{\nu-1,h})\big)^2 + 2\big(\nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})(\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star)\big)^2.
\end{aligned}$$

Summing up over $\nu \in [k]$,

$$\begin{aligned}
&\big\|\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{k,h}}^2 - \epsilon\big\|\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h^\star\big\|_2^2 \\
&\le 2\sum_{\nu=1}^{k}\big(\nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})(\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_{\nu-1,h})\big)^2 + 2\sum_{\nu=1}^{k}\big(\nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})(\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star)\big)^2.
\end{aligned}$$

Because $\widehat{\boldsymbol{\theta}}_{k,h}$ is the minimizer,

$$\sum_{\nu=1}^{k}\big(\nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})(\widehat{\boldsymbol{\theta}}_{k,h}-\boldsymbol{\theta}_{\nu-1,h})\big)^2+\frac{\epsilon}{2}\big\|\widehat{\boldsymbol{\theta}}_{k,h}\big\|_2^2$$
$$\leq\sum_{\nu=1}^{k}\big(\nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})(\boldsymbol{\theta}_h^\star-\boldsymbol{\theta}_{\nu-1,h})\big)^2+\frac{\epsilon}{2}\big\|\boldsymbol{\theta}_h^\star\big\|_2^2.$$

Thus,

$$\big\|\widehat{\boldsymbol{\theta}}_{k,h}-\boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{k,h}}^2\leq\epsilon\big(\|\widehat{\boldsymbol{\theta}}_{k,h}-\boldsymbol{\theta}_h^\star\|_2^2-\|\widehat{\boldsymbol{\theta}}_{k,h}\|_2^2+\|\boldsymbol{\theta}_h^\star\|_2^2\big)+4\sum_{\nu=1}^{k}\big(\nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})(\boldsymbol{\theta}_{\nu-1,h}-\boldsymbol{\theta}_h^\star)\big)^2.$$

By Cauchy Schwartz inequality,

$$\begin{aligned}\|\widehat{\boldsymbol{\theta}}_{k,h}-\boldsymbol{\theta}_h^\star\|_2^2-\|\widehat{\boldsymbol{\theta}}_{k,h}\|_2^2+\|\boldsymbol{\theta}_h^\star\|_2^2=&-2\boldsymbol{\theta}_h^\star\widehat{\boldsymbol{\theta}}_{k,h}+2\|\boldsymbol{\theta}_h^\star\|_2^2\\=&2\big(\boldsymbol{\theta}_h^\star-\widehat{\boldsymbol{\theta}}_{k,h}\big)^\top\boldsymbol{\theta}_h^\star\\\leq&2\big\|\boldsymbol{\theta}_h^\star-\widehat{\boldsymbol{\theta}}_{k,h}\big\|_{\widehat{\mathbf{H}}_{k,h}}\big\|\boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{k,h}^{-1}}\\\leq&2\epsilon^{-1/2}B_{\boldsymbol{\theta}}\big\|\boldsymbol{\theta}_h^\star-\widehat{\boldsymbol{\theta}}_{k,h}\big\|_{\widehat{\mathbf{H}}_{k,h}}.\end{aligned}$$

Thus, we obtain,

$$\big\|\widehat{\boldsymbol{\theta}}_{k,h}-\boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{k,h}}^2\leq2B_{\boldsymbol{\theta}}\sqrt{\epsilon}\big\|\widehat{\boldsymbol{\theta}}_{k,h}-\boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{k,h}}+4\sum_{\nu=1}^{k}\big(\nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})(\boldsymbol{\theta}_{\nu-1,h}-\boldsymbol{\theta}_h^\star)\big)^2.$$

For $a,b>0$,

$$x^2\leq2ax+b,\ \ x\geq0\Longrightarrow x\leq a+\sqrt{a^2+b}.$$

Thus,

$$\big\|\widehat{\boldsymbol{\theta}}_{k,h}-\boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{k,h}}\leq\sqrt{\epsilon}B_{\boldsymbol{\theta}}+\sqrt{\epsilon B_{\boldsymbol{\theta}}^2+4\sum_{\nu=1}^{k}\big(\nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})(\boldsymbol{\theta}_{\nu-1,h}-\boldsymbol{\theta}_h^\star)\big)^2},$$

Applying Lemma 26,

$$\begin{aligned}&\sum_{\nu=1}^{k}\big(\nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})(\boldsymbol{\theta}_{\nu-1,h}-\boldsymbol{\theta}_h^\star)\big)^2\\&\leq2C_{\boldsymbol{\Phi},\boldsymbol{\theta}}\bigg(\sum_{\nu=1}^{k}\big(\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})-\ell_{\nu,h}(\boldsymbol{\theta}_h^\star)\big)+\frac{16B_{\boldsymbol{\Phi}}^2B_{\boldsymbol{\theta}}^2}{C_{\boldsymbol{\Phi},\boldsymbol{\theta}}}\log\frac{d}{\delta}+2(e-1)(6+8B_{\boldsymbol{\Phi}}B_{\boldsymbol{\theta}})\log\frac{1}{\delta}\bigg)\\&\leq2C_{\boldsymbol{\Phi},\boldsymbol{\theta}}\bigg(\sum_{\nu=1}^{k}\big(\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})-\ell_{\nu,h}(\boldsymbol{\theta}_h^\star)\big)\bigg)+32B_{\boldsymbol{\Phi}}^2B_{\boldsymbol{\theta}}^2\log\frac{d}{\delta}+4C_{\boldsymbol{\Phi},\boldsymbol{\theta}}^2\log\frac{1}{\delta}\end{aligned}$$

By Lemma 24, we obtain

$$\begin{aligned}&\sum_{\nu=1}^{k}\big(\nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})(\boldsymbol{\theta}_{\nu-1,h}-\boldsymbol{\theta}_h^\star)\big)^2\\&\leq2C_{\boldsymbol{\Phi},\boldsymbol{\theta}}\bigg(\frac{2B_{\boldsymbol{\theta}}^2\epsilon}{\eta_{\boldsymbol{\Phi},\boldsymbol{\theta}}}+d\eta_{\boldsymbol{\Phi},\boldsymbol{\theta}}\log\frac{k+1}{d}+\frac{8B_{\boldsymbol{\Phi}}^2B_{\boldsymbol{\theta}}^2}{\eta_{\boldsymbol{\Phi},\boldsymbol{\theta}}}\log\frac{1}{\delta}\bigg)+32B_{\boldsymbol{\Phi}}^2B_{\boldsymbol{\theta}}^2\log\frac{d}{\delta}+4C_{\boldsymbol{\Phi},\boldsymbol{\theta}}^2\log\frac{1}{\delta}\\&\leq\frac{4C_{\boldsymbol{\Phi},\boldsymbol{\theta}}B_{\boldsymbol{\theta}}^2\epsilon}{\eta_{\boldsymbol{\Phi},\boldsymbol{\theta}}}+2dC_{\boldsymbol{\Phi},\boldsymbol{\theta}}\eta_{\boldsymbol{\Phi},\boldsymbol{\theta}}\log\frac{k+1}{d}+\bigg(\frac{16C_{\boldsymbol{\Phi},\boldsymbol{\theta}}B_{\boldsymbol{\Phi}}^2B_{\boldsymbol{\theta}}^2}{\eta_{\boldsymbol{\Phi},\boldsymbol{\theta}}}+4C_{\boldsymbol{\Phi},\boldsymbol{\theta}}^2\bigg)\log\frac{1}{\delta}+32B_{\boldsymbol{\Phi}}^2B_{\boldsymbol{\theta}}^2\log\frac{d}{\delta},\end{aligned}$$

where the last term is defined as γ_k. Thus, setting $\beta_k=\sqrt{\epsilon}B_{\boldsymbol{\theta}}+\sqrt{\epsilon B_{\boldsymbol{\theta}}^2+4\gamma_k}$, completes the proof. □

B.3 Proof of Lemma 3

Lemma 3 (Information Matrix Dominates the True Hessian)**.** *For any step $h \in [H]$, if the regularization parameter is set to the required value $\epsilon = B_{\boldsymbol{\Phi}}^2 + 4B_{\boldsymbol{\Phi}}^2 \log(d/\delta)$, then with probability at least $1-\delta$, we have $\widehat{\mathbf{H}}_{k,h} \succeq (1-e^{-1})\mathbf{H}_{k,h}^\star$ for all $k \in [K]$*

Proof. For each $h \in [H]$, let $\mathcal{F}_{\nu,h}$ denote the σ-algebra generated by $\{s_{\nu',h}, a_{\nu',h} : h \in [H], \nu' \in [\nu-1]\}$ and $\{s_{h'}^{(\nu)}, a_{h'}^{(\nu)} : h' \in [h]\}$. For $\nu \in [k]$ and $h \in [H]$ define $\boldsymbol{\eta}_{\nu,h}^\star := \boldsymbol{\Phi}(s_{\nu,h}, a_{\nu,h})\boldsymbol{\theta}_h^\star$ and $\boldsymbol{\eta}_{\nu,h} := \boldsymbol{\Phi}(s_{\nu,h}, a_{\nu,h})\boldsymbol{\theta}_{k-1,h}$. Following the argument in (27) and (28),

$$\lambda_{\max}\Big(\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top\Big) \le 4B_{\boldsymbol{\Phi}}^2.$$

Since

$$\min_{\mathbf{x}:\|\mathbf{x}\|_2\le 1} \mathbf{x}^\top \nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top \mathbf{x} = \min_{\mathbf{x}:\|\mathbf{x}\|_2\le 1} \Big(\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top \mathbf{x}\Big)^2 \ge 0,$$

we can use Matrix Chernoff bound (Lemma 30) to derive with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{\nu=1}^{k} \nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top \\
&\succeq (1-e^{-1})\sum_{\nu=1}^{k} \mathbb{E}\big[\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top \big| \mathcal{F}_{\nu,h}\big] - \log\frac{d}{\delta}\mathbf{I}_d.
\end{aligned}$$

for all $k \in [K]$. By the covariance inequality (Proposition 40), for each $\nu \in [k]$,

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\mathbf{e}_{s_{\nu,h+1}} - \nabla L_{\nu,h}\big(\boldsymbol{\eta}_{\nu,h}\big)\Big)\Big(\mathbf{e}_{s_{\nu,h+1}} - \nabla L_{\nu,h}\big(\boldsymbol{\eta}_{\nu,h}\big)\Big)^\top \Big| \mathcal{F}_{\nu,h}\Big] \\
&\succeq \mathbb{E}\Big[\Big(\mathbf{e}_{s_{\nu,h+1}} - \mathbb{E}\big[\mathbf{e}_{s_{\nu,h+1}}\big|\mathcal{F}_{\nu,h}\big]\Big)\Big(\mathbf{e}_{s_{\nu,h+1}} - \mathbb{E}\big[\mathbf{e}_{s_{\nu,h+1}}\big|\mathcal{F}_{\nu,h}\big]\Big)^\top \Big| \mathcal{F}_{\nu,h}\Big] \\
&= \nabla^2 L_{\nu,h}(\boldsymbol{\eta}_{\nu,h}^\star)
\end{aligned}$$

where the last equality uses Proposition 19. Because

$$\nabla^2 \ell_{\nu,h}(\boldsymbol{\theta}_h^\star) = \boldsymbol{\Phi}(s_{\nu,h}, a_{\nu,h})^\top \nabla^2 L_{\nu,h}(\boldsymbol{\eta}_{\nu,h}^\star)\boldsymbol{\Phi}(s_{\nu,h}, a_{\nu,h}),$$

we obtain,

$$\sum_{\nu=1}^{k} \nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top \succeq (1-e^{-1})\sum_{\nu=1}^{k} \nabla^2 \ell_{\nu,h}(\boldsymbol{\theta}_h^\star) - 4B_{\boldsymbol{\Phi}}^2 \log\frac{d}{\delta}\mathbf{I}_d.$$

Thus,

$$\begin{aligned}
\widehat{\mathbf{H}}_{k,h} &:= \sum_{\nu=1}^{k} \nabla \ell_{\nu,h}(\theta_{\nu-1,h})\nabla \ell_{\nu,h}(\theta_{\nu-1,h})^\top + \epsilon I_d \\
&\succeq (1-e^{-1})\sum_{\nu=1}^{k} \nabla^2 \ell_{\nu,h}(\theta_h^\star) + (\epsilon - 4B_\Phi^2 \log\frac{d}{\delta})I_d \\
&\succeq (1-e^{-1})\sum_{\nu=1}^{k} \nabla^2 \ell_{\nu,h}(\theta_h^\star) + (1-e^{-1})B_\Phi^2 I_d,
\end{aligned}$$

where the last inequality holds by $\epsilon = (1-e^{-1})B_\Phi^2 + 4B_\Phi^2 \log\frac{d}{\delta}$. □

B.4 Proof of Lemma 4

Lemma 4 (Optimism of the Second Order UCB)**.** *For any $(s,a) \in \mathcal{S}\times\mathcal{A}$, with probability at least $1-3\delta$, $Q_{\star,h}(s,a) \le \widehat{Q}_{k,h}(s,a)$, for all $k \in [K]$ and $h \in [H]$.*

Proof. For the base case $h = H$, we have

$$Q^\star_H(s,a) = r(s,a) = \widehat{Q}_H(s,a).$$

Suppose $Q^\star_{h+1}(s,a) \le \widehat{Q}_{k,h+1}(s,a)$ holds. Then, by definition of $Q^\star_h$,

$$\begin{aligned}
Q^\star_h(s,a) =& \big(r + P_h V^\star_{h+1}\big)(s,a) \\
=& r(s,a) + \mathbb{E}\Big[V^\star_{h+1}(s_{h+1})\Big|s,a\Big] \\
=& r(s,a) + \mathbb{E}\Big[\max_{a_{h+1}\in\mathcal{A}} Q^\star_{h+1}(s', a_{h+1})\Big|s,a\Big] \\
\le& r(s,a) + \mathbb{E}\Big[\max_{a_{h+1}\in\mathcal{A}} \widehat{Q}_{k,h+1}(s', a_{h+1})\Big|s,a\Big].
\end{aligned}$$

By definition $\widehat{V}_{k,h+1}(s') := \max_{a_{h+1}\in\mathcal{A}} \widehat{Q}_{k,h+1}(s', a_{h+1})$, and

$$\mathbb{E}\Big[\max_{a_{h+1}\in\mathcal{A}} \widehat{Q}_{k,h+1}(s', a_{h+1})\Big|s,a\Big] = \sum_{s'\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(s')\mathbb{P}(\mathbf{e}_{s'}|s,a,\boldsymbol{\theta}^\star_h)$$

By the second order bound for the multinomial kernel difference (Lemma 34),

$$\begin{aligned}
&\sum_{s'\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(s')\mathbb{P}(\mathbf{e}_{s'}|s,a,\boldsymbol{\theta}^\star_h) \\
&\le \sum_{s'\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(s')\mathbb{P}(\mathbf{e}_{s'}|s,a,\widehat{\boldsymbol{\theta}}_{k-1,h}) \\
&+ \sum_{s'\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(s')\mathbf{e}_{s'}^\top \nabla^2 L_{\mathcal{S}_h(s,a)}(\boldsymbol{\Phi}(s,a)\widehat{\boldsymbol{\theta}}_{k-1,h})\boldsymbol{\Phi}(s,a)\big(\boldsymbol{\theta}^\star_h - \widehat{\boldsymbol{\theta}}_{k-1,h}\big) \\
&+ \max_{s'\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(s') \max_{s'\in\mathcal{S}_{k,h}} \big((\boldsymbol{\theta}^\star_h - \widehat{\boldsymbol{\theta}}_{k-1,h})^\top \boldsymbol{\Phi}(s,a)^\top \mathbf{e}_{s'}\big)^2.
\end{aligned} \tag{12}$$

By Cauchy-Schwartz inequality,

$$\begin{aligned}
&\sum_{s'\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(s')\mathbf{e}_{s'}^\top \nabla^2 L_{\mathcal{S}_h(s,a)}(\boldsymbol{\Phi}(s,a)\widehat{\boldsymbol{\theta}}_{k-1,h})\boldsymbol{\Phi}(s,a)\big(\boldsymbol{\theta}^\star_h - \widehat{\boldsymbol{\theta}}_{k-1,h}\big) \\
&\le \Big\| \sum_{s'\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(s')\boldsymbol{\Phi}(s,a)^\top \nabla^2 L_{\mathcal{S}_h(s,a)}(\boldsymbol{\Phi}(s,a)\widehat{\boldsymbol{\theta}}_{k-1,h})\mathbf{e}_{s'}\Big\|_{\widehat{\mathbf{H}}^{-1}_{k-1,h}} \big\|\boldsymbol{\theta}^\star_h - \widehat{\boldsymbol{\theta}}_{k-1,h}\big\|_{\widehat{\mathbf{H}}_{k-1,h}} \\
&\le \beta_{k-1}\Big\| \sum_{s'\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(s')\boldsymbol{\Phi}(s,a)^\top \nabla^2 L_{\mathcal{S}_h(s,a)}(\boldsymbol{\Phi}(s,a)\widehat{\boldsymbol{\theta}}_{k-1,h})\mathbf{e}_{s'}\Big\|_{\widehat{\mathbf{H}}^{-1}_{k-1,h}},
\end{aligned}$$

where the last inequality uses the confidence bound for $\widehat{\boldsymbol{\theta}}_{k-1,h}$ (Theorem 2). Similarly,

$$\begin{aligned}
\max_{s'\in\mathcal{S}_{k,h}} \big((\boldsymbol{\theta}^\star_h - \widehat{\boldsymbol{\theta}}_{k-1,h})^\top \boldsymbol{\Phi}(s,a)^\top \mathbf{e}_{s'}\big)^2 \le& \big\|\boldsymbol{\theta}^\star_h - \widehat{\boldsymbol{\theta}}_{k-1,h}\big\|^2_{\widehat{\mathbf{H}}_{k-1,h}} \max_{s'\in\mathcal{S}_{k,h}} \big\|\boldsymbol{\Phi}(s,a)^\top \mathbf{e}_{s'}\big\|^2_{\widehat{\mathbf{H}}_{k-1,h}} \\
\le& \beta^2_{k-1} \max_{s'\in\mathcal{S}_{k,h}} \big\|\boldsymbol{\Phi}(s,a)^\top \mathbf{e}_{s'}\big\|^2_{\widehat{\mathbf{H}}_{k-1,h}},
\end{aligned}$$

Thus,

$$\begin{aligned}
&\sum_{s'\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(s')\mathbb{P}(\mathbf{e}_{s'}|s,a,\boldsymbol{\theta}^\star_h) \\
&\le \sum_{s'\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(s')\mathbb{P}(\mathbf{e}_{s'}|s,a,\widehat{\boldsymbol{\theta}}_{k-1,h}) \\
&+ \beta_{k-1}\Big\| \sum_{\tilde{s}\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(\tilde{s})\boldsymbol{\Phi}(s,a)^\top \nabla^2 L_{\mathcal{S}_h(s,a)}\big(\boldsymbol{\Phi}(s,a)\widehat{\boldsymbol{\theta}}_{k-1,h}\big)\mathbf{e}_{\tilde{s}}\Big\|_{\widehat{\mathbf{H}}^{-1}_{k-1,h}} \\
&+ \beta^2_{k-1} \sup_{s'\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(s') \max_{s'\in\mathcal{S}_{k,h}} \big\|\mathbf{e}_{s'}^\top \boldsymbol{\Phi}(s,a)\big\|^2_{\widehat{\mathbf{H}}^{-1}_{k-1,h}},
\end{aligned}$$

which implies

$$\begin{aligned}
Q_h^\star(s,a) \leq & r(s,a) + \mathbb{E}\Big[\max_{a_{h+1}\in\mathcal{A}} \widehat{Q}_{k,h+1}(s',a_{h+1})\Big|s,a\Big] \\
\leq & r(s,a) + \sum_{s'\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(s')\mathbb{P}(\mathbf{e}_{s'}|s,a,\widehat{\boldsymbol{\theta}}_{k-1,h}) \\
& + \beta_{k-1}\Big\|\sum_{\tilde{s}\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(\tilde{s})\boldsymbol{\Phi}(s,a)^\top\nabla^2 L_{\mathcal{S}_h(s,a)}\big(\boldsymbol{\Phi}(s,a)\widehat{\boldsymbol{\theta}}_{k-1,h}\big)\mathbf{e}_{\tilde{s}}\Big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}} \\
& + \beta_{k-1}^2 \sup_{s'\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(s') \max_{s'\in\mathcal{S}_{k,h}} \big\|\mathbf{e}_{s'}^\top\boldsymbol{\Phi}(s,a)\big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}}^2.
\end{aligned}$$

By definition, we obtain $Q_h^\star(s,a) \leq \widehat{Q}_{k,h}(s,a)$. By induction argument, the inequality holds for all $h \in [H]$. □

B.5 Proof of Theorem 6

Theorem 9 (Regret Bound of `VA-MNL`)**.** *Let $a_{k,h}$ denote the action selected by `VA-MNL` Algorithm at step $h \in [H]$ in episode $k \in [K]$ and*

$$\sigma_{k,h}^2 := \max_{\mathbf{x}:\|\mathbf{x}\|_\infty\leq 1} \mathbf{x}^\top\nabla^2 L_{k,h}\big(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star\big)\mathbf{x},$$

denote the maximal variance of the multinomial distribution. If the algorithm Π_K is set to `VA-MNL` then for any MNL-MDP $\mathcal{M}$, With probability at least $1-5\delta$,

$$\begin{aligned}
R(\Pi_K) \leq & H\left[\frac{1}{2} + 2\log\left(\frac{\sqrt{1+\sum_{k,h}\sigma_{k,h}^2}}{\delta}\right) + 2\log 2\right]\sqrt{1+\sum_{k=1}^K\sum_{h=1}^H \sigma_{k,h}^2} \\
& + \frac{2\beta_{K-1}\sqrt{2dH\log\frac{K+1}{d}}}{\sqrt{1-e^{-1}}}\sqrt{12eH\sum_{k=1}^K\sum_{h=1}^H\sigma_{k,h}^2 + 1200eH^4\kappa^{-2}\beta_{K-1}^4\big(d\log\frac{K+1}{d} + 3\log\frac{1}{\delta}\big)} \\
& + 16H^2\beta_{K-1}^2 d\log\frac{K+1}{d},
\end{aligned}$$

Remark 10. The trade-off between the variance-adaptive leading term and the burn-in term has practical consequences that depend on the problem scale. In short-session settings with small d and short H, such as mobile news-feed slates ($d \approx$ 10–20, $H \approx$ 3–5), the lower-order term $\mathcal{O}(\kappa^{-1}d^{5/2}H^{5/2})$ is small. The algorithm therefore exits the burn-in phase almost immediately and benefits early from the variance-adaptive rate $\widetilde{\mathcal{O}}\big(dH\sqrt{\sum_{k,h}\sigma_{k,h}^2}\big)$. In complex settings with large d and long H, such as high-dimensional catalog assortments or multi-step itinerary design ($d \geq 100$, $H \geq 20$), the $d^{5/2}H^{5/2}$ factor lengthens the burn-in period. Substantially more episodes ($K \geq K_{\text{burn-in}}$) are then needed before the variance-adaptive term dominates the regret. While mitigating this burn-in cost remains a known open challenge across the MNL-MDP literature, we establish the first variance-adaptive regret bound in this setting. By demonstrating that regret can adapt to the cumulative conditional variance rather than relying on worst-case horizon bounds, we provide a theoretical foundation for variance-aware learning in non-linear function approximation.

Proof. By definition of the action-value function and the value function, $V_h(s;\pi_n) = \mathbb{E}_{\tilde{a}_{n,h}\sim\pi_n(s,h)}[Q_h(s,\tilde{a}_{n,h})]$ for $s \in \mathcal{S}$. Let $a_{n,h} := \operatorname{argmax}_{a\in\mathcal{A}} \widehat{Q}_{n,h}(s_{n,h},a)$ denote the action chosen by Algorithm `VA-MNL` at step h and episode n. Because `VA-MNL` chooses the action deterministically, we have $V_h(s_{n,h};\pi_n) = Q_h(s_{n,h},a_{n,h};\pi_n)$. Thus the regret at episode $n \in [K]$,

$$\begin{aligned}
& V_1^\star(s_{n,1}) - V_{n,1}(s_{n,1};\pi_n) \\
& = Q_1^\star\big(s_{n,1},\pi_\star(s_{n,1})\big) - Q_{n,1}(s_{n,1},a_{n,1};\pi_n).
\end{aligned}$$

By the optimism principle for $\widehat{Q}_{n,1}$ (Lemma 4),

$$= Q_1^\star\big(s_{n,1},\pi_\star(s_{n,1})\big) - Q_{n,1}(s_{n,1},a_{n,1};\pi_n)$$

$$
\begin{aligned}
&\le \widehat{Q}_{n,1}\big(s_{n,1}, \pi_\star(s_{n,1})\big) - Q_{n,1}\big(s_{n,1}, a_{n,1}; \pi_n\big) \\
&\le \widehat{Q}_{n,1}\big(s_{n,1}, a_{n,1}\big) - Q_{n,1}\big(s_{n,1}, a_{n,1}; \pi_n\big),
\end{aligned}
$$

where the last inequality holds by $a_{n,1} := \operatorname*{argmax}_{a\in\mathcal{A}} \widehat{Q}_{n,1}(s_{n,1}, a)$. For each $n \in [K]$ and $h \in [H]$, we write the feature matrix $\boldsymbol{\Phi}_{n,h} := \boldsymbol{\Phi}(s_{n,h}, a_{n,h})$ and the transition kernels $\mathbb{P}^\star_{n,h}(\cdot) := \mathbb{P}(\cdot|s_{n,h}, a_{n,h}, \boldsymbol{\theta}^\star_h)$ and $\widehat{\mathbb{P}}_{n,h}(\cdot) := \mathbb{P}(\cdot|s_{n,h}, a_{n,h}, \widehat{\boldsymbol{\theta}}_{n-1,h})$. For simplicity, we define $\mathcal{S}_{n,h} := \mathcal{S}_h(s_{n,h}, a_{n,h})$, the set of states reachable from $(s_{n,h}, a_{n,h})$ in step h in episode n, and let $L_{n,h} := L_{\mathcal{S}_h(s_{n,h},a_{n,h})}$ the corresponding log-sum-exp function. We also write $\mathbf{V}^\star_{n,h} := \nabla^2 L_{n,h}(\boldsymbol{\Phi}_{n,h}\boldsymbol{\theta}^\star_h)$ and $\widehat{\mathbf{V}}_{n,h} := \nabla^2 L_{n,h}(\boldsymbol{\Phi}_{n,h}\widehat{\boldsymbol{\theta}}_{n,h})$. Recall that $\mathbf{A}_{n,h} := \sum_{\nu=1}^n \boldsymbol{\Phi}^\top_{\nu,h}\boldsymbol{\Phi}_{\nu,h} + B^2_{\boldsymbol{\Phi}}\mathbf{I}_d$ and $\mathbf{H}_{n,h} := \sum_{\nu=1}^n \nabla^2 \ell_{\nu,h}(\boldsymbol{\theta}^\star_h) + B^2_{\boldsymbol{\Phi}}\mathbf{I}_d$. By definition of $\widehat{Q}_{n,1}$,

$$
\begin{aligned}
&\widehat{Q}_{n,1}\big(s_{n,1}, a_{n,1}\big) - Q_{n,1}\big(s_{n,1}, a_{n,1}; \pi_n\big) \\
&\le \sum_{s'\in\mathcal{S}_{n,1}} \widehat{V}_{n,2}(s')\big(\widehat{\mathbb{P}}(\mathbf{e}_{s'}) - \mathbb{P}^\star_{n,h}(\mathbf{e}_{s'})\big) \\
&\quad + \sum_{s'\in\mathcal{S}_{n,1}} \Big(\widehat{V}_{n,2}(s') - V_{n,2}(s')\Big)\mathbb{P}^\star_{n,h}(\mathbf{e}_{s'}) \\
&\quad + \beta_{n-1}\Big\| \sum_{s'\in\mathcal{S}_{n,1}} \widehat{V}_{n,2}(s')\cdot \mathbf{e}^\top_{s'}\widehat{\mathbf{V}}_{n,1}\boldsymbol{\Phi}_{n,1}\Big\|_{\widehat{\mathbf{H}}^{-1}_{n-1,1}} \\
&\quad + \beta^2_{n-1} \max_{s'\in\mathcal{S}_{n,h}} \widehat{V}_{n,2}(s') \max_{s'\in\mathcal{S}_{n,1}} \big\|\boldsymbol{\Phi}^\top_{n,1}\mathbf{e}_{s'}\big\|^2_{\widehat{\mathbf{H}}^{-1}_{n-1,1}}.
\end{aligned}
$$

Using Lemma 25, we can decompose the regret,

$$
\begin{aligned}
&\widehat{Q}_{n,1}\big(s_{n,1}, a_{n,1}\big) - Q_{n,1}\big(s_{n,1}, a_{n,1}; \pi_n\big) \\
&\le \underbrace{\sum_{s'\in\mathcal{S}_{n,1}} \widehat{V}_{n,2}(s')\big(\widehat{\mathbb{P}}(\mathbf{e}_{s'}) - \mathbb{P}^\star_{n,h}(\mathbf{e}_{s'})\big)}_{\text{Term I}} \\
&\quad + \sum_{s'\in\mathcal{S}_{n,1}} \Big(\widehat{V}_{n,2}(s') - V_{n,2}(s')\Big)\mathbb{P}^\star_{n,h}(\mathbf{e}_{s'}) \\
&\quad + \beta_{n-1}\Big\| \sum_{s'\in\mathcal{S}_{n,1}} \widehat{V}_{n,2}(s')\cdot \mathbf{e}^\top_{s'}\mathbf{V}^\star_{n,1}\boldsymbol{\Phi}_{n,1}\Big\|_{\widehat{\mathbf{H}}^{-1}_{n-1,1}} \\
&\quad + 7\beta^2_{n-1} \max_{s'\in\mathcal{S}_{n,h}} \widehat{V}_{n,2}(s') \max_{s'\in\mathcal{S}_{n,1}} \big\|\boldsymbol{\Phi}^\top_{n,1}\mathbf{e}_{s'}\big\|^2_{\widehat{\mathbf{H}}^{-1}_{n-1,1}}
\end{aligned}
$$

We use the second order approximation bound for the transition kernel $\mathbb{P}(\cdot|s, a, \boldsymbol{\theta})$ (Lemma 34) to bound Term I as follows:

$$
\begin{aligned}
&\text{Term I} \\
&\le \sum_{s'\in\mathcal{S}_{n,1}} \widehat{V}_{n,2}(s')\mathbf{e}^\top_{s'}\mathbf{V}^\star_{n,1}\boldsymbol{\Phi}_{n,1}(\widehat{\boldsymbol{\theta}}_{n-1,1} - \boldsymbol{\theta}^\star_1) \\
&+ \max_{s'\in\mathcal{S}_{n,1}} \widehat{V}_{n,2}(s') \max_{s'\in\mathcal{S}_{n,1}} \big(\mathbf{e}^\top_{s'}\boldsymbol{\Phi}_{n,1}(\widehat{\boldsymbol{\theta}}_{n-1,1} - \boldsymbol{\theta}^\star_1)\big)^2 \\
&\le \Big\| \sum_{s'\in\mathcal{S}_{n,1}} \widehat{V}_{n,2}(s')\mathbf{e}^\top_{s'}\mathbf{V}^\star_{n,1}\boldsymbol{\Phi}_{n,1}\Big\|_{\widehat{\mathbf{H}}^{-1}_{n-1,1}} \|\widehat{\boldsymbol{\theta}}_{n-1,1} - \boldsymbol{\theta}^\star_1\|_{\widehat{\mathbf{H}}_{n-1,1}} \\
&+ \max_{s'\in\mathcal{S}_{n,1}} \widehat{V}_{n,2}(s') \max_{s'\in\mathcal{S}_{n,1}} \|\boldsymbol{\Phi}^\top_{n,1}\mathbf{e}_{s'}\|^2_{\widehat{\mathbf{H}}^{-1}_{n-1,1}} \|\widehat{\boldsymbol{\theta}}_{n-1,1} - \boldsymbol{\theta}^\star_1\|^2_{\widehat{\mathbf{H}}_{n-1,1}},
\end{aligned}
$$

where the last inequality holds by Cauchy-Schwarz inequality. Next, the confidence ellipsoid (Theorem 2) implies that

$$
\text{Term I} \le \beta_{n-1}\Big\| \sum_{s'\in\mathcal{S}_{n,1}} \widehat{V}_{n,2}(s')\mathbf{e}^\top_{s'}\mathbf{V}^\star_{n,1}\boldsymbol{\Phi}_{n,1}\Big\|_{\widehat{\mathbf{H}}^{-1}_{n-1,1}}
$$

$$+ \beta_{n-1}^2 \max_{s' \in \mathcal{S}_{n,1}} \widehat{V}_{n,2}(s') \max_{s' \in \mathcal{S}_{n,1}} \|\boldsymbol{\Phi}_{n,1}^\top \mathbf{e}_{s'}\|^2_{\widehat{\mathbf{H}}_{n-1,1}^{-1}}.$$

It follows that

$$\begin{aligned}
&\widehat{Q}_{n,1}\big(s_{n,1}, a_{n,1}\big) - Q_{n,1}\big(s_{n,1}, a_{n,1}; \pi_n\big) \\
&\le \sum_{s' \in \mathcal{S}_{n,1}} \Big(\widehat{V}_{n,2}(s') - V_{n,2}(s')\Big) \mathbb{P}^\star_{n,h}(\mathbf{e}_{s'}) \\
&\quad + 2\beta_{n-1} \Big\| \sum_{s' \in \mathcal{S}_{n,1}} \widehat{V}_{n,2}(s') \cdot \mathbf{e}_{s'}^\top \mathbf{V}^\star_{n,1} \boldsymbol{\Phi}_{n,1} \Big\|_{\widehat{\mathbf{H}}_{n-1,1}^{-1}} \\
&\quad + 8\beta_{n-1}^2 \max_{s' \in \mathcal{S}_{n,h}} \widehat{V}_{n,2}(s') \max_{s' \in \mathcal{S}_{n,1}} \big\|\boldsymbol{\Phi}_{n,1}^\top \mathbf{e}_{s'}\big\|^2_{\widehat{\mathbf{H}}_{n-1,1}^{-1}}.
\end{aligned}$$

Define $\Delta_{n,h} := \sum_{s' \in \mathcal{S}_{n,h}} \big(\widehat{V}_{n,h+1}(s') - V_{n,h+1}(s')\big) \mathbb{P}^\star_{n,h}(\mathbf{e}_{s'}) - \big(\widehat{V}_{n,h+1}(s_{n,h+1}) - V_{n,h+1}(s_{n,h+1})\big)$. Since $s_{n,h+1}$ is sampled from the distribution $\mathbb{P}^\star_{n,h}(\cdot)$, it follows that $\Delta_{n,h} = \mathbb{E}\big[\widehat{V}_{n,h+1}(s_{n,h+1}) - V_{n,h+1}(s_{n,h+1}) \mid s_{n,h}, a_{n,h}\big] - \big(\widehat{V}_{n,h+1}(s_{n,h+1}) - V_{n,h+1}(s_{n,h+1})\big)$, which is a martingale difference. Then we have the following decomposition:

$$\begin{aligned}
&\sum_{s' \in \mathcal{S}_{n,1}} \Big(\widehat{V}_{n,2}(s') - V_{n,2}(s')\Big) \mathbb{P}^\star_{n,h}(\mathbf{e}_{s'}) \\
&\le \widehat{Q}_{n,2}(s_{n,2}, a_{n,2}) - Q_2(s_{n,2}, a_{n,2}; \pi_n) + \Delta_{n,2},
\end{aligned}$$

and the regret at episode $n \in [K]$,

$$\begin{aligned}
&\widehat{Q}_{n,1}\big(s_{n,1}, a_{n,1}\big) - Q_{n,1}\big(s_{n,1}, a_{n,1}; \pi_n\big) \\
&\le \Delta_{n,2} + \widehat{Q}_{n,2}(s_{n,2}, a_{n,2}) - Q_2(s_{n,2}, a_{n,2}; \pi_n) \\
&\quad + 2\beta_{n-1} \underbrace{\Big\| \sum_{s' \in \mathcal{S}_{n,1}} \widehat{V}_{n,2}(s') \cdot \mathbf{e}_{s'}^\top \mathbf{V}^\star_{n,1} \boldsymbol{\Phi}_{n,1} \Big\|_{\widehat{\mathbf{H}}_{n-1,1}^{-1}}}_{:= \alpha_{n,1}} \\
&\quad + 8\beta_{n-1}^2 H \underbrace{\max_{s' \in \mathcal{S}_{n,1}} \big\|\boldsymbol{\Phi}_{n,1}^\top \mathbf{e}_{s'}\big\|^2_{\widehat{\mathbf{H}}_{n-1,1}^{-1}}}_{:= \chi_{n,1}},
\end{aligned}$$

where the last inequality uses $\widehat{V}_{n,h}(s) \le H$ for all $s \in \mathcal{S}_{n,h}$. Recursively, the regret,

$$\begin{aligned}
&\mathrm{R}(\mathcal{M}, \Pi_K) \\
&\le \sum_{n=1}^K \widehat{Q}_{n,1}\big(s_{n,1}, a_{n,1}\big) - Q_{n,1}\big(s_{n,1}, a_{n,1}; \pi_n\big) \\
&\le \sum_{n=1}^K \sum_{h=1}^H \big(\Delta_{n,h} + 2\beta_{n-1}\alpha_{n,h} + 8H\beta_{n-1}^2 \chi_{n,h}\big)
\end{aligned}$$

It is easy to verify $\mathbb{E}[\Delta_{n,h}^2 | s_{n,h}, a_{n,h}] \le H^2 \sigma_{n,h}^2$ and the sum $\sum_{n=1}^K \sum_{h=1}^H \Delta_{n,h} = O(H\sqrt{\sum_{n=1}^K \sum_{h=1}^H \sigma_{n,h}^2})$ by the Freedman's inequality (Lemma 28). Using simple linear algebra manipulation,

$$\alpha_{n,h} \le \sqrt{\mathrm{Tr}\big(\widehat{\mathbf{H}}_{n-1,h}^{-1} \nabla^2 \ell_{n-1,h}(\boldsymbol{\theta}_h^\star)\big)} \Big\| \sum_{\tilde{s} \in \mathcal{S}_{n,h}} \widehat{V}_{n,h+1}(\tilde{s}) \mathbf{e}_{\tilde{s}} \Big\|_{\mathbf{V}^\star_{n,h}}.$$

By the Hessian bound (Lemma 3), we have $H^\star_{n-1,h} \preceq (1 - e^{-1})^{-1} \widehat{H}_{n-1,h}$ and

$$\begin{aligned}
&\sum_{n=1}^K \sum_{h=1}^H \mathrm{Tr}\big(\widehat{\mathbf{H}}_{n-1,h}^{-1} \nabla^2 \ell_{n-1,h}(\boldsymbol{\theta}_h^\star)\big) \\
&\le (1 - e^{-1})^{-1} \sum_{n=1}^K \sum_{h=1}^H \mathrm{Tr}\big((\mathbf{H}^\star_{n-1,h})^{-1} \nabla^2 \ell_{n-1,h}(\boldsymbol{\theta}_h^\star)\big)
\end{aligned}$$

$$\leq \frac{2dH}{1-e^{-1}}\log(K+1),$$

where the last inequality uses the matrix potential lemma (Lemma 33) for each $h \in [H]$. By Cauchy-Schwarz inequality,

$$\begin{aligned}
&\sum_{n=1}^{K}\sum_{h=1}^{H} 2\beta_{n-1}\alpha_{n,h} \\
\leq &\frac{2\beta_{K-1}\sqrt{2dH\log\frac{K+1}{d}}}{\sqrt{1-e^{-1}}}\sqrt{\sum_{n=1}^{K}\sum_{h=1}^{H}\Big\|\sum_{\tilde{s}\in\mathcal{S}_{n,h}}\widehat{V}_{n,2}(\tilde{s})\mathbf{e}_{\tilde{s}}\Big\|^2_{\mathbf{V}^{\star}_{n,h}}}
\end{aligned}$$

With the identity for the covariance of the multinomial distribution (Proposition 19), we obtain the following conditional variance,

$$\Big\|\sum_{\tilde{s}\in\mathcal{S}_{n,h}}\widehat{V}_{n,h+1}(\tilde{s})\mathbf{e}_{\tilde{s}}\Big\|^2_{\mathbf{V}^{\star}_{n,h}} = \mathbb{V}\big[\widehat{V}_{n,h+1}(s_{n,h+1})\big|s_{n,h},a_{n,h}\big].$$

Note that both $\widehat{V}_{n,h}$ and $V_{n,h}$ is in $[0,H]$. Thus, $|\Delta_{n,2}| \leq 2H$, Here, we use a novel bound for the sum of variance of the estimated value function $\widehat{V}_{n,h}$. Lemma 7 implies

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} 2\beta_{k-1}\alpha_{k,h} \\
\leq &\frac{2\beta_{K-1}\sqrt{2dH\log\frac{K+1}{d}}}{\sqrt{1-e^{-1}}}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\Big\|\sum_{\tilde{s}\in\mathcal{S}_{k,h}}\widehat{V}_{k,2}(\tilde{s})\mathbf{e}_{\tilde{s}}\Big\|^2_{\mathbf{V}^{\star}_{k,h}}} \\
\leq &\frac{2\beta_{K-1}\sqrt{2dH\log\frac{K+1}{d}}}{\sqrt{1-e^{-1}}}\sqrt{12eH\sum_{k=1}^{K}\sum_{h=1}^{H}\sigma^2_{k,h}+1200eH^4\kappa^{-2}\beta^4_{K-1}\big(d\log\frac{K+1}{d}+3\log\frac{1}{\delta}\big)}.
\end{aligned}$$

Because $\widehat{H}_{n,h} \succeq (1-e^{-1})\mathbf{H}^{\star}_{n,h} \succeq \kappa\mathbf{A}_{n,h}$, we obtain

$$\chi_{n,h} \leq \kappa^{-1}\max_{s'\in\mathcal{S}_{n,h}}\big\|\boldsymbol{\Phi}^{\top}_{n,h}\mathbf{e}_{s'}\big\|^2_{\mathbf{A}_{n-1,h}}.$$

Applying the elliptical potential lemma for feature matrices (Lemma 23) for each $h \in [H]$,

$$\sum_{n=1}^{K}\sum_{h=1}^{H}\chi_{n,h} \leq 2\kappa^{-1}Hd\log\frac{K+1}{d}.$$

Thus,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}8H\beta^2_{k-1}\chi_{k,h} \leq 16H^2\beta^2_{K-1}d\log\frac{K+1}{d}.$$

Let

$$Z_{k,h+1} := \widehat{V}_{k,h+1}(s_{k,h+1}) - V^{\pi_k}_{k,h+1}(s_{k,h+1}),$$

and let $\mathcal{F}_{k,h}$ be the sigma-field immediately before observing $s_{k,h+1}$. On the optimism event,

$$0 \leq Z_{k,h+1} \leq H.$$

Define the martingale difference

$$\Delta_{k,h+1} := \mathbb{E}[Z_{k,h+1} \mid \mathcal{F}_{k,h}] - Z_{k,h+1}.$$

After flattening (k,h) into a single time index, apply Lemma 29 with dimension 1, $x_t = 1$, $\lambda = 1$, and $\epsilon_{t+1} = \Delta_{t+1}/H$. Then $|\epsilon_{t+1}| \leq 1$ and

$$q_t := \mathbb{E}[\epsilon^2_{t+1} \mid \mathcal{F}_t] = \frac{\mathrm{Var}(Z_{t+1} \mid \mathcal{F}_t)}{H^2} \leq \sigma^2_t,$$

because $Z_{t+1}/H \in [0,1] \subseteq [-1,1]$. By the self-normalized bound (Lemma 31), with probability at least $1-\delta$,

$$\left|\sum_{k,h} \Delta_{k,h+1}\right| \le H\left[\frac{1}{2} + 2\log\left(\frac{\sqrt{1+\sum_{k,h} q_{k,h}}}{\delta}\right) + 2\log 2\right]\sqrt{1+\sum_{k,h} q_{k,h}}.$$

Using $q_{k,h} \le \sigma_{k,h}^2$ gives

$$\left|\sum_{k,h} \Delta_{k,h+1}\right| \le H\left[\frac{1}{2} + 2\log\left(\frac{\sqrt{1+\sum_{k,h} \sigma_{k,h}^2}}{\delta}\right) + 2\log 2\right]\sqrt{1+\sum_{k,h} \sigma_{k,h}^2}.$$

$$\begin{aligned}&\Big|\sum_{k=1}^{K}\sum_{h=1}^{H} \Delta_{k,h+1}\Big| \\ &\le H\left[\frac{1}{2} + 2\log\left(\frac{\sqrt{1+\sum_{k,h} \sigma_{k,h}^2}}{\delta}\right) + 2\log 2\right]\sqrt{1+\sum_{k=1}^{K}\sum_{h=1}^{H} \sigma_{k,h}^2}.\end{aligned}$$

Thus,

$$\begin{aligned}\mathrm{R}(K) &\le H\left[\frac{1}{2} + 2\log\left(\frac{\sqrt{1+\sum_{k,h} \sigma_{k,h}^2}}{\delta}\right) + 2\log 2\right]\sqrt{1+\sum_{k=1}^{K}\sum_{h=1}^{H} \sigma_{k,h}^2} \\ &+ \frac{2\beta_{K-1}\sqrt{2dH\log\frac{K+1}{d}}}{\sqrt{1-e^{-1}}}\sqrt{12eH\sum_{k=1}^{K}\sum_{h=1}^{H}\sigma_{k,h}^2 + 1200eH^4\kappa^{-2}\beta_{K-1}^4\big(d\log\frac{K+1}{d} + 3\log\frac{1}{\delta}\big)} \\ &+ 16H^2\beta_{K-1}^2 d\log\frac{K+1}{d},\end{aligned}$$

which completes the proof. □

Because $\sigma_{k,h}^2 \le 1$, the variance adaptive regret bound implies the following the problem-dependent regret bound.

Corollary 11 (Problem-dependent regret bound of `VA-MNL`)**.** *If the algorithm Π_K is set to `VA-MNL` then for any MNL-MDP $\mathcal{M}$, with probability at least $1-5\delta$,*

$$R_{\Pi_K}(\mathcal{M}_{\boldsymbol{\Theta}^\star}) = O\bigg(dH^{3/2}\sqrt{K\log\frac{KH}{\delta}}\log K + \kappa^{-1}d^{\frac{5}{2}}H^{\frac{5}{2}}\log^{\frac{3}{2}}K\bigg)$$

B.6 Proof of Lemma 7

Lemma 7 (The Optimistic Law of Total Variance)**.** *The sum of variance of the value function $\hat{V}_{k,h+1}$*

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}\Big[\widehat{V}_{k,h+1}(s_{k,h+1})\Big|s_{k,h},a_{k,h}\Big] \\ &\le 12eH\sum_{k=1}^{K}\sum_{h=1}^{H}\sigma_{k,h}^2 + 1200eH^4\kappa^{-2}\beta_{K-1}^4\big(d\log\frac{K+1}{d} + 3\log\frac{1}{\delta}\big).\end{aligned}$$

Proof. Let us write $\boldsymbol{\Phi}_{k,h} := \boldsymbol{\Phi}(s_{k,h},a_{k,h})$. Define $\mathcal{F}_{k,h}$ denote the sigma-algebra generated by $\cup_{\nu=1}^{k-1}\cup_{h=1}^{H}\{s_{\nu,h},a_{\nu,h}\}$ and $\cup_{h'=1}^{h}\{s_{k,h'},a_{k,h'}\}$. For simplicity, we use $\mathbb{E}_{k,h} := \mathbb{E}[\cdot|\mathcal{F}_{k,h}]$ and $\widehat{\mathbb{E}}_{k,h} := [\cdot|\mathcal{F}_{k,h},\widehat{\boldsymbol{\theta}}_{k-1,h}]$ for each $k \in [K]$ and $h \in [H]$. Because $0 \le$

$\sum_{h=1}^{H} \mathbb{V}\Big[\widehat{V}_{k,h+1}(s_{k,h+1})\Big|s_{k,h},a_{k,h}\Big] \le H^3$, by Chernoff bound (Lemma 29),

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{V}\Big[\widehat{V}_{k,h+1}(s_{k,h+1})\Big|s_{k,h},a_{k,h}\Big] \\
&\le (e-1)\sum_{k=1}^{K}\mathbb{E}_{k,1}\Bigg[\sum_{h=1}^{H} \mathbb{V}\Big[\widehat{V}_{k,h+1}(s_{k,h+1})\Big|s_{k,h},a_{k,h}\Big]\Bigg] + H^3\log\frac{1}{\delta}.
\end{aligned} \tag{13}$$

Let us write the martingale difference $\Delta_{k,h} := \widehat{V}_{k,h+1}(s_{k,h+1}) - \mathbb{E}_{k,h}\big[\widehat{V}_{k,h+1}(s_{k,h+1})\big]$. It follows that $\mathbb{E}_{k,h}[\Delta_{k,h}] = 0$ and for $h \neq h'$ we have $\mathbb{E}_{k,1}[\Delta_{k,h}\Delta_{k,h'}] = 0$. Then,

$$\mathbb{E}_{k,1}\Bigg[\sum_{h=1}^{H} \mathbb{V}\Big[\widehat{V}_{k,h+1}(s_{k,h+1})\Big|s_{k,h},a_{k,h}\Big]\Bigg] = \mathbb{E}_{k,1}\sum_{h=1}^{H}\Delta_{k,h}^2 = \mathbb{E}_{k,1}\Bigg[\Big(\sum_{h=1}^{H}\Delta_{k,h}\Big)^2\Bigg].$$

Because $\widehat{V}_{k,H+1}(s) = 0$, for all $s \in \mathcal{S}$,

$$\begin{aligned}
&\sum_{h=1}^{H}\Delta_{k,h} \\
&= \sum_{h=1}^{H}\widehat{V}_{k,h+1}(s_{k,h+1}) - \widehat{V}_{k,h}(s_{k,h}) + \widehat{V}_{k,h}(s_{k,h}) - \mathbb{E}_{k,h}\widehat{V}_{k,h+1}(s_{k,h+1}) \\
&= -\widehat{V}_{k,1}(s_{k,1}) + \sum_{h=1}^{H}\big(\widehat{V}_{k,h}(s_{k,h}) - \mathbb{E}_{k,h}\widehat{V}_{k,h+1}(s_{k,h+1})\big)
\end{aligned}$$

By definition of $\widehat{V}_{k,h}(s_{k,h}) = \widehat{Q}_{k,h}(s_{k,h},a_{k,h})$:

$$\begin{aligned}
\widehat{V}_{k,h}(s_{k,h}) =& r(s_{k,h},a_{k,h}) + \mathbb{E}_{k,h}\big[\widehat{V}_{k,h+1}(s_{k,h+1})\big] \\
&\widehat{\mathbb{E}}_{k,h}\big[\widehat{V}_{k,h+1}(s_{k,h+1})\big] - \mathbb{E}_{k,h}\big[\widehat{V}_{k,h+1}(s_{k,h+1})\big] \\
&+ \beta_{k-1}\Bigg\|\sum_{s'\in\mathcal{S}_{k,h}}\widehat{V}_{k,h+1}(s')\cdot\boldsymbol{\Phi}_{k,h}^{\top}\nabla^2 L_{k,h}\big(\boldsymbol{\Phi}_{k,h}\widehat{\boldsymbol{\theta}}_{k-1,h}\big)\mathbf{e}_{s'}\Bigg\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}} \\
&+ 2\beta_{k-1}^2\sup_{s'\in\mathcal{S}_{k,h}}\widehat{V}_{k,h+1}(s')\max_{s'\in\mathcal{S}_{k,h}}\Big\|\boldsymbol{\Phi}_{k,h}^{\top}\mathbf{e}_{s'}\Big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}}^2.
\end{aligned}$$

Define the martingale difference $\rho_{k,h} := r(s_{k,h},a_{k,h}) - \mathbb{E}_{k,h-1}[r(s_{k,h},a_{k,h})]$. Then,

$$\begin{aligned}
\sum_{h=1}^{H}\Delta_{k,h} =& \sum_{h=1}^{H}\big(r(s_{k,h},a_{k,h}) - \rho_{k,h}\big) - \widehat{V}_{k,1}(s_{k,1}) \\
&+ \sum_{h=1}^{H}\big(\widehat{V}_{k,h}(s_{k,h}) - \mathbb{E}_{k,h}\widehat{V}_{k,h+1}(s_{k,h+1}) - r(s_{k,h},a_{k,h}) + \rho_{k,h}\big)
\end{aligned}$$

By triangular inequality and Cauchy-Schwartz inequality,

$$
\begin{aligned}
&\Big(\sum_{h=1}^{H}\Delta_{k,h}\Big)^2 \\
&\le 2\Big(\sum_{h=1}^{H}\big(r(s_{k,h},a_{k,h})-\rho_{k,h}\big)-\widehat{V}_{k,1}(s_{k,1})\Big)^2 \\
&\quad + 2\Big(\sum_{h=1}^{H}\big(\widehat{V}_{k,h}(s_{k,h})-\mathbb{E}_{k,h}\widehat{V}_{k,h+1}(s_{k,h+1})-r(s_{k,h},a_{k,h})+\rho_{k,h}\big)\Big)^2 \qquad (14)\\
&\le 2\Big(\sum_{h=1}^{H}\big(r(s_{k,h},a_{k,h})-\rho_{k,h}\big)-\widehat{V}_{k,1}(s_{k,1})\Big)^2 \\
&\quad + 2H\sum_{h=1}^{H}\Big(\widehat{V}_{k,h}(s_{k,h})-\mathbb{E}_{k,h}\widehat{V}_{k,h+1}(s_{k,h+1})-r(s_{k,h},a_{k,h})+\rho_{k,h}\Big)^2
\end{aligned}
$$

By definition of $\widehat{V}_{k,h}(s_{k,h})$:

$$
\begin{aligned}
\widehat{V}_{k,h}(s_{k,h}) &:= r(s_{k,h},a_{k,h})+\widehat{\mathbb{E}}_{k,h}\big[\widehat{V}_{k,h+1}(s_{k,h+1})\big] \\
&\quad+\beta_{k-1}\Big\|\sum_{s'\in\mathcal{S}_{k,h}}\widehat{V}_{k,h+1}(s')\cdot\boldsymbol{\Phi}_{k,h}^{\top}\nabla^2 L_{k,h}\big(\boldsymbol{\Phi}_{k,h}\widehat{\boldsymbol{\theta}}_{k-1,h}\big)\mathbf{e}_{s'}\Big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}} \\
&\quad+2\beta_{k-1}^2\sup_{s'\in\mathcal{S}_{k,h}}\widehat{V}_{k,h+1}(s')\max_{s'\in\mathcal{S}_{k,h}}\Big\|\boldsymbol{\Phi}_{k,h}^{\top}\mathbf{e}_{s'}\Big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}}^2,
\end{aligned}
$$

we obtain

$$
\begin{aligned}
&\Big|\widehat{V}_{k,h}(s_{k,h})-\mathbb{E}_{k,h}\widehat{V}_{k,h+1}(s_{k,h+1})-r(s_{k,h},a_{k,h})+\rho_{k,h}\Big| \\
&\le|\rho_{k,h}|+\Big|\widehat{\mathbb{E}}_{k,h}\widehat{V}_{k,h+1}(s_{k,h+1})-\mathbb{E}_{k,h}\widehat{V}_{k,h+1}(s_{k,h+1})\Big| \\
&\quad+\beta_{k-1}\Big\|\sum_{s'\in\mathcal{S}_{k,h}}\widehat{V}_{k,h+1}(s')\cdot\boldsymbol{\Phi}_{k,h}^{\top}\nabla^2 L_{k,h}\big(\boldsymbol{\Phi}_{k,h}\widehat{\boldsymbol{\theta}}_{k-1,h}\big)\mathbf{e}_{s'}\Big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}} \qquad (15)\\
&\quad+2\beta_{k-1}^2\sup_{s'\in\mathcal{S}_{k,h}}\widehat{V}_{k,h+1}(s')\max_{s'\in\mathcal{S}_{k,h}}\Big\|\boldsymbol{\Phi}_{k,h}^{\top}\mathbf{e}_{s'}\Big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}}^2.
\end{aligned}
$$

By the bound for second order confidence bound (Lemma 25),

$$
\begin{aligned}
&\beta_{k-1}\Big\|\sum_{s'\in\mathcal{S}_{k,h}}\widehat{V}_{k,h+1}(s')\cdot\boldsymbol{\Phi}_{k,h}^{\top}\nabla^2 L_{k,h}\big(\boldsymbol{\Phi}_{k,h}\widehat{\boldsymbol{\theta}}_{k-1,h}\big)\mathbf{e}_{s'}\Big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}} \\
&+2\beta_{k-1}^2\sup_{s'\in\mathcal{S}_{k,h}}\widehat{V}_{k,h+1}(s')\max_{s'\in\mathcal{S}_{k,h}}\Big\|\boldsymbol{\Phi}_{k,h}^{\top}\mathbf{e}_{s'}\Big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}}^2 \\
&\le\beta_{k-1}\Big\|\sum_{s'\in\mathcal{S}_{k,h}}\widehat{V}_{k,h+1}(s')\cdot\boldsymbol{\Phi}_{k,h}^{\top}\nabla^2 L_{k,h}\big(\boldsymbol{\Phi}(s_{k,h},a_{k,h})\boldsymbol{\theta}_h^{\star}\big)\mathbf{e}_{s'}\Big\|_{(\mathbf{H}_{k-1,h}^{\star})^{-1}} \\
&+8\kappa^{-1}H\beta_{k-1}^2\max_{s'\in\mathcal{S}_{k,h}}\Big\|\boldsymbol{\Phi}_{k,h}^{\top}\mathbf{e}_{s'}\Big\|_{\mathbf{A}_{k-1,h}^{-1}}^2.
\end{aligned}
$$

Because $\widehat{V}_{k,h+1}(s)\le H$,

$$
\begin{aligned}
&\Big\|\sum_{s'\in\mathcal{S}_{k,h}}\widehat{V}_{k,h+1}(s')\cdot\boldsymbol{\Phi}_{k,h}^{\top}\nabla^2 L_{k,h}\big(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^{\star}\big)\mathbf{e}_{s'}\Big\|_{(\mathbf{H}_{k-1,h}^{\star})^{-1}} \\
&=\big\|\boldsymbol{\Phi}_{k,h}^{\top}\nabla^2 L_{k,h}\big(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^{\star}\big)\Big(\sum_{s'\in\mathcal{S}_{k,h}}\widehat{V}_{k,h+1}(s')\mathbf{e}_{s'}\Big)\big\|_{(\mathbf{H}_{k-1,h}^{\star})^{-1}} \\
&\le H\max_{\mathbf{x}:\|\mathbf{x}\|_\infty\le 1}\big\|\boldsymbol{\Phi}_{k,h}^{\top}\nabla^2 L_{k,h}\big(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^{\star}\big)\mathbf{x}\big\|_{(\mathbf{H}_{k-1,h}^{\star})^{-1}}.
\end{aligned}
$$

By the definition of the Fisher information lower bound,

$$
\begin{aligned}
&H \max_{\mathbf{x}:\|\mathbf{x}\|_\infty \le 1} \left\|\boldsymbol{\Phi}_{k,h}^\top \nabla^2 L_{k,h}\big(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star\big)\mathbf{x}\right\|_{(\mathbf{H}_{k-1,h}^\star)^{-1}} \\
&\le H\kappa^{-1/2} \max_{\mathbf{x}:\|\mathbf{x}\|_\infty \le 1} \left\|\boldsymbol{\Phi}_{k,h}^\top \nabla^2 L_{k,h}\big(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star\big)\mathbf{x}\right\|_{\mathbf{A}_{k-1,h}^{-1}} \\
&= H\kappa^{-1/2} \max_{\mathbf{x}:\|\mathbf{x}\|_\infty \le 1} \sqrt{\mathbf{x}^\top \nabla^2 L_{k,h}\big(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star\big)\boldsymbol{\Phi}_{k,h}\mathbf{A}_{k-1,h}^{-1}\boldsymbol{\Phi}_{k,h}^\top \nabla^2 L_{k,h}\big(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star\big)\mathbf{x}} \\
&\le H\kappa^{-1/2} \max_{\mathbf{x}:\|\mathbf{x}\|_\infty \le 1} \|\mathbf{x}\|_{\nabla^2 L_{k,h}\left(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star\right)} \\
&\cdot \lambda_{\max}\Big(\nabla^2 L_{k,h}\big(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star\big)^{1/2}\boldsymbol{\Phi}_{k,h}\mathbf{A}_{k-1,h}^{-1}\boldsymbol{\Phi}_{k,h}^\top \nabla^2 L_{k,h}\big(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star\big)^{1/2}\Big).
\end{aligned}
$$

Because $\lambda_{\max}(\mathbf{M}\mathbf{M}^\top) = \lambda_{\max}(\mathbf{M}^\top\mathbf{M})$ for any matrix $\mathbf{M}$, we obtain

$$
\begin{aligned}
&\lambda_{\max}\Big(\nabla^2 L_{k,h}(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star)^{1/2}\boldsymbol{\Phi}_{k,h}\mathbf{A}_{k-1,h}^{-1}\boldsymbol{\Phi}_{k,h}^\top \nabla^2 L_{k,h}(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star)^{1/2}\Big) \\
&= \lambda_{\max}\Big(\mathbf{A}_{k-1,h}^{-1/2}\boldsymbol{\Phi}_{k,h}^\top \nabla^2 L_{k,h}(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star)\boldsymbol{\Phi}_{k,h}\mathbf{A}_{k-1,h}^{-1/2}\Big) \\
&\le \lambda_{\max}\Big(\mathbf{A}_{k-1,h}^{-1/2}\boldsymbol{\Phi}_{k,h}^\top \mathrm{diag}\Big(\nabla L_{k,h}(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star)\Big)\boldsymbol{\Phi}_{k,h}\mathbf{A}_{k-1,h}^{-1/2}\Big),
\end{aligned}
$$

where the last inequality uses the fact that $\nabla^2 L_{k,h}(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star) \preceq \mathrm{diag}\big(\nabla L_{k,h}(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star)\big)$. Because $\mathrm{diag}\big(\nabla L_{k,h}(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star)\big) = \sum_{s\in\mathcal{S}_{k,h}} \mathbf{e}_s^\top \nabla L_{k,h}(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star)\mathbf{e}_s\mathbf{e}_s^\top$,

$$
\begin{aligned}
&\le \lambda_{\max}\Big(\mathbf{A}_{k-1,h}^{-1/2}\boldsymbol{\Phi}_{k,h}^\top \mathrm{diag}\Big(\nabla L_{k,h}(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star)\Big)\boldsymbol{\Phi}_{k,h}\mathbf{A}_{k-1,h}^{-1/2}\Big) \\
&\le \sum_{s\in\mathcal{S}_{k,h+1}} \mathbf{e}_s^\top \nabla L_{k,h}(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star)\lambda_{\max}\Big(\mathbf{A}_{k-1,h}^{-1/2}\boldsymbol{\Phi}_{k,h}^\top \mathbf{e}_s\mathbf{e}_s^\top\boldsymbol{\Phi}_{k,h}\mathbf{A}_{k-1,h}^{-1/2}\Big) \\
&= \sum_{s\in\mathcal{S}_{k,h+1}} \mathbf{e}_s^\top \nabla L_{k,h}(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star)\left\|\boldsymbol{\Phi}_{k,h}^\top \mathbf{e}_s\right\|_{\mathbf{A}_{k-1,h}^{-1}}^2 \\
&\le \max_{s\in\mathcal{S}_{k,h+1}} \left\|\boldsymbol{\Phi}_{k,h}^\top \mathbf{e}_s\right\|_{\mathbf{A}_{k-1,h}^{-1}}^2,
\end{aligned}
$$

where the last inequality uses Hölder's inequality and the fact that $\|\nabla L_{k,h}(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star)\|_1 = 1$. Plugging in (15),

$$
\begin{aligned}
&\left|\widehat{V}_{k,h}(s_{k,h}) - \mathbb{E}_{k,h}\big[\widehat{V}_{k,h+1}(s_{k,h+1})\big] - r(s_{k,h},a_{k,h}) + \rho_{k,h}\right| \\
&\le \left|\rho_{k,h}\right| + \left|\widehat{\mathbb{E}}_{k,h}\widehat{V}_{k,h+1}(s_{k,h+1}) - \mathbb{E}_{k,h}\widehat{V}_{k,h+1}(s_{k,h+1})\right| \\
&\quad + H\kappa^{-1/2}\beta_{k-1} \max_{s\in\mathcal{S}_{k,h+1}} \left\|\boldsymbol{\Phi}_{k,h}^\top \mathbf{e}_s\right\|_{\mathbf{A}_{k-1,h}^{-1}} + 8\kappa^{-1}H\beta_{k-1}^2 \max_{s'\in\mathcal{S}_{k,h+1}} \left\|\boldsymbol{\Phi}_{k,h}^\top \mathbf{e}_{s'}\right\|_{\mathbf{A}_{k-1,h}^{-1}}^2.
\end{aligned}
$$

For simplicity, we define $U_{k,h} := \beta_{k-1}\kappa^{-1/2}\max_{s'\in\mathcal{S}_{k,h}} \|\boldsymbol{\Phi}_{k,h}^\top \mathbf{e}_{s'}\|_{\mathbf{A}_{k-1,h}^{-1}}$. Then, the above inequality is

$$
\begin{aligned}
&\left|\widehat{V}_{k,h}(s_{k,h}) - \mathbb{E}_{k,h}\big[\widehat{V}_{k,h+1}(s_{k,h+1})\big] - r(s_{k,h},a_{k,h}) + \rho_{k,h}\right| \\
&\le \left|\rho_{k,h}\right| + \left|\widehat{\mathbb{E}}_{k,h}\widehat{V}_{k,h+1}(s_{k,h+1}) - \mathbb{E}_{k,h}\widehat{V}_{k,h+1}(s_{k,h+1})\right| \\
&\quad + HU_{k,h} + 8HU_{k,h}^2.
\end{aligned}
$$

By the first order multinomial kernel difference bound (Lemma 35),

$$
\begin{aligned}
&\left|\widehat{\mathbb{E}}_{k,h}\widehat{V}_{k,h+1}(s_{k,h+1}) - \mathbb{E}_{k,h}\widehat{V}_{k,h+1}(s_{k,h+1})\right| \\
&= \left|\sum_{s'\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(s')\big(\mathbb{P}(\mathbf{e}_{s'}|s_{k,h},a_{k,h},\widehat{\boldsymbol{\theta}}_{k-1,h}) - \mathbb{P}(\mathbf{e}_{s'}|s_{k,h},a_{k,h},\boldsymbol{\theta}_{\star,h})\big)\right| \\
&\le HU_{k,h}.
\end{aligned}
$$

Thus,

$$\left|\widehat{V}_{k,h}(s_{k,h}) - \mathbb{E}_{k,h}\widehat{V}_{k,h+1}(s_{k,h+1}) - r(s_{k,h},a_{k,h}) + \rho_{k,h}\right| \le \left|\rho_{k,h}\right| + 2HU_{k,h} + 8HU_{k,h}^2.$$

Because $\mathbf{A}_{k-1,h} \succeq B_{\boldsymbol{\Phi}}^2\mathbf{I}_d$, we obtain,

$$\begin{aligned} U_{k,h} &:= \beta_{k-1}\kappa^{-1/2}\max_{s'\in\mathcal{S}_{k,h}} \|\boldsymbol{\Phi}_{k,h}^\top \mathbf{e}_{s'}\|_{\mathbf{A}_{k-1,h}^{-1}} \\ &\le \beta_{k-1}\kappa^{-1/2}B_{\boldsymbol{\Phi}}^{-1}\max_{s'\in\mathcal{S}_{k,h}} \|\boldsymbol{\Phi}_{k,h}^\top \mathbf{e}_{s'}\|_2 \\ &\le \beta_{k-1}\kappa^{-1/2}. \end{aligned}$$

Thus,

$$\begin{aligned} &\left|\widehat{V}_{k,h}(s_{k,h}) - \mathbb{E}_{k,h}\widehat{V}_{k,h+1}(s_{k,h+1}) - r(s_{k,h},a_{k,h}) + \rho_{k,h}\right| \\ &\le \left|\rho_{k,h}\right| + 2H(1+4\beta_{k-1}\kappa^{-1/2})U_{k,h} \\ &\le \left|\rho_{k,h}\right| + 10H\beta_{k-1}\kappa^{-1/2}U_{k,h}, \end{aligned} \tag{16}$$

and the squared term,

$$\begin{aligned} &\left(\widehat{V}_{k,h}(s_{k,h}) - \mathbb{E}_{k,h}\widehat{V}_{k,h+1}(s_{k,h+1}) - r(s_{k,h},a_{k,h}) + \rho_{k,h}\right)^2 \\ &\le \left(\left|\rho_{k,h}\right| + 10H\beta_{k-1}\kappa^{-1/2}U_{k,h}\right)^2 \\ &\le 2\rho_{k,h}^2 + 200H^2\beta_{k-1}^2\kappa^{-1}U_{k,h}^2, \end{aligned}$$

where the last inequality holds by $(a+b)^2 \le 2a^2+2b^2$ for $a,b\in\mathbb{R}$. Plugging in (14),

$$\begin{aligned} &\mathbb{E}_{k,1}\Big[\Big(\sum_{h=1}^H \widehat{V}_{k,h+1}(s_{k,h+1}) - \mathbb{E}_{k,h}\widehat{V}_{k,h+1}(s_{k,h+1})\Big)^2\Big] \\ &\le 2\mathbb{E}_{k,1}\Big[\Big(\sum_{h=1}^H \big(r(s_{k,h},a_{k,h}) - \rho_{k,h}\big) - \widehat{V}_{k,1}(s_{k,1})\Big)^2\Big] \\ &\quad + 2H\mathbb{E}_{k,1}\Big[\sum_{h=1}^H \left(\widehat{V}_{k,h}(s_{k,h}) - \mathbb{E}_{k,h}\widehat{V}_{k,h+1}(s_{k,h+1}) - r(s_{k,h},a_{k,h}) + \rho_{k,h}\right)^2\Big] \\ &\le 2\mathbb{E}_{k,1}\Big[\Big(\sum_{h=1}^H \big(r(s_{k,h},a_{k,h}) - \rho_{k,h}\big) - \widehat{V}_{k,1}(s_{k,1})\Big)^2\Big] \\ &\quad + 2H\mathbb{E}_{k,1}\Big[\sum_{h=1}^H 2\rho_{k,h}^2 + 200H^2\beta_{k-1}^2\kappa^{-1}U_{k,h}^2\Big], \end{aligned} \tag{17}$$

where the last inequality uses Jensen's inequality. In the first term, by (16),

$$\begin{aligned} &\Big(\sum_{h=1}^H \big(r(s_{k,h},a_{k,h}) - \rho_{k,h}\big) - \widehat{V}_{k,1}(s_{k,1})\Big)^2 \\ &= \Big(\sum_{h=2}^H \big(r(s_{k,h},a_{k,h}) - \rho_{k,h}\big) - \mathbb{E}_{k,1}\widehat{V}_{k,2}(s_{k,2}) + \mathbb{E}_{k,1}\widehat{V}_{k,2}(s_{k,2}) - \widehat{V}_{k,1}(s_{k,1}) + r(s_{k,1},a_{k,1}) - \rho_{k,1}\Big)^2 \\ &\le \Big(\Big|\sum_{h=2}^H \big(r(s_{k,h},a_{k,h}) - \rho_{k,h}\big) - \mathbb{E}_{k,1}\widehat{V}_{k,2}(s_{k,2})\Big| + \Big|\mathbb{E}_{k,1}\widehat{V}_{k,2}(s_{k,2}) - \widehat{V}_{k,1}(s_{k,1}) + r(s_{k,1},a_{k,1}) - \rho_{k,1}\Big|\Big)^2 \\ &\le \Big(\Big|\sum_{h=2}^H \big(r(s_{k,h},a_{k,h}) - \rho_{k,h}\big) - \mathbb{E}_{k,1}\widehat{V}_{k,2}(s_{k,2})\Big| + \Big|\rho_{k,1}\Big| + 10H\beta_{k-1}\kappa^{-1/2}U_{k,1}\Big)^2. \end{aligned}$$

By Jensen's inequality,

$$\begin{aligned}
&\mathbb{E}_{k,1}\Big[\Big(\sum_{h=1}^{H}\big(r(s_{k,h},a_{k,h})-\rho_{k,h}\big)-\widehat{V}_{k,1}(s_{k,1})\Big)^2\Big]\\
&\le \mathbb{E}_{k,1}\Big[\Big(\Big|\sum_{h=2}^{H}\big(r(s_{k,h},a_{k,h})-\rho_{k,h}\big)-\mathbb{E}_{k,1}\widehat{V}_{k,2}(s_{k,2})\Big|+\Big|\rho_{k,1}\Big|+10H\beta_{k-1}\kappa^{-1/2}U_{k,1}\Big)^2\Big]\\
&\le \mathbb{E}_{k,1}\Big[\Big(\Big|\sum_{h=2}^{H}\big(r(s_{k,h},a_{k,h})-\rho_{k,h}\big)-\widehat{V}_{k,2}(s_{k,2})\Big|+\Big|\rho_{k,1}\Big|+10H\beta_{k-1}\kappa^{-1/2}U_{k,1}\Big)^2\Big].
\end{aligned}$$

Again applying (16),

$$\begin{aligned}
&\Big|\sum_{h=2}^{H}\big(r(s_{k,h},a_{k,h})-\rho_{k,h}\big)-\widehat{V}_{k,2}(s_{k,2})\Big|\\
&\le \Big|\sum_{h=3}^{H}\big(r(s_{k,h},a_{k,h})-\rho_{k,h}\big)-\mathbb{E}_{k,2}\widehat{V}_{k,3}(s_{k,3})\Big|+\Big|\mathbb{E}_{k,2}\widehat{V}_{k,3}(s_{k,3})-\widehat{V}_{k,2}(s_{k,2})+r(s_{k,h},a_{k,h})-\rho_{k,h}\Big|\\
&\le \Big|\sum_{h=3}^{H}\big(r(s_{k,h},a_{k,h})-\rho_{k,h}\big)-\mathbb{E}_{k,2}\widehat{V}_{k,3}(s_{k,3})\Big|+\Big|\rho_{k,2}\Big|+10H\beta_{k-1}\kappa^{-1/2}U_{k,2},
\end{aligned}$$

it follows that

$$\begin{aligned}
&\mathbb{E}_{k,1}\Big[\Big(\sum_{h=1}^{H}\big(r(s_{k,h},a_{k,h})-\rho_{k,h}\big)-\widehat{V}_{k,1}(s_{k,1})\Big)^2\Big]\\
&\le \mathbb{E}_{k,1}\Big[\Big(\Big|\sum_{h=3}^{H}\big(r(s_{k,h},a_{k,h})-\rho_{k,h}\big)-\mathbb{E}_{k,2}\widehat{V}_{k,3}(s_{k,3})\Big|+\sum_{h=1}^{2}\Big|\rho_{k,h}\Big|+10H\beta_{k-1}\kappa^{-1/2}\sum_{h=1}^{2}U_{k,h}\Big)^2\Big]\\
&\le \mathbb{E}_{k,1}\Big[\mathbb{E}_{k,2}\Big[\Big(\Big|\sum_{h=3}^{H}\big(r(s_{k,h},a_{k,h})-\rho_{k,h}\big)-\widehat{V}_{k,3}(s_{k,3})\Big|+\sum_{h=1}^{2}\Big|\rho_{k,h}\Big|+10H\beta_{k-1}\kappa^{-1/2}\sum_{h=1}^{2}U_{k,h}\Big)^2\Big]\Big]\\
&= \mathbb{E}_{k,1}\Big[\Big(\Big|\sum_{h=3}^{H}\big(r(s_{k,h},a_{k,h})-\rho_{k,h}\big)-\widehat{V}_{k,3}(s_{k,3})\Big|+\sum_{h=1}^{2}\Big|\rho_{k,h}\Big|+10H\beta_{k-1}\kappa^{-1/2}\sum_{h=1}^{2}U_{k,h}\Big)^2\Big],
\end{aligned}$$

where the last inequality uses Jensen's inequality and the last inequality uses the tower property of the conditional expectation. Recursively,

$$\mathbb{E}_{k,1}\Big[\Big(\sum_{h=1}^{H}\big(r(s_{k,h},a_{k,h})-\rho_{k,h}\big)-\widehat{V}_{k,1}(s_{k,1})\Big)^2\Big]\le \mathbb{E}_{k,1}\Big[\Big(\sum_{h=1}^{H}\Big|\rho_{k,h}\Big|+10H\beta_{k-1}\kappa^{-1/2}\sum_{h=1}^{H}U_{k,h}\Big)^2\Big]$$

By Cauchy-Schwartz inequality,

$$\begin{aligned}
&2\mathbb{E}_{k,1}\Big[\Big(\sum_{h=1}^{H}\Big|\rho_{k,h}\Big|+10H\beta_{k-1}\kappa^{-1/2}\sum_{h=1}^{H}U_{k,h}\Big)^2\Big].\\
&\le 4\mathbb{E}_{k,1}\Big[\Big(\sum_{h=1}^{H}\Big|\rho_{k,h}\Big|\Big)^2+100H^2\beta_{k-1}^2\kappa^{-1}\Big(\sum_{h=1}^{H}U_{k,h}\Big)^2\Big]\\
&\le 4H\mathbb{E}_{k,1}\Big[\sum_{h=1}^{H}\rho_{k,h}^2+100H^2\beta_{k-1}^2\kappa^{-1}\sum_{h=1}^{H}U_{k,h}^2\Big].
\end{aligned}$$

Plugging in (17),

$$\mathbb{E}_{k,1}\Bigg[\Bigg(\sum_{h=1}^{H}\widehat{V}_{k,h+1}(s_{k,h+1})-\mathbb{E}\Big[\widehat{V}_{k,h+1}(s_{k,h+1})\Big|s_{k,h},a_{k,h}\Big]\Bigg)^2\Bigg]$$
$$\leq 8H\mathbb{E}_{k,1}\Bigg[\sum_{h=1}^{H}\rho_{k,h}^2+100H^2\beta_{k-1}^2\kappa^{-1}\sum_{h=1}^{H}U_{k,h}^2\Bigg]$$
$$=8H\mathbb{E}_{k,1}\Bigg[\sum_{h=1}^{H}\mathbb{E}_{k,h-1}[\rho_{k,h}^2]+100H^2\beta_{k-1}^2\kappa^{-1}\sum_{h=1}^{H}U_{k,h}^2\Bigg],$$

where the last equality uses the Tower property of the conditional expectation. Because $s_{k,1}$ and $a_{k,1}$ are deterministic given $\mathcal{F}_{k-1,H}$,

$$\rho_{k,1}:=r(s_{k,1},a_{k,1})-\mathbb{E}_{k-1,H}[r(s_{k,1},a_{k,1})]=0.$$

Thus,

$$\mathbb{E}_{k,1}\Bigg[\Bigg(\sum_{h=1}^{H}\widehat{V}_{k,h+1}(s_{k,h+1})-\mathbb{E}\Big[\widehat{V}_{k,h+1}(s_{k,h+1})\Big|s_{k,h},a_{k,h}\Big]\Bigg)^2\Bigg]$$
$$\leq 8H\mathbb{E}_{k,1}\Bigg[\sum_{h=1}^{H-1}\mathbb{E}_{k,h}[\rho_{k,h+1}^2]+100H^2\beta_{k-1}^2\kappa^{-1}\sum_{h=1}^{H}U_{k,h}^2\Bigg],$$

Let $\pi_{k,h}(s)=\operatorname*{argmax}_a\widehat{Q}_{k,h}(s,a)$ denote the action selected by the deterministic policy. Then,

$$\mathbb{E}_{k,h}[\rho_{k,h+1}^2]=\mathbb{E}_{k,h}[(r(s_{k,h+1},a_{k,h+1})-\mathbb{E}_{k,h}[r(s_{k,h+1},a_{k,h+1})])^2]$$
$$=\mathbb{E}_{k,h}\Big[\Big(\big(r\big(s_{k,h+1},\pi_{k,h+1}(s_{k,h+1})\big)-\mathbb{E}_{k,h}[(r(s_{k,h+1},\pi_{k,h+1}(s_{k,h+1}))]\big)^2\Big].$$

Let $v_{k,h+1}:=\sum_{s\in S_{k,h}}r(s,\pi_{k,h+1}(s))e_s$ denote the stacked vector of possible expected rewards values on $S_{k,h}$. Because the policy $\pi_{k,h+1}$ is determined at $k-1$th episode, $v_{k,h+1}$ is $\mathcal{F}_{k,h}$-measurable. Taking expectation on both sides,

$$\mathbb{E}_{k,h}\Big[\big(v_{k,h+1}^\top e_{s_{k,h+1}}-E_{k,h}[v_{k,h+1}^\top e_{s_{k,h+1}}]\big)^2\Big]$$
$$=v_{k,h+1}^\top\mathbb{E}_{k,h}\Big[\big(\mathbf{e}_{s_{k,h+1}}-\mathbb{E}_{k,h}\big[\mathbf{e}_{s_{k,h+1}}\big]\big)\big(\mathbf{e}_{s_{k,h+1}}-\mathbb{E}_{k,h}\big[\mathbf{e}_{s_{k,h+1}}\big]\big)^\top\Big]v_{k,h+1}.$$

By the identity for the covariance of Multinomial distribution (Proposition 19),

$$\mathbb{E}_{k,h}\Big[\big(\mathbf{e}_{s_{k,h+1}}-\mathbb{E}_{k,h}\big[\mathbf{e}_{s_{k,h+1}}\big]\big)\big(\mathbf{e}_{s_{k,h+1}}-\mathbb{E}_{k,h}\big[\mathbf{e}_{s_{k,h+1}}\big]\big)^\top\Big]=\nabla^2L_{k,h}(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star),$$

Thus,

$$\mathbb{E}_{k,h}[(r(s_{k,h+1},a_{k,h+1})-\mathbb{E}_{k,h}[r(s_{k,h+1},a_{k,h+1})])^2]\leq\max_{\mathbf{x}:\|\mathbf{x}\|_\infty\leq1}\mathbf{x}^\top\nabla^2L_{k,h}(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star)\mathbf{x},$$

and we define $\sigma_{k,h}^2:=\max_{\mathbf{x}:\|\mathbf{x}\|_\infty\leq1}\mathbf{x}^\top\nabla^2L_{k,h}(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star)\mathbf{x}$. Thus,

$$\mathbb{E}_{k,1}\Bigg[\Bigg(\sum_{h=1}^{H-1}\widehat{V}_{k,h+1}(s_{k,h+1})-\mathbb{E}\Big[\widehat{V}_{k,h+1}(s_{k,h+1})\Big|s_{k,h},a_{k,h}\Big]\Bigg)^2\Bigg]$$
$$\leq 8H\mathbb{E}_{k,1}\Bigg[\sum_{h=1}^{H}\sigma_{k,h}^2+100H^2\beta_{k-1}^2\kappa^{-1}\sum_{h=1}^{H}U_{k,h}^2\Bigg].$$

Summing up over $k\in[K]$,

$$\sum_{k=1}^{K}\mathbb{E}_{k,1}\Bigg[\sum_{h=1}^{H}\Big(\widehat{V}_{k,h+1}(s_{k,h+1})-\mathbb{E}\Big[\widehat{V}_{k,h+1}(s_{k,h+1})\Big|s_{k,h},a_{k,h}\Big]\Big)^2\Bigg]$$
$$\leq 12H\sum_{k=1}^{K}\mathbb{E}_{k,1}\Bigg[\underbrace{\sum_{h=1}^{H}\sigma_{k,h}^2}_{\phi_k}+100H^2\beta_{k-1}^2\kappa^{-1}\underbrace{\sum_{h=1}^{H}U_{k,h}^2}_{\psi_k}\Bigg].\tag{18}$$

By the elliptical potential lemma (Lemma 23),

$$\begin{aligned}\sum_{k=1}^{K}\psi_k :=& \sum_{h=1}^{K}\sum_{h=1}^{H} U_{k,h}^2 := \sum_{k=1}^{K}\sum_{h=1}^{H}\max_{s'\in\mathcal{S}_{k,h}} \|\boldsymbol{\Phi}_{k,h}^\top \mathbf{e}_{s'}\|_{\mathbf{A}_{k-1,h}}^2 \\ \leq& \beta_{K-1}^2\kappa^{-1}\sum_{k=1}^{K}\sum_{h=1}^{H}\max_{s'\in\mathcal{S}_{k,h}} \|\boldsymbol{\Phi}_{k,h}^\top \mathbf{e}_{s'}\|_{\mathbf{A}_{k-1,h}}^2 \\ \leq& H\beta_{K-1}^2\kappa^{-1} d\log\frac{K+1}{d}.\end{aligned} \tag{19}$$

Note that $0 \leq \psi_k \leq H\kappa^{-1}\beta_{K-1}^2$. By Chernoff bound (Lemma 29), with probability at least $1-\delta$,

$$\begin{aligned}\sum_{k=1}^{K}\mathbb{E}_{k,1}[\psi_k] \leq& \frac{e}{e-1}\sum_{k=1}^{K}\psi_k + \frac{eH\kappa^{-1}\beta_{K-1}^2}{e-1}\log\frac{1}{\delta} \\ \leq& \frac{eH\kappa^{-1}\beta_{K-1}^2}{e-1}\big(d\log\frac{K+1}{d} + \log\frac{1}{\delta}\big)\end{aligned}$$

where the last inequality uses (19). Plugging in (18),

$$\begin{aligned}&\sum_{k=1}^{K}\mathbb{E}_{k,1}\bigg[\sum_{h=1}^{H}\Big(\widehat{V}_{k,h+1}(s_{k,h+1}) - \mathbb{E}\Big[\widehat{V}_{k,h+1}(s_{k,h+1})\Big|s_{k,h},a_{k,h}\Big]\Big)^2\bigg] \\ &\leq 12H\sum_{k=1}^{K}\mathbb{E}_{k,1}[\phi_k] + \frac{1200eH^4\kappa^{-2}\beta_{K-1}^4}{e-1}\big(d\log\frac{K+1}{d} + \log\frac{1}{\delta}\big),\end{aligned} \tag{20}$$

where we use the fact that β_n is nondecreasing in k. Because $\phi_k \in [0,H]$, by Chernoff bound (Lemma 29), with probability at least $1-\delta$,

$$\sum_{k=1}^{K}\mathbb{E}_{k,1}[\phi_k] \leq \frac{e}{e-1}\big(\sum_{k=1}^{K}\phi_k + H\log\frac{1}{\delta}\big)$$

Plugging in (20),

$$\begin{aligned}&\sum_{k=1}^{K}\mathbb{E}_{k,1}\bigg[\sum_{h=1}^{H}\Big(\widehat{V}_{k,h+1}(s_{k,h+1}) - \mathbb{E}\Big[\widehat{V}_{k,h+1}(s_{k,h+1})\Big|s_{k,h},a_{k,h}\Big]\Big)^2\bigg] \\ &\leq \frac{e}{e-1}\bigg(12H\sum_{k=1}^{K}\phi_k + H\log\frac{1}{\delta} + 1200H^4\kappa^{-2}\beta_{K-1}^4\big(d\log\frac{K+1}{d} + \log\frac{1}{\delta}\big)\bigg) \\ &\leq \frac{12He}{e-1}\bigg(\sum_{k=1}^{K}\phi_k + 100H^3\kappa^{-2}\beta_{K-1}^4\big(d\log\frac{K+1}{d} + 2\log\frac{1}{\delta}\big)\bigg).\end{aligned}$$

Plugging in (13) completes the proof. □

B.7 Proof of the Variance Adaptive Lower Bound (Theorem 8)

Theorem 8 (Variance-Adaptive Lower Bound for MNL-MDP)**.** *Assume that $d \geq 2$, $H \geq 4$, and $K \geq \max\{144H(d-1)^2, \kappa^{-2}H^{-1}\}$. Given any parameter $\boldsymbol{\Theta}^\star = (\boldsymbol{\theta}_1^\star, \ldots, \boldsymbol{\theta}_H^\star) \in \mathbb{R}^{d\times H}$, there exists a MNL-MDP instance $\mathcal{M}_{\boldsymbol{\Theta}^\star} := \mathcal{M}(\mathcal{S}, \mathcal{A}, H, \boldsymbol{\Theta}^\star, \boldsymbol{\Phi}, r)$ such that with probability at least $1/e$, the regret of any algorithm Π_K,*

$$R_{\Pi_K}(\mathcal{M}_{\boldsymbol{\Theta}^\star}) \geq \frac{2Hd\sqrt{\sum_{h=1}^{H}\sum_{k=1}^{K}\sigma^2(s_{k,h},a_{k,h})}}{17},$$

where $\{s_{k,h}, a_{k,h} : k \in [K], h \in [H]\}$ denote the state and action sample generated by the interaction between Π_K and $\mathcal{M}_{\boldsymbol{\Theta}^\star}$.

We construct a hard MNL-MDP instance as follows. Let $\mathcal{A} := \{\pm 1\}^{d-1}$ denote the action space. The initial state is $s_1 = 1$ and the good state is $s = 2H+1$ where the reward is $r(2H+1, a) = 1$, for any action $\mathbf{a} \in \mathcal{A} = \{\pm 1\}^{d-1}$. Otherwise the reward $r(s, \mathbf{a}) = 0$ for any $s \neq 2H+1$ and $\mathbf{a} \in \mathcal{A}$. The reachable set is defined as $\mathcal{S}_h(2H+1, \mathbf{a}) = \{2H+1\}$ and for any $h \in [H]$ and $\mathbf{a} \in \mathcal{A}$, i.e., $2H+1$ is the absorbing state. For each $h \in [H]$, $\mathcal{S}_h(s, \mathbf{a}) = \{2H+1, 2h+1, 2h+2\}$ for $s \in \{2h-1, 2h\}$ and any $\mathbf{a} \in \mathcal{A}$.

Let us fix $\boldsymbol{\theta}_1^\star, \ldots, \boldsymbol{\theta}_H^\star$ and without loss of generality we assume that $\theta_{h,d}^\star \neq 0$ for $\boldsymbol{\theta}_h^\star := (\theta_{h,1}^\star, \ldots, \theta_{h,d}^\star)$. For $\epsilon \in (0, 1/2)$ and $\Delta \in (0, \frac{\log 2}{4(d-1)})$ to be determined later, define,

$$\tilde{\Delta} := \frac{1}{(d-1)}\Big(\frac{1}{1 + \frac{1-\epsilon}{\epsilon}\exp\Big(-4(d-1)\Delta\Big)} - \epsilon\Big), \quad \phi := \frac{1}{2}\sqrt{\frac{1-\epsilon}{\epsilon} \cdot \frac{1-\epsilon-(d-1)\tilde{\Delta}}{\epsilon + (d-1)\tilde{\Delta}}}, \tag{21}$$

and a function $p(x) := 1/(1 + 2\phi e^{-2x})$. These definitions are constructed to satisfy $p\big((d-1)\Delta\big) = \epsilon + (d-1)\tilde{\Delta}$ and $p\big(-(d-1)\Delta\big) = \epsilon$. Because $\Delta \in (0, \frac{\log 2}{4(d-1)})$, with simple algebra, for any $\epsilon \in (0, 1/2)$,

$$\Delta \leq \frac{1}{4(d-1)} \log \frac{2-2\epsilon}{1-2\epsilon} \iff (d-1)\tilde{\Delta} \leq \epsilon$$

For $s \in \{2h-1, 2h\}$ and $a \in \mathcal{A}^{d-1}$, we define the feature matrix $\Phi(s, a) \in \mathbb{R}^{3 \times d}$ as:

$$\Phi(s, a) := (\sum_{i=1}^{d-1} a_i e_i + e_d)(1, -1, -1)^\top,$$

where e_i is the Euclidean basis in R^d.

We define the perturbation set as $U_h := \{\sum_{i=1}^{d-1}(-\theta_{h,i}^\star + u_{h,i}\Delta)e_i + (-\theta_{h,d}^\star - 1/2 \log \phi)e_d : u_{h,i} \in \{\pm 1\}\}$. It directly follows that:

$$\Phi(s, a)(\theta_h^\star + u_h) = (\Delta \sum_{i=1}^{d-1} a_i u_{h,i} - \frac{1}{2}\log\phi)(1, -1, -1)^\top,$$

for $u_h \in U_h$. Then, the optimal action at step $h \in [H]$ given $\boldsymbol{\theta}_h^\star + \mathbf{u}_h \in \mathcal{D}_h$ is $a_\star(\boldsymbol{\theta}_h^\star + \mathbf{u}_h) = \mathbf{u}_h$, which maximizes the probability of reaching state $2H+1$. Under this setting, we prove a lower bound of the regret.

Lemma 12. *(Reduction to the multi-armed bandits instances) Assume that $H \geq 4$ and $\Delta \in (0, \frac{\log 2}{4(d-1)})$. For any $\epsilon \in (0, 1/H)$, nonzero parameters $\boldsymbol{\Theta}^\star := \{\boldsymbol{\theta}_h^\star : h \in [H]\}$, perturbations $\boldsymbol{\xi} := \big\{\mathbf{u}_h \in U_h : h \in [H]\big\}$, and policy Π_k, the problem constructed above has regret:*

$$\boldsymbol{R}(\boldsymbol{\Theta}^\star + \boldsymbol{\xi}, \Pi_K) \geq \frac{H\epsilon}{8(d-1)} \sum_{k=1}^{K} \sum_{h=1}^{\lfloor H/2 \rfloor} \sum_{i=1}^{d-1} \mathbb{P}_{\boldsymbol{\Theta}^\star + \boldsymbol{\xi}, \Pi_K}(\mathbf{e}_i^\top a_{h,k} \neq u_{h,i}),$$

where $a_{h,k}$ is the action chosen at step $h \in [H]$ and episode $k \in [K]$ by the policy Π_K given $\boldsymbol{\Theta}^\star + \boldsymbol{\xi}$.

Proof. Let us fix $\boldsymbol{\Theta}^\star$ throughout the proof. For each $k \in [K]$, we write the value at episode $k \in [K]$ as

$$V_k := \mathbb{E}_{\boldsymbol{\Theta}^\star + \boldsymbol{\xi}, \Pi_K}\Big[\sum_{h=1}^{H} r(s_{h,k}, a_{h,k}) \Big| s_{1,k} = 1\Big]$$

By construction of the problem,

$$V_k = \mathbb{E}_{\boldsymbol{\Theta}^\star + \boldsymbol{\xi}, \Pi_K}\Big[\sum_{h=1}^{H-1} (H-h)\mathbb{P}_{\boldsymbol{\Theta}^\star + \boldsymbol{\xi}, \Pi_K}(s_{h+1,k} = 2H+1, s_{h,k} \neq 2H+1 | s_{1,k} = 1)\Big].$$

By the law of total probability and the Markov property,

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\Theta}^\star + \boldsymbol{\xi}, \Pi_K}(s_{h+1,k} = 2H+1 | s_{h,k} \neq 2H+1, s_{1,k} = 1) \\
&\sum_{a \in \mathcal{A}} \mathbb{P}_{\boldsymbol{\Theta}^\star + \boldsymbol{\xi}, \Pi_K}(s_{h+1,k} = 2H+1, a_{h,k} = a | s_{h,k} \neq 2H+1, s_{1,k} = 1) \\
&= \sum_{a \in \mathcal{A}} p_a \mathbb{P}_{\boldsymbol{\Theta}^\star + \boldsymbol{\xi}, \Pi_K}(a_{h,k} = a | s_{h,k} \neq 2H+1, s_{1,k} = 1),
\end{aligned}$$

where $p_a := \mathbb{P}_{\boldsymbol{\Theta}^\star+\boldsymbol{\xi},\Pi_K}(s_{h+1} = 2H+1 | s_{h,k} \neq 2H+1, a_{h,k} = a)$ By construction the transition probability,

$$p_a = \frac{1}{\underbrace{1 + 2\phi \exp\big(-2\Delta \sum_{i=1}^{d-1} \text{sign}(u_{h,i})\text{sign}(a_i)\big)}_{p\left(\Delta \sum_{i=1}^{d-1} \text{sign}(u_i)\text{sign}(a_i)\right)}}.$$

Thus, for $\epsilon \in (0, 1/2)$,

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\Theta}^\star+\boldsymbol{\xi},\Pi_K}(s_{h+1,k} = 2H+1 | s_{h,k} \neq 2H+1, s_{1,k} = 1) \\
&= \sum_{a \in \mathcal{A}} p\Big(\Delta \sum_{i=1}^{d-1} \text{sign}(u_{h,i})\text{sign}(a_i)\Big) \mathbb{P}_{\boldsymbol{\Theta}^\star+\boldsymbol{\xi},\Pi_K}(a_{h,k} = a | s_{h,k} \neq 2H+1, s_{1,k} = 1) \\
&= \epsilon + \underbrace{\sum_{a \in \mathcal{A}} \left(p\Big(\Delta \sum_{i=1}^{d-1} \text{sign}(u_{h,i})\text{sign}(a_i)\Big) - \epsilon \right) \mathbb{P}_{\boldsymbol{\Theta}^\star+\boldsymbol{\xi},\Pi_K}(a_{h,k} = a | s_{h,k} \neq 2H+1, s_{1,k} = 1)}_{\psi_{h,k}}
\end{aligned}$$

Consequently,

$$\mathbb{P}_{\boldsymbol{\Theta}^\star+\boldsymbol{\xi},\Pi_K}(s_{h+1,k} \neq 2H+1 | s_{h,k} \neq 2H+1, s_{1,k} = 1) = 1 - \epsilon - \psi_{h,k}.$$

This implies,

$$\mathbb{P}_{\boldsymbol{\Theta}^\star+\boldsymbol{\xi},\Pi_K}(s_{h+1,k} = 2H+1, s_{h,k} \neq 2H+1 | s_{1,k} = 1) = (\epsilon + \psi_{h,k}) \prod_{h'=1}^{h-1} (1 - \epsilon - \psi_{h',k}).$$

Thus, the value function,

$$V_k = \mathbb{E}\Big[\sum_{h=1}^{H-1} (H-h)(\epsilon + \psi_{h,k}) \prod_{h'=1}^{h-1} (1 - \epsilon - \psi_{h',k}) \Big].$$

Because the optimal policy is deterministic $a_\star = \mathbf{u}_h$, we have

$$\begin{aligned}
V_k^\star =& \mathbb{E}_{\boldsymbol{\Theta}^\star+\boldsymbol{\xi},\pi^\star}\Big[\sum_{h=1}^{H} r(s_{h,k}, \mathbf{u}_h) \Big| s_{1,k} = 1 \Big] \\
&= \sum_{h=1}^{H-1} (H-h) \underbrace{p\big(\Delta(d-1)\big)}_{:=p^\star} \Big(1 - p\big(\Delta(d-1)\big)\Big)^{h-1} \\
&= \sum_{h=1}^{H-1} (H-h) p^\star (1-p^\star)^{h-1}
\end{aligned}$$

For $j \in [H-1]$, define,

$$\begin{aligned}
V_j^\star &:= \sum_{h=j}^{H-1} (H-h) p^\star (1-p^\star)^{h-j}, \\
\widehat{V}_j &:= \sum_{h=j}^{H-1} (H-h)(\epsilon + \psi_{h,k}) \prod_{h'=j}^{h-1} (1 - \epsilon - \psi_{h',k}).
\end{aligned}$$

Then, with some simple algebra,

$$V_j^\star - \widehat{V}_j = (H - j - V_{j+1}^\star)\Big(p^\star - \psi_{j,k}\Big) + (V_{j+1}^\star - \widehat{V}_{j+1})(1 - \epsilon - \psi_{j,k}),$$

which result is concurrent to equation (C.24) in Park et al. [2025]. Recursively the regret at episode k is

$$V_1^\star - \widehat{V}_1 = \sum_{h=1}^{H-1} (H - h - V_{h+1}^\star)(p^\star - \psi_{h,k}) \prod_{j=1}^{h-1} (1 - \epsilon - \psi_{j,k}). \tag{22}$$

Meanwhile, $V_j^\star$ can be written as,

$$V_j^\star = \frac{(1-p^\star)^{H-j}-1}{p^\star} + H - j + 1 - (1-p^\star)^{H-j},$$

which implies, for $h \le \lfloor H/2 \rfloor$

$$\begin{aligned} H-h-V_{h+1}^\star =& (1-p^\star)^{H-h-1} + \frac{1-(1-p^\star)^{H-h-1}}{p^\star} \\ =& \frac{1-(1-p^\star)^{H-h}}{p^\star} \\ \ge& \frac{1-(1-p^\star)^{H/2}}{p^\star}. \end{aligned}$$

Because $(d-1)\tilde{\Delta} \le \epsilon \le 1/H$,

$$p^\star := p(\Delta(d-1)) = \epsilon + (d-1)\tilde{\Delta} \le \frac{2}{H}.$$

Because $x \mapsto \frac{1-(1-x)^{H/2}}{x}$ is nonincreasing in $(0,1)$,

$$\begin{aligned} H-h-V_{h+1}^\star \ge& \frac{1-(1-p^\star)^{H/2}}{p^\star} \\ \ge& \frac{1-(1-\frac{2}{H})^{H/2}}{\frac{2}{H}} \\ \ge& \frac{1-e^{-1}}{2} H \\ \ge& \frac{H}{4}, \end{aligned}$$

where the second inequality holds by $(1-x^{-1})^x \le e^{-1}$ for $x \ge 1$. Plugging in (22),

$$V_1^\star - \widehat{V}_1 \ge \frac{H}{4} \sum_{h=1}^{\lfloor H/2 \rfloor} \Big(p^\star - \psi_{h,k}\Big) \prod_{j=1}^{h-1} (1-\epsilon-\psi_{j,k}).$$

Note that for any $j \in [H]$, we have

$$\begin{aligned} \epsilon + \psi_{j,k} =& \mathbb{P}_{\boldsymbol{\Theta}^\star + \boldsymbol{\xi}, \Pi_K}(s_{h+1,k} = 2H+1 | s_{h,k} \ne 2H+1, s_{1,k} = 1) \\ \le& p\big((d-1)\Delta\big) \le 1/H \end{aligned}$$

and thus,

$$\prod_{j=1}^{h-1}(1-\epsilon-\psi_{j,k}) \ge (1-\frac{1}{H})^{H/2} \ge \frac{1}{\sqrt{e}} \ge \frac{1}{2}.$$

It follows that

$$V_1^\star - \widehat{V}_1 \ge \frac{H}{8} \sum_{h=1}^{\lfloor H/2 \rfloor} \Big(p^\star - \psi_{h,k}\Big).$$

By definition of p, recall that $p\big(-(d-1)\Delta\big) = \epsilon$ and thus,

$$\begin{aligned} p\Big(\Delta \sum_{i=1}^{d-1} \mathrm{sign}(u_{h,i})\mathrm{sign}(a_i)\Big) - \epsilon =& p\Big(\Delta \sum_{i=1}^{d-1} \mathrm{sign}(u_{h,i})\mathrm{sign}(a_i)\Big) - p\big(-\Delta(d-1)\big) \\ \le& \big(\epsilon + (d-1)\tilde{\Delta}\big)\Delta\Big(\sum_{i=1}^{d-1} \mathrm{sign}(u_{h,i})\mathrm{sign}(a_i) + (d-1)\Big), \end{aligned}$$

where the last inequality holds because $\sup_{x\in[-(d-1)\Delta,(d-1)\Delta]}|p'(x)| \le \epsilon + (d-1)\tilde{\Delta}$ and the mean value theorem. By definition of $\tilde{\Delta}$, we have

$$\begin{aligned}\frac{\Delta}{\tilde{\Delta}} =& \frac{1}{4\tilde{\Delta}(d-1)}\log\frac{(\epsilon+(d-1)\tilde{\Delta})(1-\epsilon)}{(1-\epsilon-(d-1)\tilde{\Delta})\epsilon}\\ =& \frac{1}{4\tilde{\Delta}(d-1)}\log\Big(1+\frac{(d-1)\tilde{\Delta}}{(1-\epsilon-(d-1)\tilde{\Delta})\epsilon}\Big)\\ \le& \frac{1}{4\tilde{\Delta}(d-1)}\cdot\frac{(d-1)\tilde{\Delta}}{(1-\epsilon-(d-1)\tilde{\Delta})\epsilon}\\ =& \frac{1}{4\big(1-\epsilon-(d-1)\tilde{\Delta}\big)\epsilon}.\end{aligned}$$

Because $\Delta \in \big(0, \frac{\log 2}{4(d-1)}\big)$, and $(d-1)\tilde{\Delta} \le \epsilon \le 1/H$, we obtain,

$$\begin{aligned}\frac{\big(\epsilon+(d-1)\tilde{\Delta}\big)\Delta}{\tilde{\Delta}} =& \frac{\epsilon+(d-1)\tilde{\Delta}}{4\epsilon\big(1-\epsilon-(d-1)\tilde{\Delta}\big)}\\ \le& \frac{2\epsilon}{4\epsilon\big(1-2\epsilon\big)}\\ \le& \frac{1}{2(1-2/H)}\\ \le& 1,\end{aligned}$$

where the last inequality holds by $H \ge 4$. Thus,

$$\begin{aligned}p\Big(\Delta\sum_{i=1}^{d-1}\text{sign}(u_{h,i})\text{sign}(a_i)\Big)-\epsilon \le& \frac{\big(\epsilon+(d-1)\tilde{\Delta}\big)\Delta}{\tilde{\Delta}}\tilde{\Delta}\Big(\sum_{i=1}^{d-1}\text{sign}(u_{h,i})\text{sign}(a_i)+(d-1)\Big)\\ \le& \tilde{\Delta}\Big(\sum_{i=1}^{d-1}\text{sign}(u_{h,i})\text{sign}(a_i)+(d-1)\Big).\end{aligned}$$

Rearranging the terms,

$$\tilde{\Delta}(d-1)-p\Big(\Delta\sum_{i=1}^{d-1}\text{sign}(u_{h,i})\text{sign}(a_i)\Big)+\epsilon \ge -\tilde{\Delta}\sum_{i=1}^{d-1}\text{sign}(u_{h,i})\text{sign}(a_i).$$

It follows that $p^\star - \psi_{h,k} = p\big(\Delta(d-1)\big) - \psi_{h,k} = \epsilon + \tilde{\Delta}(d-1) - \psi_{h,k}$ and define $\pi_a := \mathbb{P}_{\boldsymbol{\Theta}^\star+\boldsymbol{\xi},\Pi_K}(a_{h,k}=a|s_{h,k}\neq 2H+1, s_{1,k}=s_1)$

$$\begin{aligned}&\epsilon+\tilde{\Delta}(d-1)-\psi_{h,k}\\ &=\sum_{a\in\mathcal{A}}\bigg(\epsilon+\tilde{\Delta}(d-1)-p\Big(\Delta\sum_{i=1}^{d-1}\text{sign}(u_{h,i})\text{sign}(a_i)\Big)+\epsilon\bigg)\pi_a\\ &\ge\sum_{a\in\mathcal{A}}\bigg(\epsilon-\tilde{\Delta}\sum_{i=1}^{d-1}\text{sign}(u_{h,i})\text{sign}(a_i)\bigg)\pi_a\\ &\ge\sum_{a\in\mathcal{A}}\frac{\epsilon}{d-1}\bigg(d-1-\sum_{i=1}^{d-1}\text{sign}(u_{h,i})\text{sign}(a_i)\bigg)\pi_a,\end{aligned}$$

where the second inequality holds by $(d-1)\tilde{\Delta} \leq \epsilon$. Then,

$$\begin{aligned}
&V_1^\star - \widehat{V}_1 \\
&\geq \frac{H}{8} \sum_{h=1}^{\lfloor H/2 \rfloor} \Big(p^\star - \psi_{h,k} \Big) \\
&\geq \frac{H\epsilon}{8(d-1)} \sum_{h=1}^{\lfloor H/2 \rfloor} \sum_{a \in \mathcal{A}} \left(d - 1 - \sum_{i=1}^{d-1} \text{sign}(u_{h,i}) \text{sign}(a_i) \right) \pi_a \\
&= \frac{H\epsilon}{8(d-1)} \sum_{h=1}^{\lfloor H/2 \rfloor} \sum_{a \in \mathcal{A}} \sum_{i=1}^{d-1} \left(1 - \text{sign}(u_{h,i}) \text{sign}(a_i) \right) \pi_a
\end{aligned}$$

Because $\mathrm{R}(\boldsymbol{\Theta}^\star + \boldsymbol{\xi}, \Pi_K) = \mathbb{E}\Big[\sum_{k=1}^K (V_k^\star - \widehat{V}_k) \Big]$, taking the expectation on both sides and summing up over $k \in [K]$ completes the proof. □

Theorem 13 (Variance-adaptive lower bound for MNL-MDP)**.** *Assume that $d \geq 2$, $H \geq 4$, and $K \geq \max\{144H(d-1)^2, \kappa^{-2}H^{-1}\}$. Given any parameter $\boldsymbol{\Theta}^\star = (\boldsymbol{\theta}_1^\star, \ldots, \boldsymbol{\theta}_H^\star) \in \mathbb{R}^{d \times H}$, there exists a MNL-MDP instance $\mathcal{M}_{\boldsymbol{\Theta}^\star} := \mathcal{M}(\mathcal{S}, \mathcal{A}, H, \boldsymbol{\Theta}^\star, \boldsymbol{\Phi}, r)$ such that with probability at least $\delta \in (0, 1/e)$, the regret of any algorithm Π_K,*

$$R_{\Pi_K}(\mathcal{M}_{\boldsymbol{\Theta}^\star}) \geq \frac{2Hd\sqrt{\sum_{h=1}^H \sum_{k=1}^K \sigma_{k,h}^2}\sqrt{1 - \frac{\delta}{1-\delta}\sqrt{\log\frac{1}{\delta}}}}{17}.$$

Remark 14. The term in the square root,

$$1 - \frac{\delta}{1-\delta}\sqrt{\log\frac{1}{\delta}}$$

is nonnegative and non-increasing over $\delta \in (0, 1/e)$. This result cannot be derived by the concentration inequalities and but by our novel anti-concentration inequality for nonnegative random variables (Lemma 32). Thus, Theorem 13 implies Theorem 8 for $\delta = 1/e$. Theorem 11 states that any algorithm suffers regret more than

$$\frac{2Hd\sqrt{\sum_{h=1}^H \sum_{k=1}^K \sigma_{k,h}^2}\sqrt{1 - \frac{\delta}{1-\delta}\sqrt{\log\frac{1}{\delta}}}}{17},$$

with probability at least δ. Since the lower bound event already take δ probability, no algorithm can achieve the regret less than the lower bound with probability at least $1 - \delta$. While our result carries the mild technical condition $\delta \in (0, 1/e)$, this covers the entire high-probability regime of practical interest. One limitation is that the lower bound do not match the upper bound by the factor of $\sqrt{\log\frac{1}{\delta}}$. As $\delta \to 0$ the term $\frac{\delta}{1-\delta}\sqrt{\log\frac{1}{\delta}}$ vanishes and the lower bound converges to $(2/17)Hd\sqrt{\sum_{h,k}\sigma_{k,h}^2}$. Nevertheless, our lower bound provides the first result that no algorithm cannot achieve regret lower than $(2/17)Hd\sqrt{\sum_{h,k}\sigma_{k,h}^2}$ with high probability, i.e., the variance term is not improvable with high probability.

Proof. By the reduction lemma (Lemma 12), for $\mathbf{U} \in \prod_{h=1}^H U_h$,

$$\begin{aligned}
\mathrm{R}(\boldsymbol{\Theta}^\star + \boldsymbol{\xi}, \Pi_K) \geq& \frac{H\epsilon}{8(d-1)} \sum_{k=1}^K \sum_{h=1}^{\lfloor H/2 \rfloor} \sum_{i=1}^{d-1} \mathbb{P}_{\boldsymbol{\Theta}^\star + \boldsymbol{\xi}, \Pi_K}(\mathbf{e}_i^\top a_{h,k} \neq u_{h,i}) \\
=& \frac{H\epsilon}{8(d-1)} \sum_{h=1}^{\lfloor H/2 \rfloor} \sum_{i=1}^{d-1} \mathbb{E}_{\boldsymbol{\Theta}^\star + \boldsymbol{\xi}, \Pi_K}\Big[\sum_{k=1}^K \mathbb{I}(\mathbf{e}_i^\top a_{h,k} \neq u_{h,i}) \Big].
\end{aligned}$$

Let $\tilde{u}_{h,i} := \sqrt{\tilde{\Delta}} u_{h,i}$ and $\tilde{a}_i = \sqrt{\tilde{\Delta}} a_i$ for sufficiently small $\tilde{\Delta} > 0$. Note that $\sum_{k=1}^K \mathbb{I}(\mathbf{e}_i^\top a_{h,k} \neq u_{h,i})$ is scale-invariant and the distribution of $\sum_{k=1}^K \mathbb{I}(\mathbf{e}_i^\top a_{h,k} \neq u_{h,i})$ given $\boldsymbol{\Theta}^\star + \boldsymbol{\xi}$ is equal

to the distribution of $\sum_{k=1}^K \mathbb{I}(\mathbf{e}_i^\top \tilde{a}_{h,k} \neq \tilde{u}_{h,i})$ given $\Theta^\star + \tilde{\mathbf{U}}$, where $\tilde{\mathbf{U}} \in \prod_{h=1}^H \tilde{U}_h$ and $\tilde{U}_h := \{\sum_{i=1}^{d-1}(-\theta^\star_{h,i} + \tilde{u}_{h,i}\Delta)e_i + (-\theta^\star_{h,d} - 1/2\log\phi)e_d : \tilde{u}_{h,i} \in \{\pm\sqrt{\tilde{\Delta}}\}\}$. Let $M_{h,i}\big(\text{sign}(\tilde{u}_{h,i})\big) := \sum_{k=1}^K \mathbb{I}(\mathbf{e}_i^\top \tilde{a}_{h,k} \neq \tilde{u}_{h,i})$. Then,

$$\mathrm{R}(\boldsymbol{\Theta}^\star + \boldsymbol{\xi}, \Pi_K) \geq \frac{H\epsilon}{8(d-1)} \sum_{h=1}^{\lfloor H/2 \rfloor} \sum_{i=1}^{d-1} \mathbb{E}_{\boldsymbol{\Theta}^\star + \tilde{\boldsymbol{\xi}}, \Pi_K}\Big[M_{h,i}\big(\text{sign}(\tilde{u}_{h,i})\big)\Big]$$

For each $\tilde{\mathbf{U}} \in \{\pm\sqrt{\tilde{\Delta}}\}^{(d-1)\times H} \times \{0\}^H$, we define $\bar{\mathbf{U}}(\tilde{h}, \tilde{i})$ by

$$\bar{\mathbf{U}}_{h,i}(\tilde{h}, \tilde{i}) = \begin{cases} -\tilde{u}_{h,i} & h = \tilde{h}, i = \tilde{i} \\ \tilde{u}_{h,i} & h \neq \tilde{h} \text{ or } i \neq \tilde{i} \end{cases},$$

where only the sign of $(\tilde{h}, \tilde{i})$-th coordinate differs and the corresponding perturbation is $\bar{\boldsymbol{\xi}}(\tilde{h}, \tilde{i})$. Let $\mathbb{P}_{\tilde{\mathbf{U}}}$ and $\mathbb{P}_{\bar{\mathbf{U}}(\tilde{h},\tilde{i})}$ denote the environment/learner interaction measure for $\{s_{k,h}, a_{k,h} : k \in [K], h \in [H]\}$ given $\boldsymbol{\Theta}^\star + \tilde{\boldsymbol{\xi}}$ and $\boldsymbol{\Theta}^\star + \bar{\boldsymbol{\xi}}(\tilde{h}, \tilde{i})$, respectively. Because $M_{h,i}(1) \leq K$, by Pinsker's inequality,

$$\begin{aligned} \mathbb{E}_{\tilde{\mathbf{U}}}[M_{\tilde{h},\tilde{i}}(1)] &\geq \mathbb{E}_{\bar{\mathbf{U}}(\tilde{h},\tilde{i})}[M_{\tilde{h},\tilde{i}}(1)] - K\|\mathbb{P}_{\tilde{\mathbf{U}}} - \mathbb{P}_{\bar{\mathbf{U}}(\tilde{h},\tilde{i})}\|_1 \\ &\geq \mathbb{E}_{\bar{\mathbf{U}}(\tilde{h},\tilde{i})}[M_{\tilde{h},\tilde{i}}(1)] - K\sqrt{\frac{1}{2}KL(\mathbb{P}_{\tilde{\mathbf{U}}}, \mathbb{P}_{\bar{\mathbf{U}}(\tilde{h},\tilde{i})})} \end{aligned} \tag{23}$$

Note that

$$\|\boldsymbol{\Phi}(s_{k,h}, a_{k,h})\big(\boldsymbol{\Theta}^\star + \tilde{\boldsymbol{\xi}} - \boldsymbol{\Theta}^\star + \bar{\boldsymbol{\xi}}(\tilde{h}, \tilde{i})\big)\mathbf{e}_h\|_\infty \leq 2\mathbb{I}(\tilde{h} = h)\tilde{\Delta}$$

By the divergence bound (Lemma 39),

$$KL(\mathbb{P}_{\tilde{\mathbf{U}}}, \mathbb{P}_{\bar{\mathbf{U}}(\tilde{h},\tilde{i})}) \leq 2\tilde{\Delta}^2\mathbb{E}_{\tilde{\mathbf{U}}}[\sum_{k=1}^K \max_{\mathbf{x}:\|\mathbf{x}\|_\infty \leq 1} \mathbf{x}^\top \nabla^2 L_{k,\tilde{h}}\big(\boldsymbol{\Phi}(s_{k,\tilde{h}}, a_{k,\tilde{h}})(\boldsymbol{\Theta}^\star + \tilde{\boldsymbol{\xi}})\mathbf{e}_{\tilde{h}}\big)\mathbf{x}] + \frac{8N\tilde{\Delta}^3}{\sqrt{6\kappa}}.$$

Because the log-sum function is $\sqrt{6}$-self-concordant, by the ratio bound (Lemma 15),

$$\begin{aligned} &\max_{\mathbf{x}:\|\mathbf{x}\|_\infty \leq 1} \mathbf{x}^\top \nabla^2 L_{k,\tilde{h}}\big(\boldsymbol{\Phi}(s_{k,\tilde{h}}, a_{k,\tilde{h}})(\boldsymbol{\Theta}^\star + \tilde{\boldsymbol{\xi}})\mathbf{e}_{\tilde{h}}\big)\mathbf{x} \\ &\leq \exp(\sqrt{6}\|\boldsymbol{\Phi}(s_{k,\tilde{h}}, a_{k,\tilde{h}})\tilde{\mathbf{U}}\mathbf{e}_{\tilde{h}}\|_2) \underbrace{\max_{\mathbf{x}:\|\mathbf{x}\|_\infty \leq 1} \mathbf{x}^\top \nabla^2 L_{k,\tilde{h}}\big(\boldsymbol{\Phi}(s_{k,\tilde{h}}, a_{k,\tilde{h}})\boldsymbol{\Theta}^\star\mathbf{e}_{\tilde{h}}\big)\mathbf{x}}_{\sigma^2_{k,\tilde{h}}}. \end{aligned}$$

By definition of $\boldsymbol{\Phi}$ and $\tilde{u}_H$,

$$\begin{aligned} \exp(\sqrt{6}\|\boldsymbol{\Phi}(s_{k,\tilde{h}}, a_{k,\tilde{h}})\tilde{\mathbf{U}}\mathbf{e}_{\tilde{h}}\|_2) =& e^{3\sqrt{2a_{K,H}^\top \tilde{\mathbf{u}}_H}} \\ \leq& e^{3\sqrt{2(d-1)\tilde{\Delta}}} \\ \leq& 2, \end{aligned}$$

where the last inequality uses a bound for the ratio of variance (Lemma 15) and the last inequality uses sufficiently small $\tilde{\Delta} \in (0, \frac{\log 2}{3\sqrt{2(d-1)}})$. Thus,

$$KL(\mathbb{P}_{\tilde{\mathbf{U}}}, \mathbb{P}_{\bar{\mathbf{U}}(\tilde{h},\tilde{i})}) \leq 4\tilde{\Delta}^2\mathbb{E}_{\tilde{\mathbf{U}}}[\sum_{k=1}^K \sigma^2_{k,\tilde{h}}] + \frac{8N\tilde{\Delta}^3}{\sqrt{6\kappa}}.$$

Plugging in (23),

$$\mathbb{E}_{\mathbf{U}}[M_{h,i}(1)] \geq \mathbb{E}_{\bar{\mathbf{U}}(h,i)}[M_{h,i}(1)] - K\sqrt{2\tilde{\Delta}^2 \sum_{k=1}^K \mathbb{E}_{\boldsymbol{\Theta}^\star + \tilde{\boldsymbol{\xi}}, \Pi_K}\big[\sigma^2_{k,h}\big] + \frac{4N\tilde{\Delta}^3}{\sqrt{6\kappa}}}.$$

Because $M_{h,i}(1) + M_{h,i}(-1) = K$,

$$\begin{aligned} &\mathbb{E}_{\mathbf{U}}[M_{h,i}(1)] + \mathbb{E}_{\bar{\mathbf{U}}(h,i)}[M_{h,i}(-1)] \\ &\geq \mathbb{E}_{\bar{\mathbf{U}}(h,i)}[M_{h,i}(1) + M_{h,i}(-1)] - K\sqrt{2\tilde{\Delta}^2 \sum_{k=1}^K \mathbb{E}_{\boldsymbol{\Theta}^\star + \tilde{\boldsymbol{\xi}}, \Pi_K}\big[\sigma^2_{k,h}\big] + \frac{4N\tilde{\Delta}^3}{\sqrt{6\kappa}}} \\ &\geq K - K\sqrt{2\tilde{\Delta}^2 \sum_{k=1}^K \mathbb{E}_{\boldsymbol{\Theta}^\star + \tilde{\boldsymbol{\xi}}, \Pi_K}\big[\sigma^2_{k,h}\big] + \frac{4N\tilde{\Delta}^3}{\sqrt{6\kappa}}} \end{aligned}$$

Thus, the sum of regret over $\mathcal{U} := \{\pm\sqrt{\tilde{\Delta}}\}^{H(d-1)} \times \{0\}^H$,

$$
\begin{aligned}
&\geq \sum_{\tilde{\mathbf{U}}\in\mathcal{U}} \frac{H\epsilon}{8(d-1)} \sum_{h=1}^{\lfloor H/2\rfloor}\sum_{i=1}^{d-1} \mathbb{E}_{\boldsymbol{\Theta}^\star+\tilde{\boldsymbol{\xi}},\Pi_K}\Big[M_{h,i}\big(\mathrm{sign}(\tilde{u}_{h,i})\big)\Big] \\
&\geq \frac{H\epsilon}{8(d-1)} \sum_{h=1}^{\lfloor H/2\rfloor}\sum_{i=1}^{d-1}\sum_{\tilde{\mathbf{U}}\in\mathcal{U}} \mathbb{E}_{\boldsymbol{\Theta}^\star+\tilde{\boldsymbol{\xi}},\Pi_K}\Big[M_{h,i}\big(\mathrm{sign}(\tilde{u}_{h,i})\big)\Big] \\
&= \frac{H\epsilon}{16(d-1)} \sum_{h=1}^{\lfloor H/2\rfloor}\sum_{i=1}^{d-1}\sum_{\tilde{\mathbf{U}}\in\mathcal{U}} \mathbb{E}_{\boldsymbol{\Theta}^\star+\tilde{\boldsymbol{\xi}},\Pi_K}\Big[M_{h,i}\big(-\mathrm{sign}(\tilde{u}_{h,i})\big)\Big] + \mathbb{E}_{\boldsymbol{\Theta}^\star+\tilde{\boldsymbol{\xi}},\Pi_K}\Big[M_{h,i}\big(\mathrm{sign}(\tilde{u}_{h,i})\big)\Big] \\
&\geq \frac{H\epsilon}{16(d-1)} \sum_{h=1}^{\lfloor H/2\rfloor}\sum_{i=1}^{d-1}\sum_{\tilde{\mathbf{U}}\in\mathcal{U}} \left(K - K\sqrt{2\tilde{\Delta}^2\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{\Theta}^\star+\tilde{\boldsymbol{\xi}},\Pi_K}\big[\sigma_{k,h}^2\big] + \frac{4N\tilde{\Delta}^3}{\sqrt{6\kappa}}}\right).
\end{aligned}
$$

For each $h \in [H]$, define the variance-maximizing pertubation

$$
\widehat{\mathbf{U}}_h(\Pi_K) := \operatorname*{argmax}_{\mathbf{U}\in\{\pm\sqrt{\tilde{\Delta}}\}^{H(d-1)}\times\{0\}^H} \mathbb{E}_{\boldsymbol{\Theta}^\star+\tilde{\boldsymbol{\xi}},\Pi_K}\big[\sum_{k=1}^{K}\sigma_{k,h}^2\big].
$$

Then,

$$
\begin{aligned}
&\sum_{\tilde{\mathbf{U}}\in\mathcal{U}} \frac{H\epsilon}{16(d-1)} \sum_{h=1}^{\lfloor H/2\rfloor}\sum_{i=1}^{d-1} \mathbb{E}_{\boldsymbol{\Theta}^\star+\tilde{\boldsymbol{\xi}},\Pi_K}\Big[M_{h,i}\big(\mathrm{sign}(\tilde{u}_{h,i})\big)\Big] \\
&\geq \frac{H\epsilon 2^{H(d-1)}}{16(d-1)} \sum_{h=1}^{\lfloor H/2\rfloor}\sum_{i=1}^{d-1} \left(K - K\sqrt{2\tilde{\Delta}^2\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{\Theta}^\star+\widehat{\mathbf{U}}_h(\Pi_K),\Pi_K}\big[\sigma_{k,h}^2\big] + \frac{4N\tilde{\Delta}^3}{\sqrt{6\kappa}}}\right).
\end{aligned}
$$

Using the averaging hammer,

$$
\begin{aligned}
&\max_{\tilde{\mathbf{U}}\in\mathcal{U}} \frac{H\epsilon}{8(d-1)} \sum_{h=1}^{\lfloor H/2\rfloor}\sum_{i=1}^{d-1} \mathbb{E}_{\boldsymbol{\Theta}^\star+\tilde{\boldsymbol{\xi}},\Pi_K}\Big[M_{h,i}\big(\mathrm{sign}(\tilde{u}_{h,i})\big)\Big] \\
&\geq \frac{1}{|\mathcal{U}|+1}\sum_{\tilde{\mathbf{U}}\in\mathcal{U}} \frac{H\epsilon}{8(d-1)} \sum_{h=1}^{\lfloor H/2\rfloor}\sum_{i=1}^{d-1} \mathbb{E}_{\boldsymbol{\Theta}^\star+\tilde{\boldsymbol{\xi}},\Pi_K}\Big[M_{h,i}\big(\mathrm{sign}(\tilde{u}_{h,i})\big)\Big] \\
&\geq \frac{2^{H(d-1)}}{2^{H(d-1)}+1}\cdot\frac{H\epsilon}{16(d-1)} \sum_{h=1}^{\lfloor H/2\rfloor}\sum_{i=1}^{d-1} \left(K - K\sqrt{2\tilde{\Delta}^2\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{\Theta}^\star+\tilde{\boldsymbol{\xi}},\Pi_K}\big[\sigma_{k,h}^2\big] + \frac{4N\tilde{\Delta}^3}{\sqrt{6\kappa}}}\right) \\
&\geq \frac{H\epsilon}{17(d-1)} \sum_{h=1}^{\lfloor H/2\rfloor}\sum_{i=1}^{d-1} \left(K - K\sqrt{2\tilde{\Delta}^2\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{\Theta}^\star+\tilde{\boldsymbol{\xi}},\Pi_K}\big[\sigma_{k,h}^2\big] + \frac{4N\tilde{\Delta}^3}{\sqrt{6\kappa}}}\right) \\
&= \frac{HN\epsilon}{17} \sum_{h=1}^{\lfloor H/2\rfloor} \left(1 - \sqrt{2\tilde{\Delta}^2\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{\Theta}^\star+\tilde{\boldsymbol{\xi}},\Pi_K}\big[\sigma_{k,h}^2\big] + \frac{4N\tilde{\Delta}^3}{\sqrt{6\kappa}}}\right),
\end{aligned}
$$

where the last inequality uses $2^{H(d-1)} \geq 16$. By Cauchy-Schwartz inequality,

$$\begin{aligned}
&\max_{\tilde{\mathbf{U}} \in \mathcal{U}} \frac{H\epsilon}{8(d-1)} \sum_{h=1}^{\lfloor H/2 \rfloor} \sum_{i=1}^{d-1} \mathbb{E}_{\boldsymbol{\Theta}^\star + \tilde{\boldsymbol{\xi}}, \Pi_K} \Big[M_{h,i}\big(\operatorname{sign}(\tilde{u}_{h,i})\big) \Big] \\
&\geq \frac{HN\epsilon}{17} \sum_{h=1}^{\lfloor H/2 \rfloor} \bigg(1 - \sqrt{2\tilde{\Delta}^2 \mathbb{E}_{\boldsymbol{\Theta}^\star + \widehat{\mathbf{U}}_h(\Pi_K), \Pi_K} [\sum_{k=1}^{K} \sigma_{k,h}^2] + \frac{4N\tilde{\Delta}^3}{\sqrt{6\kappa}}} \bigg). \\
&= \frac{HN\epsilon}{17} \bigg(\lfloor \frac{H}{2} \rfloor - \sum_{h=1}^{\lfloor H/2 \rfloor} \sqrt{2\tilde{\Delta}^2 \mathbb{E}_{\boldsymbol{\Theta}^\star + \widehat{\mathbf{U}}_h(\Pi_K), \Pi_K} [\sum_{k=1}^{K} \sigma_{k,h}^2] + \frac{4N\tilde{\Delta}^3}{\sqrt{6\kappa}}} \bigg) \\
&\geq \frac{HN\epsilon}{17} \bigg(\frac{H}{3} - \sqrt{\frac{H}{2}} \sqrt{2\tilde{\Delta}^2 \sum_{h=1}^{\lfloor H/2 \rfloor} \mathbb{E}_{\boldsymbol{\Theta}^\star + \widehat{\mathbf{U}}_h(\Pi_K), \Pi_K} [\sum_{k=1}^{K} \sigma_{k,h}^2] + \frac{2HN\tilde{\Delta}^3}{\sqrt{6\kappa}}} \bigg) \\
&= \frac{H^2 K\epsilon}{51} \bigg(1 - 3\sqrt{\frac{\tilde{\Delta}^2}{H} \sum_{h=1}^{\lfloor H/2 \rfloor} \mathbb{E}_{\boldsymbol{\Theta}^\star + \widehat{\mathbf{U}}_h(\Pi_K), \Pi_K} [\sum_{k=1}^{K} \sigma_{k,h}^2] + \frac{K\tilde{\Delta}^3}{\sqrt{6\kappa}}} \bigg).
\end{aligned}$$

It follows that

$$\begin{aligned}
&\max_{\tilde{\mathbf{U}} \in \mathcal{U}} \frac{H\epsilon}{8(d-1)} \sum_{h=1}^{\lfloor H/2 \rfloor} \sum_{i=1}^{d-1} \mathbb{E}_{\boldsymbol{\Theta}^\star + \tilde{\boldsymbol{\xi}}, \Pi_K} \Big[M_{h,i}\big(\operatorname{sign}(\tilde{u}_{h,i})\big) \Big] \\
&\geq \frac{H^2 K\epsilon}{51} \bigg(1 - 3\sqrt{\frac{\tilde{\Delta}^2}{H} \sum_{h=1}^{H} \mathbb{E}_{\boldsymbol{\Theta}^\star + \widehat{\mathbf{U}}_h(\Pi_K), \Pi_K} [\sum_{k=1}^{K} \sigma_{k,h}^2] + \frac{K\tilde{\Delta}^3}{\sqrt{6\kappa}}} \bigg).
\end{aligned}$$

Because $\sqrt{ax^2 + bx^3} \leq \sqrt{a}x + \frac{b}{\sqrt{a}}x^2$ for $a, b, x > 0$,

$$\begin{aligned}
&\sqrt{\frac{\tilde{\Delta}^2}{H} \sum_{h=1}^{H} \mathbb{E}_{\boldsymbol{\Theta}^\star + \widehat{\mathbf{U}}_h(\Pi_K), \Pi_K} [\sum_{k=1}^{K} \sigma_{k,h}^2] + \frac{K\tilde{\Delta}^3}{\sqrt{6\kappa}}} \\
&\leq \sqrt{\frac{1}{H} \sum_{h=1}^{H} \mathbb{E}_{\boldsymbol{\Theta}^\star + \widehat{\mathbf{U}}_h(\Pi_K), \Pi_K} [\sum_{k=1}^{K} \sigma_{k,h}^2]} \tilde{\Delta} + \frac{K\sqrt{H}\tilde{\Delta}^2}{\sqrt{6\kappa \sum_{h=1}^{H} \mathbb{E}_{\boldsymbol{\Theta}^\star + \widehat{\mathbf{U}}_h(\Pi_K), \Pi_K} [\sum_{k=1}^{K} \sigma_{k,h}^2]}}.
\end{aligned}$$

Let us write $\Xi(\Pi_K) := \sum_{h=1}^{H} \sum_{k=1}^{K} \mathbb{E}_{\tilde{\mathcal{S}}_{\boldsymbol{\Theta}^\star + \widehat{\mathbf{U}}_h(\Pi_K)}, \Pi_K} [\sigma_{k,h}^2]$. It follows that

$$\begin{aligned}
&\max_{\tilde{\mathbf{U}} \in \mathcal{U}} \frac{H\epsilon}{8(d-1)} \sum_{h=1}^{\lfloor H/2 \rfloor} \sum_{i=1}^{d-1} \mathbb{E}_{\boldsymbol{\Theta}^\star + \tilde{\boldsymbol{\xi}}, \Pi_K} \Big[M_{h,i}\big(\operatorname{sign}(\tilde{u}_{h,i})\big) \Big] \\
&\geq \frac{H^2 K\epsilon}{51} \bigg(1 - 3\sqrt{\frac{\Xi(\Pi_K)}{H}} \tilde{\Delta} - \frac{3N\sqrt{H}\kappa^{-1/2}\tilde{\Delta}^2}{\sqrt{6\Xi(\Pi_K)}} \bigg) \\
&= \frac{H^2 K\epsilon}{51} \bigg(1 - 3\sqrt{\frac{\Xi(\Pi_K)}{H}} \tilde{\Delta} - \frac{\sqrt{3}K\sqrt{H}\kappa^{-1/2}\tilde{\Delta}^2}{\sqrt{2\Xi(\Pi_K)}} \bigg)
\end{aligned}$$

Setting $\epsilon = \frac{12(d-1)\sqrt{\Xi(\Pi_K)}}{NH}$ we have

$$\begin{aligned}
&\max_{\tilde{\mathbf{U}} \in \mathcal{U}} \frac{H\epsilon}{8(d-1)} \sum_{h=1}^{\lfloor H/2 \rfloor} \sum_{i=1}^{d-1} \mathbb{E}_{\boldsymbol{\Theta}^\star + \tilde{\boldsymbol{\xi}}, \Pi_K} \Big[M_{h,i}\big(\operatorname{sign}(\tilde{u}_{h,i})\big) \Big] \\
&\geq \frac{4H^{3/2}(d-1)\sqrt{\Xi(\Pi_K)}}{17} \bigg(1 - 3\sqrt{\frac{\Xi(\Pi_K)}{H}} \tilde{\Delta} - \frac{\sqrt{3}K\sqrt{H}\kappa^{-1/2}\tilde{\Delta}^2}{\sqrt{2\Xi(\Pi_K)}} \bigg).
\end{aligned}$$

Because $K \geq 144H(d-1)^2$,

$$\epsilon = \frac{12(d-1)\sqrt{\Xi(\Pi_K)}}{NH} \leq \frac{12(d-1)}{\sqrt{NH}} \leq \frac{1}{H},$$

Thus,

$$\begin{aligned}
&\sup_{\mathbf{U}\in\{\pm\sqrt{\Delta}\}^{(d-1)H}\times\{0\}^{H\times d}} \mathrm{R}(\boldsymbol{\Theta}^\star + \boldsymbol{\xi}, \Pi_K) \\
&\geq \frac{4H(d-1)\sqrt{\Xi(\Pi_K)}}{17}\left(1 - 3\sqrt{\frac{\Xi(\Pi_K)}{H}}\tilde{\Delta} - \frac{\sqrt{3}K\sqrt{H}\kappa^{-1/2}\tilde{\Delta}^2}{\sqrt{2\Xi(\Pi_K)}}\right) \\
&\geq \frac{4H(d-1)\sqrt{\Xi(\Pi_K)}}{17}\left(1 - 3\sqrt{K}\tilde{\Delta} - \frac{\sqrt{3}\sqrt{K}\tilde{\Delta}^2}{\kappa}\right),
\end{aligned}$$

where the last inequality holds by the definition of the Fisher information lower bound (Definition 2) and $\Xi(\Pi_K) \geq \kappa NH$. By definition of $\widehat{\boldsymbol{\Theta}}_h(\Pi_K)$.

$$\begin{aligned}
&\sup_{\mathbf{U}\in\{\pm\sqrt{\Delta}\}^{(d-1)H}\times\{0\}^{H\times d}} \mathrm{R}(\boldsymbol{\Theta}^\star + \boldsymbol{\xi}, \Pi_K) \\
&\geq \frac{4H(d-1)\sqrt{\Xi(\Pi_K)}}{17}\left(1 - 3\sqrt{K}\tilde{\Delta} - \frac{\sqrt{3}\sqrt{K}\tilde{\Delta}^2}{\kappa}\right) \\
&\geq \frac{4H(d-1)\sqrt{\sum_{h=1}^H \mathbb{E}_{\widehat{\boldsymbol{\Theta}}_h(\Pi_K),\Pi_K}\left[\sum_{k=1}^K \sigma_{k,h}^2\right]}}{17}\left(1 - 3\sqrt{K}\tilde{\Delta} - \frac{\sqrt{3}\sqrt{K}\tilde{\Delta}^2}{\kappa}\right) \\
&\geq \frac{4H(d-1)\sqrt{\mathbb{E}_{\boldsymbol{\Theta}^\star,\Pi_K}[\sum_{h=1}^H \sum_{k=1}^K \sigma_{k,h}^2]}}{17}\left(1 - 3\sqrt{K}\tilde{\Delta} - \frac{\sqrt{3}\sqrt{K}\tilde{\Delta}^2}{\kappa}\right).
\end{aligned}$$

Note that

$$\begin{aligned}
&\max\{\sup_{\mathbf{U}\in\{\pm\sqrt{\Delta}\}^{(d-1)H}\times\{0\}^{H\times d}} \mathrm{R}(\boldsymbol{\Theta}^\star + \boldsymbol{\xi}, \Pi_K), \mathrm{R}(\boldsymbol{\Theta}^\star, \Pi_K)\} \\
&\geq \frac{4H(d-1)\sqrt{\mathbb{E}_{\boldsymbol{\Theta}^\star,\Pi_K}[\sum_{h=1}^H \sum_{k=1}^K \sigma_{k,h}^2]}}{17}\left(1 - 3\sqrt{K}\tilde{\Delta} - \frac{\sqrt{3}\sqrt{K}\tilde{\Delta}^2}{\kappa}\right)
\end{aligned}$$

Because $\boldsymbol{\Theta} \mapsto \mathrm{R}(\boldsymbol{\Theta}, \Pi_K)$ is Lipchitz continuous, letting $\Delta \to 0$ implies $\tilde{\Delta} \to 0$ and

$$\mathrm{R}(\boldsymbol{\Theta}^\star, \Pi_K) \geq \frac{4H(d-1)\sqrt{\mathbb{E}_{\boldsymbol{\theta}^\star,\Pi_K}[\sum_{h=1}^H \sum_{k=1}^K \sigma_{k,h}^2]}}{17}.$$

Because $\sigma_{k,h}^2 \in [0,1]$, by Lemma 32, for any $\delta \in (0, 1/2)$,

$$\mathbb{P}\Big(\sum_{h=1}^H\sum_{k=1}^K(1-\sigma_{k,h}^2) \geq \mathbb{E}_{\boldsymbol{\theta}^\star,\Pi_K}[\sum_{h=1}^H\sum_{k=1}^K(1-\sigma_{k,h}^2)] - \frac{\delta}{1-\delta}\sqrt{HN\log\frac{1}{\delta}}\Big) \geq \delta.$$

Rearranging the terms, with probability at least $\delta \in (0,1)$,

$$\mathbb{E}_{\boldsymbol{\theta}^\star,\Pi_K}[\sum_{h=1}^H\sum_{k=1}^K\sigma_{k,h}^2] \geq \sum_{h=1}^H\sum_{k=1}^K\sigma_{k,h}^2 - \frac{\delta}{1-\delta}\sqrt{HN\log\frac{1}{\delta}}$$

Because $K \geq H^{-1}\kappa^{-2}$, the right hand side is nonnegative for all $\delta \in (0, 1/e)$. Thus, with probability at least δ,

$$\begin{aligned}
\mathrm{R}(\boldsymbol{\Theta}^\star, \Pi_K) &\geq \frac{4H(d-1)\sqrt{\sum_{h=1}^H \sum_{k=1}^K \sigma_{k,h}^2 - \frac{\delta}{1-\delta}\sqrt{HN\log\frac{1}{\delta}}}}{17} \\
&\geq \frac{2Hd\sqrt{\sum_{h=1}^H \sum_{k=1}^K \sigma_{k,h}^2 - \frac{\delta}{1-\delta}\sqrt{HN\log\frac{1}{\delta}}}}{17}.
\end{aligned}$$

Because

$$\frac{\delta}{1-\delta}\sqrt{\log\frac{1}{\delta}}\sum_{h=1}^{H}\sum_{k=1}^{K}\sigma_{k,h}^2 \geq \frac{\delta}{1-\delta}\sqrt{\log\frac{1}{\delta}}\kappa HN \geq \frac{\delta}{1-\delta}\sqrt{HN\log\frac{1}{\delta}},$$

we obtain the lower bound. □

C Technical Results

C.1 Properties of Self Concordant Functions

Lemma 15. *(Ratio of the self-concordant function) Suppose f is M_f-self-concordant. Then for any $\mathbf{x}, \mathbf{y}, \mathbf{w} \in \mathbb{R}^d$,*

$$\mathbf{x}^\top \nabla^2 f(\mathbf{y}+\mathbf{w})\mathbf{x} \leq \exp(M_f\|\mathbf{w}\|_2)\mathbf{x}^\top \nabla^2 f(\mathbf{y})\mathbf{x}.$$

Proof. Given $\mathbf{x}, \mathbf{y}, \mathbf{w} \in \mathbb{R}^d$ define $\psi(t) = \mathbf{x}^\top \nabla^2 f(\mathbf{y}+t\mathbf{w})\mathbf{x}$. Then, by the mean value theorem, there exists $\tilde{t} \in [0,1]$ such that

$$\log\psi(1) - \log\psi(0) = \frac{\psi'(\tilde{t})}{\psi(\tilde{t})} = \frac{\sum_{i=1}^{S}\mathbf{x}^\top w_i D_i \nabla^2 f(\mathbf{y}+\tilde{t}\mathbf{w})\mathbf{x}}{\mathbf{x}^\top \nabla^2 f(\mathbf{y}+\tilde{t}\mathbf{w})\mathbf{x}}.$$

Because f is M_f-self-concordant,

$$\sum_{i=1}^{S}\mathbf{x}^\top w_i D_i \nabla^2 f(\mathbf{y}+\tilde{t}\mathbf{w})\mathbf{x} \leq M_f\|\mathbf{w}\|_2 \mathbf{x}^\top \nabla^2 f(\mathbf{y}+\tilde{t}\mathbf{w})\mathbf{x}.$$

Thus,

$$\log\psi(1) - \log\psi(0) \leq M_f\|\mathbf{w}\|_2,$$

which holds for any $\mathbf{x} \in \mathbb{R}^S$, Thus,

$$\psi(1) \leq \psi(0)\exp(M_f\|\mathbf{w}\|_2),$$

which completes the proof. □

Proposition 16. *Let $g : \mathbb{R}^d \to \mathbb{R}$ denote a self-concordant function with M_g. Then, for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$,*

$$\int_0^1 \nabla^2 L_{\nu,h}\big(\mathbf{x}_1 + t(\mathbf{x}_2-\mathbf{x}_1)\big)dt \succeq \frac{\nabla^2 L_{\nu,h}(\mathbf{x}_1)}{1+M_g\|\mathbf{x}_1-\mathbf{x}_2\|_2},$$

and

$$\int_0^1 \nabla^2 L_{\nu,h}\big(\mathbf{x}_1 + t(\mathbf{x}_2-\mathbf{x}_1)\big)dt \succeq \frac{\nabla^2 L_{\nu,h}(\mathbf{x}_2)}{1+M_g\|\mathbf{x}_1-\mathbf{x}_2\|_2}.$$

Proof. By Proposition 8 in Sun and Tran-Dinh 2019,

$$\nabla^2 L_{\nu,h}(\mathbf{x}_1)e^{-M_g t\|\mathbf{x}_1-\mathbf{x}_2\|_2} \preceq \nabla^2 L_{\nu,h}\big(\mathbf{x}_1 + t(\mathbf{x}_2-\mathbf{x}_1)\big).$$

Then,

$$\begin{aligned}
\int_0^1 \nabla^2 L_{\nu,h}\big(\mathbf{x}_1 + t(\mathbf{x}_2-\mathbf{x}_1)\big) &\succeq \nabla^2 L_{\nu,h}(\mathbf{x}_1)\int_0^1 e^{-M_g t\|\mathbf{x}_1-\mathbf{x}_2\|_2}dt \\
&= \nabla^2 L_{\nu,h}(\mathbf{x}_1)\frac{1-e^{-M_g\|\mathbf{x}_1-\mathbf{x}_2\|_2}}{M_g\|\mathbf{x}_1-\mathbf{x}_2\|_2} \\
&\succeq \nabla^2 L_{\nu,h}(\mathbf{x}_1)\frac{1}{1+M_g\|\mathbf{x}_1-\mathbf{x}_2\|_2},
\end{aligned}$$

where the last inequality uses $(1-e^{-x})/x \geq 1/(1+x)$ which is implied by $e^x \geq 1+x$. For the second inequality, changing the variable $u = 1-t$,

$$\begin{aligned}
\int_0^1 \nabla^2 L_{\nu,h}\big(\mathbf{x}_1 + t(\mathbf{x}_2-\mathbf{x}_1)\big)dt &= \int_0^1 \nabla^2 L_{\nu,h}\big(\mathbf{x}_1 + (1-u)(\mathbf{x}_2-\mathbf{x}_1)\big)du \\
&= \int_0^1 \nabla^2 L_{\nu,h}\big(\mathbf{x}_2 + t(\mathbf{x}_1-\mathbf{x}_2)\big)dt,
\end{aligned}$$

completes the proof. □

Lemma 17. *(Lemma 16 in Lee et al. 2024) Let $g : \mathbb{R}^d \to \mathbb{R}$ denote a function that satisfies*

$$e^{d(x_1,x_2)}\nabla^2 L_{\nu,h}(\mathbf{x}_2) \preceq \nabla^2 L_{\nu,h}(\mathbf{x}_1) \preceq e^{d(x_1,x_2)}\nabla^2 L_{\nu,h}(\mathbf{x}_2),$$

for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$ and a metric $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$. Then for any $\mathbf{x}_1, \mathbf{x}_2$ in the domain of g,

$$\int_0^1 \nabla^2 L_{\nu,h}\big(\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1)\big)(1-t)dt \succeq \frac{1}{2 + d(\mathbf{x}_1, \mathbf{x}_2)}\nabla^2 L_{\nu,h}(\mathbf{x}_1),$$

We develop a novel bound for the other side.

Lemma 18. *Let $g : \mathbb{R}^d \to \mathbb{R}$ denote a function that satisfies*

$$e^{d(x_1,x_2)}\nabla^2 L_{\nu,h}(\mathbf{x}_2) \preceq \nabla^2 L_{\nu,h}(\mathbf{x}_1) \preceq e^{d(x_1,x_2)}\nabla^2 L_{\nu,h}(\mathbf{x}_2),$$

for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$ and a metric $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$. Then for any $\mathbf{x}_1, \mathbf{x}_2$ in the domain of g,

$$\int_0^1 \nabla^2 L_{\nu,h}\big(\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1)\big)(1-t)dt \succeq \frac{1}{3 + d(\mathbf{x}_1, \mathbf{x}_2)^2}\nabla^2 L_{\nu,h}(\mathbf{x}_2).$$

Proof. Note that

$$\begin{aligned}\int_0^1 \nabla^2 L_{\nu,h}\big(\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1)\big)(1-t)dt \succeq& \bigg(\int_0^1 e^{-(1-t)d(\mathbf{x}_1,\mathbf{x}_2)}(1-t)dt\bigg)\nabla^2 L_{\nu,h}(\mathbf{x}_2)\\ =& \bigg(\frac{1 - (1 + d(\mathbf{x}_1,\mathbf{x}_2))e^{-d(\mathbf{x}_1,\mathbf{x}_2)}}{d(\mathbf{x}_1,\mathbf{x}_2)^2}\bigg)\nabla^2 L_{\nu,h}(\mathbf{x}_2).\end{aligned}$$

Note that for $x > 0$, we have $e^x \geq x^3/3 + x^2/3 + x + 1$. Rearranging the terms,

$$(x+1)e^{-x} \leq \frac{3}{3 + x^2},$$

which imples

$$\frac{1 - (1+x)e^{-x}}{x^2} \geq \frac{1}{3 + x^2}.$$

Plugging in $x = d(\mathbf{x}_1, \mathbf{x}_2)$ completes the proof. □

C.2 Properties of Log-Sum Function

Proposition 19 (Moments of state transition)**.** *Let s_{h+1} denote a random variable distributed according to the transition kernel* (1)*. For $\mathcal{S}' \subseteq \mathcal{S}$, let*

$$L_{\mathcal{S}'}(\mathbf{u}) := \log\Big(\sum_{s \in \mathcal{S}'} \exp(\mathbf{e}_s^\top \mathbf{u})\Big) \tag{24}$$

denotes the log-sum function. Then the (conditional) mean and covariance of the Euclidean basis vector $\mathbf{e}_{s_{h+1}}$ is given by

$$\begin{aligned}\mathbb{E}[\mathbf{e}_{s_{h+1}} \mid s_h, a_h, \boldsymbol{\theta}_h^\star] &= \nabla L_{\mathcal{S}_h(s_h,a_h)}\big(\boldsymbol{\Phi}(s_h,a_h)\boldsymbol{\theta}_h^\star\big),\\ \mathbb{V}[\mathbf{e}_{s_{h+1}} \mid s_h, a_h, \boldsymbol{\theta}_h^\star] &= \nabla^2 L_{\mathcal{S}_h(s_h,a_h)}\big(\boldsymbol{\Phi}(s_h,a_h)\boldsymbol{\theta}_h^\star\big).\end{aligned}$$

Proposition 20. *[Tran-Dinh et al., 2015, Lemma 4] The log-sum function $L_{\mathcal{S}}(\mathbf{u}) = \log\big(\sum_{s\in\mathcal{S}} \exp(\mathbf{e}_s^\top \mathbf{u})\big)$ is $\sqrt{6}$-self-concordant.*

Proof. For $t \in \mathbb{R}$ and $\mathbf{x}, \mathbf{u} \in \mathbb{R}^d$, let $\psi(t) := L_{\mathcal{S}}(\mathbf{x} + t\mathbf{u})$. According to Theorem 3 in [Tran-Dinh et al., 2015] it is sufficient to prove

$$|\psi'''(t)| \leq \sqrt{6}\psi''(t)\big\|\mathbf{u}\big\|_2.$$

Let us fix $t \in \mathbb{R}$ and $\mathbf{x}, \mathbf{u} \in \mathbb{R}^d$ throughout the proof. For each $i \in [d]$, let $p_i = e^{x_i+tu_i}/(\sum_{j=1}^d e^{x_j+tu_j})$. According to Section A.6 in Tran-Dinh et al. [2015],

$$\psi''(t) := \sum_{i<j}(u_i-u_j)^2 p_i p_j$$

$$\psi'''(t) := \sum_{i<j}(u_i-u_j)^2 p_i p_j \big\{\sum_{k=1}^d (u_i+u_j-2u_k)p_k\big\}.$$

Suppose $k \in [d] \setminus \{i,j\}$. Then, by Cauchy-Schwarz inequality,

$$|u_i+u_j-2u_k| \le \sqrt{1+1+4}\sqrt{u_i^2+u_j^2+u_k^2} \le \sqrt{6}\big\|\mathbf{u}\big\|_2.$$

Suppose $k = i$, then

$$|u_i+u_j-2u_k| = |u_i-u_j| \le \sqrt{2}\sqrt{u_i^2+u_j^2} \le \sqrt{6}\big\|\mathbf{u}\big\|_2.$$

Thus,

$$\begin{aligned}|\psi'''(t)| \le& \sum_{i<j}(u_i-u_j)^2 p_i p_j \big\{\sum_{k=1}^d |u_i+u_j-2u_k|p_k\big\}\\ \le& \sqrt{6}\|\mathbf{u}\|_2 \sum_{i<j}(u_i-u_j)^2 p_i p_j\\ =& \sqrt{6}\|\mathbf{u}\|_2 \psi''(t),\end{aligned}$$

which completes the proof. □

Proposition 21. *For $L_{\mathcal{S}}(\boldsymbol{\eta}) := \log\big(\sum_{s\in\mathcal{S}}\exp(\mathbf{e}_s^\top\boldsymbol{\eta})\big)$ and $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in [-C, C]^d$ for some $C > 0$, define $\psi(t) := (\boldsymbol{\eta}_2-\boldsymbol{\eta}_1)^\top\nabla^2 L_{\mathcal{S}}\big(\boldsymbol{\eta}_1+t(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1)\big)(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1)$ for $t \in [0,1]$. Then we have*

$$\int_0^1 \psi(t)(1-t)dt \ge \frac{\psi(0)}{2+2C},$$

and

$$\int_0^1 \psi(t)(1-t)dt \ge \frac{\psi(1)}{3+4C^2}.$$

Proof. The proof follows that in Lemma 4 in Jézéquel et al. 2021. Let X denote a random variable which takes value $\mathbf{e}_i^\top(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1)$ with probability

$$p_i(t) := \frac{\exp\big(\mathbf{e}_i^\top\boldsymbol{\eta}_1 + t\mathbf{e}_i^\top(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1)\big)}{\sum_{j=1}^d \exp\big(\mathbf{e}_j^\top\boldsymbol{\eta}_1 + t\mathbf{e}_j^\top(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1)\big)}.$$

By (24) in Jézéquel et al. 2021,

$$\begin{aligned}\psi'(t) =& \sum_{i_1,i_2,i_3}\Big\{\nabla^3 g\big(\boldsymbol{\eta}_1+t(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1)\big)\Big\}_{i,j,k}(\eta_{2,i_1}-\eta_{1,i_1})(\eta_{2,i_2}-\eta_{1,i_2})(\eta_{2,i_3}-\eta_{1,i_3})\\ =& \mathbb{E}\big[(X-\mathbb{E}(X))^3\big].\end{aligned}$$

By (25) in Jézéquel et al. 2021,

$$\psi(t) = \mathbb{E}\big[(X-\mathbb{E}(X))^2\big].$$

Because $\boldsymbol{\eta}_1 \in [-C, C]$,

$$|\psi'(t)| \le 2C\mathbb{E}\big[(X-\mathbb{E}(X))^2\big] \le 2C\psi(t),$$

Setting $h''(t) = \psi(t)$, we obtain an $2C$-self-concordant function h. By Proposition 8 in Sun and Tran-Dinh 2019,

$$h''(t_1)e^{-2C|t_1-t_2|} \le h''(t_2) \le h''(t_1)e^{2C|t_1-t_2|},$$

for $t_1, t_2 \in [0,1]$. Applying Lemma 17 and Lemma 18 on $h'' = \psi$ completes the proof. □

C.3 Elliptical Potential Lemmas

Lemma 22. *(Eliptical Potential Lemma; Lemma 11 in Abbasi-Yadkori et al. [2011]) Let $\{\mathbf{x}_t : t \geq 1\}$ be a sequence in $\mathbb{R}^d$ such that $\|\mathbf{x}_t\|_2 \leq L$ for some $L > 0$. Define $\bar{\mathbf{V}}_t = \sum_{s=1}^{t} \mathbf{x}_s \mathbf{x}_s^\top + V$ for a positive definite matrix V such that $\lambda_{\min}(V) \geq L^2$. Then,*

$$\sum_{t=1}^{k} \|\mathbf{x}_t\|^2_{\bar{\mathbf{V}}_{t-1}^{-1}} \leq 2d \log \frac{(k+1)}{d}.$$

Lemma 23. *(Eliptical Potential Lemma for Features) For the feature matrix $\boldsymbol{\Phi}_{k,h} := \boldsymbol{\Phi}(s_{k,h}, a_{k,h})$ let $\mathbf{A}_{k,h} := \sum_{\nu=1}^{k} \boldsymbol{\Phi}_{\nu,h}^\top \boldsymbol{\Phi}_{\nu,h} + B_{\boldsymbol{\Phi}}^2 \mathbf{I}_d$ denote the Gram matrix at episode k and step h. Then, for each $h \in [H]$,*

$$\sum_{k=1}^{K} \max_{s \in \mathcal{S}_{k,h}} \|\boldsymbol{\Phi}_{k,h}^\top \mathbf{e}_s\|^2_{\mathbf{A}_{k-1,h}^{-1}} \leq 2d \log \frac{K+1}{d}.$$

Proof. For each $k \in [K]$ and $h \in [H]$ let $\tilde{s}_{k,h} := \operatorname*{argmax}_{s \in \mathcal{S}_{k,h}} \|\boldsymbol{\Phi}_{k,h}^\top \mathbf{e}_s\|^2_{\mathbf{A}_{k-1,h}^{-1}}$. Define $\tilde{A}_{k-1,h} := \sum_{\nu=1}^{k-1} \boldsymbol{\Phi}_{\nu,h}^\top \mathbf{e}_{\tilde{s}_{\nu,h}} \mathbf{e}_{\tilde{s}_{\nu,h}}^\top \boldsymbol{\Phi}_{\nu,h} + B_{\boldsymbol{\Phi}}^2 \mathbf{I}_d$. Then $\mathbf{A}_{k-1,h} \succeq \tilde{\mathbf{A}}_{k-1,h}$ and

$$\sum_{k=1}^{K} \max_{s \in \mathcal{S}_{k,h}} \|\boldsymbol{\Phi}_{k,h}^\top \mathbf{e}_s\|^2_{\mathbf{A}_{k-1,h}^{-1}} = \sum_{k=1}^{K} \|\boldsymbol{\Phi}_{k,h}^\top \mathbf{e}_{\tilde{s}_{k,h}}\|^2_{\mathbf{A}_{k-1,h}^{-1}} \leq \sum_{k=1}^{K} \|\boldsymbol{\Phi}_{k,h}^\top \mathbf{e}_{\tilde{s}_{k,h}}\|^2_{\tilde{\mathbf{A}}_{k-1,h}^{-1}}.$$

Because $\|\boldsymbol{\Phi}_{k,h}^\top \mathbf{e}_{\tilde{s}_{k,h}}\|_2 \leq B_{\boldsymbol{\Phi}}$, applying Lemma 22, gives,

$$\sum_{k=1}^{K} \|\boldsymbol{\Phi}_{k,h}^\top \mathbf{e}_{\tilde{s}_{k,h}}\|^2_{\tilde{\mathbf{A}}_{k-1,h}^{-1}} \leq 2d \log \frac{K+1}{d},$$

which completes the proof. □

C.4 Regret Bound for Online Newton Step

Lemma 24 (Regret of the ONS estimate $\boldsymbol{\theta}_{k,h}$)**.** *Let $\boldsymbol{\theta}_{k,h}$ be the ONS estimate of $\boldsymbol{\theta}_h^\star$ obtained with the initial matrix $\widehat{\mathbf{H}}_{0,h} := \epsilon I_d$, where $\epsilon > 4B_{\boldsymbol{\Phi}}^2$. Set the learning rate $\eta_{\boldsymbol{\Phi},\boldsymbol{\theta}} := (e-1)(3 + 4B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2)/2$. Then, with probability at least $1-\delta$,*

$$\sum_{\nu=1}^{k} \big(\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) - \ell_{\nu,h}(\boldsymbol{\theta}_h^\star)\big) \leq \frac{2B_{\boldsymbol{\theta}}^2 \epsilon}{\eta_{\boldsymbol{\Phi},\boldsymbol{\theta}}} + d\eta_{\boldsymbol{\Phi},\boldsymbol{\theta}} \log \frac{k+1}{d} + \frac{8B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2}{\eta_{\boldsymbol{\Phi},\boldsymbol{\theta}}} \log \frac{1}{\delta}$$

for all $h \in [H]$ and $k \in [K]$.

Lemma 24 provides, to our knowledge, the first κ-*free* regret bound for the multinomial ONS estimator. Earlier κ-free results in this setting (e.g., Cho et al., 2024, Li et al., 2024) are derived for Online Mirror Descent based on Zhang and Sugiyama [2024], while the ONS bound for generalized linear models in Jun et al. [2017] retains a κ-dependence. The regret bound in Hazan et al. [2007] is for κ^{-1}-exp concave loss functions, and the resulting regret bound has κ^{-1}; the κ-free regret guarantees on MNL regression have not been established yet.

Proof. By definition of the Bregman divergence, for $\nu \in [k]$ and $h \in [H]$

$$\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) - \ell_{\nu,h}(\boldsymbol{\theta}_h^\star) = \nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top (\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star) - B_{\ell_{\nu,h}}(\boldsymbol{\theta}_h^\star, \boldsymbol{\theta}_{\nu-1,h}). \tag{25}$$

For brevity, let $L_{\mathcal{S}_h(s_{\nu,h}, a_{\nu,h})} := L_{\nu,h}$ denote the log-sum function for $\mathcal{S}_{\nu,h} := \mathcal{S}_h(s_{\nu,h}, a_{\nu,h})$ and let $\boldsymbol{\Phi}_{\nu,h} := \boldsymbol{\Phi}(s_{\nu,h}, a_{\nu,h})$. Then, by definition

$$\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) = \boldsymbol{\Phi}_{\nu,h}^\top \big(\nabla L_{\nu,h}(\boldsymbol{\Phi}_{\nu,h} \boldsymbol{\theta}_{\nu-1,h}) - \mathbf{e}_{s_{\nu,h+1}}\big).$$

By the information matrix bound (Lemma 27)

$$\begin{aligned}
&\sum_{\nu=1}^{k} \left(\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top (\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star)\right)^2 \\
&\leq 2(e-1) B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2 \sum_{\nu=1}^{k} B_{\ell_{\nu,h}}(\boldsymbol{\theta}_h^\star, \boldsymbol{\theta}_{\nu-1,h}) \\
&+ (e-1) \|\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star\|^2_{\nabla^2 \ell_{\nu,h}(\boldsymbol{\theta}_h^\star)} \\
&+ 16 B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2 \log \frac{1}{\delta}.
\end{aligned}$$

Let $\psi_{\nu,h}(t) := L_{\nu,h}\big(\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_{\nu-1,h} + t\boldsymbol{\Phi}_{\nu,h}(\boldsymbol{\theta}_h^\star - \boldsymbol{\theta}_{\nu-1,h})\big)$. By definition of Bregman divergence,

$$\begin{aligned}
B_{\ell_{\nu,h}}(\boldsymbol{\theta}_h^\star, \boldsymbol{\theta}_{\nu-1,h}) =& \ell_{\nu,h}(\boldsymbol{\theta}_h^\star) - \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) - \nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})(\boldsymbol{\theta}_h^\star - \boldsymbol{\theta}_{\nu-1,h}) \\
=& \frac{1}{2}\int_0^1 \psi''_{\nu,h}(t)(1-t)dt.
\end{aligned}$$

Because $\|\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_{\nu-1,h}\|_\infty \leq B_{\boldsymbol{\Phi}} B_{\boldsymbol{\theta}}$ and $\|\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_h^\star\|_\infty \leq B_{\boldsymbol{\Phi}} B_{\boldsymbol{\theta}}$, by the lower bound for the log-sum function (Proposition 21),

$$\begin{aligned}
&\int_0^1 \psi''_{\nu,h}(t)(1-t)dt \\
&\geq \frac{\|\boldsymbol{\Phi}_{\nu,h}(\boldsymbol{\theta}_h^\star - \boldsymbol{\theta}_{\nu-1,h})\|_{\nabla^2 L_{\nu,h}(\boldsymbol{\theta}_h^\star)}}{3 + 4B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2} \\
&= \frac{\|\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star\|_{\nabla^2 \ell_{\nu,h}(\boldsymbol{\theta}_h^\star)}}{3 + 4B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2}.
\end{aligned}$$

Thus, $\|\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star\|_{\nabla^2 \ell_{\nu,h}(\boldsymbol{\theta}_h^\star)} \leq (3 + 4B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2) B_{\ell_{\nu,h}}(\boldsymbol{\theta}_h^\star, \boldsymbol{\theta}_{\nu-1,h})$. Now we have

$$\begin{aligned}
&\sum_{\nu=1}^{k} \left(\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top (\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star)\right)^2 \\
&\leq (e-1)(3 + 4B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2) \sum_{\nu=1}^{k} B_{\ell_{\nu,h}}(\boldsymbol{\theta}_h^\star, \boldsymbol{\theta}_{\nu-1,h}) + 16 B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2 \log \frac{1}{\delta}.
\end{aligned}$$

Rearranging the terms,

$$\begin{aligned}
&-\sum_{\nu=1}^{k} B_{\ell_{\nu,h}}(\boldsymbol{\theta}_h^\star, \boldsymbol{\theta}_{\nu-1,h}) \\
&\leq \frac{-\sum_{\nu=1}^{k} \left(\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top (\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star)\right)^2 + 16 B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2 \log \frac{1}{\delta}}{(e-1)(3 + 4B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2)}.
\end{aligned}$$

From (25),

$$\begin{aligned}
&\sum_{\nu=1}^{k} \left(\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) - \ell_{\nu,h}(\boldsymbol{\theta}_h^\star)\right) \\
&= \sum_{\nu=1}^{k} \nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top (\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star) - \sum_{\nu=1}^{k} B_{\ell_{\nu,h}}(\boldsymbol{\theta}_h^\star, \boldsymbol{\theta}_{\nu-1,h}) \\
&\leq \sum_{\nu=1}^{k} \nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top (\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star) \\
&\quad + \frac{-\sum_{\nu=1}^{k} \left(\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top (\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star)\right)^2 + 16 B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2 \log \frac{1}{\delta}}{(e-1)(3 + 4B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2)}.
\end{aligned} \tag{26}$$

By definition of $\tilde{\boldsymbol{\theta}}_{\nu,h}$ in Algorithm 1,

$$\tilde{\boldsymbol{\theta}}_{\nu,h} - \boldsymbol{\theta}_h^\star = \boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star - \eta \widehat{\mathbf{H}}_{\nu,h}^{-1} \nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}),$$

which implies

$$\begin{aligned}\big\|\tilde{\boldsymbol{\theta}}_{\nu,h} - \boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{\nu,h}}^2 =& \big\|\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{k,h}}^2 - 2\eta\big(\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})\big)^\top (\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star) \\ &+ \eta^2 \big\|\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})\big\|_{\widehat{\mathbf{H}}_{\nu,h}^{-1}}^2.\end{aligned}$$

By the property of the generalized projection (see e.g., Lemma 8 in Hazan et al. 2007),

$$\big\|\tilde{\boldsymbol{\theta}}_{\nu,h} - \boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{\nu,h}}^2 \geq \big\|\boldsymbol{\theta}_{\nu,h} - \boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{\nu,h}}^2.$$

Thus,

$$\begin{aligned}\big\|\boldsymbol{\theta}_{\nu,h} - \boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{\nu,h}}^2 \leq& \big\|\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{\nu,h}}^2 - 2\eta\big(\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})\big)^\top (\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star) \\ &+ \eta^2 \big\|\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})\big\|_{(\widehat{\mathbf{H}}_{\nu,h})^{-1}}^2.\end{aligned}$$

For $\eta > 0$, rearranging the terms,

$$\begin{aligned}&\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top (\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star) \\ &\leq \frac{\big\|\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{\nu,h}}^2 - \big\|\boldsymbol{\theta}_{\nu,h} - \boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{\nu,h}}^2}{2\eta} + \frac{\eta}{2}\big\|\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})\big\|_{\widehat{\mathbf{H}}_{\nu,h}^{-1}}^2.\end{aligned}$$

Plugging in (26),

$$\begin{aligned}\sum_{\nu=1}^k \Big(\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) - \ell_{\nu,h}(\boldsymbol{\theta}_h^\star)\Big) \leq& \frac{\eta}{2}\big\|\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})\big\|_{\widehat{\mathbf{H}}_{\nu,h}^{-1}}^2 + \frac{16 B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2 \log\frac{1}{\delta}}{(e-1)(3+4B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2)} \\ &+ \sum_{\nu=1}^k \frac{\big\|\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{\nu,h}}^2 - \big\|\boldsymbol{\theta}_{\nu,h} - \boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{\nu,h}}^2}{2\eta} \\ &- \frac{1}{(e-1)(3+4B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2)} \sum_{\nu=1}^k \big(\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top (\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star)\big)^2.\end{aligned}$$

Setting $\eta = (e-1)(3+4B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2)/2$, we obtain,

$$\begin{aligned}&\sum_{\nu=1}^k \frac{\big\|\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{k,h}}^2 - \big\|\boldsymbol{\theta}_{\nu,h} - \boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{k,h}}^2}{2\eta} - \frac{1}{2\eta}\sum_{\nu=1}^k \big(\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top (\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star)\big)^2 \\ &= \sum_{\nu=1}^k \frac{\big\|\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{\nu-1,h}}^2 - \big\|\boldsymbol{\theta}_{\nu,h} - \boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_h^{(\nu)}}^2}{2\eta}.\end{aligned}$$

Thus,

$$\begin{aligned}&\sum_{\nu=1}^k \Big(\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) - \ell_{\nu,h}(\boldsymbol{\theta}_h^\star)\Big) \\ &\leq \sum_{\nu=1}^k \frac{\big\|\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{\nu-1,h}}^2 - \big\|\boldsymbol{\theta}_{\nu,h} - \boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{\nu,h}}^2}{2\eta} + \frac{\eta}{2}\sum_{\nu=1}^k \big\|\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})\big\|_{\widehat{\mathbf{H}}_{\nu,h}^{-1}}^2 + \frac{16 B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2 \log\frac{1}{\delta}}{2\eta} \\ &= \frac{\big\|\boldsymbol{\theta}_{0,h} - \boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{0,h}}^2 - \big\|\boldsymbol{\theta}_{k,h} - \boldsymbol{\theta}_h^\star\big\|_{\widehat{\mathbf{H}}_{k,h}}^2}{2\eta} + \frac{\eta}{2}\sum_{\nu=1}^k \big\|\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})\big\|_{\widehat{\mathbf{H}}_{\nu,h}^{-1}}^2 + \frac{16 B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2 \log\frac{1}{\delta}}{2\eta} \\ &\leq \frac{4 B_{\boldsymbol{\theta}}^2 \epsilon}{2\eta} + \frac{\eta}{2}\sum_{\nu=1}^k \big\|\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})\big\|_{\widehat{\mathbf{H}}_{\nu,h}^{-1}}^2 + \frac{16 B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2 \log\frac{1}{\delta}}{2\eta}\end{aligned}$$

Now, to bound the self-normalized gradient, recall that

$$\widehat{\mathbf{H}}_{\nu,h} := \sum_{\nu'=1}^{\nu} \nabla \ell_{\nu',h}(\boldsymbol{\theta}_{\nu'-1,h}) \nabla \ell_{\nu',h}(\boldsymbol{\theta}_{\nu'-1,h})^\top + \epsilon \mathbf{I}_d.$$

Because $\|\mathbf{e}_{s_{\nu,h+1}} - \nabla L_{\nu,h}(\boldsymbol{\eta}_{\nu,h})\|_1 \le 2$,

$$\begin{aligned}
\lambda_{\max}\big(\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) \nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top\big) =& \Big\| \nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) \Big\|_2^2 \\
=& \Big\| \big(\mathbf{e}_{s_{\nu,h+1}} - \nabla L_{\nu,h}(\boldsymbol{\eta}_{\nu,h})\big)^\top \boldsymbol{\Phi}(s_{\nu,h}, a_{\nu,h}) \Big\|_2^2 \\
\le& 4 \max_{\mathbf{x}:\|\mathbf{x}\|_1 \le 1} \Big\| \mathbf{x}^\top \boldsymbol{\Phi}(s_{\nu,h}, a_{\nu,h}) \Big\|_2^2.
\end{aligned} \tag{27}$$

Writing $\mathbf{x} = (x_1, \ldots, x_{|\mathcal{S}_{\nu,h}|})$, since $\max_{s' \in \mathcal{S}_{\nu,h}} \|\boldsymbol{\Phi}(s_{\nu,h}, a_{\nu,h})^\top \mathbf{e}_{s'}\|_2 \le B_{\boldsymbol{\Phi}}$,

$$\begin{aligned}
& 4 \max_{\mathbf{x}:\|\mathbf{x}\|_1 \le 1} \Big\| \mathbf{x}^\top \boldsymbol{\Phi}(s_{\nu,h}, a_{\nu,h}) \Big\|_2^2 \\
&= 4 \max_{\mathbf{x}:\|\mathbf{x}\|_1 \le 1} \Big\| \sum_{s' \in \mathcal{S}_{\nu,h}} x_{s'} \mathbf{e}_{s'}^\top \boldsymbol{\Phi}(s_{\nu,h}, a_{\nu,h}) \Big\|_2^2 \\
&= 4 \max_{\mathbf{x}:\|\mathbf{x}\|_1 \le 1} \sum_{s' \in \mathcal{S}_{\nu,h}} \sum_{s'' \in \mathcal{S}_{\nu,h}} x_{s'} x_{s''} \mathbf{e}_{s'}^\top \boldsymbol{\Phi}(s_{\nu,h}, a_{\nu,h}) \boldsymbol{\Phi}(s_{\nu,h}, a_{\nu,h})^\top \mathbf{e}_{s''} \\
&\le 4 B_{\boldsymbol{\Phi}}^2 \max_{\mathbf{x}:\|\mathbf{x}\|_1 \le 1} \Big(\sum_{s' \in \mathcal{S}_{\nu,h}} |x_{s'}| \Big)^2 \\
&\le 4 B_{\boldsymbol{\Phi}}^2.
\end{aligned} \tag{28}$$

Thus, $\|\nabla \ell_{\nu',h}(\boldsymbol{\theta}_{\nu'-1,h})\|_2 \le 2B_{\boldsymbol{\Phi}}$. Since $\lambda_{\min}(\widehat{\mathbf{H}}_{0,h}) = \epsilon \ge 4B_{\boldsymbol{\Phi}}^2$. Thus, by the elliptical potential bound (Lemma 22),

$$\sum_{\nu=1}^{k} \big\| \nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) \big\|_{\widehat{\mathbf{H}}_{\nu,h}^{-1}}^2 \le \sum_{\nu=1}^{k} \big\| \nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) \big\|_{\widehat{\mathbf{H}}_{\nu-1,h}^{-1}}^2 \le 2d \log \frac{k+1}{d},$$

which completes the proof. □

C.5 Hessian Approximation Error Bound

Lemma 25 (Hessian Approximation Error Bound)**.** *For each $h \in [H]$ and $k \in [K]$, with probability at least $1 - \delta$,*

$$\begin{aligned}
\Big\| \sum_{s' \in \mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(s') \cdot \mathbf{e}_{s'}^\top \widehat{\mathbf{V}}_{k,h} \boldsymbol{\Phi}_{k,h} \Big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}} \le& \Big\| \sum_{s' \in \mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(s') \cdot \mathbf{e}_{s'}^\top \mathbf{V}_{k,h}^\star \boldsymbol{\Phi}_{k,h} \Big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}} \\
&+ \frac{6\beta_{k-1} H}{(1-e^{-1})\kappa} \max_{s \in \mathcal{S}_{k,h}} \big\| \boldsymbol{\Phi}_{k,h} \mathbf{e}_s \big\|_{\mathbf{A}_{k-1,h}^{-1}}^2
\end{aligned}$$

Proof. Define $\mathbf{D}_{k-1,h} := \nabla^2 L_{k,h}\big(\boldsymbol{\Phi}(s_{k,h}, a_{k,h}) \widehat{\boldsymbol{\theta}}_{k-1,h}\big) - \nabla^2 L_{k,h}\big(\boldsymbol{\Phi}(s_{k,h}, a_{k,h}) \boldsymbol{\theta}_h^\star\big)$. For simplicity we write $\boldsymbol{\Phi}_{k,h} := \boldsymbol{\Phi}(s_{k,h}, a_{k,h})$. By triangular inequality,

$$\begin{aligned}
& \big\| \sum_{\tilde{s} \in \mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(\tilde{s}) \boldsymbol{\Phi}_{k,h}^\top \nabla^2 L_{\nu,h}\big(\boldsymbol{\Phi}_{k,h} \widehat{\boldsymbol{\theta}}_{k-1,h}\big) \mathbf{e}_{\tilde{s}} \big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}} \\
& \le \big\| \sum_{\tilde{s} \in \mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(\tilde{s}) \boldsymbol{\Phi}_{k,h}^\top \nabla^2 L_{\nu,h}\big(\boldsymbol{\Phi}_{k,h} \boldsymbol{\theta}_h^\star\big) \mathbf{e}_{\tilde{s}} \big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}} + \big\| \sum_{\tilde{s} \in \mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(\tilde{s}) \boldsymbol{\Phi}_{k,h}^\top \mathbf{D}_{k-1,h} \mathbf{e}_{\tilde{s}} \big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}}
\end{aligned}$$

For the second term,

$$\big\| \sum_{\tilde{s} \in \mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(\tilde{s}) \boldsymbol{\Phi}_{k,h}^\top \mathbf{D}_{k-1,h} \mathbf{e}_{\tilde{s}} \big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}} \le \big\| \widehat{\mathbf{H}}_{k-1,h}^{-1/2} \boldsymbol{\Phi}_{k,h}^\top \big\|_{1,2} \big\| \sum_{\tilde{s} \in \mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(\tilde{s}) \mathbf{D}_{k-1,h} \mathbf{e}_{\tilde{s}} \big\|_1, \tag{29}$$

where $\|\mathbf{A}\|_{1,2} = \sup\{\|\mathbf{A}\mathbf{x}\|_2 : \|\mathbf{x}\|_1 \le 1\}$ is the matrix norm. We carefully bound the ℓ_1 norm using the third derivative of $L_{k,h}$. By the mean value theorem, there exists $\bar{\boldsymbol{\eta}}$ such that

$$
\begin{aligned}
&\Big\| \sum_{\tilde{s}\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(\tilde{s}) \mathbf{D}_{k-1,h} \mathbf{e}_{\tilde{s}} \Big\|_1 \\
&= \Big\| \sum_{\tilde{s}\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(\tilde{s}) \mathbf{D}_{k-1,h} \big(\nabla(\nabla^2 L_{k,h}(\bar{\boldsymbol{\eta}})\mathbf{e}_{\tilde{s}})\big)^\top \boldsymbol{\Phi}_{k,h}(\widehat{\boldsymbol{\theta}}_{k-1,h} - \boldsymbol{\theta}^\star_h) \Big\|_1 \\
&= \sum_{s\in\mathcal{S}_{k,h}} \left| \sum_{\tilde{s}\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(\tilde{s}) \mathbf{e}_s^\top \big(\nabla(\nabla^2 L_{k,h}(\bar{\boldsymbol{\eta}})\mathbf{e}_{\tilde{s}})\big)^\top \boldsymbol{\Phi}_{k,h}(\widehat{\boldsymbol{\theta}}_{k-1,h} - \boldsymbol{\theta}^\star_h) \right| \\
&\le \sum_{s\in\mathcal{S}_{k,h}} \Big\| \sum_{\tilde{s}\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(\tilde{s}) \big(\nabla(\nabla^2 L_{k,h}(\bar{\boldsymbol{\eta}})\mathbf{e}_{\tilde{s}})\big)\mathbf{e}_s \Big\|_1 \Big\| \boldsymbol{\Phi}_{k,h}(\widehat{\boldsymbol{\theta}}_{k-1,h} - \boldsymbol{\theta}^\star_h) \Big\|_\infty,
\end{aligned}
$$

where the last inequality uses Hölder's inequality. For each $s \in \mathcal{S}_{k,h}$,

$$
\begin{aligned}
\Big\| \sum_{\tilde{s}\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(\tilde{s}) \big(\nabla(\nabla^2 L_{k,h}(\bar{\boldsymbol{\eta}})\mathbf{e}_{\tilde{s}})\big)\mathbf{e}_s \Big\|_1 &= \Big\| \sum_{\tilde{s}\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(\tilde{s}) D_s \nabla^2 L_{k,h}(\bar{\boldsymbol{\eta}})\mathbf{e}_{\tilde{s}} \Big\|_1 \\
&= \sum_{s'\in\mathcal{S}_{k,h}} \left| \sum_{\tilde{s}\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(\tilde{s}) \mathbf{e}_{s'}^\top D_s \nabla^2 L_{k,h}(\bar{\boldsymbol{\eta}})\mathbf{e}_{\tilde{s}} \right| \\
&\le \max_{s\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(s) \sum_{s',\tilde{s}\in\mathcal{S}_{k,h}} \left| \mathbf{e}_{\tilde{s}}^\top D_s \nabla^2 L_{k,h}(\bar{\boldsymbol{\eta}})\mathbf{e}_{s'} \right|.
\end{aligned}
$$

Let us write $\nabla L_{k,h}(\bar{\boldsymbol{\eta}}) = (p_1, \dots, p_{|\mathcal{S}_{k,h}|})^\top$. Then, (23) in Jézéquel et al. [2021] implies

$$
\begin{aligned}
\mathbf{e}_{\tilde{s}}^\top D_s \nabla^2 L_{k,h}(\bar{\boldsymbol{\eta}})\mathbf{e}_{s'} =& p_s \mathbb{I}(s = \tilde{s} = s') \\
&- p_s p_{\tilde{s}} \mathbb{I}(s = s') - p_s p_{s'} \mathbb{I}(s = \tilde{s}) - p_s p_{s'} \mathbb{I}(s' = \tilde{s}) \\
&+ 2 p_s p_{s'} p_{\tilde{s}}.
\end{aligned}
$$

Thus,

$$
\sum_{s',\tilde{s}\in\mathcal{S}_{k,h}} \left| \mathbf{e}_{\tilde{s}}^\top D_s \nabla^2 L_{k,h}(\bar{\boldsymbol{\eta}})\mathbf{e}_{s'} \right| = p_s + 3p_s + 2p_s = 6p_s.
$$

Gathering the terms,

$$
\begin{aligned}
\Big\| \sum_{\tilde{s}\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(\tilde{s}) \mathbf{D}_{k-1,h} \mathbf{e}_{\tilde{s}} \Big\|_1 &\le 6 \max_{s\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(s) \sum_{s\in\mathcal{S}_{k,h}} p_s \Big\| \boldsymbol{\Phi}_{k,h}(\widehat{\boldsymbol{\theta}}_{k-1,h} - \boldsymbol{\theta}^\star_h) \Big\|_\infty \\
&= 6 \max_{s\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(s) \Big\| \boldsymbol{\Phi}_{k,h}(\widehat{\boldsymbol{\theta}}_{k-1,h} - \boldsymbol{\theta}^\star_h) \Big\|_\infty
\end{aligned}
$$

By Cauchy-Schwartz inequality,

$$
\begin{aligned}
\Big\| \boldsymbol{\Phi}_{k,h}(\widehat{\boldsymbol{\theta}}_{k-1,h} - \boldsymbol{\theta}^\star_h) \Big\|_\infty &= \max_{s\in\mathcal{S}_{k,h}} \left| \mathbf{e}_s^\top \boldsymbol{\Phi}_{k,h}(\widehat{\boldsymbol{\theta}}_{k-1,h} - \boldsymbol{\theta}^\star_h) \right| \\
&\le \max_{s\in\mathcal{S}_{k,h}} \Big\| \widehat{\mathbf{H}}_{k-1,h}^{-1/2} \boldsymbol{\Phi}_{k,h} \mathbf{e}_s \Big\|_2 \Big\| \widehat{\boldsymbol{\theta}}_{k-1,h} - \boldsymbol{\theta}^\star_h \Big\|_{\widehat{\mathbf{H}}_{k-1,h}} \\
&\le \max_{s\in\mathcal{S}_{k,h}} \Big\| \boldsymbol{\Phi}_{k,h} \mathbf{e}_s \Big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}} \beta_{k-1},
\end{aligned}
$$

where the second inequality uses the confidence ellipsoid (Theorem 2). Plugging in (29)

$$
\begin{aligned}
&\Big\| \sum_{\tilde{s}\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(\tilde{s}) \boldsymbol{\Phi}_{k,h}^\top \mathbf{D}_{k-1,h} \mathbf{e}_{\tilde{s}} \Big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}} \\
&\le 6 \max_{s\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(s) \Big\| \widehat{\mathbf{H}}_{k-1,h}^{-1/2} \boldsymbol{\Phi}_{k,h}^\top \Big\|_{1,2} \max_{s\in\mathcal{S}_{k,h}} \Big\| \widehat{\mathbf{H}}_{k-1,h}^{-1/2} \boldsymbol{\Phi}_{k,h} \mathbf{e}_s \Big\|_2 \beta_{k-1}.
\end{aligned}
$$

By the bound for the matrix norm (Lemma 38),

$$
\Big\| \sum_{\tilde{s}\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(\tilde{s}) \boldsymbol{\Phi}_{k,h}^\top \mathbf{D}_{k-1,h} \mathbf{e}_{\tilde{s}} \Big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}} \le 6 \max_{s\in\mathcal{S}_{k,h}} \widehat{V}_{k,h+1}(s) \max_{s\in\mathcal{S}_{k,h}} \Big\| \boldsymbol{\Phi}_{k,h} \mathbf{e}_s \Big\|^2_{\widehat{\mathbf{H}}_{k-1,h}^{-1}} \beta_{k-1}.
$$

By the Hessian lower bound (Lemma 3), and the definition of the Fisher information lower bound,

$$\left\|\boldsymbol{\Phi}_{k,h}\mathbf{e}_s\right\|^2_{\widehat{\mathbf{H}}^{-1}_{k-1,h}} \leq \frac{1}{(1-e^{-1})^2}\left\|\boldsymbol{\Phi}_{k,h}\mathbf{e}_s\right\|^2_{(\mathbf{H}^{\star}_{k-1,h})^{-1}} \leq \frac{1}{(1-e^{-1})^2\kappa}\left\|\boldsymbol{\Phi}_{k,h}\mathbf{e}_s\right\|^2_{\mathbf{A}^{-1}_{k-1,h}}$$

Thus,

$$\Big\|\sum_{\tilde{s}\in\mathcal{S}_{k,h}}\widehat{V}_{k,h+1}(\tilde{s})\boldsymbol{\Phi}^{\top}_{k,h}\mathbf{D}_{k-1,h}\mathbf{e}_{\tilde{s}}\Big\|_{\widehat{\mathbf{H}}^{-1}_{k-1,h}} \leq \frac{6\beta_{k-1}\max_{s\in\mathcal{S}_{k,h}}\widehat{V}_{k,h+1}(s)}{(1-e^{-1})\kappa^1}\max_{s\in\mathcal{S}_{k,h}}\left\|\boldsymbol{\Phi}_{k,h}\mathbf{e}_s\right\|^2_{\mathbf{A}^{-1}_{k-1,h}}.$$

Since $\widehat{V}_{k,h+1}(s) \leq H$ for all $s \in \mathcal{S}_{k,h}$, the proof is completed. □

C.6 Online regret to self-normalized bound

To prove the ellipsoid confidence for our proposed estimator $\widehat{\boldsymbol{\theta}}_{k,h}$, we first prove the following result which relates the online regret to the self-normalized bound.

Lemma 26. *(Online regret to self-normalized bound) Define* $C_{\boldsymbol{\Phi},\boldsymbol{\theta}} := (e-1)(6+8B_{\boldsymbol{\Phi}}B_{\boldsymbol{\theta}}+2B^2_{\boldsymbol{\Phi}}B^2_{\boldsymbol{\theta}})$. *For each* $h \in [H]$, *with probability at least* $1-\delta$,

$$\begin{aligned}\sum_{\nu=1}^{k}\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) - \ell_{\nu,h}(\boldsymbol{\theta}^{\star}_h) \geq& \frac{1}{2C_{\boldsymbol{\Phi},\boldsymbol{\theta}}}\sum_{\nu=1}^{k}\Big(\nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^{\top}(\boldsymbol{\theta}_{\nu-1,h}-\boldsymbol{\theta}^{\star}_h)\Big)^2 \\ &- \frac{16B^2_{\boldsymbol{\Phi}}B^2_{\boldsymbol{\theta}}}{C_{\boldsymbol{\Phi},\boldsymbol{\theta}}}\log\frac{d}{\delta} - 2(e-1)(6+8B_{\boldsymbol{\Phi}}B_{\boldsymbol{\theta}})\log\frac{1}{\delta},\end{aligned}$$

for any $k \in [K]$ *and* $h \in [H]$.

Proof. By the definition of Bregman divergence,

$$\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) = \ell_{\nu,h}(\boldsymbol{\theta}^{\star}_h) + \nabla\ell_{\nu,h}(\boldsymbol{\theta}^{\star}_h)^{\top}(\boldsymbol{\theta}_{\nu-1,h}-\boldsymbol{\theta}^{\star}_h) + B_{\ell_{\nu,h}}(\boldsymbol{\theta}_{\nu-1,h},\boldsymbol{\theta}^{\star}_h)$$

Let us write the residual

$$\xi_{\nu,h} := \left\{\mathbf{e}_{s_{\nu,h+1}} - \mathbb{E}\left[\mathbf{e}_{s_{\nu,h+1}}\middle|s_{\nu,h},a_{\nu,h}\right]\right\}^{\top}\boldsymbol{\Phi}(s_{\nu,h},a_{\nu,h})(\boldsymbol{\theta}_{\nu-1,h}-\boldsymbol{\theta}^{\star}_h)$$

Then rearranging the terms,

$$\ell_{\nu,h}(\boldsymbol{\theta}_{\nu,h}) - \ell_{\nu,h}(\boldsymbol{\theta}^{\star}_h) = -\xi_{\nu,h} + B_{\ell_{\nu,h}}(\boldsymbol{\theta}_{\nu-1,h},\boldsymbol{\theta}^{\star}_h)$$

By Freedman's concentration inequality (Lemma 28), for any $v \in [0, \frac{1}{2B_{\boldsymbol{\Phi}}B_{\boldsymbol{\theta}}}]$, with probability at least $1-\delta$

$$\sum_{\nu=1}^{k}\xi_{\nu,h} \leq (e-2)v\sum_{\nu=1}^{k}\mathbb{E}\left[\xi^2_{\nu,h}\middle|s_{\nu,h},a_{\nu,h}\right] + \frac{1}{v}\log\frac{1}{\delta},$$

for all k and for each $h \in [H]$. Because $\|\boldsymbol{\theta}_{\nu-1,h}-\boldsymbol{\theta}^{\star}_h\|^2_{\nabla^2\ell_{\nu,h}(\boldsymbol{\theta}^{\star}_h)} = \mathbb{E}[\xi^2_{\nu,h}|s_{\nu,h},a_{\nu,h}]$, summing up,

$$\begin{aligned}&\sum_{\nu=1}^{k}\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) - \ell_{\nu,h}(\boldsymbol{\theta}^{\star}_h) \\ &\geq -(e-2)v\sum_{\nu=1}^{k}\|\boldsymbol{\theta}_{\nu-1,h}-\boldsymbol{\theta}^{\star}_h\|^2_{\nabla^2\ell_{\nu,h}(\boldsymbol{\theta}^{\star}_h)} \\ &\quad + \sum_{\nu=1}^{k}B_{\ell_{\nu,h}}(\boldsymbol{\theta}_{\nu-1,h},\boldsymbol{\theta}^{\star}_h) - \frac{1}{v}\log\frac{1}{\delta}.\end{aligned}$$

By the second order of the Taylor expansion,

$$B_{\ell_{\nu,h}}(\boldsymbol{\theta}_{\nu-1,h},\boldsymbol{\theta}^{\star}_h) = \frac{1}{2}\int_0^1\psi''_{\nu,h}(t)t dt,$$

where $\psi_{\nu,h}(t) := L_{\nu,h}\big(\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_{\nu-1,h} + t\boldsymbol{\Phi}_{\nu,h}(\boldsymbol{\theta}^\star_h - \boldsymbol{\theta}_{\nu-1,h})\big)$.By Proposition 21,

$$\frac{1}{2}\int_0^1 \psi''_{\nu,h}(t)tdt \geq \frac{\psi''_{\nu,h}(1)}{6+8B_{\boldsymbol{\Phi}}B_{\boldsymbol{\theta}}} = \frac{\|\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}^\star_h\|^2_{\nabla^2\ell_{\nu,h}(\boldsymbol{\theta}^\star_h)}}{6+8B_{\boldsymbol{\Phi}}B_{\boldsymbol{\theta}}}$$

Thus,

$$\sum_{\nu=1}^k \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) - \ell_{\nu,h}(\boldsymbol{\theta}^\star_h) \geq \big(1-(e-2)(6+8B_{\boldsymbol{\Phi}}B_{\boldsymbol{\theta}})v\big)\sum_{\nu=1}^k B_{\ell_{\nu,h}}(\boldsymbol{\theta}_{\nu-1,h}, \boldsymbol{\theta}^\star_h) - \frac{1}{v}\log\frac{1}{\delta}. \quad (30)$$

By Lemma 27,

$$\begin{aligned}
&\sum_{\nu=1}^k \big(\nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top(\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}^\star_h)\big)^2 \\
&\leq 2(e-1)B^2_{\boldsymbol{\Phi}}B^2_{\boldsymbol{\theta}}\sum_{\nu=1}^k B_{\ell_{\nu,h}}(\boldsymbol{\theta}_{\nu-1,h}, \boldsymbol{\theta}^\star_h) \\
&+ (e-1)\sum_{\nu=1}^k \big\|\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}^\star_h\big\|^2_{\nabla^2\ell_{\nu,h}(\boldsymbol{\theta}^\star_h)} + 16B^2_{\boldsymbol{\Phi}}B^2_{\boldsymbol{\theta}}\log\frac{d}{\delta} \\
&\leq (e-1)(6+8B_{\boldsymbol{\Phi}}B_{\boldsymbol{\theta}} + 2B^2_{\boldsymbol{\Phi}}B^2_{\boldsymbol{\theta}})\sum_{\nu=1}^k B_{\ell_{\nu,h}}(\boldsymbol{\theta}_{\nu-1,h}, \boldsymbol{\theta}^\star_h) + 16B^2_{\boldsymbol{\Phi}}B^2_{\boldsymbol{\theta}}\log\frac{d}{\delta}
\end{aligned}$$

Rearranging the terms,

$$\sum_{\nu=1}^k B_{\ell_{\nu,h}}(\boldsymbol{\theta}_{\nu-1,h}, \boldsymbol{\theta}^\star_h) \geq \frac{\sum_{\nu=1}^k \big(\nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top(\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}^\star_h)\big)^2}{(e-1)(6+8B_{\boldsymbol{\Phi}}B_{\boldsymbol{\theta}} + 2B^2_{\boldsymbol{\Phi}}B^2_{\boldsymbol{\theta}})} - \frac{16B^2_{\boldsymbol{\Phi}}B^2_{\boldsymbol{\theta}}\log\frac{d}{\delta}}{(e-1)(6+8B_{\boldsymbol{\Phi}}B_{\boldsymbol{\theta}} + 2B^2_{\boldsymbol{\Phi}}B^2_{\boldsymbol{\theta}})}.$$

Plugging in (30),

$$\begin{aligned}
&\sum_{\nu=1}^k \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) - \ell_{\nu,h}(\boldsymbol{\theta}^\star_h) \\
&\geq \frac{\big(1-(e-2)(6+8B_{\boldsymbol{\Phi}}B_{\boldsymbol{\theta}})v\big)}{(e-1)(6+8B_{\boldsymbol{\Phi}}B_{\boldsymbol{\theta}} + 2B^2_{\boldsymbol{\Phi}}B^2_{\boldsymbol{\theta}})}\sum_{\nu=1}^k \big(\nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top(\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}^\star_h)\big)^2 \\
&\quad - \frac{16B^2_{\boldsymbol{\Phi}}B^2_{\boldsymbol{\theta}}\log\frac{d}{\delta}}{(e-1)(6+8B_{\boldsymbol{\Phi}}B_{\boldsymbol{\theta}} + 2B^2_{\boldsymbol{\Phi}}B^2_{\boldsymbol{\theta}})} - \frac{1}{v}\log\frac{1}{\delta}
\end{aligned}$$

Setting $v = (1/2)(e-2)^{-1}(6+8B_{\boldsymbol{\Phi}}B_{\boldsymbol{\theta}})^{-1}$,

$$\begin{aligned}
\sum_{\nu=1}^k \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) - \ell_{\nu,h}(\boldsymbol{\theta}^\star_h) \geq &\frac{\sum_{\nu=1}^k \big(\nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top(\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}^\star_h)\big)^2}{2(e-1)(6+8B_{\boldsymbol{\Phi}}B_{\boldsymbol{\theta}} + 2B^2_{\boldsymbol{\Phi}}B^2_{\boldsymbol{\theta}})} \\
&- \frac{16B^2_{\boldsymbol{\Phi}}B^2_{\boldsymbol{\theta}}\log\frac{d}{\delta}}{(e-1)(6+8B_{\boldsymbol{\Phi}}B_{\boldsymbol{\theta}} + 2B^2_{\boldsymbol{\Phi}}B^2_{\boldsymbol{\theta}})} - 2(e-1)(6+8B_{\boldsymbol{\Phi}}B_{\boldsymbol{\theta}})\log\frac{1}{\delta}
\end{aligned}$$

which completes the proof. □

C.7 Information Matrix Bound

The following key concentration result for our analysis in an immediate consequence of (5).

Lemma 27 (Information matrix bound)**.** *For each step $h \in [H]$ and episode $\nu \in [K]$, let $\ell_{\nu,h}$ denote the negative log-likelihood defined in* (2)*. Then with probability at least $1-\delta$,*

$$\sum_{\nu=1}^k \big(\nabla\ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top(\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}^\star_h)\big)^2$$

$$\leq 2(e-1)B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2 \sum_{\nu=1}^{k} B_{\ell_{\nu,h}}(\boldsymbol{\theta}_h^\star, \boldsymbol{\theta}_{\nu-1,h}) + (e-1)\sum_{\nu=1}^{k} \|\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star\|^2_{\nabla^2 \ell_{\nu,h}(\boldsymbol{\theta}_h^\star)} + 16 B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2 \log\frac{1}{\delta}.$$

Lemma 27 furnishes a *lower bound* on $\sum_{\nu=1}^{k} \|\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star\|^2_{\nabla^2 \ell_{\nu,h}(\boldsymbol{\theta}_h^\star)}$ that is independent of κ. The proof of Lemma 24 also involves creatively combining standard argument for the online step update with a new lower bound for self-concordant function (Lemma 17).

Proof. For $\nu \in [k]$ and $h \in [H]$,

$$\begin{aligned}
&|\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top (\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star)| \\
&\leq \left|\mathbf{e}_{s_{\nu,h+1}}^\top \boldsymbol{\Phi}_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star)\right| + \left|\nabla L_{\nu,h}(\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_{\nu-1,h})^\top \boldsymbol{\Phi}_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star)\right| \\
&\leq \left(\|\mathbf{e}_{s_{\nu,h+1}}^\top \boldsymbol{\Phi}_{\nu,h}\|_2 + \|\nabla L_{\nu,h}(\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_{\nu-1,h})^\top \boldsymbol{\Phi}_{\nu,h}\|_2\right)\|\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star\|_2 \\
&\leq 4 B_{\boldsymbol{\Phi}} B_{\boldsymbol{\theta}}.
\end{aligned}$$

By the concentration inequality (Lemma 29), for any $k \geq 1$, with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{\nu=1}^{k} \left(\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top (\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star)\right)^2 \\
&\leq (e-1)\sum_{\nu=1}^{k} \mathbb{E}\left[\left(\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top (\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star)\right)^2 \middle| \mathcal{F}_{\nu,h}\right] + 16 B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2 \log\frac{1}{\delta}.
\end{aligned}$$

For each $\nu \in [k]$ and $h \in [H]$, we have $\mathbb{E}[\nabla \ell_{\nu,h}(\boldsymbol{\theta}_h^\star)|\mathcal{F}_{\nu,h}] = 0$ and $\mathbb{E}[\nabla \ell_{\nu,h}(\boldsymbol{\theta}_h^\star)\nabla \ell_{\nu,h}(\boldsymbol{\theta}_h^\star)^\top|\mathcal{F}_{\nu,h}] = \nabla^2 \ell_{\nu,h}(\boldsymbol{\theta}_h^\star)$. Thus,

$$\begin{aligned}
&\mathbb{E}\left[\left(\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h})^\top (\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star)\right)^2 \middle| \mathcal{F}_{\nu,h}\right] \\
&= \mathbb{E}\left[\left(\left(\nabla \ell_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h}) - \nabla \ell_{\nu,h}(\boldsymbol{\theta}_h^\star) + \nabla \ell_{\nu,h}(\boldsymbol{\theta}_h^\star)\right)^\top (\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star)\right)^2 \middle| \mathcal{F}_{\nu,h}\right] \\
&= \mathbb{E}\left[\left(\left(\boldsymbol{\Phi}_{\nu,h}^\top\left(\nabla L_{\nu,h}(\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_{\nu-1,h}) - \nabla L_{\nu,h}(\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_h^\star)\right) + \nabla \ell_{\nu,h}(\boldsymbol{\theta}_h^\star)\right)^\top (\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star)\right)^2 \middle| \mathcal{F}_{\nu,h}\right] \\
&= \left(\left(\nabla L_{\nu,h}(\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_{\nu-1,h}) - \nabla L_{\nu,h}(\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_h^\star)\right)^\top \boldsymbol{\Phi}_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star)\right)^2 + \|\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star\|^2_{\nabla^2 \ell_{\nu,h}(\boldsymbol{\theta}_h^\star)}.
\end{aligned}$$

By the property of the mean-variance property, $\nabla L_{\nu,h}(\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_{\nu-1,h}) = \mathbb{E}_{\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_{\nu-1,h}}[\mathbf{e}_s]$. Thus,

$$\begin{aligned}
&\left(\nabla L_{\nu,h}(\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_{\nu-1,h}) - \nabla L_{\nu,h}(\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_h^\star)\right)^\top \boldsymbol{\Phi}_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star) \\
&= \mathbb{E}_{\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_{\nu-1,h}}\left[\mathbf{e}_s^\top \boldsymbol{\Phi}_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star)\right] - \mathbb{E}_{\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_h^\star}\left[\mathbf{e}_s^\top \boldsymbol{\Phi}_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star)\right] \\
&\leq 2 B_{\boldsymbol{\Phi}} B_{\boldsymbol{\theta}} \cdot TV(\mathbb{P}_{\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_{\nu-1,h}}, \mathbb{P}_{\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_h^\star}),
\end{aligned}$$

By Pinsker's inequality,

$$\begin{aligned}
&\left|\left(\nabla L_{\nu,h}(\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_{\nu-1,h}) - \nabla L_{\nu,h}(\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_h^\star)\right)^\top \boldsymbol{\Phi}_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star)\right| \\
&\leq B_{\boldsymbol{\Phi}} B_{\boldsymbol{\theta}} \sqrt{2KL(\mathbb{P}_{\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_{\nu-1,h}}, \mathbb{P}_{\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_h^\star})} \\
&= B_{\boldsymbol{\Phi}} B_{\boldsymbol{\theta}} \sqrt{2 B_{\ell_{\nu,h}}(\boldsymbol{\theta}_h^\star, \boldsymbol{\theta}_{\nu-1,h})},
\end{aligned}$$

where the last inequality holds by the relationship between the Bregman divergence and KL-divergence for multinomial distribution. Thus,

$$\left(\left(\nabla L_{\nu,h}(\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_{\nu-1,h}) - \nabla L_{\nu,h}(\boldsymbol{\Phi}_{\nu,h}\boldsymbol{\theta}_h^\star)\right)^\top \boldsymbol{\Phi}_{\nu,h}(\boldsymbol{\theta}_{\nu-1,h} - \boldsymbol{\theta}_h^\star)\right)^2 \leq 2 B_{\boldsymbol{\Phi}}^2 B_{\boldsymbol{\theta}}^2 B_{\ell_{\nu,h}}(\boldsymbol{\theta}_h^\star, \boldsymbol{\theta}_{\nu-1,h}),$$

which completes the proof. □

C.8 Concentration Inequalities

Lemma 28. *(Freedman's inequality) [Lee et al., 2024, Lemma 3] Let $\{D_s : s \geq 1\}$ denote a martingale difference sequence adapted to $\mathcal{F}_s$ with $\max|D_s| \leq R$ almost surely. Then for any $\delta \in (0,1)$ and any $\eta \in [0, 1/R]$, with probability at least $1-\delta$,*

$$\sum_{s=1}^{t} D_s \leq (e-2)\eta \sum_{s=1}^{t} \mathbb{E}\big[D_s^2 \big| \mathcal{F}_{s-1}\big] + \frac{1}{\eta}\log\frac{1}{\delta},$$

for all $t \geq 1$.

Lemma 29. *(Chernoff Bound) Let $\{X_s : s \geq 1\}$ denote a sequence of nonnegative random variables adapted to $\mathcal{F}_s$ with $\max X_s \leq R$ almost surely. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sum_{s=1}^{t} X_s \leq (e-1)\sum_{s=1}^{t} \mathbb{E}\big[X_s \big| \mathcal{F}_{s-1}\big] + R\log\frac{1}{\delta},$$

and

$$\sum_{s=1}^{t} X_s \geq (1-e^{-1})\sum_{s=1}^{t} \mathbb{E}\big[X_s \big| \mathcal{F}_{s-1}\big] - R\log\frac{1}{\delta},$$

for all $t \geq 1$.

Proof. For any $x > 0$ and $y > 0$, by Markov inequality,

$$\mathbb{P}\Big(\sum_{s=1}^{t} X_s > (e-1)\sum_{s=1}^{t}\mathbb{E}\big[X_s|\mathcal{F}_{s-1}\big] + R\log\frac{1}{\delta}\Big) \leq \delta\mathbb{E}\exp\Big(\sum_{s=1}^{t}\frac{X_s}{R} - (e-1)\sum_{s=1}^{t}\frac{\mathbb{E}\big[X_s|\mathcal{F}_{s-1}\big]}{R}\Big)$$

For $x \in [0,1]$, we have $1 + x \leq e^x \leq 1 + (e-1)x$. For each $s \in [t]$,

$$\begin{aligned}\mathbb{E}\Big[\exp\Big(\frac{X_s}{R}\Big)\Big|\mathcal{F}_{s-1}\Big] \leq& 1 + (e-1)\frac{\mathbb{E}\Big[X_s\Big|\mathcal{F}_{s-1}\Big]}{R} \\ \leq& \exp\left((e-1)\frac{\mathbb{E}\Big[X_s\Big|\mathcal{F}_{s-1}\Big]}{R}\right).\end{aligned}$$

Thus, the sequence

$$\left\{\exp\left(\sum_{s=1}^{t}\frac{X_s}{R} - (e-1)\sum_{s=1}^{t}\frac{\mathbb{E}\big[X_s|\mathcal{F}_{s-1}\big]}{R}\right) : t \geq 1\right\},$$

is a super-martingale. Following the stopping time argument (see e.g., proof of Theorem 1 in Abbasi-Yadkori et al. 2011) we have

$$\mathbb{P}\left(\bigcup_{t\geq 1}\left\{\sum_{s=1}^{t} X_s > (e-1)\sum_{s=1}^{t}\mathbb{E}\big[X_s|\mathcal{F}_{s-1}\big] + R\log\frac{1}{\delta}\right\}\right) \leq \delta.$$

□

Lemma 30 (Matrix Chernoff Bound [Tropp, 2012])**.** *Let $\{\mathbf{X}_s : s \geq 1\}$ be a sequence of nonnegative semi-definite random matrices in $\mathbb{R}^{d\times d}$, adapted to $\mathcal{F}_s$, with $\lambda_{\max}(\mathbf{X}_s) \leq R$ almost surely. Then, for any $\delta \in (0,1)$, the following inequalities hold with probability at least $1-\delta$:*

$$\sum_{s=1}^{t}\mathbf{X}_s \preceq (e-1)\sum_{s=1}^{t}\mathbb{E}\big[\mathbf{X}_s\big|\mathcal{F}_{s-1}\big] + R\log\frac{d}{\delta}$$

and

$$\sum_{s=1}^{t}\mathbf{X}_s \succeq (1-e^{-1})\sum_{s=1}^{t}\mathbb{E}\big[\mathbf{X}_s\big|\mathcal{F}_{s-1}\big] - R\log\frac{d}{\delta}$$

for all $t \geq 1$.

Lemma 31 (Theorem 1 in Faury et al. 2020). *Let $\{\mathcal{F}_t\}_{t=1}^{\infty}$ be a filtration. Let $\{x_t\}_{t=1}^{\infty}$ be a stochastic process in $B_2(d)$ such that x_t is $\mathcal{F}_t$ measurable. Let $\{\epsilon_{t+1}\}_{t=2}^{\infty}$ be a martingale difference sequence such that ϵ_{t+1} is $\mathcal{F}_{t+1}$ measurable. Furthermore, assume that conditionally on $\mathcal{F}_t$ we have $|\epsilon_{t+1}| \leq 1$ almost surely, and note $\sigma_t^2 := \mathbb{E}[\epsilon_{t+1}^2|\mathcal{F}_t]$. Let $\lambda > 0$ and for any $t \geq 1$ define:*

$$\mathbf{H}_t := \sum_{s=1}^{t-1} \sigma_s^2 x_s x_s^\top + \lambda I_d, \qquad S_t := \sum_{s=1}^{t-1} \epsilon_s x_s.$$

Then for any $\delta \in (0, 1]$:

$$\mathbb{P}\left(\exists t \geq 1,\ \|S_t\|_{\mathbf{H}_t^{-1}} \geq \frac{\sqrt{\lambda}}{2} + \frac{2}{\sqrt{\lambda}} \log\left(\frac{\det(\mathbf{H}_t)^{1/2} \lambda^{-d/2}}{\delta} \right) + \frac{2}{\sqrt{\lambda}} d \log(2) \right) \leq \delta.$$

Lemma 32 (Anti-concentration inequality for nonnegative bounded random variables). *Let $\{X_i : i \geq 1\}$ denote the sequence of nonnegative random variables adapted to filtration $\{\mathcal{F}_i : i \geq 1\}$ such that $0 \leq X_i \leq b$. Let $\mu_k := \mathbb{E}[\sum_{i=1}^k X_i]$. Then for any $\delta \in (0, 1)$,*

$$\mathbb{P}\left(S_k \geq \mu_k - \frac{\delta}{1-\delta} \sqrt{2kb^2 \log \frac{1}{\delta}} \right) \geq \delta.$$

Proof. Let us write $S_k := \sum_{i=1}^k X_i$ and $\mu_k := \mathbb{E}[S_k]$. Then for any $k \geq 1$, for $\theta \in \mathbb{R}$ to be determined later,

$$\begin{aligned} \mu_k =& \mathbb{E}[S_k \mathbb{I}(S_k \leq \theta)] + \mathbb{E}[S_k \mathbb{I}(S_k > \theta)] \\ \leq& \theta \mathbb{P}(S_k \leq \theta) + \mathbb{E}[S_k \mathbb{I}(S_k > \theta)]. \end{aligned}$$

Rearranging the terms,

$$\mu_k - \theta \leq -\theta \mathbb{P}(S_k > \theta) + \mathbb{E}[S_k \mathbb{I}(S_k > \theta)].$$

By the definition of the conditional expectation,

$$\mu_k - \theta \leq \mathbb{P}(S_k > \theta) \big(\mathbb{E}[S_k | S_k > \theta] - \theta \big) \tag{31}$$

For any $\lambda > 0$, by the entropic variational principle (Lemma 41)

$$\mathbb{E}\big[S_k \big| S_k > \theta\big] \leq \frac{\log \mathbb{E}[e^{\lambda S_k}] + KL(\mathbb{P}_{S_k | S_k > \theta}, \mathbb{P}_{S_k})}{\lambda}.$$

The Kullback-Libeler divergence,

$$KL(\mathbb{P}_{S_k | S_k > \theta}, \mathbb{P}_{S_k}) = \mathbb{E}\big[\log \frac{\mathbb{I}(S_k > \theta)}{\mathbb{P}(S_k > \theta)} \big| S_k \geq \mu_k \big] = \log \frac{1}{\mathbb{P}(S_k > \theta)}.$$

It follows that

$$\mathbb{E}\big[S_k \big| S_k > \theta\big] \leq \frac{\log \mathbb{E}[e^{\lambda S_k}] + \log \frac{1}{\mathbb{P}(S_k > \theta)}}{\lambda}.$$

Note that $\log \mathbb{E}[e^{\lambda S_k}] = \log \mathbb{E}[e^{\lambda S_k - \lambda \mu_k}] + \lambda \mu_k$. For each $i = 1, \ldots, k$ let Y_i denote a random variable such that the distribution of $Y_i | \mathcal{F}_{i-1}$ is equal to $X_i | \mathcal{F}_{i-1}$ and X_i, Y_i are conditionally independent given $\mathcal{F}_{i-1}$. Let us write the conditional expectation $\mathbb{E}_i := \mathbb{E}[\cdot | \mathcal{F}_{i-1}]$. By the tower property of the conditional expectation, $\mu_k = \mathbb{E}\big[\sum_{i=1}^k \mathbb{E}_i[X_i]\big] = \mathbb{E}[\sum_{i=1}^k \mathbb{E}_i[Y_i]] = \mathbb{E}[\sum_{i=1}^k Y_i]$ and by Jensen's inequality,

$$\mathbb{E}[e^{\lambda S_k - \lambda \mu_k}] \leq \mathbb{E}[e^{\lambda \sum_{i=1}^k (X_i - Y_i)}].$$

Define $M_i(\lambda) = \exp\big(\lambda \sum_{j=1}^i (X_j - Y_j)\big)$ and $M_0(\lambda) = 1$. By the tower property of the conditional expectation, because $Y_1, \ldots, Y_{k-1}$ are independent of X_k and Y_k

$$\begin{aligned} \mathbb{E}[M_k(\lambda)] =& \mathbb{E}\Big[\mathbb{E}[M_k(\lambda) | \mathcal{F}_{k-1}] \Big] \\ =& \mathbb{E}\Big[\mathbb{E}[M_k(\lambda) | Y_1, \ldots, Y_{k-1}, \mathcal{F}_{k-1}] \Big] \end{aligned}$$

$$
\begin{aligned}
=&\mathbb{E}\Big[M_{k-1}(\lambda)\mathbb{E}[\exp(\lambda X_k - \lambda Y_k)|Y_1,\dots,Y_{k-1},\mathcal{F}_{k-1}]\Big]\\
=&\mathbb{E}\Big[M_{k-1}(\lambda)\mathbb{E}[\exp(\lambda X_k - \lambda Y_k)|\mathcal{F}_{k-1}]\Big].
\end{aligned}
$$

Since X_k and Y_k are conditionally identically distributed and independent given $\mathcal{F}_{k-1}$, for a Radamacher random variable ϵ_k,

$$
\begin{aligned}
\mathbb{E}[\exp(\lambda X_k - \lambda Y_k)|\mathcal{F}_{k-1}] =&\mathbb{E}\big[\exp\big(\lambda\epsilon_k(X_k - Y_k)\big)\big|\mathcal{F}_{k-1}\big]\\
\le&\exp(\frac{\lambda b^2}{2}),
\end{aligned}
$$

where the last inequality uses Hoeffing's inequality with the fact that $\epsilon_k(X_k - Y_k) \in [-b, b]$. Thus,

$$
\begin{aligned}
\mathbb{E}[M_k(\lambda)] \le&\mathbb{E}\big[M_{k-1}(\lambda)\big]\exp(\frac{\lambda^2 b^2}{2})\\
&\vdots\\
\le&\mathbb{E}\big[M_0(\lambda)\big]\exp(\frac{\lambda^2 k b^2}{2}).
\end{aligned}
$$

It follows that

$$
\mathbb{E}\big[S_k\big|S_k > \theta\big] \le \mu_k + \frac{\lambda k b^2}{2} + \frac{1}{\lambda}\log\frac{1}{\mathbb{P}(S_k > \theta)}.
$$

Setting $\lambda = \sqrt{\frac{2\log\frac{1}{\mathbb{P}(S_k>\theta)}}{kb^2}}$,

$$
\mathbb{E}\big[S_k\big|S_k > \theta\big] \le \mu_k + \sqrt{2kb^2\log\frac{1}{\mathbb{P}(S_k > \theta)}}
$$

Plugging in (31),

$$
\mu_k - \theta \le \mathbb{P}(S_k > \theta)\big(\mu_k + \sqrt{2kb^2\log\frac{1}{\mathbb{P}(S_k > \theta)}} - \theta\big)
$$

Rearranging the terms,

$$
\theta \ge \mu_k - \frac{\mathbb{P}(S_k > \theta)}{1 - \mathbb{P}(S_k > \theta)}\sqrt{2kb^2\log\frac{1}{\mathbb{P}(S_k > \theta)}}.
$$

Given $\delta \in (0, 1)$, setting $\theta = \mu_k - \frac{\delta}{(1-\delta)}\sqrt{2kb^2\log\frac{1}{\delta}}$ yields

$$
\frac{\delta}{(1-\delta)}\sqrt{2kb^2\log\frac{1}{\delta}} \le \frac{\mathbb{P}(S_k > \theta)}{1 - \mathbb{P}(S_k > \theta)}\sqrt{2kb^2\log\frac{1}{\mathbb{P}(S_k > \theta)}}.
$$

Because $p \mapsto \frac{p}{1-p}\sqrt{2kb^2\log\frac{1}{p}}$ is nondecreasing in $p \in (0, 1)$, we conclude $\delta \le \mathbb{P}(S_k > \theta)$. □

C.9 Matrix Potential Lemma

Lemma 33 (Matrix Elliptical Potential Lemma)**.** *Let $\{\mathbf{D}_\nu \in \mathbb{R}^{d\times d} : \nu \ge 0\}$ denote a sequence of positive definite symmetric matrices with $Tr(\mathbf{D}_\nu) \le b$ for all $\nu \ge 1$ for some $b > 0$. Let $\mathbf{M}_n = \sum_{\nu=1}^n \mathbf{D}_\nu + \mathbf{D}_0$ such that $\lambda_{\min}(\mathbf{D}_0) \ge \max\{1, \lambda_{\max}(\mathbf{D}_\nu)\}$. Then, for any $n \ge 1$*

$$
\sum_{\nu=1}^{n} Tr(\mathbf{M}_{n-1}^{-1/2}\mathbf{D}_n\mathbf{M}_{n-1}^{-1/2}) \le 2d\log(n+1).
$$

Proof. Let us fix $n \geq 1$ throughout the proof. Then

$$\begin{aligned}\det(\mathbf{M}_n) &= \det(\mathbf{M}_{n-1} + \mathbf{D}_n) \\ &= \det(\mathbf{M}_{n-1}) \det(\mathbf{I}_d + \mathbf{M}_{n-1}^{-1/2} \mathbf{D}_n \mathbf{M}_{n-1}^{-1/2}) \\ &\vdots \\ &= \det(\mathbf{D}_0) \prod_{\nu=1}^{n} \det(\mathbf{I}_d + \mathbf{M}_{\nu-1}^{-1/2} \mathbf{D}_\nu \mathbf{M}_{\nu-1}^{-1/2}).\end{aligned}$$

Taking logarithm on both sides,

$$\log \frac{\det(\mathbf{M}_n)}{\det(\mathbf{D}_0)} = \sum_{\nu=1}^{n} \log \det(\mathbf{I}_d + \mathbf{M}_{\nu-1}^{-1/2} \mathbf{D}_\nu \mathbf{M}_{\nu-1}^{-1/2})$$

Let $\lambda_{\nu,1}, \ldots, \lambda_{\nu,d}$ denote the eigenvalues of $\mathbf{M}_{\nu-1}^{-1/2} \mathbf{D}_\nu \mathbf{M}_{\nu-1}^{-1/2}$. For each $\nu \in [n]$, because $\mathbf{D}_0 \succeq \sup_{\nu \geq 1} \max\{1, \lambda_{\max}(\mathbf{D}_\nu)\} \succeq \mathbf{D}_\nu$

$$\begin{aligned}(\mathbf{M}_{\nu-1}^{-1/2} \mathbf{D}_\nu \mathbf{M}_{\nu-1}^{-1/2})^2 =& \mathbf{M}_{\nu-1}^{-1/2} \mathbf{D}_\nu \mathbf{M}_{\nu-1}^{-1} \mathbf{D}_\nu \mathbf{M}_{\nu-1}^{-1/2} \\ \preceq& \mathbf{M}_{\nu-1}^{-1/2} \mathbf{D}_\nu \mathbf{D}_0^{-1} \mathbf{D}_\nu \mathbf{M}_{\nu-1}^{-1/2} \\ \preceq& \mathbf{M}_{\nu-1}^{-1/2} \mathbf{D}_\nu \mathbf{D}_\nu^{-1} \mathbf{D}_\nu \mathbf{M}_{\nu-1}^{-1/2} \\ =& \mathbf{M}_{\nu-1}^{-1/2} \mathbf{D}_\nu \mathbf{M}_{\nu-1}^{-1/2}.\end{aligned}$$

By the property of contraction matrix (Lemma 37), $\lambda_{\nu,i} \in [0, 1]$ for $i \in [d]$. Thus,

$$\begin{aligned}\log \frac{\det(\mathbf{M}_n)}{\det(\mathbf{D}_0)} &= \sum_{\nu=1}^{n} \log \det(\mathbf{I}_d + \mathbf{M}_{\nu-1}^{-1/2} \mathbf{D}_\nu \mathbf{M}_{\nu-1}^{-1/2}) \\ &= \sum_{\nu=1}^{n} \log \prod_{i=1}^{d} (1 + \lambda_{\nu,i}) \\ &= \sum_{\nu=1}^{n} \sum_{i=1}^{d} \log(1 + \lambda_{\nu,i}) \\ &\geq \frac{1}{2} \sum_{\nu=1}^{n} \sum_{i=1}^{d} \lambda_{\nu,i}.\end{aligned}$$

where the last inequality holds because $\log(1 + x) \geq x/2$ for $x \in [0, 1]$. Equating $\sum_{i=1}^{d} \lambda_{\nu,i} = \mathrm{Tr}(\mathbf{M}_{\nu-1}^{-1/2} \mathbf{D}_\nu \mathbf{M}_{\nu-1}^{-1/2})$,

$$\sum_{\nu=1}^{n} \mathrm{Tr}(\mathbf{M}_{\nu-1}^{-1/2} \mathbf{D}_\nu \mathbf{M}_{\nu-1}^{-1/2}) \leq \log \frac{\det(\mathbf{M}_n)}{\det(\mathbf{D}_0)}$$

By GM-AM inequality, $\{\det(\mathbf{M}_n)\}^{1/d} \leq \mathrm{Tr}(\mathbf{M}_n)/d$ and,

$$\begin{aligned}\log \frac{\det(\mathbf{M}_n)}{\det(\mathbf{D}_0)} &= \log \det(\mathbf{D}_0^{-1/2} \mathbf{M}_n \mathbf{D}_0^{-1/2}) \\ &\leq d \log \frac{\mathrm{Tr}(\mathbf{D}_0^{-1/2} \mathbf{M}_n \mathbf{D}_0^{-1/2})}{d}.\end{aligned}$$

Since $\mathbf{D}_0 \succeq \mathbf{D}_\nu$ for all $\nu \geq 1$,

$$\begin{aligned}\mathrm{Tr}(\mathbf{D}_0^{-1/2} \mathbf{M}_n \mathbf{D}_0^{-1/2}) &= d + \sum_{\nu=1}^{n} \mathrm{Tr}(\mathbf{D}_0^{-1/2} \mathbf{D}_\nu \mathbf{D}_0^{-1/2}) \\ &\leq d(n+1),\end{aligned}$$

which completes the proof. □

C.10 Bounds for the Multinomial Transition Kernel Difference

Lemma 34 (The Second Order Difference Bound for Transition Kernel Difference)**.** *(Second Order Difference between the two transition kernels) Let $\mathbb{P}(\mathbf{e}_{s'}|s,a,\boldsymbol{\theta})$ denote the multinomial transition kernel defined in* (1)*. Then for any sequence $v_s \in \mathbb{R}$ for $s \in \tilde{\mathcal{S}} \subseteq \mathcal{S}$ and $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$,*

$$\begin{aligned}
&\sum_{s'\in\tilde{\mathcal{S}}} v_{s'}\big(\mathbb{P}(s'|s,a,\boldsymbol{\theta}_1) - \mathbb{P}(s'|s,a,\boldsymbol{\theta}_2)\big)\\
&\leq \sum_{\tilde{s}\in\tilde{\mathcal{S}}} v_{\tilde{s}}\mathbf{e}_{\tilde{s}}^\top \nabla^2 L_{\tilde{\mathcal{S}}}\big(\boldsymbol{\Phi}(s,a)\boldsymbol{\theta}_2\big)\boldsymbol{\Phi}(s,a)(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)\\
&+ \max_{\tilde{s}\in\tilde{\mathcal{S}}}|v_{\tilde{s}}| \max_{\tilde{s}\in\tilde{\mathcal{S}}}\big((\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)^\top\boldsymbol{\Phi}(s,a)^\top\mathbf{e}_{\tilde{s}}\big)^2
\end{aligned}$$

Proof. For $\boldsymbol{\theta}\in\mathbb{R}^d$ define the residual $\boldsymbol{\xi}_{s',\boldsymbol{\theta}} := \mathbf{e}_{s'} - \mathbb{E}[\mathbf{e}_{s'}|s,a,\boldsymbol{\theta}]$ of the multinomial distribution given $\boldsymbol{\theta}$. Applying the second order Taylor expansion on

$$\mathbb{P}(s'|s,a,\boldsymbol{\theta}) = \exp\big(\mathbf{e}_{s'}^\top\boldsymbol{\Phi}(s,a)\boldsymbol{\theta} - g(\boldsymbol{\Phi}(s,a)\boldsymbol{\theta})\big),$$

for any $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\mathbb{R}^d$, there exists $v\in[0,1]$ such that $\bar{\boldsymbol{\theta}} := v\boldsymbol{\theta}_1 + (1-v)\boldsymbol{\theta}_2$ and

$$\begin{aligned}
&\mathbb{P}(s'|s,a,\boldsymbol{\theta}_1) - \mathbb{P}(s'|s,a,\boldsymbol{\theta}_2)\\
&= \boldsymbol{\xi}_{s',\boldsymbol{\theta}_2}^\top\boldsymbol{\Phi}(s,a)(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)\mathbb{P}(s'|s,a,\boldsymbol{\theta}_2)\\
&+ \frac{1}{2}(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)^\top\boldsymbol{\Phi}(s,a)^\top\big\{\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}^\top - \nabla^2 L_{\nu,h}\big(\boldsymbol{\Phi}(s,a)^\top\bar{\boldsymbol{\theta}}\big)\big\}\boldsymbol{\Phi}(s,a)(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)\mathbb{P}(s'|s,a,\bar{\boldsymbol{\theta}}).
\end{aligned}$$

Summing up over $s'\in\mathcal{S}$,

$$\begin{aligned}
&\sum_{s'\in\mathcal{S}} v_{s'}\big(\mathbb{P}(s'|s,a,\boldsymbol{\theta}_1) - \mathbb{P}(s'|s,a,\boldsymbol{\theta}_2)\big)\\
&= \sum_{s'\in\mathcal{S}} v_{s'}\boldsymbol{\xi}_{s',\boldsymbol{\theta}_2}^\top\boldsymbol{\Phi}(s,a)(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)\mathbb{P}(s'|s,a,\boldsymbol{\theta}_2)\\
&+ \frac{1}{2}(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)^\top\boldsymbol{\Phi}(s,a)^\top\bigg(\sum_{s'\in\mathcal{S}} v_{s'}\Big(\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}^\top - v_{s'}\nabla^2 L_{\nu,h}\big(\boldsymbol{\Phi}(s,a)^\top\bar{\boldsymbol{\theta}}\big)\Big)\bigg)\boldsymbol{\Phi}(s,a)(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)\mathbb{P}(s'|s,a,\bar{\boldsymbol{\theta}}).
\end{aligned}$$

For the first term, because $\mathbb{E}[\boldsymbol{\xi}_{s',\boldsymbol{\theta}}|s,a,\boldsymbol{\theta}] = \mathbf{0}$.

$$\begin{aligned}
&\sum_{s'\in\mathcal{S}} v_{s'}\boldsymbol{\xi}_{s',\boldsymbol{\theta}_2}^\top\boldsymbol{\Phi}(s,a)(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)\mathbb{P}(s'|s,a,\boldsymbol{\theta}_2)\\
&= \sum_{\tilde{s}\in\mathcal{S}} v_{\tilde{s}}\mathbf{e}_{\tilde{s}}^\top\mathbb{E}\big[\mathbf{e}_{s'}\boldsymbol{\xi}_{s',\boldsymbol{\theta}_2}^\top|s,a,\boldsymbol{\theta}_2\big]\boldsymbol{\Phi}(s,a)(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)\\
&= \sum_{\tilde{s}\in\mathcal{S}} v_{\tilde{s}}\mathbf{e}_{\tilde{s}}^\top\mathbb{E}\big[\boldsymbol{\xi}_{s',\boldsymbol{\theta}_2}\boldsymbol{\xi}_{s',\boldsymbol{\theta}_2}^\top|s,a,\boldsymbol{\theta}_2\big]\boldsymbol{\Phi}(s,a)(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)\\
&= \sum_{\tilde{s}\in\mathcal{S}} v_{\tilde{s}}\mathbf{e}_{\tilde{s}}^\top\nabla^2 L_{\nu,h}\big(\boldsymbol{\Phi}(s,a)\boldsymbol{\theta}_2\big)\boldsymbol{\Phi}(s,a)(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2).
\end{aligned}$$

For the second term, because $\mathbb{E}[\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}^\top|s,a,\bar{\boldsymbol{\theta}}] - \nabla^2 L_{\nu,h}\big(\boldsymbol{\Phi}(s,a)^\top\bar{\boldsymbol{\theta}}\big) = \mathbf{O}$,

$$\begin{aligned}
&\sum_{s'\in\mathcal{S}} v_{s'}\Big(\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}^\top - \nabla^2 L_{\nu,h}\big(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}}\big)\Big)\mathbb{P}(s'|s,a,\bar{\boldsymbol{\theta}})\\
&= \sum_{\tilde{s}\in\mathcal{S}} v_{\tilde{s}}\mathbf{e}_{\tilde{s}}^\top\mathbb{E}\big[\mathbf{e}_{s'}\big(\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}^\top - \nabla^2 L_{\nu,h}\big(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}}\big)\big)\big|s,a,\bar{\boldsymbol{\theta}}\big]\\
&= \sum_{\tilde{s}\in\mathcal{S}} v_{\tilde{s}}\mathbf{e}_{\tilde{s}}^\top\mathbb{E}\big[\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}\big(\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}^\top - \nabla^2 L_{\nu,h}\big(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}}\big)\big)\big|s,a,\bar{\boldsymbol{\theta}}\big]\\
&= \sum_{\tilde{s}\in\mathcal{S}} v_{\tilde{s}}\mathbf{e}_{\tilde{s}}^\top\mathbb{E}\big[\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}^\top\big|s,a,\bar{\boldsymbol{\theta}}\big]\\
&\preceq \max_{\tilde{s}\in\mathcal{S}}|v_{\tilde{s}}|\mathbb{E}\big[\|\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}\|_1\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}^\top\big|s,a,\bar{\boldsymbol{\theta}}\big],
\end{aligned}$$

where the last matrix ordering holds by Hölder's inequality. Because $\|\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}\|_1 \leq 2$ almost surely,

$$\begin{aligned}
&\sum_{s' \in \mathcal{S}} v_{s'} \Big(\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}} \boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}^\top - \nabla^2 L_{\nu,h}\big(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}}\big) \Big) \mathbb{P}(s'|s,a,\bar{\boldsymbol{\theta}}) \\
&\preceq 2 \max_{\tilde{s} \in \mathcal{S}} |v_{\tilde{s}}| \mathbb{E}[\boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}} \boldsymbol{\xi}_{s',\bar{\boldsymbol{\theta}}}^\top \big| s,a,\bar{\boldsymbol{\theta}}] \\
&= 2 \max_{\tilde{s} \in \mathcal{S}} |v_{\tilde{s}}| \nabla^2 L_{\nu,h}\big(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}}\big).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\sum_{s' \in \mathcal{S}} v_{s'} \big(\mathbb{P}(s'|s,a,\boldsymbol{\theta}_1) - \mathbb{P}(s'|s,a,\boldsymbol{\theta}_2) \big) \\
&\leq \sum_{\tilde{s} \in \mathcal{S}} v_{\tilde{s}} \mathbf{e}_{\tilde{s}}^\top \nabla^2 L_{\nu,h}\big(\boldsymbol{\Phi}(s,a)\boldsymbol{\theta}_2\big) \boldsymbol{\Phi}(s,a)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) + \max_{\tilde{s} \in \mathcal{S}} |v_{\tilde{s}}| \big\| \boldsymbol{\Phi}(s,a)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \big\|^2_{\nabla^2 L_{\nu,h}\left(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}}\right)}.
\end{aligned}$$

Because

$$\begin{aligned}
\nabla^2 L_{\nu,h}\big(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}}\big) &= \mathrm{diag}\Big(\nabla L_{\nu,h}\big(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}}\big) \Big) - \nabla L_{\nu,h}\big(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}}\big) \nabla L_{\nu,h}\big(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}}\big)^\top \\
&\preceq \mathrm{diag}\Big(\nabla L_{\nu,h}\big(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}}\big) \Big),
\end{aligned}$$

we obtain,

$$\begin{aligned}
&\big\| \boldsymbol{\Phi}(s,a)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \big\|^2_{\nabla^2 L_{\nu,h}\left(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}}\right)} \\
&= (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top \boldsymbol{\Phi}(s,a)^\top \nabla^2 L_{\nu,h}\big(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}}\big) \boldsymbol{\Phi}(s,a)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \\
&= \sum_{s_1,s_2 \in \mathcal{S}} (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top \boldsymbol{\Phi}(s,a)^\top \mathbf{e}_{s_1} \mathbf{e}_{s_1}^\top \nabla^2 L_{\nu,h}\big(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}}\big) \mathbf{e}_{s_2} \mathbf{e}_{s_2}^\top \boldsymbol{\Phi}(s,a)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \\
&\leq \max_{s_1 \in \mathcal{S}} \big((\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top \boldsymbol{\Phi}(s,a)^\top \mathbf{e}_{s_1} \big)^2 \sum_{s_1,s_2 \in \mathcal{S}} |\mathbf{e}_{s_1}^\top \nabla^2 L_{\nu,h}\big(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}}\big) \mathbf{e}_{s_2}| \\
&\leq \max_{s_1 \in \mathcal{S}} \big((\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top \boldsymbol{\Phi}(s,a)^\top \mathbf{e}_{s_1} \big)^2 \sum_{s_1,s_2 \in \mathcal{S}} \Big| \mathbf{e}_{s_1}^\top \mathrm{diag}\Big(\nabla L_{\nu,h}\big(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}}\big) \Big) \mathbf{e}_{s_2} \Big| \\
&= \max_{s_1 \in \mathcal{S}} \big((\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top \boldsymbol{\Phi}(s,a)^\top \mathbf{e}_{s_1} \big)^2 \big\| \nabla L_{\nu,h}\big(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}}\big) \big\|_1 \\
&= \max_{s_1 \in \mathcal{S}} \big((\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top \boldsymbol{\Phi}(s,a)^\top \mathbf{e}_{s_1} \big)^2,
\end{aligned}$$

which completes the proof. □

Lemma 35. *(First Order Difference between the two transition kernels) For $k \in [K]$ and $h \in [H]$, for the estimator $\widehat{\boldsymbol{\theta}}_{k,h}$, we have Then for any sequence $v_{\tilde{s}} \in \mathbb{R}$ for $\tilde{s} \in \mathcal{S}$ and $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$,*

$$\sum_{s' \in \mathcal{S}} v_{s'} \big(\mathbb{P}(s'|s,a,\boldsymbol{\theta}_1) - \mathbb{P}(s'|s,a,\boldsymbol{\theta}_2) \big) \leq \frac{\max_{\tilde{s} \in \mathcal{S}} |v_{\tilde{s}}|}{2} \max_{\tilde{s} \in \mathcal{S}} (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top \boldsymbol{\Phi}(s,a)^\top \mathbf{e}_{\tilde{s}}.$$

Proof. By the definition of total variation,

$$\sum_{s' \in \mathcal{S}} v_{s'} \big(\mathbb{P}(s'|s,a,\boldsymbol{\theta}_1) - \mathbb{P}(s'|s,a,\boldsymbol{\theta}_2) \big) \leq \max_{\tilde{s} \in \mathcal{S}} |v_{\tilde{s}}| \|\mathbb{P}(\cdot|s,a,\boldsymbol{\theta}_1) - \mathbb{P}(\cdot|s,a,\boldsymbol{\theta}_2)\|_1.$$

Applying Pinsker's inequality,

$$\|\mathbb{P}(\cdot|s,a,\boldsymbol{\theta}_1) - \mathbb{P}(\cdot|s,a,\boldsymbol{\theta}_2)\|_1 \leq \sqrt{\frac{1}{2} KL\big(\mathbb{P}(\cdot|s,a,\boldsymbol{\theta}_1), \mathbb{P}(\cdot|s,a,\boldsymbol{\theta}_2)\big)}.$$

Because $\mathbb{P}(\cdot|s,a,\boldsymbol{\theta})$ is multinomial probability measure,

$$\begin{aligned}
&KL\big(\mathbb{P}(\cdot|s,a,\boldsymbol{\theta}_1), \mathbb{P}(\cdot|s,a,\boldsymbol{\theta}_2)\big) \\
&= \mathbb{E}\big[\mathbf{e}_{s'}^\top \boldsymbol{\Phi}(s,a)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \big| s,a,\boldsymbol{\theta}_1 \big] - g\big(\boldsymbol{\Phi}(s,a)\boldsymbol{\theta}_1\big) + g\big(\boldsymbol{\Phi}(s,a)\boldsymbol{\theta}_2\big) \\
&= \nabla L_{\nu,h}\big(\boldsymbol{\Phi}(s,a)\boldsymbol{\theta}_1\big)^\top \boldsymbol{\Phi}(s,a)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) - g\big(\boldsymbol{\Phi}(s,a)\boldsymbol{\theta}_1\big) + g\big(\boldsymbol{\Phi}(s,a)\boldsymbol{\theta}_2\big),
\end{aligned}$$

where the last equality uses the identity for the mean of multinomial (Proposition 19). By second order Taylor series, there exists $\bar{\boldsymbol{\theta}}$ such that

$$\begin{aligned}
&KL\big(\mathbb{P}(\cdot|s,a,\boldsymbol{\theta}_1),\mathbb{P}(\cdot|s,a,\boldsymbol{\theta}_2)\big)\\
&=\frac{1}{2}(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)^\top\boldsymbol{\Phi}(s,a)^\top\nabla^2 L_{\nu,h}(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}})\boldsymbol{\Phi}(s,a)(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)\\
&\le\frac{1}{2}(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)^\top\boldsymbol{\Phi}(s,a)^\top\mathrm{diag}\big(\nabla L_{\nu,h}(\boldsymbol{\Phi}(s_{k,h},a_{k,h})\bar{\boldsymbol{\theta}}\big)\boldsymbol{\Phi}(s,a)(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)\\
&=\frac{1}{2}\sum_{\tilde{s}\in\mathcal{S}}\mathbf{e}_{\tilde{s}}^\top\nabla L_{\nu,h}(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}})\big((\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)^\top\boldsymbol{\Phi}(s,a)^\top\mathbf{e}_{\tilde{s}}\big)^2
\end{aligned}$$

where the first inequality uses the fact that $\nabla^2 L_{\nu,h}(\boldsymbol{\eta})=\mathrm{diag}\big(\nabla L_{\nu,h}(\boldsymbol{\eta})\big)-\nabla L_{\nu,h}(\boldsymbol{\eta})\nabla L_{\nu,h}(\boldsymbol{\eta})^\top\preceq\mathrm{diag}\big(\nabla L_{\nu,h}(\boldsymbol{\eta})\big)$. Because $\|\nabla L_{\nu,h}(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}})\|_1=1$,by Hölder's inequality,

$$\begin{aligned}
&\frac{1}{2}\sum_{\tilde{s}\in\mathcal{S}}\mathbf{e}_{\tilde{s}}^\top\nabla L_{\nu,h}(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}})\big((\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)^\top\boldsymbol{\Phi}(s,a)^\top\mathbf{e}_{\tilde{s}}\big)^2\\
&\le\frac{1}{2}\max_{\tilde{s}\in\mathcal{S}}\big((\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)^\top\boldsymbol{\Phi}(s,a)^\top\mathbf{e}_{\tilde{s}}\big)^2\|\nabla L_{\nu,h}(\boldsymbol{\Phi}(s,a)\bar{\boldsymbol{\theta}})\|_1\\
&=\frac{1}{2}\max_{\tilde{s}\in\mathcal{S}}\big((\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)^\top\boldsymbol{\Phi}(s,a)^\top\mathbf{e}_{\tilde{s}}\big)^2.
\end{aligned}$$

which completes the proof. □

Proof. For the first term,

$$\begin{aligned}
&\big\|\boldsymbol{\Phi}_{k,h}^\top\nabla^2 L_{\nu,h}\big(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star\big)\Big(\sum_{\tilde{s}\in\mathcal{S}_{k,h}}\widehat{V}_{k,h+1}(\tilde{s})\mathbf{e}_{\tilde{s}}\Big)\big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}}\\
&\le H\max_{\mathbf{x}:\|\mathbf{x}\|_\infty\le 1}\big\|\boldsymbol{\Phi}_{k,h}\nabla^2 L_{\nu,h}\big(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star\big)\mathbf{x}\big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}}\\
&\le H\|\widehat{\mathbf{H}}_{k-1,h}^{-1/2}\boldsymbol{\Phi}_{k,h}\big(\nabla^2 L_{k,h}(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star)\big)^{1/2}\|_2\max_{\mathbf{x}:\|\mathbf{x}\|_\infty\le 1}\|\Big(\nabla^2 L_{k,h}(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star)\Big)^{1/2}\mathbf{x}\|_2\\
&= H\sigma_{k,h}\|\widehat{\mathbf{H}}_{k-1,h}^{-1/2}\boldsymbol{\Phi}_{k,h}\big(\nabla^2 L_{k,h}(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star)\big)^{1/2}\|_2\\
&\le H\sigma_{k,h}\|\widehat{\mathbf{H}}_{k-1,h}^{-1/2}\boldsymbol{\Phi}_{k,h}\big(\nabla^2 L_{k,h}(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star)\big)^{1/2}\|_F\\
&= H\sigma_{k,h}\sqrt{\mathrm{Tr}\Big(\nabla^2\ell_{k,h}(\boldsymbol{\theta}_h^\star)\widehat{\mathbf{H}}_{k-1,h}^{-1}\Big)}
\end{aligned}$$

where the last equality uses the fact that $\nabla^2\ell_{k,h}(\boldsymbol{\theta}_h^\star)=\boldsymbol{\Phi}_{k,h}^\top\nabla^2 L_{k,h}(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star)\boldsymbol{\Phi}_{k,h}$. By Lemma 3, we have $\widehat{\mathbf{H}}_{k,h}\succeq\mathbf{H}_{k,h}^\star:=\sum_{\nu=1}^k\nabla^2\ell_{\nu,h}(\boldsymbol{\theta}_h^\star)+4B_{\boldsymbol{\Phi}}^2\mathbf{I}_d$ and

$$\big\|\sum_{\tilde{s}\in\mathcal{S}_{k,h}}\widehat{V}_{k,h+1}(\tilde{s})\boldsymbol{\Phi}_{k,h}^\top\nabla^2 L_{\nu,h}\big(\boldsymbol{\Phi}_{k,h}\boldsymbol{\theta}_h^\star\big)\mathbf{e}_{\tilde{s}}\big\|_{\widehat{\mathbf{H}}_{k-1,h}^{-1}}\le H\sqrt{\sigma_{k,h}^2\mathrm{Tr}\Big(\nabla^2\ell(\boldsymbol{\theta}_h^\star)\big(\mathbf{H}_{k-1,h}^\star\big)^{-1}\Big)}.$$

□

C.11 Matrix Algebras

Proposition 36. *(Inverse Matrix Differentiation) Let $\mathbf{F}:\mathbb{R}^d\to\mathbb{R}^{d\times d}$ denote an invertible matrix map. Then for any $\mathbf{x}\in\mathbb{R}^d$ and $i\in[d]$,*

$$D_i\big\{\mathbf{F}(\mathbf{x})^{-1}\big\}=-\mathbf{F}(\mathbf{x})^{-1}D_i\mathbf{F}(\mathbf{x})\mathbf{F}(\mathbf{x})^{-1}.$$

Proof. Note that $D_i\{\mathbf{F}(\mathbf{x})^{-1}\mathbf{F}(\mathbf{x})\}=\mathbf{O}$. By the chain rule,

$$D_i\big\{\mathbf{F}(\mathbf{x})^{-1}\big\}\mathbf{F}(\mathbf{x})+\mathbf{F}(\mathbf{x})^{-1}D_i\mathbf{F}(\mathbf{x})=\mathbf{O},$$

and rearranging the term completes the proof. □

Lemma 37 (A Property of a contraction matrix)**.** *Let $\mathbf{M} \in \mathbb{R}^{d\times d}$ denote a positive definite symmetric matrix such that $\mathbf{M}^2 \preceq \mathbf{M}$. Then the eigenvalues $\lambda_1, \ldots, \lambda_d$ of $\mathbf{M}$ are in $[0, 1]$.*

Proof. Since $\mathbf{M}^{-1/2}$ exists, we have $\mathbf{M}^{-1/2}\mathbf{M}^2\mathbf{M}^{-1/2} \preceq \mathbf{M}^{-1/2}\mathbf{M}\mathbf{M}^{-1/2}$ and thus $\mathbf{M} \preceq \mathbf{I}_d$ which completes the proof. □

Lemma 38 (A matrix norm bound)**.** *Let $\mathbf{A} \in \mathbb{R}^{d\times d}$ denote a positive definite symmetric matrix and $\mathbf{B} \in \mathbb{R}^{d\times s}$ denote an arbitral matrix. Then,*

$$\left\|\mathbf{A}^{-1/2}\mathbf{B}\right\|_{1,2} := \sup_{\mathbf{x}:\|\mathbf{x}\|_1\le 1} \left\|\mathbf{A}^{-1/2}\mathbf{B}\mathbf{x}\right\|_2 \le \max_{s'\in[s]} \left\|\mathbf{A}^{-1/2}\mathbf{B}\mathbf{e}_{s'}\right\|_2$$

Proof. Writing $\mathbf{x} = (x_1, \ldots, x_s)$,

$$\begin{aligned}
\sup_{\mathbf{x}:\|\mathbf{x}\|_1\le 1} \left\|\mathbf{A}^{-1/2}\mathbf{B}\mathbf{x}\right\|_2^2 &= \sup_{\mathbf{x}:\|\mathbf{x}\|_1\le 1} \mathbf{x}^\top\mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B}\mathbf{x} \\
&= \sup_{\mathbf{x}:\|\mathbf{x}\|_1\le 1} \sum_{s_1,s_2\in[s]} x_{s_1}x_{s_2}\mathbf{e}_{s_1}^\top\mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B}\mathbf{e}_{s_2}^\top \\
&\le \max_{s_1,s_2\in[s]} \left|\mathbf{e}_{s_1}^\top\mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B}\mathbf{e}_{s_2}^\top\right| \sup_{\mathbf{x}:\|\mathbf{x}\|_1\le 1} \sum_{s_1,s_2\in[s]} |x_{s_1}x_{s_2}| \\
&\le \max_{s_1,s_2\in[s]} \left|\mathbf{e}_{s_1}^\top\mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B}\mathbf{e}_{s_2}^\top\right|.
\end{aligned}$$

By Cauchy-Schwarz inequality,

$$\begin{aligned}
\max_{s_1,s_2\in[s]} \left|\mathbf{e}_{s_1}^\top\mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B}\mathbf{e}_{s_2}^\top\right| &\le \max_{s_1,s_2\in[s]} \left\|\mathbf{A}^{-1/2}\mathbf{B}\mathbf{e}_{s_1}\right\|_2 \left\|\mathbf{A}^{-1/2}\mathbf{B}\mathbf{e}_{s_2}\right\|_2 \\
&= \max_{s'\in[s]} \left\|\mathbf{A}^{-1/2}\mathbf{B}\mathbf{e}_{s'}\right\|_2^2,
\end{aligned}$$

which completes the proof. □

C.12 Divergence Between the Two Interaction Measures

Lemma 39. *(Kullback-Liebler Divergence Between the Two Interaction Measures) Given an MNL-MDP instance $\mathcal{M}(\mathcal{S}, \mathcal{A}, H, \boldsymbol{\Theta}, \boldsymbol{\Phi}, r)$ and an algorithm Π_K, let*

$$\begin{aligned}
\mathbf{y}_{k,h} &:= (s_{1,1}, a_{1,1}, s_{1,2}, a_{1,2}, \ldots, s_{1,H+1}, s_{2,1}, a_{2,1}, \ldots, s_{k,h}, a_{k,h}), \\
\mathbf{z}_{k,h} &:= (s_{1,1}, a_{1,1}, s_{1,2}, a_{1,2}, \ldots, s_{1,H+1}, s_{2,1}, a_{2,1}, \ldots, s_{k,h}),
\end{aligned}$$

denote the sequence of action-state random variables generated by the interaction measure $\mathbb{P}_{\boldsymbol{\Theta},\Pi_K}$. Then for any parameters $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \in \{(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_h) \in \mathbb{R}^{d\times H} : \max_{h\in[H]} \|\boldsymbol{\theta}_h\|_2 \le B_{\boldsymbol{\theta}}\}$, the Kullback-Leibler divergence,

$$\begin{aligned}
&KL(\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}, \mathbb{P}_{\boldsymbol{\Theta}_2,\Pi_K}) \\
&\le \frac{\mathbb{E}_{\boldsymbol{\Theta}_1,\Pi_K}\left[\sum_{k=1}^K \sum_{h=1}^H \|\boldsymbol{\Phi}(s_{k,h}, a_{k,h})\big(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\big)\mathbf{e}_h\|_\infty^2 \sigma_{k,h}^2\right]}{2} \\
&\quad + \frac{\mathbb{E}_{\boldsymbol{\Theta}_1,\Pi_K}\left[\sum_{k=1}^K \sum_{h=1}^H \|\boldsymbol{\Phi}(s_{k,h}, a_{k,h})\big(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\big)\mathbf{e}_h\|_\infty^3\right]}{\sqrt{6\kappa}},
\end{aligned}$$

where $\sigma_{k,h}^2 := \max_{\mathbf{x}:\|\mathbf{x}\|_\infty\le 1} \mathbf{x}^\top\nabla^2 L_{k,h}\big(\boldsymbol{\Phi}(s_{k,h}, a_{k,h})\boldsymbol{\Theta}_1\mathbf{e}_H\big)\mathbf{x}$ and $L_{k,h}(\mathbf{u}) := \log\big(\sum_{s\in\mathcal{S}_h(s_{k,h},a_{k,h})} \exp(\mathbf{e}_s^\top\mathbf{u})\big)$ is the log-sum function.

Proof. By definition of $KL(\cdot,\cdot)$,

$$\begin{aligned}
&KL(\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}, \mathbb{P}_{\boldsymbol{\Theta}_2,\Pi_K}) \\
&= \int \log \frac{d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{z}_{K,H+1})}{d\mathbb{P}_{\boldsymbol{\Theta}_2,\Pi_K}(\mathbf{z}_{K,H+1})} d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{z}_{K,H+1}) \\
&= \int \Big(\int \log \frac{d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(s_{K,H+1}|\mathbf{y}_{K,H})}{d\mathbb{P}_{\boldsymbol{\Theta}_2,\Pi_K}(s_{K,H+1}|\mathbf{y}_{K,H})} + \log \frac{d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{y}_{K,H})}{d\mathbb{P}_{\boldsymbol{\Theta}_2,\Pi_K}(\mathbf{y}_{K,H})} d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(s_{K,H+1}|\mathbf{y}_{K,H})\Big) d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{y}_{K,H}).
\end{aligned}$$

For each $k \in [K]$ and $h \in [H]$, we write the log-sum function $L_{k,h} := L_{\mathcal{S}_h(s_{k,h},a_{k,h})}$ and the feature $\boldsymbol{\Phi}_{k,h} := \boldsymbol{\Phi}(s_{k,h}, a_{k,h})$. Because $\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(s_{k,h+1}|\mathbf{y}_{k,h}) = \mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(s_{k,h+1}|s_{k,h},a_{k,h})$ is a multinomial distribution,

$$\begin{aligned}
&\int \log \frac{d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(s_{K,H+1}|\mathbf{y}_{K,H})}{d\mathbb{P}_{\boldsymbol{\Theta}_2,\Pi_K}(s_{K,H+1}|\mathbf{y}_{K,H})} d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(s_{K,H+1}|\mathbf{y}_{K,H}) \\
&= \mathbb{E}_{\boldsymbol{\Theta}_1,\Pi_K}[\mathbf{e}_{s_{K,H+1}}]^\top \boldsymbol{\Phi}_{K,H}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\mathbf{e}_H + L_{K,H}\big(\boldsymbol{\Phi}_{K,H}\boldsymbol{\Theta}_2\mathbf{e}_H\big) - L_{K,H}\big(\boldsymbol{\Phi}_{K,H}\boldsymbol{\Theta}_1\mathbf{e}_H\big) \\
&= \nabla L_{\nu,h}\big(\boldsymbol{\Phi}_{K,H}\boldsymbol{\Theta}_1\mathbf{e}_H\big)^\top \boldsymbol{\Phi}_{K,H}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\mathbf{e}_H + L_{K,H}\big(\boldsymbol{\Phi}_{K,H}\boldsymbol{\Theta}_2\mathbf{e}_H\big) - L_{K,H}\big(\boldsymbol{\Phi}_{K,H}\boldsymbol{\Theta}_1\mathbf{e}_H\big)
\end{aligned}$$

where the last equality uses Proposition 19. By the third order Taylor expansion, there exists $\bar{\boldsymbol{\theta}}_H \in \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta}\|_2 \le B_{\boldsymbol{\theta}}\}$ such that

$$\begin{aligned}
&\nabla L_{\nu,h}\big(\boldsymbol{\Phi}_{K,H}\boldsymbol{\Theta}_1\mathbf{e}_H\big)^\top \boldsymbol{\Phi}_{K,H}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\mathbf{e}_H + L_{K,H}\big(\boldsymbol{\Phi}_{K,H}\boldsymbol{\Theta}_2\mathbf{e}_H\big) - L_{K,H}\big(\boldsymbol{\Phi}_{K,H}\boldsymbol{\Theta}_1\mathbf{e}_H\big) \\
&= \frac{1}{2}\Big\|\boldsymbol{\Phi}_{K,H}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\mathbf{e}_H\Big\|^2_{\nabla^2 L_{K,H}\left(\boldsymbol{\Phi}_{K,H}\boldsymbol{\Theta}_1\mathbf{e}_H\right)} \\
&+ \frac{1}{6}\sum_{s\in\mathcal{S}_{K,H}} \mathbf{e}_s^\top \boldsymbol{\Phi}_{K,H}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\mathbf{e}_H \Big\|\boldsymbol{\Phi}_{K,H}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\mathbf{e}_H\Big\|^2_{D_s\nabla^2 L_{K,H}(\boldsymbol{\Phi}_{K,H}\bar{\boldsymbol{\theta}}_H)}.
\end{aligned}$$

In the first term,

$$\begin{aligned}
&\Big\|\boldsymbol{\Phi}_{K,H}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\mathbf{e}_H\Big\|^2_{\nabla^2 L_{K,H}\left(\boldsymbol{\Phi}_{K,H}\boldsymbol{\Theta}_1\mathbf{e}_H\right)} \\
&\le \Big(\boldsymbol{\Phi}_{K,H}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\mathbf{e}_H\Big)^\top \nabla^2 L_{K,H}\big(\boldsymbol{\Phi}_{K,H}\boldsymbol{\Theta}_1\mathbf{e}_H\big)\Big(\boldsymbol{\Phi}_{K,H}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\mathbf{e}_H\Big) \\
&\le \big\|\boldsymbol{\Phi}_{K,H}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\mathbf{e}_H\big\|^2_\infty \underbrace{\max_{\mathbf{x}:\|\mathbf{x}\|_\infty\le 1} \mathbf{x}^\top \nabla^2 L_{K,H}\big(\boldsymbol{\Phi}_{K,H}\boldsymbol{\Theta}_1\mathbf{e}_H\big)\mathbf{x}}_{\sigma^2_{K,H}}
\end{aligned}$$

Because $L_{K,H}$ is $\sqrt{6}$-self concordant, the third order term,

$$\begin{aligned}
&\sum_{s\in\mathcal{S}_{K,H}} \mathbf{e}_s^\top \boldsymbol{\Phi}_{K,H}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\mathbf{e}_H \Big\|\boldsymbol{\Phi}_{K,H}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\mathbf{e}_H\Big\|^2_{D_s\nabla^2 L_{K,H}(\boldsymbol{\Phi}_{K,H}\bar{\boldsymbol{\theta}}_H)} \\
&\le \sqrt{6}\|\boldsymbol{\Phi}_{K,H}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\mathbf{e}_H\|_2 \|\boldsymbol{\Phi}_{K,H}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\mathbf{e}_H\|^2_{\nabla^2 L_{K,H}(\boldsymbol{\Phi}_{K,H}\bar{\boldsymbol{\theta}}_H)} \\
&\le \sqrt{6\kappa^{-1}}\|\boldsymbol{\Phi}_{K,H}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\mathbf{e}_H\|^3_{\nabla^2 L_{K,H}(\boldsymbol{\Phi}_{K,H}\bar{\boldsymbol{\theta}}_H)},
\end{aligned}$$

where the last inequality uses Definition 2. Because $\max_{\mathbf{x}:\|\mathbf{x}\|_\infty\le 1} \mathbf{x}^\top \nabla^2 L_{\nu,h}\big(\boldsymbol{\Phi}(s_{K,H}, a_{K,H})\bar{\boldsymbol{\theta}}_H\big)\mathbf{x} \le 1$,

$$\begin{aligned}
&\sqrt{6\kappa^{-1}}\|\boldsymbol{\Phi}_{K,H}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\mathbf{e}_H\|^3_{\nabla^2 L_{K,H}(\boldsymbol{\Phi}_{K,H}\bar{\boldsymbol{\theta}}_H)} \\
&\le \sqrt{6\kappa^{-1}}\|\boldsymbol{\Phi}_{K,H}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\mathbf{e}_H\|^3_\infty \bigg(\max_{\mathbf{x}:\|\mathbf{x}\|_\infty\le 1} \mathbf{x}^\top \nabla^2 L_{\nu,h}\big(\boldsymbol{\Phi}(s_{K,H}, a_{K,H})\bar{\boldsymbol{\theta}}_H\big)\mathbf{x}\bigg)^{3/2} \\
&\le \sqrt{6\kappa^{-1}}\|\boldsymbol{\Phi}_{K,H}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\mathbf{e}_H\|^3_\infty.
\end{aligned}$$

Recall that $\Delta_{k,h} := \|\boldsymbol{\Phi}_{k,h}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\mathbf{e}_h\|_\infty$. Gathering the terms,

$$\int \log \frac{d\mathbb{P}_{\tilde{\mathbf{U}}}(s_{K,H+1}|\mathbf{y}_{K,H})}{d\mathbb{P}_{\tilde{\mathbf{U}}(h,i)}(s_{K,H+1}|\mathbf{y}_{K,H})} d\mathbb{P}_{\tilde{\mathbf{U}}}(s_{K,H+1}|\mathbf{y}_{K,H}) \le \frac{\Delta^2_{K,H}\sigma^2_{K,H}}{2} + \frac{\Delta^3_{K,H}}{\sqrt{6\kappa}}.$$

Thus,

$$\begin{aligned}
&KL(\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}, \mathbb{P}_{\boldsymbol{\Theta}_2,\Pi_K}) \\
&\le \int\Big(\frac{\Delta^2_{K,H}\sigma^2_{K,H}}{2} + \frac{\Delta^3_{K,H}}{\sqrt{6\kappa}} + \int \log \frac{d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{y}_{K,H})}{d\mathbb{P}_{\boldsymbol{\Theta}_2,\Pi_K}(\mathbf{y}_{K,H})} d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(s_{K,H+1}|\mathbf{y}_{K,H})\Big) d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{y}_{K,H}) \\
&\le \frac{\mathbb{E}_{\boldsymbol{\Theta}_1,\Pi_K}\big[\Delta^2_{K,H}\sigma^2_{K,H}\big]}{2} + \frac{\mathbb{E}_{\boldsymbol{\Theta}_1,\Pi_K}\big[\Delta^3_{K,H}\big]}{\sqrt{6\kappa}} + \int \log \frac{d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{y}_{K,H})}{d\mathbb{P}_{\boldsymbol{\Theta}_2,\Pi_K}(\mathbf{y}_{K,H})} d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{y}_{K,H}).
\end{aligned}$$

Note that the last term,

$$\begin{aligned}
&\int \log \frac{d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{y}_{K,H})}{d\mathbb{P}_{\boldsymbol{\Theta}_2,\Pi_K}(\mathbf{y}_{K,H})} d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{y}_{K,H}) \\
&= \int\int \log \frac{d\pi_{K,H}(a_{K,H}|\mathbf{z}_{K,H})}{d\pi_{K,H}(a_{K,H}|\mathbf{z}_{K,H})} + \log \frac{d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{z}_{K,H})}{d\mathbb{P}_{\boldsymbol{\Theta}_2,\Pi_K}(\mathbf{z}_{K,H})} d\pi_{K,H}(a_{K,H}|\mathbf{z}_{K,H}) d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{z}_{K,H}) \\
&= \int \log \frac{d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{z}_{K,H})}{d\mathbb{P}_{\boldsymbol{\Theta}_2,\Pi_K}(\mathbf{z}_{K,H})} \Big(\int d\pi_{K,H}(a_{K,H}|\mathbf{z}_{K,H}) \Big) d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{z}_{K,H}) \\
&= \int \log \frac{d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{z}_{K,H})}{d\mathbb{P}_{\boldsymbol{\Theta}_2,\Pi_K}(\mathbf{z}_{K,H})} d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{z}_{K,H})
\end{aligned}$$

Recursively,

$$\begin{aligned}
&KL(\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}, \mathbb{P}_{\boldsymbol{\Theta}_2,\Pi_K}) \\
&= \int \log \frac{d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{z}_{K,H+1})}{d\mathbb{P}_{\boldsymbol{\Theta}_2,\Pi_K}(\mathbf{z}_{K,H+1})} d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{z}_{K,H+1}) \\
&\leq \frac{\mathbb{E}_{\boldsymbol{\Theta}_1,\Pi_K}\left[\Delta_{K,H}^2 \sigma_{K,H}^2\right]}{2} + \frac{\mathbb{E}_{\boldsymbol{\Theta}_1,\Pi_K}\left[\Delta_{K,H}^3\right]}{\sqrt{6\kappa}} + \int \log \frac{d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{z}_{K,H})}{d\mathbb{P}_{\boldsymbol{\Theta}_2,\Pi_K}(\mathbf{z}_{K,H})} d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{z}_{K,H}) \\
&\leq \frac{\mathbb{E}_{\boldsymbol{\Theta}_1,\Pi_K}\left[\Delta_{K,H-1}^2 \sigma_{K,H-1}^2 + \Delta_{K,H} \sigma_{K,H}^2\right]}{2} + \frac{\mathbb{E}_{\boldsymbol{\Theta}_1,\Pi_K}\left[\Delta_{K,H-1}^3 + \Delta_{K,H}^3\right]}{\sqrt{6\kappa}} \\
&\quad + \int \log \frac{d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{z}_{K,H-1})}{d\mathbb{P}_{\boldsymbol{\Theta}_2,\Pi_K}(\mathbf{z}_{K,H-1})} d\mathbb{P}_{\boldsymbol{\Theta}_1,\Pi_K}(\mathbf{z}_{K,H-1}) \\
&\vdots \\
&\leq \frac{\mathbb{E}_{\boldsymbol{\Theta}_1,\Pi_K}\left[\sum_{k=1}^K \sum_{h=1}^H \Delta_{k,h}^2 \sigma_{k,h}^2\right]}{2} + \frac{\mathbb{E}_{\boldsymbol{\Theta}_1,\Pi_K}\left[\sum_{k=1}^K \sum_{h=1}^H \Delta_{k,h}^3\right]}{\sqrt{6\kappa}},
\end{aligned}$$

which completes the proof. □

C.13 Covariance Inequality

Proposition 40. *(Covariance inequality) Let $\mathbf{y} \in \mathbb{R}^d$ denote a vector valued random variable. Then for any deterministic vector $\mathbf{a} \in \mathbb{R}^d$*

$$\mathbb{E}\Big[\big(\mathbf{y} - \mathbb{E}[\mathbf{y}]\big)\big(\mathbf{y} - \mathbb{E}[\mathbf{y}]\big)^\top\Big] \preceq \mathbb{E}\Big[\big(\mathbf{y} - \mathbf{a}\big)\big(\mathbf{y} - \mathbf{a}\big)^\top\Big].$$

Proof. For any $\mathbf{a} \in \mathbb{R}^d$,

$$\begin{aligned}
\mathbb{E}\Big[\big(\mathbf{y} - \mathbf{a}\big)\big(\mathbf{y} - \mathbf{a}\big)^\top\Big] - \mathbb{E}\Big[\big(\mathbf{y} - \mathbb{E}[\mathbf{y}]\big)\big(\mathbf{y} - \mathbb{E}[\mathbf{y}]\big)^\top\Big] &= \mathbf{a}\mathbf{a}^\top - \mathbb{E}[\mathbf{y}]\mathbf{a}^\top - \mathbf{a}\mathbb{E}[\mathbf{y}]^\top + \mathbb{E}[\mathbf{y}]\mathbb{E}[\mathbf{y}]^\top \\
&= \big(\mathbf{a} - \mathbb{E}[\mathbf{y}]\big)\big(\mathbf{a} - \mathbb{E}[\mathbf{y}]\big)^\top \\
&\succeq \mathbf{O},
\end{aligned}$$

which completes the proof. □

C.14 Entropic Variational Principle

Proposition 41 (Entropic Variational Principle)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $X : \Omega \to \mathbb{R}$ be a measurable function such that $\mathbb{E}_P[e^X] < \infty$. Let $\mathcal{Q}$ be the set of all probability measures on $(\Omega, \mathcal{F})$ that are absolutely continuous with respect to P. Then,*

$$\log \mathbb{E}_P[e^X] = \sup_{Q \in \mathcal{Q}} \{\mathbb{E}_Q[X] - D_{KL}(Q\|P)\} \tag{32}$$

where $KL(Q\|P)$ denotes the Kullback-Leibler divergence. Furthermore, the supremum is uniquely attained by the Gibbs measure P^* defined by the Radon-Nikodym derivative:

$$\frac{dP^*}{dP} = \frac{e^X}{\mathbb{E}_P[e^X]} \tag{33}$$