Provably Efficient Reinforcement Learning with Multinomial Logit Function Approximation

Long-Fei Li[1, 2], **Yu-Jie Zhang**[3], **Peng Zhao**[1, 2], **Zhi-Hua Zhou**[1, 2]
[1]National Key Laboratory for Novel Software Technology, Nanjing University, China
[2]School of Artificial Intelligence, Nanjing University, China
[3]The University of Tokyo, Chiba, Japan
{lilf, zhaop, zhouzh}@lamda.nju.edu.cn, yujie.zhang@ms.k.u-tokyo.ac.jp

Abstract

We study a new class of MDPs that employs multinomial logit (MNL) function approximation to ensure valid probability distributions over the state space. Despite its benefits, introducing the non-linear function raises significant challenges in both *computational* and *statistical* efficiency. The best-known result of Hwang and Oh [2023] has achieved an $\widetilde{\mathcal{O}}(\kappa^{-1}dH^2\sqrt{K})$ regret, where κ is a problem-dependent quantity, d is the feature dimension, H is the episode length, and K is the number of episodes. While this result attains the same rate in K as linear cases, the method requires storing all historical data and suffers from an $\mathcal{O}(K)$ computation cost per episode. Moreover, the quantity κ can be exponentially small in the worst case, leading to a significant gap for the regret compared to linear function approximation. In this work, we first address the computational and storage issue by proposing an algorithm that achieves the same regret with only $\mathcal{O}(1)$ cost. Then, we design an enhanced algorithm that leverages local information to enhance statistical efficiency. It not only maintains an $\mathcal{O}(1)$ computation and storage cost per episode but also achieves an improved regret of $\widetilde{\mathcal{O}}(dH^2\sqrt{K} + d^2H^2\kappa^{-1})$, nearly closing the gap with linear function approximation. Finally, we establish the first lower bound for MNL function approximation, justifying the optimality of our results in d and K.

1 Introduction

Reinforcement Learning (RL) with function approximation has achieved remarkable success in various applications involving large state and action spaces, such as games [Silver et al., 2016], algorithm discovery [Fawzi et al., 2022] and large language models [Ouyang et al., 2022]. Therefore, establishing the theoretical foundation for RL with function approximation is of great importance. Recently, there have been many efforts devoted to understanding the linear function approximation, yielding numerous valuable results [Yang and Wang, 2019, Jin et al., 2020, Ayoub et al., 2020].

While these studies make important steps toward understanding RL with function approximation, there are still some challenges to be solved. In linear function approximation, transitions are assumed to be linear in specified feature mappings, such as $\mathbb{P}(s'|s,a) = \phi(s'|s,a)^\top\theta^*$ for linear mixture MDPs and $\mathbb{P}(s'|s,a) = \phi(s,a)^\top\mu^*(s')$ for linear MDPs. Here $\mathbb{P}(s'|s,a)$ is the probability from state s to state s' taking action a, $\phi(s'|s,a)$ and $\phi(s,a)$ are feature mappings, θ^* and $\mu^*(s')$ are unknown parameters. However, a transition function is a *probability distribution* over states, i.e., the magnitude falls within the range of $[0,1]$, and the aggregation equals 1. For certain feature mappings, linear transitions may not yield a valid probability distribution with arbitrary θ and μ as show by [Hwang and Oh, 2023]. While there are also some works exploring generalized linear [Wang et al., 2021] and general function approximation [Russo and Roy, 2013, Foster et al., 2021, Chen et al., 2023], they made assumptions over value functions rather than transitions, hence do not tackle this challenge.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Table 1: Comparison between previous works and ours in terms of the regret, computation cost and storage cost. Here κ and κ^* are exponentially small problem-dependent quantities defined in Assumption 1, d is the feature dimension, H is the episode length and K is the number of episodes. The computational cost and storage cost per episode indicate the dependence on episode count k.

Reference	Regret	Com.	Sto.	MDP model
Hwang and Oh [2023]	$\widetilde{\mathcal{O}}(\kappa^{-1}dH^2\sqrt{K})$	$\mathcal{O}(k)$	$\mathcal{O}(k)$	homogeneous
UCRL-MNL-OL (Theorem 1)	$\widetilde{\mathcal{O}}(\kappa^{-1}dH^2\sqrt{K})$	$\mathcal{O}(1)$	$\mathcal{O}(1)$	inhomogeneous
UCRL-MNL-LL (Theorem 2)	$\widetilde{\mathcal{O}}(dH^2\sqrt{K} + d^2H^2\kappa^{-1})$	$\mathcal{O}(1)$	$\mathcal{O}(1)$	inhomogeneous
Lower Bound (Corollary 1)	$\Omega(dH\sqrt{K\kappa^*})$	/	/	infinite action space

Towards addressing the limitation of linear function approximation, a new class of MDPs that utilizes multinomial logit function approximation has been proposed by Hwang and Oh [2023] recently. Despite its benefits, the introduction of non-linear functions raises significant challenges in both *computational* and *statistical* efficiency. Specifically, the best-known approach of Hwang and Oh [2023] has achieved an $\widetilde{\mathcal{O}}(\kappa^{-1}dH^2\sqrt{K})$ regret. While matching the optimal regret of $\widetilde{\mathcal{O}}(d\sqrt{H^3K})$ as linear function approximation in terms of K [Zhou et al., 2021, He et al., 2023], their method requires storing all historical data and the computational cost per episode grows *linearly* with the episode count, which is expensive. Moreover, the quantity κ can be exponentially small, leading to a significant gap for the regret compared to the linear cases. Thus, a natural question arises:

Is MNL function approximation more difficult than linear function approximation for RL?

In this paper, we answer this question by making significant advancements in both *computational* and *statistical* efficiency for MDPs with MNL function approximation, nearly closing the gap with linear function approximation. Table 1 presents a comparison between previous studies and our work. Specifically, our contributions are summarized as follows:

- We first propose the UCRL-MNL-OL algorithm based on a variant of Online Newton Step, attaining an $\widetilde{\mathcal{O}}(\kappa^{-1}dH^2\sqrt{K})$ regret with $\mathcal{O}(1)$ computational and storage costs per episode. This result matches the best-known regret by Hwang and Oh [2023], yet achieves the same *computational* and *storage* efficiency as linear function approximation [Jin et al., 2020, Zhou et al., 2021].
- As the quantity κ can be exponentially small, we propose an enhanced algorithm UCRL-MNL-LL, which leverages the local information to improve the *statistical* efficiency. It not only maintain $\mathcal{O}(1)$ computational and storage costs but achieve an improved regret of $\widetilde{\mathcal{O}}(dH^2\sqrt{K} + d^2H^2\kappa^{-1})$. The higher-order term matches the optimal regret $\widetilde{\mathcal{O}}(d\sqrt{H^3K})$ for linear mixture MDPs [Zhou et al., 2021] and linear MDPs [He et al., 2023] in d and K, differing only by an $\mathcal{O}(H^{1/2})$ factor.
- We establish the *first* lower bound for MDPs with MNL function approximation by introducing a reduction to the logistic bandit problem. We prove a lower bound of $\Omega(dH\sqrt{K\kappa^*})$ for MDPs with infinite action space. Though this does not constitute a strict lower bound for the finite actions case studied in this work, it suggests that our result may be optimal with respect to d and K. [1]

From a technical perspective, inspired by recent advances in logistic bandits [Zhang et al., 2016, Faury et al., 2020, Zhang and Sugiyama, 2023], we observe that the negative log-likelihood function is exponentially concave, which motivates us to apply online Newton step to estimate the unknown transition parameter in an online manner. Moreover, we employ the Bernstein-like inequalities and the self-concordant-like property [Bach, 2010] of the log-loss to achieve the better dependence on κ.

Organization. We introduce the related work in Section 2 and present the setup in Section 3. Then, we design an algorithm with constant computational and storage costs in Section 4. Next, we present an algorithm with improved regret guarantee in Section 5. Finally, we establish the lower bound in Section 6. Section 7 concludes the paper. Due to space limits, we defer all proofs to the appendixes.

[1] After the submission of our work to arXiv [Li et al., 2024a], a follow up work by Park et al. [2024] improved the lower bound to $\Omega(dH^{3/2}\sqrt{K})$. This confirms that our result is indeed optimal with respect to d and K.

Notations. Denote by $[n]$ the set $\{1, \ldots, n\}$ and use $[x]_{[a,b]}$ to denote $\min(\max(x, a), b)$. For $X, Y \in \mathbb{R}^{d\times d}$, $X \succeq Y$ means $X - Y$ is positive semi-definite. For a vector $\mathbf{x} \in \mathbb{R}^d$ and positive semi-definite matrix $A \in \mathbb{R}^{d\times d}$, denote $\|\mathbf{x}\|_A = \sqrt{\mathbf{x}^\top A\mathbf{x}}$. The $\widetilde{\mathcal{O}}(\cdot)$ hides polylogarithmic factors.

2 Related Work

In this section, we review related works from both setup and technical perspectives.

RL with Generalized Linear Function Approximation. There are recent efforts devoted to investigating function approximation beyond the linear models. Wang et al. [2021] investigated RL with generalized linear function approximation. Notably, unlike our approach which models transitions using a generalized linear model, they apply this approximation directly to the value function. Another line of works [Chowdhury et al., 2021, Li et al., 2022, Ouhamma et al., 2023] has studied RL with exponential function approximation and also aimed to ensure that transitions constitute valid probability distributions. The MDP model can be viewed as an extension of bilinear MDPs in their work while our setting extends linear mixture MDPs. These studies are complementary to ours and not directly comparable. Moreover, these works also enter the computational and statistical challenges arising from non-linear function approximation that remain to be addressed. The most relevant work to ours is the recent work by Hwang and Oh [2023], which firstly explored a similar setting to ours, where the transition is characterized using a multinomial logit model. We significantly improve upon their results by providing computationally and statistically more efficient algorithms.

RL with General Function Approximation. There have also been some works studies RL with general function approximation. Russo and Roy [2013] and Osband and Roy [2014] initiated the study on the minimal structural assumptions that render sample-efficient learning by proposing a structural condition called Eluder dimension. Recently, several works have investigated different conditions for sample-efficient interactive learning, such as Bellman Eluder (BE) dimension [Jin et al., 2021], Bilinear classes [Du et al., 2021], Decision-Estimation Coefficient (DEC) [Foster et al., 2021], and Admissible Bellman Characterization (ABC) [Chen et al., 2023]. A notable difference is that they impose assumptions on the value functions while we study function approximation on the transitions to ensure valid probability distributions. Moreover, the goal of these works is to study the conditions for sample-efficient reinforcement learning, but not focus on the computational efficiency.

Multinomial Logit Bandits. There are two types of multinomial logit bandits studied in the literature: the single-parameter model, where the parameter is a vector [Cheung and Simchi-Levi, 2017] and multiple-parameter model, where the parameter is a matrix [Amani and Thrampoulidis, 2021]. We focus on the single-parameter model, which are more relevant to our setting. The pioneering work by Cheung and Simchi-Levi [2017] achieved a Bayesian regret of $\widetilde{\mathcal{O}}(\kappa^{-1}d\sqrt{K})$, where K denotes the number of rounds in bandits. This result was further enhanced by subsequent studies [Oh and Iyengar, 2019, 2021, Agrawal et al., 2023]. In particular, Périvier and Goyal [2022] significantly improved the dependence on κ, obtaining a regret of $\widetilde{\mathcal{O}}(d\sqrt{\kappa K} + \kappa^{-1})$ in the uniform revenue setting. Most prior methods required storing all historical data and faced computational challenge. To address this issue, the most recent work by Lee and Oh [2024] proposed an algorithm with constant computational and storage costs building on recent advances in multiple-parameter model [Zhang and Sugiyama, 2023]. Their algorithm achieves the optimal regret of $\widetilde{\mathcal{O}}(d\sqrt{\kappa K} + \kappa^{-1})$ and $\widetilde{\mathcal{O}}(d\sqrt{K} + \kappa^{-1})$ under uniform and non-uniform rewards respectively. However, although the underlying models of MNL bandits and MDPs share similarities, the challenges they present differ substantially, and techniques developed for MNL bandits cannot be directly applied to MNL MDPs. For example, in MNL bandits, the objective is to select a series of assortments with *varying* sizes that maximize the expected revenue, whereas in MNL MDPs, the goal is to choose *one* action at each stage to maximize the cumulative reward. Thus, it is necessary to design new algorithms tailored for MDPs to address these unique challenges.

3 Problem Setup

In this section, we present the problem setup of RL with multinomial logit function approximation.

Inhomogeneous, Episodic MDPs. An inhomogeneous, episodic MDP instance can be denoted by a tuple $M = (\mathcal{S}, \mathcal{A}, H, \{\mathbb{P}_h\}_{h=1}^H, \{r_h\}_{h=1}^H)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, H is the length of each episode, $\mathbb{P}_h : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition kernel with $\mathbb{P}_h(s' \mid s, a)$ is being the probability of transferring to state s' from state s and taking action a at stage h, $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the deterministic reward function. A policy $\pi = \{\pi_h\}_{h=1}^H$ is a collection of mapping π_h, where each $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ is a function maps a state s to distributions over $\mathcal{A}$ at stage h. For any policy π and $(s, a) \in \mathcal{S} \times \mathcal{A}$, we define the action-value function Q_h^π and value function V_h^π as follows:

$$Q_h^\pi(s,a) = \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}\left(s_{h'}, a_{h'}\right) \Big| s_h = s, a_h = a\right], \quad V_h^\pi(s) = \mathbb{E}_{a\sim\pi_h(\cdot|s)}\left[Q_h^\pi(s,a)\right],$$

where the expectation of Q_h^π is taken over the randomness of the transition $\mathbb{P}$ and policy π. The optimal value function V_h^* and action-value function Q_h^* given by $V_h^*(s) = \sup_\pi V_h^\pi(s)$ and $Q_h^*(s,a) = \sup_\pi Q_h^\pi(s,a)$. For any function $V : \mathcal{S} \to \mathbb{R}$, we define $[\mathbb{P}_h V](s,a) = \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)} V(s')$. The Bellman equation for policy π and Bellman optimality equation are given by

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

Learning Protocol. In the online MDP setting, the learner interacts with the environment without the knowledge of the transition $\{\mathbb{P}_h\}_{h=1}^H$. As learning rewards is no harder than transitions, we assume the reward $\{r_h\}_{h=1}^H$ is known. The interaction proceeds in K episodes. At the beginning of episode k, the learner chooses a policy $\pi_k = \{\pi_{k,h}\}_{h=1}^H$. At each stage $h \in [H]$, starting from the initial state $s_{k,1}$, the learner observes the state $s_{k,h}$, chooses an action $a_{k,h}$ sampled from $\pi_{k,h}(\cdot \mid s_{k,h})$, obtains reward $r_h(s_{k,h}, a_{k,h})$ and transits to the next state $s_{k,h+1} \sim \mathbb{P}_h(\cdot \mid s_{k,h}, a_{k,h})$ for $h \in [H]$. The episode ends when s_{H+1} is reached. The goal of the learner is to minimize regret, defined as

$$\text{Reg}(K) = \sum_{k=1}^K V_1^*(s_{k,1}) - \sum_{k=1}^K V_1^{\pi_k}(s_{k,1}),$$

which is the difference between the cumulative reward of the optimal policy and the learner's policy.

Multinomial Logit (MNL) Mixture MDPs. Although significant advances have been achieved for MDPs with linear function approximation, Hwang and Oh [2023] show that there exists a set of features such that no linear transition model (including bilinear and low-rank MDPs) can induce a valid probability distribution over the state space, which limits the expressiveness of function approximation. To overcome this limitation, they propose a new class of MDPs with multinomial logit function approximation. We introduce the definition of MNL mixture MDPs below.

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

Remark 1. There are many cases that even when the state space is very large, the maximum number of reachable states U is small. This phenomenon is common in situations where the next state is close to the current state. An illustrative example is the RiverSwim problem [Strehl and Littman, 2004].

Definition 2 (MNL Mixture MDP)**.** An MDP instance $M = (\mathcal{S}, \mathcal{A}, H, \{\mathbb{P}_h\}_{h=1}^H, \{r_h\}_{h=1}^H)$ is called an inhomogeneous, episodic B-bounded MNL mixture MDP if there exist a *known* feature mapping $\phi(s' \mid s,a) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ with $\|\phi(s' \mid s,a)\|_2 \le 1$ and *unknown* vectors $\{\theta_h^*\}_{h=1}^H \in \Theta$ with $\Theta = \{\theta \in \mathbb{R}^d, \|\theta\|_2 \le B\}$, such that for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and $s' \in \mathcal{S}_{h,s,a}$, it holds that

$$\mathbb{P}_h(s' \mid s,a) = \frac{\exp(\phi(s' \mid s,a)^\top \theta_h^*)}{\sum_{\widetilde{s}\in\mathcal{S}_{h,s,a}} \exp(\phi(\widetilde{s} \mid s,a)^\top \theta_h^*)}.$$

Remark 2. While the work of Hwang and Oh [2023] focuses on the *homogeneous* setting, where the transitions remain the same across all stages (i.e., $\mathbb{P}_1 = \ldots = \mathbb{P}_H$), we address the more general *inhomogeneous* setting, allowing transitions to vary across different stages.

For any $\theta \in \mathbb{R}^d$, we define the induced transition as $p_{s,a}^{s'}(\theta) = \frac{\exp(\phi(s' \mid s,a)^\top \theta)}{\sum_{\widetilde{s}\in\mathcal{S}_{s,a}} \exp(\phi(\widetilde{s} \mid s,a)^\top \theta)}$. We introduce the key problem-dependent quantities for this problem firstly introduced by Hwang and Oh [2023].

Assumption 1. There exists $0 < \kappa \le \kappa^* < 1$ such that for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and $s', s'' \in \mathcal{S}_{h,s,a}$, it holds that $\inf_{\theta\in\Theta} p_{s,a}^{s'}(\theta) p_{s,a}^{s''}(\theta) \ge \kappa$ and $p_{s,a}^{s'}(\theta^*) p_{s,a}^{s''}(\theta^*) > \kappa^*$.

Assumption 1 is similar to the standard assumption in generalized linear model literature [Filippi et al., 2010, Wang et al., 2021] and logistic bandits literature [Faury et al., 2020, Abeille et al., 2021, Zhang and Sugiyama, 2023] to guarantee the Hessian matrix is non-singular in Property 1.

4 Computationally Efficient Algorithm by Online Learning

In this part, we design an algorithm that achieves $\mathcal{O}(1)$ computational and storage costs per episode.

Since the transition parameter θ_h^* is unknown, we need to estimate it using the historical data. At episode k, we collect a trajectory $\{(s_{k,h}, a_{k,h})\}_{h=1}^H$, then define the variable: $y_{k,h} \in \{0,1\}^{N_{k,h}}$ where $y_{k,h}^{s'} = \mathbb{1}(s_{k,h+1} = s')$ for $s' \in \mathcal{S}_{k,h} \triangleq \mathcal{S}_{s_{k,h},a_{k,h}}$ and $N_{k,h} = |\mathcal{S}_{k,h}|$. We denote by $p_{k,h}^{s'}(\theta) = p_{s_{k,h},a_{k,h}}^{s'}(\theta)$. Then $y_{k,h}$ is a sample from the following multinomial distribution:

$$y_{k,h} \sim \text{multinomial}(1, [p_{k,h}^{s_1}(\theta^*), \dots, p_{k,h}^{s_{N_{k,h}}}(\theta^*)]),$$

where the parameter 1 indicates that $y_{k,h}$ is a single-trial sample. Furthermore, we define the noise $\epsilon_{k,h}^{s'} = y_{k,h}^{s'} - p_{k,h}^{s'}(\theta_h^*)$. It is clear that $\epsilon_{k,h} \in [-1,1]^{N_{k,h}}$, $\mathbb{E}[\epsilon_{k,h}] = \mathbf{0}$ and $\sum_{s' \in \mathcal{S}_{k,h}} \epsilon_{i,h}^{s'} = 0$.

Efficiency Concern. Hwang and Oh [2023] made the first step by proposing an approach using the maximum likelihood estimation (MLE). Specifically, their estimator $\theta_{k,h}$ is defined as

$$\theta_{k,h} = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{k-1} \sum_{s' \in \mathcal{S}_{i,h}} -y_{i,h}^{s'} \log p_{i,h}^{s'}(\theta) + \frac{\lambda}{2}\|\theta\|_2^2. \tag{1}$$

Despite favorable theoretical guarantee, the computational and storage cost of this method is expensive. First, in episode k, the estimator $\theta_{k,h}$ in (1) is computed using all samples collected in previous episodes, resulting an $\mathcal{O}(k)$ storage cost. Second, to solve this optimization problem, the projected gradient descent [Boyd and Vandenberghe, 2004] are usually applied. However, as discussed in Faury et al. [2022], the optimization of the MLE problem typically require $\mathcal{O}(k \log(1/\epsilon))$ iterations to achieve an ϵ-accurate solution, which is computationally expensive.

To address this issue, we introduce a novel algorithm named UCRL-MNL with Online Learning (UCRL-MNL-OL), which achieves both the same regret with $\mathcal{O}(1)$ computational and storage cost. At a high level, our algorithm can be divided into two phases: (i) efficient online estimation, and (ii) efficient optimistic value function construction. We introduce the details in the following.

Efficient Online Estimation. Instead of using all historical data, our algorithm updates the estimator in an online manner. Specifically, inspired by the works [Hazan et al., 2014, Zhang et al., 2016] on logistic bandit, we find the negative log-likelihood function is exponentially concave, which motivates us to apply a variant of online Newton step [Zhang et al., 2016, Oh and Iyengar, 2021].

First, we define per-episodic loss function $f_{k,h}(\theta)$, gradient $g_{k,h}(\theta)$ and Hessian matrix $H_{k,h}(\theta)$ as

$$f_{k,h}(\theta) = -\sum_{s' \in \mathcal{S}_{k,h}} y_{k,h}^{s'} \log p_{k,h}^{s'}(\theta), \quad g_{k,h}(\theta) = \sum_{s' \in \mathcal{S}_{k,h}} (p_{k,h}^{s'}(\theta) - y_{k,h}^{s'})\phi_{k,h}^{s'},$$

$$H_{k,h}(\theta) = \sum_{s' \in \mathcal{S}_{k,h}} p_{k,h}^{s'}(\theta)\phi_{k,h}^{s'}(\phi_{k,h}^{s'})^\top - \sum_{s' \in \mathcal{S}_{k,h}} \sum_{s'' \in \mathcal{S}_{k,h}} p_{k,h}^{s'}(\theta) p_{k,h}^{s''}(\theta)\phi_{k,h}^{s'}(\phi_{k,h}^{s''})^\top. \tag{2}$$

We introduce the following key property of the loss function $f_{k,h}(\theta)$ by a second-order Taylor expansion and mean value theorem, whose proof can be found in Lemma 5 in Appendix A.

Property 1. *(i). For any $\theta_1, \theta_2 \in \mathbb{R}^d$, $\exists\, \bar\theta = \nu\theta_1 + (1-\nu)\theta_2$ such that*

$$f_{k,h}(\theta_2) = f_{k,h}(\theta_1) + \langle g_{k,h}(\theta_1), \theta_2 - \theta_1\rangle + \frac{1}{2}\|\theta_2 - \theta_1\|_{H_{k,h}(\bar\theta)}^2.$$

(ii). For any $\theta \in \Theta$, it holds that

$$H_{k,h}(\theta) \succeq \kappa \sum_{s' \in \mathcal{S}_{k,h}} \phi_{k,h}^{s'}(\phi_{k,h}^{s'})^\top \triangleq \kappa W_{k,h}. \tag{3}$$

Property 1 implies that $f_{k,h}(\theta)$ is exponentially concave, which enables us to apply the online Newton step to construct an efficient estimator. Specifically, we construct $\widehat\theta_{k+1,h}$ by solving the problem:

$$\widehat\theta_{k+1,h} = \arg\min_{\theta \in \Theta} \langle g_{k,h}(\widehat\theta_{k,h}), \theta - \widehat\theta_{k,h}\rangle + \frac{1}{2}\|\theta - \widehat\theta_{k,h}\|_{\widehat\Sigma_{k+1,h}}^2, \tag{4}$$

where $\widehat\Sigma_{k+1,h} = \widehat\Sigma_{k,h} + \frac{\kappa}{2}\sum_{s' \in \mathcal{S}_{k,h}} \phi_{k,h}^{s'}(\phi_{k,h}^{s'})^\top$ is the feature covariance matrix.

Algorithm 1 UCRL-MNL-OL

Input: Regularization parameter λ, confidence width $\widehat{\beta}_k$, confidence parameter δ.

Initialization: $\widehat{\Sigma}_{1,h} = \lambda I, \widehat{\theta}_{1,h} = \mathbf{0}$ for all $h \in [H]$.
for $k = 1, \ldots, K$ **do**
  Compute $\widehat{Q}_{k,h}(\cdot,\cdot)$ in a backward way as in (7).
  for $h = 1, \ldots, H$ **do**
    Observe current state $s_{k,h}$, select action $a_{k,h} = \arg\max_{a \in \mathcal{A}} \widehat{Q}_{k,h}(s_{k,h}, a)$.
    Set $\widehat{\Sigma}_{k+1,h} = \widehat{\Sigma}_{k,h} + \frac{\kappa}{2} \sum_{s' \in \mathcal{S}_{k,h}} \phi_{k,h}^{s'} (\phi_{k,h}^{s'})^\top$.
    Compute $\widehat{\theta}_{k+1,h} = \arg\min_{\theta \in \Theta} \langle g_{k,h}(\widehat{\theta}_{k,h}), \theta - \widehat{\theta}_{k,h} \rangle + \frac{1}{2} \|\theta - \widehat{\theta}_{k,h}\|^2_{\widehat{\Sigma}_{k+1,h}}$.
  end for
end for

We show the online estimator $\widehat{\theta}_{k,h}$ in (4) enjoys computational and storage efficiency simultaneously. As the optimization problem exhibits a standard online mirror descent formulation, it can be solved with a projected gradient step with the following equivalent formulation by

$$\theta'_{k+1,h} = \widehat{\theta}_{k,h} - \widehat{\Sigma}^{-1}_{k+1,h} g_{k,h}(\widehat{\theta}_{k,h}), \quad \text{and} \quad \widehat{\theta}_{k+1,h} = \arg\min_{\theta \in \Theta} \|\theta - \theta'_{k+1,h}\|_{\widehat{\Sigma}_{k+1,h}}.$$

This update enjoys a computational cost of only $\mathcal{O}(d^3 U)$, independent of episode count k [Zhang and Sugiyama, 2023, Lee and Oh, 2024]. As for storage costs, it avoids the need to store all historical data by updating the feature covariance matrix $\widehat{\Sigma}_{k,h}$ incrementally, requiring only $\mathcal{O}(d^2)$ storage cost. We note that a dependence on U is introduced in the computational cost. However, as we have discussed in Remark 1, U can be much smaller than the state space size S, which is acceptable. This dependency is typical for model-based methods that directly learn transitions, as it need to control the estimation error of transitions, which typically involves a total of U elements. Similar dependencies have been observed in the literature [Hwang and Oh, 2023]. Additionally, even for the model-free method [Yang and Wang, 2020, Zhou et al., 2021] which learn value functions, a common assumption is the value $\sum_{s' \in \mathcal{S}_{k,h}} \phi_{k,h}^{s'} V(s')$ can be obtained by an Oracle, which depends on U implicitly.

Then, we show the estimator $\widehat{\theta}_{k,h}$ is close to the true parameter θ_h^* by the following lemma.

Lemma 1. *For any $\delta \in (0,1)$, define the confidence set as*

$$\widehat{\mathcal{C}}_{k,h} = \left\{ \theta \in \Theta \mid \|\theta - \widehat{\theta}_{k,h}\|_{\widehat{\Sigma}_{k,h}} \leq \sqrt{\lambda} B + \sqrt{\frac{8}{\kappa} d \log\left(1 + \frac{kUH}{d\lambda\delta}\right)} \triangleq \widehat{\beta}_k \right\}. \tag{5}$$

Then, we have $\Pr[\theta_h^* \in \widehat{\mathcal{C}}_{k,h}] \geq 1 - \delta, \forall k \in [K], h \in [H]$.

Remark 3. Notably, a similar confidence set is achieved by Hwang and Oh [2023] by using the MLE defined in (1). In contrast, we obtain the same results by using an online estimator which only suffers $\mathcal{O}(1)$ computation and storage cost per round. Besides, we identify a technical issue in their analysis. Specifically, they bound the confidence set using the self-normalized concentration for vector-valued martingales in Lemma 15. However, it is crucial to recognize that the noise is not independent and satisfies $\sum_{s' \in \mathcal{S}_{i,h}} \epsilon_{i,h}^{s'} = 0$ since the learner visits each stage h exactly once per episode. Thus, the noise $\epsilon_{i,h}^{s'}$ becomes deterministic and non-zero given the remaining noise $\epsilon_{i,h}^{s''}$ for $s'' \neq s'$. This contravenes the *zero-mean* sub-Gaussian condition in Lemma 15. We observe similar oversights also appear in the works on multinomial logit contextual bandits [Oh and Iyengar, 2019, 2021, Agrawal et al., 2023]. To our knowledge, this issue has not been explicitly identified in previous studies. We note that this issue can be resolved by a new self-normalized concentration with dependent noises in Lemma 1 of Li et al. [2024b] with only slight modifications in constant factors.

Efficient Optimistic Value Function Construction. Given the confidence set $\widehat{\mathcal{C}}_{k,h}$, the most direct way is to adopt the principle of "optimism in the face of uncertainty" and construct the optimistic value function as the maximum expected reward over the confidence set. Specifically, we define the optimistic value function $\widehat{Q}_{k,h}(s,a)$ and $\widehat{V}_{k,h}(s)$ as

$$\widehat{Q}_{k,h}(s,a) = \left[r_h(s,a) + \max_{\theta \in \widehat{\mathcal{C}}_{k,h}} \sum_{s' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\theta) \widehat{V}_{k,h+1}(s') \right]_{[0,H]}, \widehat{V}_{k,h}(s) = \max_{a \in \mathcal{A}} \widehat{Q}_{k,h}(s,a). \tag{6}$$

However, this construction is not efficient as it requires solving a maximization problem over the confidence set. Actually, the algorithm only requires the estimated action-value function to be an upper bound for the true value functions. Thus, we can use a closed-form confidence bound instead of computing the maximal value over the confidence set. To this end, we present the following lemma that enables us to construct the optimistic value function more efficiently.

Lemma 2. *Suppose Lemma 1 holds. For any $V : \mathcal{S} \to [0, H]$ and $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$, it holds*

$$\left| \sum_{s' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\widehat{\theta}_{k,h}) V(s') - \sum_{s' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\theta_h^*) V(s') \right| \leq H \widehat{\beta}_k \max_{s' \in \mathcal{S}_{h,s,a}} \|\phi_{s,a}^{s'}\|_{\widehat{\Sigma}_{k,h}^{-1}}.$$

Based on this lemma, we can replace the maximization problem in (6) with closed-form confidence bound and construct the optimistic value function $\widehat{Q}_{k,h}(s, a)$ as

$$\widehat{Q}_{k,h}(s, a) = \left[r_h(s, a) + \sum_{s' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\widehat{\theta}_{k,h}) \widehat{V}_{k,h+1}(s') + H \widehat{\beta}_k \max_{s' \in \mathcal{S}_{h,s,a}} \|\phi_{s,a}^{s'}\|_{\widehat{\Sigma}_{k,h}^{-1}} \right]_{[0,H]}. \quad (7)$$

At state $s_{k,h}$, our algorithm chooses $a_{k,h} = \arg\max_{a \in \mathcal{A}} \widehat{Q}_{k,h}(s_{k,h}, a)$. Algorithm 1 presents the detailed procedure. We show that our algorithm achieves the following regret guarantee.

Theorem 1. *For any $\delta \in (0, 1)$, set $\widehat{\beta}_k$ as in Lemma 1 and $\lambda = 1$, with probability at least $1 - \delta$, Algorithm 1 ensures the following regret guarantee*

$$\mathrm{Reg}(K) \leq \widetilde{\mathcal{O}}\big(\kappa^{-1} d H^2 \sqrt{K}\big).$$

Remark 4. Our algorithm attains the same regret of Hwang and Oh [2023], yet for the more general inhomogeneous MDPs. Importantly, our algorithm only requires constant computational and storage costs, matching the computational efficiency of the linear cases [Jin et al., 2020, Zhou et al., 2021].

5 Statistically Improved Algorithm by Local Learning

In this section, we present an enhanced algorithm, UCRL-MNL-LL, that leverages local information to improve the statistical efficiency for MDPs with MNL function approximation.

While the UCRL-MNL-OL algorithm in Section 4 offers favorable computational and storage efficiency, it suffers from a dependence on κ^{-1} in the regret. By Assumption 1, the quantity κ may be exponentially small in the worst case, exhibiting an exponential dependence on the radius of the parameter set and linear in the number of reachable states, i.e., $\kappa^{-1} = \mathcal{O}((U \exp(B))^2)$. This creates a significant gap between MNL and linear function approximation.

Recently, the improved dependence on κ has been achieved in the logistic bandit literature [Faury et al., 2020, Abeille et al., 2021, Périvier and Goyal, 2022] by the use of generalization of the Bernstein-like tail inequality [Faury et al., 2020] and the self-concordant-like property of the log loss [Bach, 2010]. Thus, a natural question then arises: *Can we improve the statistical efficiency of MNL function approximation for MDPs?* We answer this question affirmatively by proposing an enhanced algorithm that reduces the dependence on κ significantly through the use of local information. Though the achievement has been made in the logistic bandit, the extension to MDPs with MNL function approximation is non-trivial and new techniques specific to MDPs are required.

We first analyze where the dependence on κ comes from. By the analysis of Theorem 1, we can see that the regret of the algorithm can be upper-bounded as follows,

$$\mathrm{Reg}(K) \leq 2H\widehat{\beta}_K \sum_{k=1}^{K} \sum_{h=1}^{H} \max_{s' \in \mathcal{S}_{k,h}} \|\phi_{k,h}^{s'}\|_{\widehat{\Sigma}_{k,h}^{-1}} + \sum_{k=1}^{K} \sum_{h=1}^{H} \mathcal{M}_{k,h}. \quad (8)$$

Here, the first term corresponds to the overestimation of the value function, and the second term represents the martingale difference sequence arising from the stochastic transition dynamics. The second term can be bounded using the Azuma-Hoeffding inequality, which is independent of κ. For the first term, by the construction of the confidence set in Lemma 1, $\widehat{\beta}_k$ represents the width of the confidence set, contributing to a $\kappa^{-1/2}$ dependence. The feature covariance matrix $\widehat{\Sigma}_{k,h}$ defined in (4) reflects the degree of exploration across different states, which also results in a $\kappa^{-1/2}$ dependence.

To mitigate these dependencies, we design a local learning algorithm that: (i) constructs a confidence set independent of κ; and (ii) builds a κ-independent feature covariance matrix that effectively captures the exploration degree across different states. We introduce the details in the following.

Algorithm 2 UCRL-MNL-LL

Input: Step size η, regularization parameter λ, confidence width $\widetilde{\beta}_k$, confidence parameter δ.

Initialization: $\mathcal{H}_{1,h} = \lambda I, \widehat{\theta}_{1,h} = \mathbf{0}$ for all $h \in [H]$.
for $k = 1, \ldots, K$ **do**
  Compute $\widetilde{Q}_{k,h}(\cdot, \cdot)$ in a backward way as in (11).
  for $h = 1, \ldots, H$ **do**
    Observe state $s_{k,h}$, select action $a_{k,h} = \arg\max_{a \in \mathcal{A}} \widetilde{Q}_{k,h}(s_{k,h}, a)$.
    Update $\widetilde{\mathcal{H}}_{k,h} = \mathcal{H}_{k,h} + \eta H_{k,h}(\widetilde{\theta}_{k,h})$.
    Compute $\widetilde{\theta}_{k+1,h} = \arg\min_{\theta \in \Theta} \langle g_{k,h}(\widetilde{\theta}_{k,h}), \theta - \widetilde{\theta}_{k,h} \rangle + \frac{1}{2\eta} \|\theta - \widetilde{\theta}_{k,h}\|_{\widetilde{\mathcal{H}}_{k,h}}$.
    Update $\mathcal{H}_{k,h} = \mathcal{H}_{k,h} + H_{k,h}(\widetilde{\theta}_{k+1,h})$.
  end for
end for

5.1 Improved Online Estimation

In Property 1, we show the Hessian matrix $H_{k,h}$ is lower bounded by a positive definite matrix $\kappa \sum_{s' \in \mathcal{S}_{k,h}} \phi_{k,h}^{s'} (\phi_{k,h}^{s'})^\top$ uniformly over the parameter domain, as in (3). This quantity measures the exploration degree across different states and is used to update the parameter estimation and construct the confidence set. However, the bound is not tight, as the Hessian matrix can be significantly larger in certain regions, away from the global minimum. This observation motivates us to design a local learning algorithm that can adaptively leverage local information for improved guarantee.

Inspired by recent advances in multinomial logistic bandit (MLogB) [Zhang and Sugiyama, 2023] and multinomial logit contextual bandit (MNL) [Lee and Oh, 2024], we run an online mirror descent algorithm to estimate the parameter θ_h^*. Differently from (4), we use the local Hessian matrix $\widetilde{\mathcal{H}}_{k,h}$ to update the estimation instead of the global lower bound. Specifically, we estimate $\widetilde{\theta}_{k,h}$ as follows:

$$\widetilde{\theta}_{k+1,h} = \arg\min_{\theta \in \Theta} \langle g_{k,h}(\widetilde{\theta}_{k,h}), \theta - \widetilde{\theta}_{k,h} \rangle + \frac{1}{2\eta} \|\theta - \widetilde{\theta}_{k,h}\|_{\widetilde{\mathcal{H}}_{k,h}}. \tag{9}$$

where η is the step size, $\mathcal{H}_{k,h} = \sum_{i=1}^{k-1} H_{i,h}(\widetilde{\theta}_{i+1,h}) + \lambda I_d$ and $\widetilde{\mathcal{H}}_{k,h} = \mathcal{H}_{k,h} + \eta H_{k,h}(\widetilde{\theta}_{k,h})$. Note that both $\mathcal{H}_{k,h}$ and $\widetilde{\mathcal{H}}_{k,h}$ can be updated incrementally. Similar to the update in (4), the optimization problem in (9) can be efficiently solved using the following two-step update:

$$\theta'_{k+1,h} = \theta_{k,h} - \eta \widetilde{\mathcal{H}}_{k,h}^{-1} g_{k,h}(\theta_{k,h}), \quad \text{and} \quad \widetilde{\theta}_{k+1,h} = \arg\min_{\theta \in \Theta} \|\theta - \theta'_{k+1,h}\|_{\widetilde{\mathcal{H}}_{k,h}}.$$

This two-step update procedure incurs a computational cost of $\mathcal{O}(d^3 U)$ and a storage cost of $\mathcal{O}(d^2)$ per episode, both independent of the episode count k.

Based on this estimator, we can construct the κ-independent confidence set as follows.

Lemma 3. *For any $\delta \in (0, 1)$, define the confidence set as*

$$\widetilde{\mathcal{C}}_{k,h} = \left\{ \theta \in \Theta \mid \|\theta - \widetilde{\theta}_{k,h}\|_{\mathcal{H}_{k,h}} \leq \widetilde{\beta}_k \right\},$$

where $\widetilde{\beta}_k = \mathcal{O}(\sqrt{d} \log U \log(KH/\delta))$. Then, we have $\Pr[\theta_h^* \in \widetilde{\mathcal{C}}_{k,h}] \geq 1 - \delta, \forall k \in [K], h \in [H]$.

Remark 5. Zhang and Sugiyama [2023] studied the multiple-parameter MLogB model, where the unknown parameter is a matrix. Consequently, the confidence set in Theorem 3 of their work exhibits a polynomial dependence on the number of possible outcomes, which corresponds to the number of reachable states U in our setting. This dependence is acceptable in the bandit setting while is not suitable for the MDP setting. In contrast, Lee and Oh [2024] focused on the single-parameter MNL model and revisited the self-concordant-like property, demonstrating that the log-loss of the single-parameter MNL model is $3\sqrt{2}$-self-concordant-like (Proposition B.1 in Lee and Oh [2024]). This property is crucial for the improved confidence set in Lemma 3 that is independent of κ and U.

5.2 Improved Optimistic Value Function Construction

Based on the confidence set in Lemma 3, a natural choice for the optimistic value function construction is analogous to (7) and can be expressed as:

$$\bar{Q}_{k,h}(s,a) = \left[r_h(s,a) + \sum_{s' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\widehat{\theta}_{k,h}) \bar{V}_{k,h+1}(s') + H\widetilde{\beta}_k \max_{s' \in \mathcal{S}_{h,s,a}} \|\phi_{s,a}^{s'}\|_{\mathcal{H}_{k,h}^{-1}} \right]_{[0,H]}. \quad (10)$$

However, though $\widetilde{\beta}_K$ now is independent of κ and $\mathcal{H}_{k,h}^{-1}$ preserves the local information in (10), the norm $\max_{s' \in \mathcal{S}_{k,h}} \|\phi_{k,h}^{s'}\|_{\mathcal{H}_{k,h}^{-1}}$ is still in a global manner due to the maximum operation, leading to a $\kappa^{-1/2}$ dependence. To address this issue, we propose a new construction of the optimistic value function. Specifically, we employ a second-order Taylor expansion to more accurately bound the value difference arising from transition estimation errors.

Although the idea of using second-order Taylor expansion has been explored in bandits [Périvier and Goyal, 2022, Zhang and Sugiyama, 2023, Lee and Oh, 2024], fundamental differences arise in the MDP setting. Specifically, Périvier and Goyal [2022] studied the *uniform* revenue setting, where the reward is identical for all actions. Zhang and Sugiyama [2023] and Lee and Oh [2024] focused on the non-uniform setting, but the rewards for different actions are *known* a priori. However, in the MDP setting, the value function is *state-dependent* and *unknown* to the learner, leading to a more challenging problem. Moreover, in MNL bandits, the objective is to select a series of assortments with *varying* sizes that maximize the expected revenue, whereas in MNL MDPs, the goal is to choose *one* action at each stage to maximize the cumulative reward. Due to these differences, the analyses used in the bandit setting cannot be directly applied to the MDP setting.

For MDPs, we show the value difference arising from the transition estimation error as follows.

Lemma 4. *Suppose Lemma 3 holds. For any $V : \mathcal{S} \to [0,H]$ and $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, it holds*

$$\left| \sum_{s' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\widehat{\theta}_{k,h}) V(s') - \sum_{s' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\theta_h^*) V(s') \right| \leq \epsilon_{s,a}^{\texttt{fst}} + \epsilon_{s,a}^{\texttt{snd}}.$$

where

$$\epsilon_{s,a}^{\texttt{fst}} = H\widetilde{\beta}_k \sum_{s' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\widetilde{\theta}_{k,h}) \Big\| \phi_{s,a}^{s'} - \sum_{s'' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s''}(\widetilde{\theta}_{k,h}) \phi_{s,a}^{s''} \Big\|_{\mathcal{H}_{k,h}^{-1}}, \epsilon_{s,a}^{\texttt{snd}} = \frac{5}{2} H \widetilde{\beta}_k^2 \max_{s' \in \mathcal{S}_{h,s,a}} \|\phi_{s,a}^{s'}\|_{\mathcal{H}_{k,h}^{-1}}^2.$$

Based on Lemma 4, we construct the optimistic value function as follows:

$$\widetilde{Q}_{k,h}(s,a) = \left[r_h(s,a) + \sum_{s' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\widehat{\theta}_{k,h}) \widetilde{V}_{k,h+1}(s') + \epsilon_{s,a}^{\texttt{fst}} + \epsilon_{s,a}^{\texttt{snd}} \right]_{[0,H]}. \quad (11)$$

In contrast to the value function specified in (10), where the term $\max_{s' \in \mathcal{S}_{h,s,a}} \|\phi_{s,a}^{s'}\|_{\mathcal{H}_{k,h}^{-1}}$ is utilized, the refined value function introduced in (11) substitutes this term with $\epsilon_{s,a}^{\texttt{fst}} + \epsilon_{s,a}^{\texttt{snd}}$. This adjustment better preserves local information, offering a more precise and κ-independent estimation error bound.

At state $s_{k,h}$, our algorithm chooses action $a_{k,h} = \max \widetilde{Q}_{k,h}(s_{k,h}, \cdot)$. The detailed algorithm is presented in Algorithm 2. We show the regret guarantee of UCRL-MNL-LL in the following theorem.

Theorem 2. *For any $\delta \in (0,1)$, set $\widetilde{\beta}_k = \mathcal{O}(\sqrt{d} \log U \log(KH/\delta))$, $\eta = \frac{1}{2}\log(1+U) + (B+1)$ and $\lambda = 84\sqrt{2}\eta(B+d)$, with probability at least $1-\delta$, UCRL-MNL-LL algorithm ensures the following regret guarantee*

$$\mathrm{Reg}(K) \leq \widetilde{\mathcal{O}}\big(dH^2\sqrt{K} + \kappa^{-1}d^2H^2\big).$$

Remark 6. The high-order term in Theorem 2 is now independent κ, significantly improving the statistical efficiency compared with Theorem 1. In comparison with the optimal regret of $\widetilde{\mathcal{O}}(d\sqrt{H^3K})$ for linear cases [Zhou et al., 2021, He et al., 2023], the higher-order term in Theorem 2 only differs by a factor of $H^{1/2}$, almost matching the same computational and statistical efficiency simultaneously.

6 Lower Bound

In this section, we establish a lower bound for MNL mixture MDPs by presenting a novel reduction, which connects MNL mixture MDPs and the logistic bandit problem.

Consider the following logistic bandit problem [Faury et al., 2020, Abeille et al., 2021]: at each round $k \in [K]$, the learner selects an action $x_k \in \mathcal{X}$ and receives a reward r_k sampled from Bernoulli distribution with mean $\mu(x^\top \theta^*) = (1 + \exp(-x^\top \theta^*))^{-1}$, where $\theta^* \in \{\theta \in \mathbb{R}^d, \|\theta\|_2 \le B\}$ is the unknown parameter. The learner aims to to minimize the regret:

$$\mathrm{Reg}^{\mathtt{LogB}}(K) = \max_{x \in \mathcal{X}} \sum_{k=1}^{K} \mu(x^\top \theta^*) - \sum_{k=1}^{K} \mu(x_k^\top \theta^*).$$

Theorem 3 (Lower Bound)**.** *For any logistic bandit problem $\mathcal{B}$, there exists an MNL mixture MDP $\mathcal{M}$ such that learning $\mathcal{M}$ is as hard as learning $H/2$ independent instances of $\mathcal{B}$ simultaneously.*

Corollary 1. For any problem instance $\{\theta_h^*\}_{h=1}^H$ and for $K \ge d^2\kappa^*$, there exists an MNL mixture MDP with *infinite* action space such that $\mathrm{Reg}(K) \ge \Omega(dH\sqrt{K\kappa^*})$.

Remark 7. Corollary 1 can be proved by combining Theorem 3 and the $\Omega(d\sqrt{K\kappa^*})$ lower bound for logistic bandits with *infinite* arms by Abeille et al. [2021]. To the best of our knowledge, a lower bound for logistic bandits with finite arms has not been established, which is beyond the scope of this work. This absence leaves the lower bound for MNL mixture MDPs with a finite action space open through this reduction. However, after the submission of our work to arXiv [Li et al., 2024a], a follow up work by Park et al. [2024] proposed a new reduction that bridges MNL mixture MDPs with linear mixture MDPs by approximating multinomial logit functions to linear functions, employing the mean value theorem. Leveraging this new reduction, they established an $\Omega(dH^{3/2}\sqrt{K})$ lower bound for MNL mixture MDPs with the finite action space. This achievement confirms that our result is indeed optimal with respect to the dependence on d and K, only differing by a $\mathcal{O}(H^{1/2})$ factor.

Dependence on H. By the discussion in Remark 7, we note that our result is optimal with respect to d and K, but loosing by a $\mathcal{O}(H^{1/2})$ factor. We discuss the challenges in improving the dependence on H. Notably, MNL mixture MDPs can be viewed as a generalization of linear mixture MDPs [Ayoub et al., 2020, Zhou et al., 2021]. The pioneering work by Ayoub et al. [2020] achieved a regret bound of $\widetilde{\mathcal{O}}(dH^2\sqrt{K})$ for linear mixture MDPs, which matches our results in Theorem 2, differing only on the lower order term. Later, Zhou et al. [2021] enhanced the dependence on H and attained an optimal regret bound of $\widetilde{\mathcal{O}}(d\sqrt{H^3K})$. This was made possible by recognizing that the value function in linear mixture MDPs is linear, allowing for direct learning of the value function while incorporating *variance information*. In contrast, the value function for MNL mixture MDPs does not conform to a specific structure, posing a significant challenge in using the variance information of value functions. Thus, it remains open whether similar improvements on H are attainable for MNL mixture MDPs.

7 Conclusion and Future Work

In this work, we study MNL mixture MDPs that employ multinomial logit function approximation to ensure valid probability distributions over the state space. We address both the computational and statistical challenges for this problem. Specifically, we first propose an algorithm based on the online Newton step that attains the $\widetilde{\mathcal{O}}(\kappa^{-1}dH^2\sqrt{K})$ regret with $\mathcal{O}(1)$ computational and storage costs per episode. Next, we propose an enhanced algorithm that leverages local information to improve the statistical efficiency. It not only maintains $\mathcal{O}(1)$ computational and storage costs but also achieves an improved regret of $\widetilde{\mathcal{O}}(dH^2\sqrt{K} + d^2H^2\kappa^{-1})$, nearly matching the result of linear function approximation from both computational and statistical perspectives. Finally, we establish the first lower bound for MNL mixture MDPs, justifying the optimality of our results in d and K.

There are several interesting directions for future work. First, there still exists a gap between the upper and lower bounds and how to close this gap is an open problem. Besides, while this work focuses on stochastic rewards, extending this model to the non-stationary settings and studying the dynamic regret [Wei and Luo, 2021, Zhao et al., 2022, Li et al., 2023] is an another important direction.

Acknowledgments

This research was supported by National Science and Technology Major Project (2022ZD0114800) and NSFC (U23A20382, 62206125). Peng Zhao was supported in part by the Xiaomi Foundation.

References

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

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

Priyank Agrawal, Theja Tulabandhula, and Vashist Avadhanula. A tractable online learning algorithm for the multinomial logit contextual bandit. *European Journal of Operational Research*, 310(2): 737–750, 2023.

Sanae Amani and Christos Thrampoulidis. Ucb-based algorithms for multinomial logistic regression bandits. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 2913–2924, 2021.

Alex Ayoub, Zeyu Jia, Csaba Szepesvári, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 463–474, 2020.

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4: 384–414, 2010.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Zixiang Chen, Chris Junchi Li, Huizhuo Yuan, Quanquan Gu, and Michael I. Jordan. A general framework for sample-efficient function approximation in reinforcement learning. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

Wang Chi Cheung and David Simchi-Levi. Thompson sampling for online personalized assortment optimization problems with multinomial logit choice models. *Available at SSRN 3075658*, 2017.

Sayak Ray Chowdhury, Aditya Gopalan, and Odalric-Ambrym Maillard. Reinforcement learning in parametric mdps with exponential families. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1855–1863, 2021.

Simon S. Du, Sham M. Kakade, Jason D. Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in RL. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 2826–2836, 2021.

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

Louis Faury, Marc Abeille, Kwang-Sung Jun, and Clément Calauzènes. Jointly efficient and optimal algorithms for logistic bandits. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 546–580, 2022.

Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J R Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53, 2022.

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

Dylan J. Foster, Sham M. Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *ArXiv preprint*, 2112.13487, 2021.

Elad Hazan, Tomer Koren, and Kfir Y. Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *Proceedings of The 27th Conference on Learning Theory (COLT)*, pages 197–209, 2014.

Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for linear Markov decision processes. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 12790–12822, 2023.

Taehyun Hwang and Min-hwan Oh. Model-based reinforcement learning with multinomial logistic function approximation. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*, pages 7971–7979, 2023.

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

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 13406–13418, 2021.

Joongkyu Lee and Min-hwan Oh. Nearly minimax optimal regret for multinomial logistic bandit. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, page to appear, 2024.

Gene Li, Junbo Li, Anmol Kabra, Nati Srebro, Zhaoran Wang, and Zhuoran Yang. Exponential family model-based reinforcement learning via score matching. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Long-Fei Li, Peng Zhao, and Zhi-Hua Zhou. Dynamic regret of adversarial linear mixture MDPs. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pages 60685–60711, 2023.

Long-Fei Li, Yu-Jie Zhang, Peng Zhao, and Zhi-Hua Zhou. Provably efficient reinforcement learning with multinomial logit function approximation. *ArXiv preprint*, 2405.17061, 2024a. URL https://arxiv.org/abs/2405.17061v1.

Long-Fei Li, Peng Zhao, and Zhi-Hua Zhou. Improved algorithm for adversarial linear mixture MDPs with bandit feedback and unknown transition. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3061–3069, 2024b.

Min-hwan Oh and Garud Iyengar. Thompson sampling for multinomial logit contextual bandits. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 3145–3155, 2019.

Min-hwan Oh and Garud Iyengar. Multinomial logit contextual bandits: Provable optimality and practicality. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, pages 9205–9213, 2021.

Francesco Orabona. A modern introduction to online learning. *ArXiv preprint*, 1912.13213, 2019.

Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 1466–1474, 2014.

Reda Ouhamma, Debabrota Basu, and Odalric Maillard. Bilinear exponential family of mdps: Frequentist regret bound with tractable exploration & planning. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*, pages 9336–9344, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 27730–27744, 2022.

Jaehyun Park, Junyeop Kwon, and Dabeen Lee. Infinite-horizon reinforcement learning with multinomial logistic function approximation. *ArXiv preprint*, 2406.13633, 2024.

Noémie Périvier and Vineet Goyal. Dynamic pricing and assortment under a contextual MNL demand. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 3461–3474, 2022.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 2256–2264, 2013.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, pages 484–489, 2016.

Alexander L. Strehl and Michael L. Littman. An empirical evaluation of interval estimation for markov decision processes. In *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 128–135, 2004.

Yining Wang, Ruosong Wang, Simon Shaolei Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

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

Lin Yang and Mengdi Wang. Sample-optimal parametric Q-learning using linearly additive features. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6995–7004, 2019.

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 10746–10756, 2020.

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

Yu-Jie Zhang and Masashi Sugiyama. Online (multinomial) logistic bandit: Improved regret and constant computation cost. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pages 29741–29782, 2023.

Canzhe Zhao, Ruofeng Yang, Baoxiang Wang, and Shuai Li. Learning adversarial linear mixture Markov decision processes with bandit feedback and unknown transition. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

Peng Zhao, Long-Fei Li, and Zhi-Hua Zhou. Dynamic regret of online Markov decision processes. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 26865–26894, 2022.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvári. Nearly minimax optimal reinforcement learning for linear mixture Markov decision processes. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, pages 4532–4576, 2021.

A Properties of the Multinomial Logit Function

This section collects several key properties of the multinomial logit function used in the paper.

Without loss of generality, we assume $\forall \mathcal{S}_{h,s,a}, \exists s' \in \mathcal{S}_{h,s,a}$ such that $\phi(s' \mid s,a) = \mathbf{0}$. Otherwise, we can always define a new feature mapping $\phi'(s'' \mid s,a) = \phi(s' \mid s,a) - \phi(s'' \mid s,a)$ for any $s'' \in \mathcal{S}_{h,s,a}$ such that $\phi'(s' \mid s,a) = \mathbf{0}$ and the transition induced by ϕ' is the same as that induced by ϕ. We denote this state as $\dot{s}_{h,s,a}$ and $\dot{\mathcal{S}}_{h,s,a} = \mathcal{S}_{h,s,a} \backslash \{\dot{s}_{h,s,a}\}$.

Recall the following definitions in the paper:

$$
\begin{aligned}
f_{k,h}(\theta) &= -\sum_{s' \in \mathcal{S}_{k,h}} y_{k,h}^{s'} \log p_{k,h}^{s'}(\theta), \\
g_{k,h}(\theta) &= \sum_{s' \in \mathcal{S}_{k,h}} (p_{k,h}^{s'}(\theta) - y_{k,h}^{s'}) \phi_{k,h}^{s'} \\
H_{k,h}(\theta) &= \sum_{s' \in \mathcal{S}_{k,h}} p_{k,h}^{s'}(\theta) \phi_{k,h}^{s'} (\phi_{k,h}^{s'})^\top - \sum_{s' \in \mathcal{S}_{k,h}} \sum_{s'' \in \mathcal{S}_{k,h}} p_{k,h}^{s'}(\theta) p_{k,h}^{s''}(\theta) \phi_{k,h}^{s'} (\phi_{k,h}^{s''})^\top. \\
\widehat{\Sigma}_{k,h} &= \frac{\kappa}{2} \sum_{i=1}^{k-1} \sum_{s' \in \mathcal{S}_{i,h}} \phi_{i,h}^{s'} (\phi_{i,h}^{s'})^\top + \lambda I_d, \\
\mathcal{H}_{k,h} &= \sum_{i=1}^{k-1} H_{i,h}(\widetilde{\theta}_{i+1,h}) + \lambda I_d,
\end{aligned}
$$

Lemma 5. *The following statements hold for any $k \in [K], h \in [H]$:*

$$
H_{k,h}(\theta) \succeq \sum_{s' \in \dot{\mathcal{S}}_{k,h}} p_{k,h}^{s'}(\theta) p_{k,h}^{\dot{s}_{k,h}}(\theta) \phi_{k,h}^{s'} (\phi_{k,h}^{s'})^\top \succeq \kappa \sum_{s' \in \dot{\mathcal{S}}_{k,h}} \phi_{k,h}^{s'} (\phi_{k,h}^{s'})^\top.
$$

Proof. First, note that

$$
\forall x, y \in \mathbb{R}^d, (x-y)(x-y)^\top = xx^\top + yy^\top - xy^\top - yx^\top \succeq 0 \Longrightarrow xx^\top + yy^\top \succeq xy^\top + yx^\top.
$$

Then, we have

$$
\begin{aligned}
& H_{k,h}(\theta) \\
=& \sum_{s' \in \mathcal{S}_{k,h}} p_{k,h}^{s'}(\theta) \phi_{k,h}^{s'} (\phi_{k,h}^{s'})^\top - \sum_{s' \in \mathcal{S}_{k,h}} \sum_{s'' \in \mathcal{S}_{k,h}} p_{k,h}^{s'}(\theta) p_{k,h}^{s''}(\theta) \phi_{k,h}^{s'} (\phi_{k,h}^{s''})^\top \\
=& \sum_{s' \in \dot{\mathcal{S}}_{k,h}} p_{k,h}^{s'}(\theta) \phi_{k,h}^{s'} (\phi_{k,h}^{s'})^\top - \sum_{s' \in \dot{\mathcal{S}}_{k,h}} \sum_{s'' \in \dot{\mathcal{S}}_{k,h}} p_{k,h}^{s'}(\theta) p_{k,h}^{s''}(\theta) \phi_{k,h}^{s'} (\phi_{k,h}^{s''})^\top \\
=& \sum_{s' \in \dot{\mathcal{S}}_{k,h}} p_{k,h}^{s'}(\theta) \phi_{k,h}^{s'} (\phi_{k,h}^{s'})^\top - \frac{1}{2} \sum_{s' \in \dot{\mathcal{S}}_{k,h}} \sum_{s'' \in \dot{\mathcal{S}}_{k,h}} p_{k,h}^{s'}(\theta) p_{k,h}^{s''}(\theta) \left(\phi_{k,h}^{s'} (\phi_{k,h}^{s''})^\top + \phi_{k,h}^{s''} (\phi_{k,h}^{s'})^\top \right) \\
\succeq& \sum_{s' \in \dot{\mathcal{S}}_{k,h}} p_{k,h}^{s'}(\theta) \phi_{k,h}^{s'} (\phi_{k,h}^{s'})^\top - \frac{1}{2} \sum_{s' \in \dot{\mathcal{S}}_{k,h}} \sum_{s'' \in \dot{\mathcal{S}}_{k,h}} p_{k,h}^{s'}(\theta) p_{k,h}^{s''}(\theta) \left(\phi_{k,h}^{s'} (\phi_{k,h}^{s'})^\top + \phi_{k,h}^{s''} (\phi_{k,h}^{s''})^\top \right) \\
=& \sum_{s' \in \dot{\mathcal{S}}_{k,h}} p_{k,h}^{s'}(\theta) \phi_{k,h}^{s'} (\phi_{k,h}^{s'})^\top - \sum_{s' \in \dot{\mathcal{S}}_{k,h}} \sum_{s'' \in \dot{\mathcal{S}}_{k,h}} p_{k,h}^{s'}(\theta) p_{k,h}^{s''}(\theta) \phi_{k,h}^{s'} (\phi_{k,h}^{s'})^\top \\
=& \sum_{s' \in \dot{\mathcal{S}}_{k,h}} p_{k,h}^{s'}(\theta) \Big(1 - \sum_{s'' \in \dot{\mathcal{S}}_{k,h}} p_{k,h}^{s''}(\theta) \Big) \phi_{k,h}^{s'} (\phi_{k,h}^{s'})^\top \\
=& \sum_{s' \in \dot{\mathcal{S}}_{k,h}} p_{k,h}^{s'}(\theta) p_{k,h}^{\dot{s}_{k,h}}(\theta) \phi_{k,h}^{s'} (\phi_{k,h}^{s'})^\top \\
\succeq& \kappa \sum_{s' \in \dot{\mathcal{S}}_{k,h}} \phi_{k,h}^{s'} (\phi_{k,h}^{s'})^\top,
\end{aligned}
$$

where the last inequality holds by the definition of κ in Assumption 1. This finishes the proof. ■

Lemma 6. *Suppose $\lambda \geq 1$, define $\widetilde{\phi}_{k,h}^{s'} = \phi_{k,h}^{s'} - \sum_{s'' \in \mathcal{S}_{k,h}} p_{k,h}^{s''}(\widetilde{\theta}_{k+1,h})\phi_{k,h}^{s''}$, for any $k \in [K], h \in [H]$, the following statements hold:*

$$\text{(I)} \quad \sum_{i=1}^{k} \max_{s' \in \mathcal{S}_{i,h}} \|\phi_{i,h}^{s'}\|_{\widehat{\Sigma}_{i,h}^{-1}}^2 \leq \frac{4}{\kappa} d \log\left(1 + \frac{kU}{\lambda d}\right)$$

$$\text{(II)} \quad \sum_{i=1}^{k} \sum_{s' \in \mathcal{S}_{i,h}} p_{i,h}^{s'}(\widetilde{\theta}_{i+1,h}) p_{i,h}^{\dot{s}_{k,h}}(\widetilde{\theta}_{i+1,h}) \|\phi_{i,h}^{s'}\|_{\mathcal{H}_{i,h}^{-1}}^2 \leq 2d \log\left(1 + \frac{k}{\lambda d}\right)$$

$$\text{(III)} \quad \sum_{i=1}^{k} \max_{s' \in \mathcal{S}_{i,h}} \|\phi_{i,h}^{s'}\|_{\mathcal{H}_{i,h}^{-1}}^2 \leq \frac{2}{\kappa} d \log\left(1 + \frac{k}{\lambda d}\right)$$

$$\text{(IV)} \quad \sum_{i=1}^{k} \sum_{s' \in \mathcal{S}_{i,h}} p_{i,h}^{s'}(\widetilde{\theta}_{i+1,h}) \|\widetilde{\phi}_{i,h}^{s'}\|_{\mathcal{H}_{i,h}^{-1}}^2 \leq 2d \log\left(1 + \frac{k}{\lambda d}\right)$$

$$\text{(V)} \quad \sum_{i=1}^{k} \max_{s' \in \mathcal{S}_{i,h}} \|\widetilde{\phi}_{i,h}^{s'}\|_{\mathcal{H}_{i,h}^{-1}}^2 \leq \frac{2}{\kappa} d \log\left(1 + \frac{k}{\lambda d}\right)$$

Proof. We prove the five statements individually.

Proof of statement (I). By the definition that $\widehat{\Sigma}_{k+1,h} = \widehat{\Sigma}_{k,h} + \frac{\kappa}{2} \sum_{s' \in \mathcal{S}_{k,h}} \phi_{k,h}^{s'} (\phi_{k,h}^{s'})^\top$, we have

$$\det(\widehat{\Sigma}_{k+1,h}) = \det(\widehat{\Sigma}_{k,h}) \Big(1 + \frac{\kappa}{2} \sum_{s' \in \mathcal{S}_{k,h}} \|\phi_{k,h}^{s'}\|_{\widehat{\Sigma}_{k,h}^{-1}}^2\Big).$$

Then, we get

$$\sum_{i=1}^{k} \log\left(1 + \frac{\kappa}{2} \sum_{s' \in \mathcal{S}_{i,h}} \|\phi_{i,h}^{s'}\|_{\widehat{\Sigma}_{i,h}^{-1}}^2\right) \leq \log\left(\frac{\det(\widehat{\Sigma}_{k+1,h})}{\det(\widehat{\Sigma}_{1,h})}\right) \leq d \log\left(1 + \frac{kU}{\lambda d}\right), \tag{12}$$

where the last inequality holds by determinant trace inequality in Lemma 16. Since $\lambda \geq 1$, we have $\max_{s' \in \mathcal{S}_{i,h}} \|\phi_{i,h}^{s'}\|_{\widehat{\Sigma}_{i,h}^{-1}} \leq 1$, thus, we have

$$\begin{aligned}
\sum_{i=1}^{k} \max_{s' \in \mathcal{S}_{i,h}} \|\phi_{i,h}^{s'}\|_{\widehat{\Sigma}_{i,h}^{-1}} &\leq \frac{2}{\kappa} \sum_{i=1}^{k} \min\left\{1, \frac{\kappa}{2} \sum_{s' \in \mathcal{S}_{i,h}} \|\phi_{i,h}^{s'}\|_{\widehat{\Sigma}_{i,h}^{-1}}\right\} \\
&\leq \frac{4}{\kappa} \sum_{i=1}^{k} \log\left(1 + \frac{\kappa}{2} \sum_{s' \in \mathcal{S}_{i,h}} \|\phi_{i,h}^{s'}\|_{\widehat{\Sigma}_{i,h}^{-1}}^2\right) \\
&\leq \frac{4}{\kappa} d \log\left(1 + \frac{kU}{\lambda d}\right),
\end{aligned}$$

where the first inequality holds by the fact that $z \leq 2\log(1+z)$ for any $z \in [0,1]$ and the last inequality holds by (12).

Proof of statement (II). By Lemma 5, we have $H_{k,h}(\theta) \succeq \sum_{s' \in \dot{\mathcal{S}}_{k,h}} p_{k,h}^{s'}(\theta) p_{k,h}^{\dot{s}_{k,h}}(\theta) \phi_{k,h}^{s'} (\phi_{k,h}^{s'})^\top$. Thus, we have

$$\mathcal{H}_{k+1,h} \succeq \mathcal{H}_{k,h} + \sum_{s' \in \dot{\mathcal{S}}_{k,h}} p_{k,h}^{s'}(\widetilde{\theta}_{k+1,h}) p_{k,h}^{\dot{s}_{k,h}}(\widetilde{\theta}_{k+1,h}) \phi_{k,h}^{s'} (\phi_{k,h}^{s'})^\top$$

Then, we get

$$\det(\mathcal{H}_{i+1,h}) \geq \det(\mathcal{H}_{i,h}) \Big(1 + \sum_{s' \in \dot{\mathcal{S}}_{i,h}} p_{i,h}^{s'}(\widetilde{\theta}_{i+1,h}) p_{i,h}^{\dot{s}_{i,h}}(\widetilde{\theta}_{i+1,h}) \|\phi_{i,h}^{s'}\|_{\mathcal{H}_{i,h}^{-1}}^2\Big).$$

Since $\lambda \geq 1$, we have $\sum_{s' \in \dot{\mathcal{S}}_{i,h}} p_{i,h}^{s'}(\widetilde{\theta}_{i+1,h}) p_{i,h}^{\dot{s}_{i,h}}(\widetilde{\theta}_{i+1,h}) \|\phi_{i,h}^{s'}\|_{\mathcal{H}_{i,h}^{-1}}^2 \leq 1$. Using the fact that $z \leq 2\log(1+z)$ for any $z \in [0,1]$, we get

$$
\begin{aligned}
&\sum_{i=1}^{k} \sum_{s' \in \dot{\mathcal{S}}_{i,h}} p_{i,h}^{s'}(\widetilde{\theta}_{i+1,h}) p_{i,h}^{\dot{s}_{i,h}}(\widetilde{\theta}_{i+1,h}) \|\phi_{i,h}^{s'}\|_{\mathcal{H}_{i,h}^{-1}}^2 \\
&\leq 2\sum_{i=1}^{k} \log\left(1 + \sum_{s' \in \dot{\mathcal{S}}_{i,h}} p_{i,h}^{s'}(\widetilde{\theta}_{i+1,h}) p_{i,h}^{\dot{s}_{i,h}}(\widetilde{\theta}_{i+1,h}) \|\phi_{i,h}^{s'}\|_{\mathcal{H}_{i,h}^{-1}}^2\right) \\
&\leq 2\log\left(\frac{\det(\mathcal{H}_{k+1,h})}{\det(\mathcal{H}_{1,h})}\right) \\
&\leq 2d\log\left(1 + \frac{k}{\lambda d}\right).
\end{aligned}
$$

Proof of statement (III). By Lemma 5, we have

$$
\mathcal{H}_{k+1,h} \succeq \mathcal{H}_{k+1,h} + \kappa \sum_{s' \in \dot{\mathcal{S}}_{k,h}} \phi_{k,h}^{s'} (\phi_{k,h}^{s'})^\top
$$

Since $\lambda \geq 1$, we have $\kappa \max_{s' \in \mathcal{S}_{i,h}} \|\phi_{i,h}^{s'}\|_{\mathcal{H}_{i,h}^{-1}} \leq \kappa$. Using the fact that $z \leq 2\log(1+z)$ for any $z \in [0,1]$. By a similar analysis as the statement (2), we have

$$
\begin{aligned}
\sum_{i=1}^{k} \max_{s' \in \mathcal{S}_{i,h}} \|\phi_{i,h}^{s'}\|_{\mathcal{H}_{i,h}^{-1}}^2 &\leq \frac{2}{\kappa} \sum_{i=1}^{k} \log\left(1 + \max_{s' \in \mathcal{S}_{i,h}} \|\phi_{i,h}^{s'}\|_{\mathcal{H}_{i,h}^{-1}}\right) \\
&\leq \frac{2}{\kappa} \log\left(\frac{\det(\mathcal{H}_{k+1,h})}{\det(\mathcal{H}_{1,h})}\right) \\
&\leq \frac{2}{\kappa} d \log\left(1 + \frac{k}{\lambda d}\right).
\end{aligned}
$$

This finishes the proof.

Proof of statement (IV). By the definition of $H_{k,h}(\theta)$, we have

$$
\begin{aligned}
H_{i,h}(\theta) &= \sum_{s' \in \mathcal{S}_{i,h}} p_{i,h}^{s'}(\theta) \phi_{i,h}^{s'} (\phi_{i,h}^{s'})^\top - \sum_{s' \in \mathcal{S}_{i,h}} \sum_{s'' \in \mathcal{S}_{i,h}} p_{i,h}^{s'}(\theta) p_{i,h}^{s''}(\theta) \phi_{i,h}^{s'} (\phi_{i,h}^{s''})^\top \\
&= \mathbb{E}_{s' \in p_{i,h}(\theta)}[\phi_{i,h}^{s'} (\phi_{i,h}^{s'})^\top] - \mathbb{E}_{s' \in p_{i,h}(\theta)}[\phi_{i,h}^{s'}] \big(\mathbb{E}_{s' \in p_{i,h}(\theta)}[\phi_{i,h}^{s'}]\big)^\top \\
&= \mathbb{E}_{s' \in p_{i,h}(\theta)}\big[(\phi_{i,h}^{s'} - \mathbb{E}_{s'' \in p_{i,h}(\theta)} \phi_{i,h}^{s''})(\phi_{i,h}^{s'} - \mathbb{E}_{s'' \in p_{i,h}(\theta)} \phi_{i,h}^{s''})^\top\big]
\end{aligned}
$$

Thus, we have

$$
H_{i,h}(\widetilde{\theta}_{i+1,h}) \succeq \sum_{s' \in \mathcal{S}_{i,h}} p_{i,h}(\widetilde{\theta}_{i+1,h}) (\widetilde{\phi}_{i,h}^{s'}) (\widetilde{\phi}_{i,h}^{s'})^\top. \tag{13}
$$

Then, we get

$$
\mathcal{H}_{k+1,h} \succeq \mathcal{H}_{k,h} + \sum_{s' \in \mathcal{S}_{k,h}} p_{k,h}(\widetilde{\theta}_{k+1,h}) (\widetilde{\phi}_{k,h}^{s'}) (\widetilde{\phi}_{k,h}^{s'})^\top
$$

The remaining proof is similar to the proof of statement (2).

Proof of statement (V). By (13), we have

$$
H_{i,h}(\widetilde{\theta}_{i+1,h}) \succeq \sum_{s' \in \mathcal{S}_{i,h}} p_{i,h}(\widetilde{\theta}_{i+1,h}) (\widetilde{\phi}_{i,h}^{s'}) (\widetilde{\phi}_{i,h}^{s'})^\top \succeq \kappa \sum_{s' \in \mathcal{S}_{i,h}} (\widetilde{\phi}_{i,h}^{s'}) (\widetilde{\phi}_{i,h}^{s'})^\top.
$$

Then, the remaining proof is similar to the proof of statement (3). ■

B Proof of Lemma 1

B.1 Main Proof

Proof. The main proof follows the proof of Theorem 1 in Zhang et al. [2016]. We define

$$\bar{f}_{k,h}(\theta) = \mathbb{E}_{y_{k,h}}[f_{k,h}(\theta) \mid \mathcal{F}_{k,h}], \quad \bar{g}_{k,h}(\theta) = \mathbb{E}_{y_{k,h}}[g_{k,h}(\theta) \mid \mathcal{F}_{k,h}].$$

By Property 1, we have $f_{k,h}(\theta)$ is exponential concave such that

$$f_{k,h}(\widehat{\theta}_{k,h}) \leq f_{k,h}(\theta_h^*) + \langle g_{k,h}(\theta_h^*), \widehat{\theta}_{k,h} - \theta_h^* \rangle - \frac{\kappa}{2} \|(\widehat{\theta}_{k,h} - \theta_h^*)\|_{W_{k,h}}.$$

Taking expectations on both sides, we have

$$\bar{f}_{k,h}(\widehat{\theta}_{k,h}) \leq \bar{f}_{k,h}(\theta_h^*) + \langle \bar{g}_{k,h}(\theta_h^*), \widehat{\theta}_{k,h} - \theta_h^* \rangle - \frac{\kappa}{2} \|(\widehat{\theta}_{k,h} - \theta_h^*)\|_{W_{k,h}}. \tag{14}$$

Based the property of KL diverge, we ensure the true parameter θ_h^* is the minimizer of the expected loss function by the following lemma.

Lemma 7. *For any $\theta \in \Theta$, we have $\bar{f}_{k,h}(\theta_h^*) \leq \bar{f}_{k,h}(\theta)$.*

Combining Lemma 7 and Equation (14), we have

$$\begin{aligned} 0 &\leq \bar{f}_{k,h}(\widehat{\theta}_{k,h}) - \bar{f}_{k,h}(\theta_h^*) \\ &\leq \langle \bar{g}_{k,h}(\widehat{\theta}_{k,h}), \widehat{\theta}_{k,h} - \theta_h^* \rangle - \frac{\kappa}{2} \|\theta_h^* - \widehat{\theta}_{k,h}\|_{W_{k,h}}^2 \\ &\leq \langle g_{k,h}(\widehat{\theta}_{k,h}), \widehat{\theta}_{k,h} - \theta_h^* \rangle - \frac{\kappa}{2} \|\theta_h^* - \widehat{\theta}_{k,h}\|_{W_{k,h}}^2 + \langle \bar{g}_{k,h}(\widehat{\theta}_{k,h}) - g_{k,h}(\widehat{\theta}_{k,h}), \widehat{\theta}_{k,h} - \theta_h^* \rangle. \end{aligned} \tag{15}$$

By standard analysis of OMD in Lemma 17, it holds

$$2g_{k,h}(\widehat{\theta}_{k,h})(\widehat{\theta}_{k,h} - \theta_h^*) \leq \|g_{k,h}(\widehat{\theta}_{k,h})\|_{\widehat{\Sigma}_{k+1,h}} + \|\widehat{\theta}_{k,h} - \theta_h^*\|_{\widehat{\Sigma}_{k+1,h}} - \|\widehat{\theta}_{k+1,h} - \theta_h^*\|_{\widehat{\Sigma}_{k+1,h}}. \tag{16}$$

Combining Equation (15) and Equation (16), we have

$$\begin{aligned} 0 \leq \frac{1}{2} \|g_{k,h}(\widehat{\theta}_{k,h})\|_{\widehat{\Sigma}_{k+1,h}^{-1}}^2 + \frac{1}{2} \|\widehat{\theta}_{k,h} - \theta_h^*\|_{\widehat{\Sigma}_{k+1,h}}^2 - \frac{1}{2} \|\widehat{\theta}_{k+1,h} - \theta_h^*\|_{\widehat{\Sigma}_{k+1,h}}^2 \\ - \frac{\kappa}{2} \|\theta_h^* - \widehat{\theta}_{k,h}\|_{W_{k,h}}^2 + \langle \bar{g}_{k,h}(\widehat{\theta}_{k,h}) - g_{k,h}(\widehat{\theta}_{k,h}), \widehat{\theta}_{k,h} - \theta_h^* \rangle. \end{aligned} \tag{17}$$

First, we consider the first term . We show the gradient can be bounded by the following lemma.

Lemma 8. *For any positive semi-definite matrix Z, it holds that*

$$\|g_{k,h}(\theta)\|_Z^2 \leq 4 \max_{s' \in \mathcal{S}_{k,h}} \|\phi_{k,h}^{s'}\|_Z^2. \tag{18}$$

Then, we consider the second term. Note that $\widehat{\Sigma}_{k+1,h} = \widehat{\Sigma}_{k,h} + \frac{\kappa}{2} W_{k,h}$, we have

$$\|\widehat{\theta}_{k,h} - \theta^*\|_{\widehat{\Sigma}_{k+1,h}}^2 = \|\widehat{\theta}_{k,h} - \theta^*\|_{\widehat{\Sigma}_{k,h}}^2 + \frac{\kappa}{2} \|\widehat{\theta}_{k,h} - \theta^*\|_{W_{k,h}}^2. \tag{19}$$

Next, we bound $\langle \bar{g}_{k,h}(\widehat{\theta}_{k,h}) - g_{k,h}(\widehat{\theta}_{k,h}), \widehat{\theta}_{k,h} - \theta_h^* \rangle$, which is a martingale difference sequence.

Lemma 9. *For any $\delta \in (0, 1)$ and $\theta_1, \ldots, \theta_k, \theta^* \in [0, B]^d$, with probability at least $1 - \delta$, for any $k \in [K], h \in [H]$ it holds that*

$$\sum_{i=1}^{k} \langle \bar{g}_{i,h}(\theta_i) - g_{i,h}(\theta_i), \theta_i - \theta_h^* \rangle \leq \frac{\kappa}{4} \sum_{i=1}^{k} \|\theta_i - \theta^*\|_{W_{i,h}}^2 + (\frac{4}{\kappa} + 8B)\gamma_k. \tag{20}$$

where $\gamma_k = \log \frac{2k^2 H \log(kU)}{\delta}$.

Combining (17), (18), and (19), we have

$$\begin{aligned}\|\widehat{\theta}_{k+1,h} - \theta_h^*\|_{\widehat{\Sigma}_{k+1,h}}^2 &\le \|\widehat{\theta}_{k,h} - \theta_h^*\|_{\widehat{\Sigma}_{k+1,h}}^2 - \kappa\|\theta_h^* - \widehat{\theta}_{k,h}\|_{W_{k,h}}^2 + \|g_{k,h}(\widehat{\theta}_{k,h})\|_{\widehat{\Sigma}_{k+1,h}^{-1}}^2 \\ &\quad + 2\langle \bar{g}_{k,h}(\widehat{\theta}_{k,h}) - g_{k,h}(\widehat{\theta}_{k,h}), \widehat{\theta}_{k,h} - \theta_h^*\rangle \\ &\le \|\widehat{\theta}_{k,h} - \theta_h^*\|_{\widehat{\Sigma}_{k,h}}^2 + 4\max_{s'\in\mathcal{S}_{k,h}} \|\phi_{k,h}^{s'}\|_{\widehat{\Sigma}_{k+1,h}^{-1}}^2 + (\frac{8}{\kappa} + 16B)\gamma_k \qquad (21)\end{aligned}$$

where the first inequality holds by rearranging the terms in (17), the second holds by (18), (19) and (20). Summing (21) from $i = 1$ to k, we have

$$\begin{aligned}\|\widehat{\theta}_{k+1,h} - \theta_h^*\|_{\widehat{\Sigma}_{k+1,h}}^2 &\le \lambda B^2 + 4\sum_{i=1}^{k} \max_{s'\in\mathcal{S}_{i,h}} \|\phi_{i,h}^{s'}\|_{\widehat{\Sigma}_{i+1,h}^{-1}}^2 + (\frac{8}{\kappa} + 16B)\gamma_k \\ &\le \lambda B^2 + \frac{8}{\kappa} d\log\left(1 + \frac{kU}{\lambda d}\right) + (\frac{8}{\kappa} + 16B)\gamma_k,\end{aligned}$$

where the second inequality holds by $\|\phi_{i,h}^{s'}\|_2 \le 1$ and $\widehat{\Sigma}_{i,h} \ge I, \forall i \in [K]$, and the last inequality is by Lemma 6. This finishes the proof. ■

B.2 Proof of Auxiliary Lemmas

In this section, we provide the proofs of the lemmas used in Appendix B.1.

B.2.1 Proof of Lemma 7

Proof. By the definition of $\bar{f}_{k,h}(\theta)$, we have

$$\bar{f}_{k,h}(\theta) - \bar{f}_{k,h}(\theta_h^*) = \sum_{s'\in\mathcal{S}_{k,h}} p_{k,h}^{s'}(\theta_h^*) \log \frac{p_{k,h}^{s'}(\theta_h^*)}{p_{k,h}^{s'}(\theta)} \ge 0,$$

where the last inequality is due to $\sum_{s'\in\mathcal{S}_{k,h}} p_{k,h}^{s'}(\theta_h^*) \log \frac{p_{k,h}^{s'}(\theta_h^*)}{p_{k,h}^{s'}(\theta)}$ is the Kullback-Leibler divergence between $p_{k,h}(\theta_h^*)$ and $p_{k,h}(\theta)$, which always is non-negative. ■

B.2.2 Proof of Lemma 8

Proof. For any positive semi-definite matrix Z,

$$(x_i - x_j)^\top Z(x_i - x_j) = x_i^\top Z x_i - x_i^\top Z x_j - x_j^\top Z x_i + x_j^\top Z x_j \ge 0, \quad \forall x_i, x_j \in \mathbb{R}^d. \qquad (22)$$

which implies $x_i^\top Z x_i + x_j^\top Z x_j \ge x_i^\top Z x_j + x_j^\top Z x_i, \quad \forall x_i, x_j \in \mathbb{R}^d$.

Let $x_i = (p_{k,h}^{s'}(\theta) - y_{k,h}^{s'})\phi_{k,h}^{s'}$, we have

$$\begin{aligned}&\|g_{k,h}(\theta)\|_Z^2 \\ &= \sum_{s'\in\mathcal{S}_{k,h}} \sum_{s''\in\mathcal{S}_{k,h}} (p_{k,h}^{s'}(\theta) - y_{k,h}^{s'})(p_{k,h}^{s''}(\theta) - y_{k,h}^{s''})\phi_{k,h}^{s'} Z \phi_{k,h}^{s''} \\ &= \sum_{s'\in\mathcal{S}_{k,h}} (p_{k,h}^{s'}(\theta) - y_{k,h}^{s'})^2 \phi_{k,h}^{s'} Z \phi_{k,h}^{s'} \\ &\qquad + \frac{1}{2}\sum_{s'\in\mathcal{S}_{k,h}} \sum_{s''\in\mathcal{S}_{k,h}} (p_{k,h}^{s'}(\theta) - y_{k,h}^{s'})(p_{k,h}^{s''}(\theta) - y_{k,h}^{s''})(\phi_{k,h}^{s'} Z \phi_{k,h}^{s''} + \phi_{k,h}^{s''} Z \phi_{k,h}^{s'}) \\ &\le \sum_{s'\in\mathcal{S}_{k,h}} (p_{k,h}^{s'}(\theta) - y_{k,h}^{s'})^2 \phi_{k,h}^{s'} Z \phi_{k,h}^{s'} + \sum_{s'\in\mathcal{S}_{k,h}} (p_{k,h}^{s'}(\theta) - y_{k,h}^{s'})^2 \phi_{k,h}^{s'} Z \phi_{k,h}^{s'} \\ &= 2\sum_{s'\in\mathcal{S}_{k,h}} (p_{k,h}^{s'}(\theta) - y_{k,h}^{s'})^2 \phi_{k,h}^{s'} Z \phi_{k,h}^{s'} \\ &\le 4\max_{s'\in\mathcal{S}_{k,h}} \|\phi_{k,h}^{s'}\|_Z^2.\end{aligned}$$

This finishes the proof. ■

B.2.3 Proof of Lemma 9

Proof. First, notice that $\left(\bar{g}_{i,h}(\theta_i) - g_{i,h}(\theta_i)\right)^\top (\theta_i - \theta^*)$ is a martingale difference sequence. Also, we have

$$
\begin{aligned}
&\left|\left(\bar{g}_{i,h}(\theta_i) - g_{i,h}(\theta_i)\right)^\top (\theta_i - \theta^*)\right| \\
&\leq \left|\left(\bar{g}_{i,h}(\theta_i)\right)^\top (\theta_i - \theta^*)\right| + \left|\left(g_{i,h}(\theta_i)\right)^\top (\theta_i - \theta^*)\right| \\
&\leq \|\bar{g}_{i,h}(\theta_i)\|_2 \|\theta_i - \theta^*\|_2 + \|g_{i,h}(\theta_i)\|_2 \|\theta_i - \theta^*\|_2 \\
&\leq 4\sqrt{2}B,
\end{aligned}
$$

where the last inequality holds by $\|\bar{g}_{i,h}(\theta_i)\|_2 = \|\sum_{s' \in \mathcal{S}_{i,h}} (p_{i,h}^{s'}(\theta) - y_{i,h}^{s'})\phi_{i,h}^{s'}\|_2 \leq \sqrt{2}$.

We define the martingale $M_{k,h} = \sum_{i=1}^k \left(\bar{g}_{i,h}(\theta_i) - g_{i,h}(\theta_i)\right)^\top (\theta_i - \theta^*)$, and define the conditional variance σ_i^2 as

$$
\begin{aligned}
\sigma_{k,h}^2 &= \mathbb{E}[M_{k,h}^2 \mid \mathcal{F}_{i-1}] \\
&= \sum_{i=1}^k \mathbb{E}_{y_{k,h}} \left[\left(\bar{g}_{i,h}(\theta_i) - g_{i,h}(\theta_i)\right)^\top (\theta_i - \theta^*)^2\right] \\
&\leq \sum_{i=1}^k \mathbb{E}_{y_{k,h}} \left[\left(g_{i,h}(\theta_i)\right)^\top (\theta_i - \theta^*)^2\right] \\
&\leq \sum_{i=1}^k \sum_{s' \in \mathcal{S}_{i,h}} (\phi_{i,h}^{s'})^\top (\theta_i - \theta^*)^2 \\
&= \sum_{i=1}^k \|\theta_i - \theta^*\|_{W_{i,h}}^2 \\
&\triangleq A_{k,h}.
\end{aligned}
$$

where the first inequality is due to the fact that $\mathbb{E}[(\xi - \mathbb{E}[\xi])^2] \leq \mathbb{E}[\xi^2]$ for any random variable ξ. Note that $A_{k,h}$ is a random variable, so we cannot directly apply the Bernstein inequality to $M_{k,h}$. Instead, we consider the following two cases:(i) $A_{k,h} \leq \frac{4B^2}{kU}$ and (ii) $A_{k,h} > \frac{4B^2}{kU}$.

Case (i): $A_{k,h} = \sum_{i=1}^k \|\theta_i - \theta^*\|_{W_{i,h}}^2 \leq \frac{4B^2}{kU}$. Then, we have

$$
\begin{aligned}
M_{k,h} &= \sum_{i=1}^k \left(\bar{g}_{i,h}(\theta_i) - g_{i,h}(\theta_i)\right)^\top (\theta_i - \theta^*) \\
&= \sum_{i=1}^k \sum_{s' \in \mathcal{S}_{i,h}} (p_{i,h}^{s'}(\theta_i) - y_{i,h}^{s'})(\phi_{i,h}^{s'})^\top (\theta_i - \theta^*) \\
&\leq \sum_{i=1}^k \sum_{s' \in \mathcal{S}_{i,h}} |(\phi_{i,h}^{s'})^\top (\theta_i - \theta^*)| \\
&\leq \sqrt{kU \sum_{i=1}^k \sum_{s' \in \mathcal{S}_{i,h}} ((\phi_{i,h}^{s'})^\top (\theta_i - \theta^*))^2} \\
&\leq 2B.
\end{aligned}
$$

where the second equality is due to the definition of $\bar{g}_{i,h}(\theta_i)$ and $g_{i,h}(\theta_i)$, the first inequality holds by $p_{i,h}^{s'}(\theta_i) - y_{i,h}^{s'} \in [-1, 1]$, the second inequality holds by the Cauchy-Schwarz inequality, and the last inequality holds by the condition of case (i).

Case (ii): $A_{k,h} = \sum_{i=1}^{k} \|\theta_i - \theta^*\|_{W_{j,h}}^2 > \frac{4B^2}{kU}$. We have both a lower and upper bound for $A_{k,h}$, i.e., $\frac{4B^2}{kU} < A_{k,h} \le 4B^2 kU$. Then, we can use the peeling process to bound $M_{k,h}$ as follows:

$$
\begin{aligned}
&\Pr\left[M_{k,h} \ge 2\sqrt{\tau_{k,h} A_{k,h}} + \frac{8B\tau_{k,h}}{3}\right] \\
=\ &\Pr\left[M_{k,h} \ge 2\sqrt{\tau_{k,h} A_{k,h}} + \frac{8B\tau_{k,h}}{3}, \frac{4B^2}{kU} < A_{k,h} \le 4kUB^2\right] \\
=\ &\Pr\left[M_{k,h} \ge 2\sqrt{\tau_{k,h} A_{k,h}} + \frac{8B\tau_{k,h}}{3}, \frac{4B^2}{kU} < A_{k,h} \le 4kUB^2, \sigma_{k,h} \le A_{k,h}\right] \\
\le &\sum_{i=1}^{m} \Pr\left[M_{k,h} \ge 2\sqrt{\tau_{k,h} A_{k,h}} + \frac{8B\tau_{k,h}}{3}, \frac{4B^2 2^{i-1}}{kU} < A_{k,h} \le \frac{4B^2 2^i}{kU}, \sigma_{k,h} \le A_{k,h}\right] \\
\le &\sum_{i=1}^{m} \Pr\left[M_{k,h} \ge \sqrt{\frac{8B^2 2^i}{3kU}\tau_{k,h}} + \frac{8B\tau_{k,h}}{3}, \sigma_{k,h} \le \frac{4B^2 2^i}{kU}\right] \\
\le &\, m\exp(-\tau_{k,h}).
\end{aligned}
$$

where $m = 2\log_2(kU)$, and the last inequality follows the Bernstein inequality for martingales.

Combining above two cases, letting $\tau = \log \frac{mk^2}{\delta/H}$ and taking the union bound over k and $h \in [H]$, we have with probability at least $1-\delta$, for any $k \in [K], h \in [H]$ it holds that

$$
\begin{aligned}
M_{k,h} &= \sum_{i=1}^{k} \left(\bar{g}_{i,h}(\theta_i) - g_{i,h}(\theta_i)\right)^\top (\theta_i - \theta^*) \\
&\le 2\sqrt{\tau_{k,h} A_{k,h}} + \frac{8B\tau_{k,h}}{3} + 4\sqrt{2}B \\
&\le 2\sqrt{\sum_{i=1}^{k} \|\theta_i - \theta^*\|_{W_{i,h}}^2 \log \frac{2k^2 H \log(kU)}{\delta}} + 8B\left(1 + \log \frac{2k^2 H \log(kU)}{\delta}\right) \\
&= 2\sqrt{\gamma_k \sum_{i=1}^{k} \|\theta_i - \theta^*\|_{W_{i,h}}^2} + 8B\left(1 + \gamma_k\right),
\end{aligned} \tag{23}
$$

where $\gamma_k = \log \frac{2k^2 H \log(kU)}{\delta}$.

Then, applying $uv \le cu^2 + v^2/(4c)$ for any $c, u, v > 0$ with $c = 2/\kappa$, we have

$$
\sqrt{\gamma_k \sum_{i=1}^{k} \|\theta_i - \theta^*\|_{W_{i,h}}^2} \le \frac{2\gamma_k}{\kappa} + \frac{\kappa}{8}\sum_{i=1}^{k} \|\theta_i - \theta^*\|_{W_{i,h}}^2. \tag{24}
$$

Combining (23) and (24), we have

$$
\sum_{i=1}^{k} \left(\bar{g}_{i,h}(\theta_i) - g_{i,h}(\theta_i)\right)^\top (\theta_i - \theta^*) \le \frac{\kappa}{4}\sum_{i=1}^{k} \|\theta_i - \theta^*\|_{W_{i,h}}^2 + (\frac{4}{\kappa} + 8B)\gamma_k.
$$

This finishes the proof. ■

C Proof of Lemma 2

Proof. The gradient of $p_{s,a}(\theta)$ is given by

$$
\nabla p_{s,a}^{s'}(\theta) = p_{s,a}^{s'}(\theta)\phi_{s,a}^{s'} - p_{s,a}^{s'}(\theta) \sum_{s'' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s''}(\theta)\phi_{s,a}^{s''}.
$$

By the mean value theorem, there exists $\bar{\theta} = \nu\theta_h^* + (1-\nu)\widehat{\theta}_{k,h}$ for some $\nu \in [0,1]$, such that

$$
\begin{aligned}
&\bigg|\sum_{s' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\widehat{\theta}_{k,h})V(s') - \sum_{s' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\theta_h^*)V(s')\bigg| \\
&= \bigg|\sum_{s' \in \mathcal{S}_{h,s,a}} \nabla p_{s,a}^{s'}(\bar{\theta})(\widehat{\theta}_{k,h} - \theta_h^*)V(s')\bigg| \\
&= \bigg|\sum_{s' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\bar{\theta})\phi_{s,a}^{s'}(\widehat{\theta}_{k,h} - \theta_h^*)V(s') - \sum_{s' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\bar{\theta}) \sum_{s'' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s''}(\bar{\theta})\phi_{s,a}^{s''}(\widehat{\theta}_{k,h} - \theta_h^*)V(s')\bigg| \\
&= \bigg|\sum_{s' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\bar{\theta})\phi_{s,a}^{s'}(\widehat{\theta}_{k,h} - \theta_h^*)\Big(V(s') - \sum_{s'' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s''}(\bar{\theta})V(s'')\Big)\bigg| \\
&\le H \sum_{s' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\bar{\theta})|\phi_{s,a}^{s'}(\widehat{\theta}_{k,h} - \theta_h^*)| \\
&\le H \max_{s' \in \mathcal{S}_{h,s,a}} \phi_{s,a}^{s'}(\widehat{\theta}_{k,h} - \theta_h^*) \qquad (25) \\
&\le H \max_{s' \in \mathcal{S}_{h,s,a}} \|\phi_{s,a}^{s'}\|_{\widehat{\Sigma}_{k,h}^{-1}} \|\widehat{\theta}_{k,h} - \theta_h^*\|_{\widehat{\Sigma}_{k,h}} \\
&\le H\beta_k \max_{s' \in \mathcal{S}_{h,s,a}} \|\phi_{s,a}^{s'}\|_{\widehat{\Sigma}_{k,h}^{-1}},
\end{aligned}
$$

where the first inequality is by $V : \mathcal{S} \to [0, H]$, the second inequality is by the definition of $\phi_{k,h}^{s'}$, the third inequality is by the Holder's inequality, the last inequality is by Lemma 1. ■

D Proof of Theorem 1

D.1 Main Proof

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

Lemma 10. *Suppose for all $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$, it holds that*

$$
\bigg|\sum_{s' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\widehat{\theta}_{k,h})V(s') - \sum_{s' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\theta_h^*)V(s')\bigg| \le \Gamma_{h,s,a}. \qquad (26)
$$

Define

$$
\widehat{Q}_{k,h}(s,a) = \bigg[r_h(s,a) + \sum_{s' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\widehat{\theta}_{k,h})\widehat{V}_{k,h+1}(s') + \Gamma_{h,s,a}\bigg]_{[0,H]}. \qquad (27)
$$

Then, for any $\delta \in (0, 1]$, with probability at least $1 - \delta$, it holds that

$$
\mathrm{Reg}(K) \le 2\sum_{k=1}^{K}\sum_{h=1}^{H} \Gamma_{h,s_{k,h},a_{k,h}} + H\sqrt{2KH\log(2/\delta)}.
$$

Proof of Theorem 1. Substituting $\Gamma_{h,s,a} = H\widehat{\beta}_k \max_{s' \in \mathcal{S}_{h,s,a}} \|\phi_{s,a}^{s'}\|_{\widehat{\Sigma}_{k,h}^{-1}}$ into Lemma 10, we have

$$
\begin{aligned}
\mathrm{Reg}(K) &\le 2H\widehat{\beta}_k \sum_{k=1}^{K}\sum_{h=1}^{H} \max_{s' \in \mathcal{S}_{k,h}} \|\phi_{k,h}^{s'}\|_{\widehat{\Sigma}_{k,h}^{-1}} + H\sqrt{2KH\log(2/\delta)} \\
&\le 2H\widehat{\beta}_k \sum_{h=1}^{H} \sqrt{K \sum_{k=1}^{K} \max_{s' \in \mathcal{S}_{k,h}} \|\phi_{k,h}^{s'}\|_{\widehat{\Sigma}_{k,h}^{-1}}^2} + H\sqrt{2KH\log(2/\delta)} \\
&\le 2H^2\widehat{\beta}_k \sqrt{\frac{4dK}{\kappa}\log\left(1 + \frac{KU}{d\lambda}\right)} + H\sqrt{2KH\log(2/\delta)} \le \widetilde{\mathcal{O}}(\kappa^{-1}dH^2\sqrt{K}),
\end{aligned}
$$

where the second inequality holds by the Cauchy-Schwarz inequality and the third inequality holds by Lemma 6. This finishes the proof. ■

D.2 Proof of Auxiliary Lemmas

In this section, we provide the proofs of the lemmas used in Appendix D.1.

First, we introduce the following lemma.

Lemma 11. *Suppose* (26) *and* (27) *in Lemma 10 holds for all* $k \in [K], h \in [H]$. *Then, for any* $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, *it holds that*

$$Q_h^*(s,a) \le \widehat{Q}_{k,h}(s,a) \le r_h(s,a) + \mathbb{P}_h V_{k,h+1}(s,a) + 2\Gamma_{h,s,a}.$$

Proof. First, we prove the left-hand side of the lemma. We prove this by backward induction on h. For the stage $h = H$, by definition, we have

$$\widehat{Q}_{k,H}(s,a) = r_H(s,a) = Q_H^*(s,a), \quad \widehat{V}_{k,H+1}(s) = 0 = V_{H+1}^*(s).$$

Suppose the statement holds for $h+1$, we show it holds for h. By definition, if $\widehat{Q}_{k,h}(s,a) = H$, this holds trivially. Otherwise, we have

$$\begin{aligned}\widehat{Q}_{k,h}(s,a) &= r_h(s,a) + p_{s,a}(\widehat{\theta}_{k,h})\widehat{V}_{k,h+1} + \Gamma_{h,s,a} \\ &\ge r_h(s,a) + p_{s,a}(\widehat{\theta}_{k,h})V_k^* + \Gamma_{h,s,a} \ge r_h(s,a) + p_{s,a}(\theta_h^*)V_k^* = Q_h^*(s,a).\end{aligned}$$

where the first inequality is by the induction hypothesis, and the second inequality is by (26).

Then, we prove the right-hand side of the lemma. By the definition of $\widehat{Q}_{k,h}(s,a)$, we have

$$\widehat{Q}_{k,h}(s,a) = r_h(s,a) + p_{s,a}(\widehat{\theta}_{k,h})\widehat{V}_{k,h+1} + \Gamma_{h,s,a} \le r_h(s,a) + p_{s,a}(\theta_h^*)\widehat{V}_{k,h+1} + 2\Gamma_{h,s,a},$$

where the inequality is by (26). ■

D.2.1 Proof of Lemma 10

Proof. By the definition of $\mathrm{Reg}(K) = \sum_{k=1}^K V_1^*(s_{k,1}) - \sum_{k=1}^K V_1^{\pi_k}(s_{k,1})$, we have

$$\begin{aligned}\sum_{k=1}^K V_1^*(s_{k,1}) - \sum_{k=1}^K V_1^{\pi_k}(s_{k,1}) &= \sum_{k=1}^K Q_1^*(s_{k,1}, \pi^*(s_{k,1})) - \sum_{k=1}^K V_1^{\pi_k}(s_{k,1}) \\ &\le \sum_{k=1}^K \widehat{Q}_1(s_{k,1}, \pi^*(s_{k,1})) - \sum_{k=1}^K V_1^{\pi_k}(s_{k,1}) \\ &\le \sum_{k=1}^K \widehat{Q}_1(s_{k,1}, a_{k,1}) - \sum_{k=1}^K V_1^{\pi_k}(s_{k,1}),\end{aligned}$$

where the first inequality is by Lemma 11 and the second is by $a_{k,h} = \arg\max_{a \in \mathcal{A}} \widehat{Q}_{k,h}(s_{k,h}, a)$.

By the right-hand side of Lemma 11, we have

$$\begin{aligned}&\widehat{Q}_1(s_{k,1}, a_{k,1}) - V_1^{\pi_k}(s_{k,1}) \\ &= r(s_{k,1}, a_{k,1}) + \mathbb{P}_1 \widehat{V}_{k,2}(s_{k,1}, a_{k,1}) + +2\Gamma_{h,s_{k,1},a_{k,1}} - r(s_{k,1}, a_{k,1}) - \mathbb{P}_1 V_2^{\pi_k}(s_{k,1}, a_{k,1}) \\ &\le \mathbb{P}_1(\widehat{V}_{k,2} - V_2^{\pi_k})(s_{k,1}, a_{k,1}) - (\widehat{V}_{k,2} - V_2^{\pi_k})(s_{k,2}) + (\widehat{V}_{k,2} - V_2^{\pi_k})(s_{k,2}) + 2\Gamma_{h,s_{k,1},a_{k,1}} \\ &\le \mathbb{P}_1(\widehat{V}_{k,2} - V_2^{\pi_k})(s_{k,1}, a_{k,1}) - (\widehat{V}_{k,2} - V_2^{\pi_k})(s_{k,2}) + \Big(\widehat{Q}_2(s_{k,2}, a_{k,2}) - V_2^{\pi_k}(s_{k,2})\Big) + 2\Gamma_{h,s_{k,1},a_{k,1}}.\end{aligned}$$

Define $\mathcal{M}_{k,h} = \mathbb{P}_h(\widehat{V}_{k,h+1} - V_{h+1}^{\pi_k})(s_{k,h}, a_{k,h}) - (\widehat{V}_{k,h+1} - V_{h+1}^{\pi_k})(s_{k,h+1})$. Applying this recursively, we have

$$\widehat{Q}_1(s_{k,1}, a_{k,1}) - V_1^{\pi_k}(s_{k,1}) \le 2\sum_{h=1}^H \Gamma_{h,s_{k,h},a_{k,h}} + \sum_{h=1}^H \mathcal{M}_{k,h}$$

Summing over k, we have for any $\delta \in (0,1]$, with probability at least $1-\delta$, it holds that

$$\mathrm{Reg}(K) \le 2\sum_{k=1}^K \sum_{h=1}^H \Gamma_{h,s_{k,h},a_{k,h}} + \sum_{k=1}^K \sum_{h=1}^H \mathcal{M}_{k,h} \le 2\sum_{k=1}^K \sum_{h=1}^H \Gamma_{h,s_{k,h},a_{k,h}} + H\sqrt{2KH\log(2/\delta)}$$

where the inequality holds by the Azuma-Hoeffding inequality as $\mathcal{M}_{k,h}$ is a martingale difference sequence with $\mathcal{M}_{k,h} \le 2H$. This finishes the proof. ■

E Proof of Lemma 3

Proof. The proof is similar to the proof of Theorem 3 in Zhang and Sugiyama [2023] and Lemma 1 of Lee and Oh [2024]. Define

$$\widetilde{f}_{k,h}(\theta) = f_{k,h}(\theta_{k,h}) + \langle g_{k,h}(\widetilde{\theta}_{k,h}), \theta - \widetilde{\theta}_{k,h}\rangle + \frac{1}{2}\|\theta - \widetilde{\theta}_{k,h}\|_{H_{k,h}(\widetilde{\theta}_{k,h})},$$

which is a second order approximation of the original function $f_{k,h}(\theta)$ at $\widetilde{\theta}_{k,h}$. Then, the update rule in (9) can be rewritten as

$$\widetilde{\theta}_{k+1,h} = \arg\min_{\theta\in\Theta} \widetilde{f}_{k,h}(\theta) + \frac{1}{2\eta}\|\theta - \widetilde{\theta}_{k,h}\|_{\mathcal{H}_{k,h}}. \tag{28}$$

Then, we present the following lemma that bounds the estimation error of the update rule in (28).

Lemma 12 (Lemma E.1 of Lee and Oh [2024]). *For the update rule defined in* (28)*, it holds that*

$$\begin{aligned}\|\widetilde{\theta}_{k+1,h} - \theta^*_h\|^2_{\mathcal{H}_{k+1,h}} \le & 2\eta\left(\sum_{i=1}^k f_{i,h}(\theta^*_h) - \sum_{i=1}^k f_{i,h}(\widetilde{\theta}_{i+1,h})\right) + 4\lambda B \\ & + 12\sqrt{2}B\eta\sum_{i=1}^k \|\widetilde{\theta}_{i+1,h} - \widetilde{\theta}_{i,h}\|_2^2 - \sum_{i=1}^k\|\widetilde{\theta}_{i+1,h} - \widetilde{\theta}_{i,h}\|^2_{\mathcal{H}_{i,h}}.\end{aligned}$$

We bound the right-hand side of the above lemma separately in the following.

First, we bound the first term. Define the softmax function as follows.

$$[\sigma_{k,h}(z)]_s = \frac{\exp([z])_s}{1 + \sum_{s\in\dot{\mathcal{S}}_{k,h}}\exp([z])_s}, \quad \forall s \in \mathcal{S}_{k,h}.$$

Then, the loss function in (2) can be rewritten as

$$f_{k,h}(z_{k,h}, y_{k,h}) = \sum_{s'\in\mathcal{S}_{k,h}} \mathbb{1}[y^{s'}_{k,h} = 1]\log\left(\frac{1}{[\sigma_{k,h}(z_{k,h})]_{s'}}\right).$$

Define a pseudo-inverse function of $\sigma_{k,h}(\cdot)$ as

$$[\sigma^{-1}_{k,h}(p)]_{s'} = \log\left(\frac{[p]_{s'}}{1 - \|p\|_1}\right), \forall p \in \{p \in [0,1]^{N_{k,h}} \mid \|p\|_1 < 1\}.$$

Then, we decompose the first term as follows.

$$\begin{aligned}&\sum_{i=1}^k f_{i,h}(\theta^*_h) - \sum_{i=1}^k f_{i,h}(\widetilde{\theta}_{i+1,h}) \\ =& \underbrace{\sum_{i=1}^k f_{i,h}(\theta^*_h) - \sum_{i=1}^k f_{i,h}(z_{i,h}, y_{i,h})}_{(a)} + \underbrace{\sum_{i=1}^k f_{i,h}(z_{i,h}, y_{i,h}) - \sum_{i=1}^k f_{i,h}(\widetilde{\theta}_{i+1,h})}_{(b)}\end{aligned}$$

where $z_{k,h} = \sigma^{-1}_{k,h}(\mathbb{E}_{\theta\sim P_{k,h}}[\sigma_{k,h}((\phi^{s'}_{k,h})^\top\theta)_{s'\in\dot{\mathcal{S}}_{k,h}}])$, $P_{k,h} \triangleq \mathcal{N}(\widetilde{\theta}_{k,h}, (1 + c\mathcal{H}^{-1}_{k,h}))$ is the Gaussian distribution with mean $\widetilde{\theta}_{k,h}$ and covariance $(1 + c\mathcal{H}^{-1}_{k,h})$ where c is a constant to be specified later.

First, we bound the term (a) as follows.

Lemma 13 (Lemma E.2 of Lee and Oh [2024]). *Let $\delta \in (0,1]$ and $\lambda \ge 1$, for all $k \in [K], h \in [H]$, with probability at least $1-\delta$, we have*

$$\begin{aligned}&\sum_{i=1}^k f_{i,h}(\theta^*_h) - \sum_{i=1}^k f_{i,h}(z_{i,h}, y_{i,h}) \\ \le& (3\log(1 + (U+1)k) + 3)\left(\frac{17}{16}\lambda + 2\sqrt{\lambda}\log\left(\frac{2H\sqrt{1+2k}}{\delta}\right) + 16\left(\log\left(\frac{2H\sqrt{1+2k}}{\delta}\right)\right)^2\right) + 2.\end{aligned}$$

Then, we bound the term (b) as follows.

Lemma 14 (Lemma E.3 of Lee and Oh [2024])**.** *For any $c \geq 0$, let $\lambda \geq \max\{2, 72cd\}$, then for all $k \in [K], h \in [H]$, we have*

$$\sum_{i=1}^{k} f_{i,h}(z_{i,h}, y_{i,h}) - \sum_{i=1}^{k} f_{i,h}(\widetilde{\theta}_{i+1,h}) \leq \frac{1}{2c} \sum_{i=1}^{k} \left\| \widetilde{\theta}_{i,h} - \theta_{i+1,h} \right\|_{\mathcal{H}_{i,h}}^{2} + \sqrt{6}cd \log\left(1 + \frac{k+1}{2\lambda}\right).$$

Now, we are ready to prove Lemma 3. Combining Lemma 12, Lemma 13, and Lemma 14, we have

$$\begin{aligned}
&\|\widetilde{\theta}_{k+1,h} - \theta_h^*\|_{\mathcal{H}_{k+1,h}}^2 \\
&\leq 2\eta \Bigg[(3\log(1 + (U+1)k) + 3) \left(\frac{17}{16}\lambda + 2\sqrt{\lambda} \log\left(\frac{2H\sqrt{1+2k}}{\delta} \right) + 16 \left(\log\left(\frac{2H\sqrt{1+2k}}{\delta} \right) \right)^2 \right) \\
&+2 + \sqrt{6}cd \log\left(1 + \frac{k+1}{2\lambda}\right) \Bigg] + 4\lambda B + 12\sqrt{2}B\eta \sum_{i=1}^{k} \|\widetilde{\theta}_{i+1,h} - \widetilde{\theta}_{i,h}\|_2^2 + (\frac{\eta}{c} - 1) \sum_{i=1}^{k} \|\widetilde{\theta}_{i+1,h} + \widetilde{\theta}_{i,h}\|_{\mathcal{H}_{i,h}}^2 \\
&\leq 2\eta \Bigg[(3\log(1 + (U+1)k) + 3) \left(\frac{17}{16}\lambda + 2\sqrt{\lambda} \log\left(\frac{2H\sqrt{1+2k}}{\delta} \right) + 16 \left(\log\left(\frac{2H\sqrt{1+2k}}{\delta} \right) \right)^2 \right) \\
&\quad + 2 + \sqrt{6}cd \log\left(1 + \frac{k+1}{2\lambda}\right) \Bigg] + 4\lambda B,
\end{aligned}$$

where the second inequality holds by setting $c = 7\eta/6$ and $\lambda \geq \max\{84\sqrt{2}\eta B, 84d\eta\}$, we have

$$\begin{aligned}
&12\sqrt{2}B\eta \sum_{i=1}^{k} \|\widetilde{\theta}_{i+1,h} - \widetilde{\theta}_{i,h}\|_2^2 + \left(\frac{\eta}{c} - 1\right) \sum_{i=1}^{k} \|\widetilde{\theta}_{i+1,h} + \widetilde{\theta}_{i,h}\|_{\mathcal{H}_{i,h}}^2 \\
&= 12\sqrt{2}B\eta \sum_{i=1}^{k} \|\widetilde{\theta}_{i+1,h} - \widetilde{\theta}_{i,h}\|_2^2 - \frac{1}{7} \sum_{i=1}^{k} \|\widetilde{\theta}_{i+1,h} + \widetilde{\theta}_{i,h}\|_{\mathcal{H}_{i,h}}^2 \\
&\leq \left(12\sqrt{2}B\eta - \frac{\lambda}{7}\right) \sum_{i=1}^{k} \|\widetilde{\theta}_{i+1,h} - \widetilde{\theta}_{i,h}\|_2^2 \\
&\leq 0.
\end{aligned}$$

Thus, by setting $\eta = \frac{1}{2}\log(U+1) + (B+1)$, $\lambda = 84\sqrt{2}\eta(B+d)$, we have

$$\|\widetilde{\theta}_{k+1,h} - \theta_h^*\|_{\mathcal{H}_{k+1,h}} \leq \mathcal{O}\big(\sqrt{d} \log U \log(KH/\delta)\big) \triangleq \widetilde{\beta}_k.$$

This finishes the proof. ■

F Proof of Lemma 4

Proof. By the second-order Taylor expansion, there exists $\bar{\theta} = \nu\theta_h^* + (1-\nu)\widehat{\theta}_{k,h}$ for some $\nu \in [0,1]$, such that

$$\begin{aligned}
&\sum_{s' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\widehat{\theta}_{k,h}) V(s') - \sum_{s' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\theta_h^*) V(s') \\
&= \sum_{s' \in \mathcal{S}_{h,s,a}} \nabla p_{s,a}^{s'}(\widetilde{\theta}_{k,h})^\top (\theta_h^* - \widetilde{\theta}_{k,h}) V(s') + \frac{1}{2} \sum_{s' \in \mathcal{S}_{h,s,a}} (\widehat{\theta}_{k,h} - \theta_h^*)^\top \nabla^2 p_{s,a}^{s'}(\bar{\theta}) (\widehat{\theta}_{k,h} - \theta_h^*) V(s')
\end{aligned}$$

The gradient of $p_{s,a}(\theta)$ is given by

$$\nabla p_{s,a}^{s'}(\theta) = p_{s,a}^{s'}(\theta)\phi_{s,a}^{s'} - p_{s,a}^{s'}(\theta) \sum_{s'' \in \mathcal{S}_{h,s,a}} p_{s,a}^{s''}(\theta)\phi_{s,a}^{s''}.$$

For the first-order term, we have

$$
\begin{aligned}
&\sum_{s'\in\mathcal{S}_{h,s,a}} \nabla p_{s,a}^{s'}(\widetilde{\theta}_{k,h})^\top(\theta_h^* - \widetilde{\theta}_{k,h})V(s') \\
&= \sum_{s'\in\mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\widetilde{\theta}_{k,h})\phi_{s,a}^{s'}(\theta_h^* - \widetilde{\theta}_{k,h})V(s') - \sum_{s'\in\mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\widetilde{\theta}_{k,h}) \sum_{s''\in\mathcal{S}_{h,s,a}} p_{s,a}^{s''}(\widetilde{\theta}_{k,h})\phi_{s,a}^{s''}(\theta_h^* - \widetilde{\theta}_{k,h})V(s') \\
&\le H \sum_{s'\in\mathcal{S}_{h,s,a}^+} p_{s,a}^{s'}(\widetilde{\theta}_{k,h}) \left(\phi_{s,a}^{s'}(\theta_h^* - \widetilde{\theta}_{k,h}) - \sum_{s''\in\mathcal{S}_{h,s,a}} p_{s,a}^{s''}(\widetilde{\theta}_{k,h})\phi_{s,a}^{s''}(\theta_h^* - \widetilde{\theta}_{k,h})\right) \\
&\le H \sum_{s'\in\mathcal{S}_{h,s,a}^+} p_{s,a}^{s'}(\widetilde{\theta}_{k,h}) \left(\left\|\phi_{s,a}^{s'} - \sum_{s''\in\mathcal{S}_{h,s,a}} p_{s,a}^{s''}(\widetilde{\theta}_{k,h})\phi_{s,a}^{s''}\right\|_{\mathcal{H}_{k,h}^{-1}} \left\|\theta_h^* - \widetilde{\theta}_{k,h}\right\|_{\mathcal{H}_{k,h}}\right) \\
&\le H\widetilde{\beta}_k \sum_{s'\in\mathcal{S}_{h,s,a}^+} p_{s,a}^{s'}(\widetilde{\theta}_{k,h}) \left(\left\|\phi_{s,a}^{s'} - \sum_{s''\in\mathcal{S}_{h,s,a}} p_{s,a}^{s''}(\widetilde{\theta}_{k,h})\phi_{s,a}^{s''}\right\|_{\mathcal{H}_{k,h}^{-1}}\right) \\
&\le H\widetilde{\beta}_k \sum_{s'\in\mathcal{S}_{h,s,a}} p_{s,a}^{s'}(\widetilde{\theta}_{k,h}) \left(\left\|\phi_{s,a}^{s'} - \sum_{s''\in\mathcal{S}_{h,s,a}} p_{s,a}^{s''}(\widetilde{\theta}_{k,h})\phi_{s,a}^{s''}\right\|_{\mathcal{H}_{k,h}^{-1}}\right) \qquad (29)
\end{aligned}
$$

where in the first inequality, we denote $\mathcal{S}_{h,s,a}^+$ as the subset of $\mathcal{S}_{h,s,a}$ such that $\phi_{s,a}^{s'}(\theta_h^* - \widetilde{\theta}_{k,h}) - \sum_{s''\in\mathcal{S}_{h,s,a}} p_{s,a}^{s''}(\widetilde{\theta}_{k,h})\phi_{s,a}^{s''}(\theta_h^* - \widetilde{\theta}_{k,h})$ is non-negative, the second inequality holds by the Holder's inequality, and the third inequality is by the confidence set in Lemma 3.

For the second-order term, we define $u_{s,a}^{s'}(\theta) = \phi_{s,a}^{s'}\theta$, and $p_{s,a}^{s'}(u) = \frac{\exp(u_{s,a}^{s'})}{1+\sum_{s''}\exp(u_{s,a}^{s''})}$, further define

$$
F(u) = \sum_{s'\in\mathcal{S}_{h,s,a}} \frac{\exp(u_{s,a}^{s'})}{1+\sum_{s''\in\mathcal{S}_{h,s,a}}\exp(u_{s,a}^{s''})}, \widetilde{F}(u) = \sum_{s'\in\mathcal{S}_{h,s,a}} \frac{\exp(u_{s,a}^{s'})V(s')}{1+\sum_{s''\in\mathcal{S}_{h,s,a}}\exp(u_{s,a}^{s''})}.
$$

Then, we have

$$
\begin{aligned}
&\frac{1}{2}\sum_{s'\in\mathcal{S}_{h,s,a}} (\widehat{\theta}_{k,h} - \theta_h^*)^\top \nabla^2 p_{s,a}^{s'}(\bar{\theta})(\widehat{\theta}_{k,h} - \theta_h^*)V(s') \\
&= \frac{1}{2}\big(u(\widehat{\theta}_{k,h}) - u(\theta_h^*)\big)^\top \nabla^2 \widetilde{F}(u(\bar{\theta}))\big(u(\widehat{\theta}_{k,h}) - u(\theta_h^*)\big) \\
&= \frac{1}{2}\sum_{s'\in\mathcal{S}_{h,s,a}}\sum_{s''\in\mathcal{S}_{h,s,a}} \big(u_{s,a}^{s'}(\widehat{\theta}_{k,h}) - u_{s,a}^{s'}(\theta_h^*)\big)^\top \frac{\partial^2 \widetilde{F}(u(\bar{\theta}))}{\partial s'\partial s''}\big(u_{s,a}^{s''}(\widehat{\theta}_{k,h}) - u_{s,a}^{s''}(\theta_h^*)\big) \\
&\le \frac{H}{2}\sum_{s'\in\mathcal{S}_{h,s,a}}\sum_{s''\in\mathcal{S}_{h,s,a}} \big|u_{s,a}^{s'}(\widehat{\theta}_{k,h}) - u_{s,a}^{s'}(\theta_h^*)\big| \frac{\partial^2 F(u(\bar{\theta}))}{\partial s'\partial s''}\big|u_{s,a}^{s''}(\widehat{\theta}_{k,h}) - u_{s,a}^{s''}(\theta_h^*)\big|
\end{aligned}
$$

where the inequality holds by $V(s) \in [0, H], \forall s$.

According to Lemma 18, we have (omit the subscript $\mathcal{S}_{h,s,a}$ for simplicity):

$$
\begin{aligned}
&\frac{H}{2}\sum_{s'}\sum_{s''}\big|u_{s,a}^{s'}(\widehat{\theta}_{k,h}) - u_{s,a}^{s'}(\theta_h^*)\big| \frac{\partial^2 F(u(\bar{\theta}))}{\partial s'\partial s''}\big|u_{s,a}^{s''}(\widehat{\theta}_{k,h}) - u_{s,a}^{s''}(\theta_h^*)\big| \\
&\le H\sum_{s'}\sum_{s''\neq s'}\big|u_{s,a}^{s'}(\widehat{\theta}_{k,h}) - u_{s,a}^{s'}(\theta_h^*)\big| p_{s,a}^{s'}\big(u(\bar{\theta})\big)p_{s,a}^{s''}\big(u(\bar{\theta})\big)\big|u_{s,a}^{s''}(\widehat{\theta}_{k,h}) - u_{s,a}^{s''}(\theta_h^*)\big| \\
&\quad + \frac{3H}{2}\sum_{s'}\big(u_{s,a}^{s'}(\widehat{\theta}_{k,h}) - u_{s,a}^{s'}(\theta_h^*)\big)^2 p_{s,a}^{s'}(u(\bar{\theta})). \qquad (30)
\end{aligned}
$$

To bound the first term, by applying the AM-GM inequality, we obtain

$$
H\sum_{s'}\sum_{s''\neq s'}\big|u_{s,a}^{s'}(\widehat{\theta}_{k,h}) - u_{s,a}^{s'}(\theta_h^*)\big| p_{s,a}^{s'}\big(u(\bar{\theta})\big)p_{s,a}^{s''}\big(u(\bar{\theta})\big)\big|u_{s,a}^{s''}(\widehat{\theta}_{k,h}) - u_{s,a}^{s''}(\theta_h^*)\big|
$$

$$
\begin{aligned}
&\le H \sum_{s'} \sum_{s''} \big| u_{s,a}^{s'}(\widehat{\theta}_{k,h}) - u_{s,a}^{s'}(\theta_h^*) \big| p_{s,a}^{s'}\big(u(\bar{\theta})\big) p_{s,a}^{s''}\big(u(\bar{\theta})\big) \big| u_{s,a}^{s''}(\widehat{\theta}_{k,h}) - u_{s,a}^{s''}(\theta_h^*) \big| \\
&\le \frac{H}{2} \sum_{s'} \sum_{s''} \big(u_{s,a}^{s'}(\widehat{\theta}_{k,h}) - u_{s,a}^{s'}(\theta_h^*) \big)^2 p_{s,a}^{s'}\big(u(\bar{\theta})\big) p_{s,a}^{s''}\big(u(\bar{\theta})\big) \\
&\qquad\qquad + \frac{H}{2} \sum_{s'} \sum_{s''} \big(u_{s,a}^{s''}(\widehat{\theta}_{k,h}) - u_{s,a}^{s''}(\theta_h^*) \big)^2 p_{s,a}^{s'}\big(u(\bar{\theta})\big) p_{s,a}^{s''}\big(u(\bar{\theta})\big) \\
&\le H \sum_{s'} \big(u_{s,a}^{s'}(\widehat{\theta}_{k,h}) - u_{s,a}^{s'}(\theta_h^*) \big)^2 p_{s,a}^{s'}\big(u(\bar{\theta})\big) \qquad (31)
\end{aligned}
$$

Plugging (31) into (30), we have

$$
\begin{aligned}
&\frac{H}{2} \sum_{s'} \sum_{s''} \big| u_{s,a}^{s'}(\widehat{\theta}_{k,h}) - u_{s,a}^{s'}(\theta_h^*) \big| \frac{\partial^2 F(u(\bar{\theta}))}{\partial s' \partial s''} \big| u_{s,a}^{s''}(\widehat{\theta}_{k,h}) - u_{s,a}^{s''}(\theta_h^*) \big| \\
&\le \frac{5H}{2} \sum_{s'} \big(u_{s,a}^{s'}(\widehat{\theta}_{k,h}) - u_{s,a}^{s'}(\theta_h^*) \big)^2 p_{s,a}^{s'}(u(\bar{\theta})) \\
&= \frac{5H}{2} \sum_{s'} \big((\phi_{s,a}^{s'})^\top (\widehat{\theta}_{k,h} - \theta_h^*) \big)^2 p_{s,a}^{s'}(u(\bar{\theta})) \\
&\le \frac{5}{2} H \widetilde{\beta}_k^2 \max_{s'} \|\phi_{s,a}^{s'}\|_{\mathcal{H}_{k,h}^{-1}}^2, \qquad (32)
\end{aligned}
$$

where the last inequality holds by Lemma 3. Combining (29) and (32) finishes the proof. ∎

G Proof of Theorem 2

Proof. Combining Lemma 4 and Lemma 10, we have

$$
\sum_{k=1}^K V_1^*(s_{k,1}) - \sum_{k=1}^K V_1^{\pi_k}(s_{k,1}) \le 2 \sum_{k=1}^K \sum_{h=1}^H (\epsilon_{k,h}^{\texttt{fst}} + \epsilon_{k,h}^{\texttt{snd}}) + H\sqrt{2KH \log(2/\delta)}
$$

where

$$
\begin{aligned}
\epsilon_{k,h}^{\texttt{fst}} &= H \widetilde{\beta}_k \sum_{s' \in \mathcal{S}_{k,h}} p_{k,h}^{s'}(\widetilde{\theta}_{k,h}) \Big\| \phi_{k,h}^{s'} - \sum_{s'' \in \mathcal{S}_{k,h}} p_{k,h}^{s''}(\widetilde{\theta}_{k,h}) \phi_{k,h}^{s''} \Big\|_{\mathcal{H}_{k,h}^{-1}}, \\
\epsilon_{k,h}^{\texttt{snd}} &= \frac{5}{2} H \widetilde{\beta}_k^2 \max_{s' \in \mathcal{S}_{k,h}} \|\phi_{k,h}^{s'}\|_{\mathcal{H}_{k,h}^{-1}}^2.
\end{aligned}
$$

Next, we bound $\epsilon_{k,h}^{\texttt{fst}}$ and $\epsilon_{k,h}^{\texttt{snd}}$ respectively.

Bounding $\epsilon_{k,h}^{\texttt{fst}}$. For simplicity, we denote

$$
\mathbb{E}_\theta[\phi_{k,h}^{s'}] = \mathbb{E}_{s' \sim p_{k,h}^s(\theta)}[\phi_{k,h}^{s'}], \quad \bar{\phi}_{s,a}^{s'} = \phi_{s,a}^{s'} - \mathbb{E}_{\widetilde{\theta}_{k,h}}[\phi_{k,h}^{s'}], \quad \widetilde{\phi}_{s,a}^{s'} = \phi_{s,a}^{s'} - \mathbb{E}_{\widetilde{\theta}_{k+1,h}}[\phi_{k,h}^{s'}]
$$

Then, we have

$$
\begin{aligned}
&\sum_{s' \in \mathcal{S}_{k,h}} p_{k,h}^{s'}(\widetilde{\theta}_{k,h}) \Big\| \phi_{k,h}^{s'} - \sum_{s'' \in \mathcal{S}_{k,h}} p_{k,h}^{s''}(\widetilde{\theta}_{k,h}) \phi_{k,h}^{s''} \Big\|_{\mathcal{H}_{k,h}^{-1}} = \sum_{s' \in \mathcal{S}_{k,h}} p_{k,h}^{s'}(\widetilde{\theta}_{k,h}) \|\bar{\phi}_{k,h}^{s'}\|_{\mathcal{H}_{k,h}^{-1}} \\
&\le \sum_{s' \in \mathcal{S}_{k,h}} p_{k,h}^{s'}(\widetilde{\theta}_{k,h}) \|\bar{\phi}_{k,h}^{s'} - \widetilde{\phi}_{k,h}^{s'}\|_{\mathcal{H}_{k,h}^{-1}} + \sum_{s' \in \mathcal{S}_{k,h}} p_{k,h}^{s'}(\widetilde{\theta}_{k,h}) \|\widetilde{\phi}_{k,h}^{s'}\|_{\mathcal{H}_{k,h}^{-1}} \\
&= \underbrace{\sum_{s' \in \mathcal{S}_{k,h}} p_{k,h}^{s'}(\widetilde{\theta}_{k,h}) \|\bar{\phi}_{k,h}^{s'} - \widetilde{\phi}_{k,h}^{s'}\|_{\mathcal{H}_{k,h}^{-1}}}_{(c)} + \underbrace{\sum_{s' \in \mathcal{S}_{k,h}} (p_{k,h}^{s'}(\widetilde{\theta}_{k,h}) - p_{k,h}^{s'}(\widetilde{\theta}_{k+1,h})) \|\widetilde{\phi}_{k,h}^{s'}\|_{\mathcal{H}_{k,h}^{-1}}}_{(d)} \\
&\quad + \underbrace{\sum_{s' \in \mathcal{S}_{k,h}} p_{k,h}^{s'}(\widetilde{\theta}_{k+1,h}) \|\widetilde{\phi}_{k,h}^{s'}\|_{\mathcal{H}_{k,h}^{-1}}}_{(e)}.
\end{aligned}
$$

We bound these terms separately in the following.

For the first term (c), we have

$$
\begin{aligned}
&\|\bar{\phi}_{k,h}^{s'} - \widetilde{\phi}_{k,h}^{s'}\|_{\mathcal{H}_{k,h}^{-1}} \\
&= \Big\| \sum_{s''\in\mathcal{S}_{k,h}} \Big(p_{k,h}^{s''}(\widetilde{\theta}_{k+1,h}) - p_{k,h}^{s''}(\widetilde{\theta}_{k,h})\Big)\,\phi_{k,h}^{s''}\Big\|_{\mathcal{H}_{k,h}^{-1}} \\
&= \Big\| \sum_{s''\in\mathcal{S}_{k,h}} \Big(\nabla p_{k,h}^{s''}(\xi_{k,h})^\top(\widetilde{\theta}_{k+1,h} - \widetilde{\theta}_{k,h})\Big)\,\phi_{k,h}^{s''}\Big\|_{\mathcal{H}_{k,h}^{-1}} \\
&\le \sum_{s''\in\mathcal{S}_{k,h}} \Big|\nabla p_{k,h}^{s''}(\xi_{k,h})^\top(\widetilde{\theta}_{k+1,h} - \widetilde{\theta}_{k,h})\Big| \big\|\phi_{k,h}^{s''}\big\|_{\mathcal{H}_{k,h}^{-1}} \\
&= \sum_{s''\in\mathcal{S}_{k,h}} \Big|\Big(p_{k,h}^{s''}(\xi_{k,h})\phi_{k,h}^{s''} - p_{k,h}^{s''}(\xi_{k,h})\sum_{s'''\in\mathcal{S}_{k,h}} p_{k,h}^{s'''}(\xi_{k,h})\phi_{k,h}^{s'''}\Big)^\top(\widetilde{\theta}_{k+1,h} - \widetilde{\theta}_{k,h})\Big| \big\|\phi_{k,h}^{s''}\big\|_{\mathcal{H}_{k,h}^{-1}} \\
&\le \sum_{s''\in\mathcal{S}_{k,h}} p_{k,h}^{s''}(\xi_{k,h})\Big|\big(\phi_{k,h}^{s''}\big)^\top(\widetilde{\theta}_{k+1,h} - \widetilde{\theta}_{k,h})\Big| \big\|\phi_{k,h}^{s''}\big\|_{\mathcal{H}_{k,h}^{-1}} \\
&\qquad + \sum_{s''\in\mathcal{S}_{k,h}} p_{k,h}^{s''}(\xi_{k,h})\big\|\phi_{k,h}^{s''}\big\|_{\mathcal{H}_{k,h}^{-1}} \sum_{s'''\in\mathcal{S}_{k,h}} p_{k,h}^{s'''}(\xi_{k,h})\big|(\phi_{k,h}^{s'''})^\top(\widetilde{\theta}_{k+1,h} - \widetilde{\theta}_{k,h})\big| \\
&\le \sum_{s''\in\mathcal{S}_{k,h}} p_{k,h}^{s''}(\xi_{k,h})\big\|\widetilde{\theta}_{k+1,h} - \widetilde{\theta}_{k,h}\big\|_{\mathcal{H}_{k,h}} \big\|\phi_{k,h}^{s''}\big\|_{\mathcal{H}_{k,h}^{-1}}^2 \\
&\qquad + \sum_{s''\in\mathcal{S}_{k,h}} p_{k,h}^{s''}(\xi_{k,h})\big\|\phi_{k,h}^{s''}\big\|_{\mathcal{H}_{k,h}^{-1}} \sum_{s'''\in\mathcal{S}_{k,h}} p_{k,h}^{s'''}(\xi_{k,h})\big\|\phi_{k,h}^{s'''}\big\|_{\mathcal{H}_{k,h}^{-1}} \big\|\widetilde{\theta}_{k+1,h} - \widetilde{\theta}_{k,h}\big\|_{\mathcal{H}_{k,h}} \\
&\le \frac{2\eta}{\sqrt{\lambda}} \sum_{s''\in\mathcal{S}_{k,h}} p_{k,h}^{s''}(\xi_{k,h})\big\|\phi_{k,h}^{s''}\big\|_{\mathcal{H}_{k,h}^{-1}}^2 + \frac{2\eta}{\sqrt{\lambda}}\Big(\sum_{s''\in\mathcal{S}_{k,h}} p_{k,h}^{s''}(\xi_{k,h})\big\|\phi_{k,h}^{s''}\big\|_{\mathcal{H}_{k,h}^{-1}}\Big)^2 \\
&\le \frac{4\eta}{\sqrt{\lambda}} \sum_{s''\in\mathcal{S}_{k,h}} p_{k,h}^{s''}(\xi_{k,h})\big\|\phi_{k,h}^{s''}\big\|_{\mathcal{H}_{k,h}^{-1}}^2 \\
&\le \frac{4\eta}{\sqrt{\lambda}} \max_{s''\in\mathcal{S}_{k,h}} \big\|\phi_{k,h}^{s''}\big\|_{\mathcal{H}_{k,h}^{-1}}^2
\end{aligned}
$$

where and the fifth inequality is by Cauchy-Schwarz inequality and the fourth inequality is because by Lemma 17, since $\widetilde{\mathcal{H}}_{k,h} \succeq \mathcal{H}_{k,h} \succeq \lambda I_d$, we have

$$
\big\|\widetilde{\theta}_{k+1,h} - \widetilde{\theta}_{k,h}\big\|_{\mathcal{H}_{k,h}} \le \big\|\widetilde{\theta}_{k+1,h} - \widetilde{\theta}_{k,h}\big\|_{\widetilde{\mathcal{H}}_{k,h}} \le 2\eta\|g_{k,h}(\widetilde{\theta}_{k,h})\|_{\widetilde{\mathcal{H}}_{k,h}^{-1}} \le \frac{2\eta}{\sqrt{\lambda}}\|g_{k,h}(\widetilde{\theta}_{k,h})\|_2,
$$

and since $g_{k,h}(\theta) = \sum_{s'\in\mathcal{S}_{k,h}}(p_{k,h}^{s'}(\theta) - y_{k,h}^{s'})\phi_{k,h}^{s'}$, we have

$$
\|g_{k,h}(\widetilde{\theta}_{k,h})\|_2 \le \Big\|\sum_{s'\in\mathcal{S}_{k,h}} p_{k,h}^{s'}(\widetilde{\theta}_{k,h})\phi_{k,h}^{s'}\Big\|_2 + \Big\|\sum_{s'\in\mathcal{S}_{k,h}} y_{k,h}^{s'}\phi_{k,h}^{s'}\Big\|_2 \le 2\max_{s'\in\mathcal{S}_{k,h}}\big\|\phi_{k,h}^{s'}\big\|_2 \le 2.
$$

Therefore, we have

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s'\in\mathcal{S}_{k,h}} p_{k,h}^{s'}(\widetilde{\theta}_{k,h})\big\|\bar{\phi}_{k,h}^{s'} - \widetilde{\phi}_{k,h}^{s'}\big\|_{\mathcal{H}_{k,h}^{-1}} &\le \frac{4\eta}{\sqrt{\lambda}}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s'\in\mathcal{S}_{k,h}} p_{k,h}^{s'}(\widetilde{\theta}_{k,h}) \max_{s''\in\mathcal{S}_{k,h}}\big\|\phi_{k,h}^{s''}\big\|_{\mathcal{H}_{k,h}^{-1}}^2 \\
&\le \frac{4\eta}{\sqrt{\lambda}}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s'\in\mathcal{S}_{k,h}} \max_{s''\in\mathcal{S}_{k,h}}\big\|\phi_{k,h}^{s''}\big\|_{\mathcal{H}_{k,h}^{-1}}^2 \\
&\le \frac{8H\eta}{\kappa\sqrt{\lambda}} d\log\left(1 + \frac{K}{d\lambda}\right). \qquad (33)
\end{aligned}
$$

where the last inequality holds by Lemma 6.

For the term (d), by similar analysis, we have

$$
\begin{aligned}
&(p^{s'}_{k,h}(\widetilde{\theta}_{k,h}) - p^{s'}_{k,h}(\widetilde{\theta}_{k+1,h}))\|\widetilde{\phi}^{s'}_{k,h}\|_{\mathcal{H}^{-1}_{k,h}} \\
&= \nabla p^{s'}_{k,h}(\xi_{k,h})^\top(\widetilde{\theta}_{k,h} - \widetilde{\theta}_{k+1,h})\|\widetilde{\phi}^{s'}_{k,h}\|_{\mathcal{H}^{-1}_{k,h}} \\
&= \Big(p^{s'}_{k,h}(\xi_{k,h})\phi^{s'}_{k,h} - p^{s'}_{k,h}(\xi_{k,h})\sum_{s''\in\mathcal{S}_{k,h}} p^{s''}_{k,h}(\xi_{k,h})\phi^{s''}_{k,h}\Big)^\top(\widetilde{\theta}_{k+1,h} - \widetilde{\theta}_{k,h})\|\widetilde{\phi}^{s'}_{k,h}\|_{\mathcal{H}^{-1}_{k,h}} \\
&\leq \frac{2\eta}{\sqrt{\lambda}}\left(p^{s'}_{k,h}(\xi_{k,h})\|\phi^{s'}_{k,h}\|_{\mathcal{H}^{-1}_{k,h}}\|\widetilde{\phi}^{s'}_{k,h}\|_{\mathcal{H}^{-1}_{k,h}} + p^{s'}_{k,h}(\xi_{k,h})\|\widetilde{\phi}^{s'}_{k,h}\|_{\mathcal{H}^{-1}_{k,h}}\sum_{s''\in\mathcal{S}_{k,h}} p^{s''}_{k,h}(\xi_{k,h})\|\phi^{s''}_{k,h}\|_{\mathcal{H}^{-1}_{k,h}}\right) \\
&\leq \frac{2\eta}{\sqrt{\lambda}}\left(\max_{s''\in\mathcal{S}_{k,h}}\|\phi^{s''}_{k,h}\|_{\mathcal{H}^{-1}_{k,h}}\|\widetilde{\phi}^{s''}_{k,h}\|_{\mathcal{H}^{-1}_{k,h}} + \max_{s''\in\mathcal{S}_{k,h}}\|\widetilde{\phi}^{s''}_{k,h}\|_{\mathcal{H}^{-1}_{k,h}}\max_{s'''\in\mathcal{S}_{k,h}}\|\phi^{s'''}_{k,h}\|_{\mathcal{H}^{-1}_{k,h}}\right) \\
&\leq \frac{2\eta}{\sqrt{\lambda}}\max_{s''\in\mathcal{S}_{k,h}}\frac{\|\phi^{s''}_{k,h}\|^2_{\mathcal{H}^{-1}_{k,h}} + \|\widetilde{\phi}^{s''}_{k,h}\|^2_{\mathcal{H}^{-1}_{k,h}}}{2} \\
&\qquad + \frac{2\eta}{\sqrt{\lambda}}\frac{\left(\max_{s''\in\mathcal{S}_{k,h}}\|\widetilde{\phi}^{s''}_{k,h}\|_{\mathcal{H}^{-1}_{k,h}}\right)^2 + \left(\max_{s'''\in\mathcal{S}_{k,h}}\|\phi^{s'''}_{k,h}\|_{\mathcal{H}^{-1}_{k,h}}\right)^2}{2} \\
&\leq \frac{4\eta}{\sqrt{\lambda}}\max\left\{\max_{s''\in\mathcal{S}_{k,h}}\|\phi^{s''}_{k,h}\|^2_{\mathcal{H}^{-1}_{k,h}}, \max_{s''\in\mathcal{S}_{k,h}}\|\widetilde{\phi}^{s''}_{k,h}\|^2_{\mathcal{H}^{-1}_{k,h}}\right\},
\end{aligned}
$$

where the third inequality holds by the AM-GM inequality. Thus, we have

$$
\begin{aligned}
&\sum_{k=1}^K\sum_{h=1}^H\sum_{s'\in\mathcal{S}_{k,h}}(p^{s'}_{k,h}(\widetilde{\theta}_{k,h}) - p^{s'}_{k,h}(\widetilde{\theta}_{k+1,h}))\|\widetilde{\phi}^{s'}_{k,h}\|_{\mathcal{H}^{-1}_{k,h}} \\
&\leq \frac{4\eta}{\sqrt{\lambda}}\sum_{k=1}^K\sum_{h=1}^H\max\left\{\max_{s''\in\mathcal{S}_{k,h}}\|\phi^{s''}_{k,h}\|^2_{\mathcal{H}^{-1}_{k,h}}, \max_{s''\in\mathcal{S}_{k,h}}\|\widetilde{\phi}^{s''}_{k,h}\|^2_{\mathcal{H}^{-1}_{k,h}}\right\} \\
&\leq \frac{8H\eta}{\kappa\sqrt{\lambda}}d\log\left(1 + \frac{K}{d\lambda}\right),
\end{aligned} \tag{34}
$$

where the last inequality holds by Lemma 6.

Finally, we bound the term (e) as follows.

$$
\begin{aligned}
&\sum_{k=1}^K\sum_{s'\in\mathcal{S}_{k,h}} p^{s'}_{k,h}(\widetilde{\theta}_{k+1,h})\|\widetilde{\phi}^{s'}_{k,h}\|_{\mathcal{H}^{-1}_{k,h}} \\
&\leq \sqrt{\sum_{k=1}^K\sum_{s'\in\mathcal{S}_{k,h}} p^{s'}_{k,h}(\widetilde{\theta}_{k+1,h})}\sqrt{\sum_{k=1}^K\sum_{s'\in\mathcal{S}_{k,h}} p^{s'}_{k,h}(\widetilde{\theta}_{k+1,h})\|\widetilde{\phi}^{s'}_{k,h}\|^2_{\mathcal{H}^{-1}_{k,h}}} \\
&\leq \sqrt{K}\sqrt{2d\log\left(1 + \frac{K}{d\lambda}\right)},
\end{aligned} \tag{35}
$$

where the first inequality holds by the Cauchy-Schwarz inequality and the last holds by Lemma 6.

Thus, combining (33), (34), and (35), we have

$$
\begin{aligned}
\sum_{k=1}^K\sum_{h=1}^H \epsilon^{\mathtt{fst}}_{k,h} &= H\widetilde{\beta}_K\sum_{k=1}^K\sum_{h=1}^H\sum_{s'\in\mathcal{S}_{k,h}} p^{s'}_{k,h}(\widetilde{\theta}_{k,h})\Big\|\phi^{s'}_{k,h} - \sum_{s''\in\mathcal{S}_{k,h}} p^{s''}_{k,h}(\widetilde{\theta}_{k,h})\phi^{s''}_{k,h}\Big\|_{\mathcal{H}^{-1}_{k,h}} \\
&\leq H^2\widetilde{\beta}_K\left(\sqrt{2dK\log\left(1 + \frac{K}{d\lambda}\right)} + \frac{16\eta}{\kappa\sqrt{\lambda}}d\log(1 + \frac{K}{d\lambda})\right)
\end{aligned} \tag{36}
$$

For the second-order term, by Lemma 6, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\epsilon_{k,h}^{\texttt{snd}} = \frac{5}{2}H\widetilde{\beta}_k^2\sum_{k=1}^{K}\sum_{h=1}^{H}\max_{s'\in\mathcal{S}_{k,h}}\|\phi_{k,h}^{s'}\|_{\mathcal{H}_{k,h}^{-1}}^2$$
$$\leq \frac{5}{\kappa}H^2\widetilde{\beta}_K^2 d\log\left(1+\frac{K}{d\lambda}\right). \tag{37}$$

where the last inequality holds by Lemma 6.

Combining (36) and (37), we have

$$\begin{aligned}
\mathrm{Reg}(K) &\leq 2\sum_{k=1}^{K}\sum_{h=1}^{H}(\epsilon_{k,h}^{\texttt{fst}}+\epsilon_{k,h}^{\texttt{snd}}) + H\sqrt{2KH\log(2/\delta)}\\
&\leq H^2\widetilde{\beta}_K\left(\sqrt{2dK\log\left(1+\frac{K}{d\lambda}\right)}+\frac{16\eta}{\kappa\sqrt{\lambda}}d\log(1+\frac{K}{d\lambda})\right)+\frac{5}{\kappa}H^2\widetilde{\beta}_K^2 d\log\left(1+\frac{K}{d\lambda}\right)\\
&\quad + H\sqrt{2KH\log\left(\frac{2}{\delta}\right)}\\
&\leq \widetilde{\mathcal{O}}\big(dH^2\sqrt{K}+d^2H^2\kappa^{-1}\big)
\end{aligned}$$

This finishes the proof. ■

H Proof of Theorem 3

Proof. Our proof is similar to the proof of the lower bound for adversarial linear mixture MDPs with the unknown transition in Zhao et al. [2023]. We prove this lemma by reducing the MNL mixture MDP problem to a sequence of binary logistic bandit problems.

Specifically, we use each three layers to construct a block. Note that the third layer of block i is also the first layer of block $i+1$ and hence there are total $H/2$ blocks. In each block, both the first and third layers of this block only have one state, and the second layer has two states. Here we take block i as an example. The first two layers of this block are associated with transition probability $\mathbb{P}_{i,1}$ and $\mathbb{P}_{i,2}$. Denote by $s_{i,1}$ the only state in the first layer of this block. In the second layer of the block i, we assume there exist two states $s_{i,2}^*$ and $s_{i,2}$. Let $s_{i,3}$ be the only state in the third layer of this block. Further, for any $a\in\mathcal{A}$, let $\phi(s_{i,1},a,s_{i,2})=0$. The probability of transferring to state $s_{i,2}^*$ when executing action a at state $s_{i,1}$ is ρ_a, in particular,

$$\mathbb{P}_{i,1}(s_{i,2}^*\mid s_{i,1},a)=\frac{\exp(\phi(s_{i,2}^*\mid s_{i,1},a)^\top\theta_{i,1}^*)}{1+\exp(\phi(s_{i,2}^*\mid s_{i,1},a))^\top\theta_{i,1}^*)}=\rho_a,$$
$$\mathbb{P}_{i,1}(s_{i,2}\mid s_{i,1},a)=\frac{1}{1+\exp(\phi(s_{i,2}^*\mid s_{i,1},a)^\top\theta_{i,1}^*)}=1-\rho_a.$$

For the second layer, the MDP instance satisfies that $\forall s=s_{i,2}^*,s_{i,2}$, and $a\in\mathcal{A}$, $\mathbb{P}_{i,2}(s_{i,3}\mid s,a)=1$. The reward function satisfies $r_k(s_{i,1},a)=0$ for the first layer and $r_k(s_{i,2}^*,a)=1$, $r_k(s_{i,2},a)=0$ for the second layer for all $a\in\mathcal{A}$.

Consider the logistic bandit problem where a learner selects action $x\in\mathbb{R}^d$ and receives a reward r_k sampled from Bernoulli distributed with mean

$$\mu(x^\top\theta^*)=\frac{1}{1+\exp(x^\top\theta^*)}.$$

By this configuration, we can see that learning in this block of MDP can be regarded as learning a d-dimensional logistic bandit problem with A arms, where the arm set is $\phi(s_{i,2}^*\mid s_{i,1},a)$ and the expected reward of each arm is ρ_a. It is also clear that the optimal policy at state $s_{i,1}$ is to select action $a_{i,1}^*=\arg\max_{a\in\mathcal{A}}\phi(s_{i,2}^*\mid s_{i,1},a)$. Thus, learning this MNL mixture MDP equals to learning $H/2$ logistic bandit problems. This finishes the proof. ■

I Supporting Lemmas

In this section, we provide several supporting lemmas used in the proofs of the main results.

Lemma 15 (Abbasi-Yadkori et al. [2011, Theorem 1])**.** *Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\eta_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that η_t is $\mathcal{F}_t$-measurable and η_t is conditionally zero-mean R-sub-Gaussian for some $R \geq 0$ i.e. $\forall \lambda \in \mathbb{R}$, $\mathbb{E}\left[e^{\lambda \eta_t} \mid F_{t-1}\right] \leq \exp\left(\lambda^2 R^2/2\right)$. Let $\{X_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that X_t is $\mathcal{F}_{t-1}$-measurable. Assume that V is a $d \times d$ positive definite matrix. For any $t \geq 0$, define*

$$V_t = V + \sum_{s=1}^{t-1} X_s X_s^\top \quad S_t = \sum_{s=1}^{t-1} \eta_s X_s.$$

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

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

Lemma 16 (Abbasi-Yadkori et al. [2011, Lemma 10])**.** *Suppose $x_1, \ldots, x_t \in \mathbb{R}^d$ and for any $1 \leq s \leq t$, $\|x_s\|_2 \leq L$. Let $V_t = \lambda I_d + \sum_{s=1}^{t} x_s x_s^\top$ for $\lambda \geq 0$. Then, for any $1 \leq s \leq t$, we have*

$$\det(V_t) \leq \left(\lambda + \frac{tL^2}{d}\right)^d.$$

Lemma 17 (Orabona [2019, Lemma 6.9])**.** *Let Z be a positive define matrix and $\mathcal{X}$ be a convex set,*

$$\widetilde{\mathbf{x}} = \arg\min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{g}, \mathbf{x} \rangle + \frac{1}{2\eta}\|\mathbf{x} - \widehat{\mathbf{x}}\|_Z^2.$$

Then, for all $\mathbf{x}^* \in \mathcal{X}$, we have

$$\|\widetilde{\mathbf{x}} - \widehat{\mathbf{x}}\|_Z \leq 2\eta\|\mathbf{g}\|_{Z^{-1}}.$$

and

$$\langle \mathbf{g}, \widehat{\mathbf{x}} - \mathbf{x}^* \rangle \leq \frac{\eta}{2}\|\mathbf{g}\|_{Z^{-1}}^2 + \frac{1}{2\eta}\|\widehat{\mathbf{x}} - \mathbf{x}^*\|_Z^2 - \frac{1}{2\eta}\|\widetilde{\mathbf{x}} - \mathbf{x}^*\|_Z^2.$$

Lemma 18 (Lee and Oh [2024, Lemma D.3])**.** *Define $Q : \mathbb{R}^K \to \mathbb{R}$, such that for any $\mathbf{u} = (u_1, \ldots, u_K) \in \mathbb{R}^K, Q(\mathbf{u}) = \sum_{i=1}^K \frac{\exp(u_i)}{v + \sum_{k=1}^K \exp(u_k)}$. Let $p_i(\mathbf{u}) = \frac{\exp(u_i)}{v + \sum_{k=1}^K \exp(u_k)}$. Then, for all $i \in [K]$, we have*

$$\left|\frac{\partial^2 Q}{\partial i \partial j}\right| \leqslant \begin{cases} 3p_i(\mathbf{u}) & \text{if } i = j, \\ 2p_i(\mathbf{u})p_j(\mathbf{u}) & \text{if } i \neq j. \end{cases}$$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of this work in the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We present the assumption in Assumption 1 and provide detailed proofs for all theoretical results in the appendices.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: This is a theoretical paper and does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: This is a theoretical paper and does not include experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: This is a theoretical paper and does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: This is a theoretical paper and does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: This is a theoretical paper and does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This paper adheres fully to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This is a theoretical paper and there is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

 Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

 Answer: [NA]

 Justification: This is a theoretical paper and does not release any data or models that have a high risk for misuse.

 Guidelines:

 - The answer NA means that the paper poses no such risks.
 - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
 - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
 - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

 Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

 Answer: [NA]

 Justification: This is a theoretical paper and does not use existing assets.

 Guidelines:

 - The answer NA means that the paper does not use existing assets.
 - The authors should cite the original paper that produced the code package or dataset.
 - The authors should state which version of the asset is used and, if possible, include a URL.
 - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
 - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
 - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
 - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
 - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This is a theoretical paper and does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.