Randomized Exploration for Reinforcement Learning with Multinomial Logistic Function Approximation

Wooseong Cho[*]
Seoul National University
Seoul, South Korea
wooseong_cho@snu.ac.kr

Taehyun Hwang[*]
Seoul National University
Seoul, South Korea
th.hwang@snu.ac.kr

Joongkyu Lee
Seoul National University
Seoul, South Korea
jklee0717@snu.ac.kr

Min-hwan Oh[†]
Seoul National University
Seoul, South Korea
minoh@snu.ac.kr

Abstract

We study reinforcement learning with *multinomial logistic* (MNL) function approximation where the underlying transition probability kernel of the *Markov decision processes* (MDPs) is parametrized by an unknown transition core with features of state and action. For the finite horizon episodic setting with inhomogeneous state transitions, we propose provably efficient algorithms with randomized exploration having frequentist regret guarantees. For our first algorithm, RRL-MNL, we adapt optimistic sampling to ensure the optimism of the estimated value function with sufficient frequency. We establish that RRL-MNL achieves a $\widetilde{\mathcal{O}}(\kappa^{-1}d^{\frac{3}{2}}H^{\frac{3}{2}}\sqrt{T})$ frequentist regret bound with constant-time computational cost per episode. Here, d is the dimension of the transition core, H is the horizon length, T is the total number of steps, and κ is a problem-dependent constant. Despite the simplicity and practicality of RRL-MNL, its regret bound scales with κ^{-1}, which is potentially large in the worst case. To improve the dependence on κ^{-1}, we propose ORRL-MNL, which estimates the value function using the local gradient information of the MNL transition model. We show that its frequentist regret bound is $\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H^{\frac{3}{2}}\sqrt{T}+\kappa^{-1}d^2H^2)$. To the best of our knowledge, these are the first randomized RL algorithms for the MNL transition model that achieve statistical guarantees with constant-time computational cost per episode. Numerical experiments demonstrate the superior performance of the proposed algorithms.

1 Introduction

Reinforcement learning (RL) is a sequential decision-making problem in which an agent tries to maximize its expected cumulative reward by interacting with an unknown environment over time. Despite significant empirical progress in RL algorithms for various applications [47, 52, 65, 66, 25], the theoretical understanding of RL algorithms had long been limited to tabular methods [40, 56, 10, 77, 79], which explicitly enumerate the entire state and action spaces and learn the value (or the policy) for each state and action. Recently, there has been an increasing body of research in RL with function approximation to extend beyond the tabular problem setting. In particular, *linear function approximation* has served as a foundational model [43, 73, 22, 9, 37]. On the other hand,

---

[*]Equal contribution
[†]Corresponding author

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

the linear transition model assumption poses significant constraints: 1) the output of the function must be within $[0, 1]$, and 2) the sum of the probabilities for all possible next states must be exactly 1. These constraints make it challenging to apply RL with linear function approximation to real-world applications [35]. To overcome such challenges, there has been literature on RL with general function approximation [21, 28, 37, 44, 4, 18]. Despite the guarantee of sample efficiency achieved by their algorithms, this accomplishment might be impeded by computational intractability or the necessity to rely on stronger assumptions. As a result, the resulting methods may not be as general or practical.

On the other hand, Hwang and Oh [35] introduce specific non-linear parametric MDPs called MNL-MDPs (Assumption 1) where the transition probability of MDPs is given by an MNL model. They consider an *upper confidence bound* (UCB) approach to balance exploration and exploitation. Since it is costly or even intractable to compute UCB explicitly, randomized exploration methods such as *Thompson Sampling* (TS) are widely studied in RL with linear function approximation as well as tabular MDPs. This is because, in various decision-making problems ranging from multi-armed bandits to RL, randomized exploration algorithms have been shown to perform better than UCB methods in empirical evaluations [16, 57, 64, 49]. Furthermore, randomized exploration can be easily integrated with linear function approximation. This is because the value function in linear MDPs can be linearly parameterized, allowing perturbations of the estimator to directly control the perturbations of the value function. However, although there has been some literature aiming to propose randomized algorithms for general function classes [37, 4, 5, 75], these methods do not discuss how to define the posterior distribution supported by the given function class and how to draw the optimistic sample from the posterior [4, 5, 75], or they require stronger assumptions on stochastic optimism [37], which is one of the most challenging elements in frequentist regret analysis. Thus, the design of a tractable randomized exploration RL algorithm and the feasibility of frequentist regret analysis for randomized exploration remain open challenges. Hence, the following question arises:

Can we design a provably efficient and tractable randomized algorithm for RL with MNL function approximation?

We answer the above question by proposing the first randomized algorithm, `RRL-MNL`, achieving $\widetilde{\mathcal{O}}(\kappa^{-1} d^{\frac{3}{2}} H^{\frac{3}{2}} \sqrt{T})$ frequentist regret with constant-time computational cost per episode. `RRL-MNL` is not only the first algorithm with randomized exploration for MNL-MDPs, but also, to the best of our knowledge, it provides the first frequentist regret analysis for a *non-linear model-based* algorithm with randomized exploration without assuming stochastic optimism [37].

While `RRL-MNL` is *statistically* efficient, the current method used to analyze the regret of MNL function approximation introduces a problem-dependent constant κ (Assumption 4), which reflects the level of non-linearity of the MNL transition model. This constant κ originates from the use of generalized linear models (GLMs) for contextual bandit settings [26, 51, 45] and MNL bandit settings [54, 17, 55]. The magnitude of the constant κ can be exponentially small with respect to the size of the decision set, hence the regret bound scaling with κ^{-1} could be prohibitively large in the worst case [23]. Worse yet, the situation is even more challenging in RL, as in the worst case, κ^{-1} can be much larger than in the case of bandits. To overcome the prohibitive dependence on κ, algorithms based on new Bernstein-like inequalities and the self-concordant-like property of the log-loss have been proposed for logistic bandits [23, 3, 24] and for MNL bandits [61, 6, 50]. As an extension of these works, the following fundamental question remains open:

Is it possible for RL algorithms with MNL function approximation to have a sharper dependence on the problem-dependent constant κ?

For the above question, we propose the second randomized algorithm referred to as `ORRL-MNL`, which establishes a regret bound of $\widetilde{\mathcal{O}}(d^{\frac{3}{2}} H^{\frac{3}{2}} \sqrt{T} + \kappa^{-1} d^2 H^2)$ with constant-time computational cost per episode. We summarize our main contributions as follows:

- We propose computationally tractable randomized algorithms for RL with MNL function approximation: `RRL-MNL` and `ORRL-MNL`. To the best of our knowledge, these are the first randomized model-based RL algorithms with MNL function approximation that achieve the frequentist regret bounds with constant-time computational cost per episode.

- We establish that `RRL-MNL` enjoys $\widetilde{\mathcal{O}}(\kappa^{-1} d^{\frac{3}{2}} H^{\frac{3}{2}} \sqrt{T})$ frequentist regret bound with constant-time computational cost per episode, where d is the dimension of the transition core, H is horizon length, T is the total number of rounds, and κ is a problem-dependent constant. We

2

derive the stochastic optimism of `RRL-MNL`, and to our knowledge, this is the first frequentist regret analysis for a non-linear model-based algorithm with randomized exploration without assuming stochastic optimism.

- To achieve a regret bound with improved dependence on κ, we introduce `ORRL-MNL`, which constructs the optimistic randomized value functions by taking into account the effects of the local gradient information for the MNL transition model at each reachable state. We prove that `ORRL-MNL` enjoys an $\widetilde{\mathcal{O}}(d^{\frac{3}{2}} H^{\frac{3}{2}} \sqrt{T} + \kappa^{-1} d^2 H^2)$ regret with constant-time computational cost per episode, significantly improving the regret of `RRL-MNL` without requiring prior knowledge of κ.

- We evaluate our algorithms on tabular MDPs and demonstrate the superior performance of our proposed algorithms compared to the existing state-of-the-art MNL-MDP algorithm [35]. The experiments provide evidence that our proposed algorithms are both computationally and statistically efficient.

Related works on RL with function approximation and MNL contextual bandits are provided in Appendix A.

2 Problem Setting

We consider the episodic *Markov decision processes* (MDPs) denoted by $\mathcal{M}(\mathcal{S}, \mathcal{A}, H, \{P\}_{h=1}^H, r)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, H is the horizon length of each episode, $\{P\}_{h=1}^H$ is the collection of probability distributions, and r is the reward function. Every episodes start from the initial state s_1 and for every step $h \in [H] := \{1, ..., H\}$ in an episode, the learning agent interacts with the environment represented as $\mathcal{M}$. The agent observes the state $s_h \in \mathcal{S}$, chooses an action $a_h \in \mathcal{A}$, receives a reward $r(s_h, a_h) \in [0, 1]$ and the next state s_{h+1} is given by the transition probability distribution $P_h(\cdot|s_h, a_h)$. Then this process is repeated throughout the episode. A policy $\pi : \mathcal{S} \times [H] \to \mathcal{A}$ is a function that determines the action of the agent at state s_h, i.e., $a_h = \pi(s_h, h) := \pi_h(s_h)$.

We define the value function of the policy π, denoted by $V_h^\pi(s)$, as the expected sum of rewards under the policy π until the end of the episode starting from $s_h = s$, i.e., $V_h^\pi(s) = \mathbb{E}_\pi \left[\sum_{h'=h}^H r(s_{h'}, \pi_{h'}(s_{h'})) \mid s_h = s \right]$. Similarly, we define the action-value function $Q_h^\pi(s, a) = r(s, a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)} \left[V_{h+1}^\pi(s') \right]$. We define an optimal policy π^* to be a policy that achieves the highest possible value at every $(s, h) \in \mathcal{S} \times [H]$. We denote the optimal value function by $V_h^*(s) = V_h^{\pi^*}(s)$ and the optimal action-value function by $Q_h^*(s, a) = Q_h^{\pi^*}(s, a)$. To simplify, we introduce the notation $P_h V_{h+1}(s, a) = \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[V_{h+1}(s')]$. Recall that the Bellman equations are,
$$Q_h^\pi(s, a) = r(s, a) + P_h V_{h+1}^\pi(s, a), \quad Q_h^*(s, a) = r(s, a) + P_h V_{h+1}^*(s, a),$$
where $V_{H+1}^\pi(s) = V_{H+1}^*(s) = 0$ and $V_h^*(s) = \max_{a \in \mathcal{A}} Q_h^*(s, a)$ for all $s \in \mathcal{S}$.

The goal of the agent is to maximize the sum of rewards for K episodes. In other words, the goal is to minimize the cumulative regret of the policy π over K episodes where $\pi = \{\pi^k\}_{k=1}^K$ is a collection of policies π^k at k-th episode. The regret is defined as

$$\mathbf{Regret}_\pi(K) := \sum_{k=1}^K (V_1^* - V_1^{\pi^k})(s_1^k)$$

where s_1^k is the initial state at the k-th episode.

2.1 Multinomial Logistic Markov Decision Processes (MNL-MDPs)

Even though a lot of provable RL algorithms for linear MDPs are proposed, there is a simple but fundamental problem with the linear transition model assumption on the linear MDPs. In other words, the output of a linear function approximating the transition model must be in $[0, 1]$ and the probability of all possible following states must sum to 1 exactly. Such restrictive assumption can affect the regret performances of algorithm suggested under the linearity assumption. To resolve

3

these challenges, Hwang and Oh [35] propose a setting of a *multinomial logistic Markov decision processes* (MNL-MDPs), where the state transition model is given by a multinomial logistic model. We introduce the formal definition for MNL-MDP as follows:

Assumption 1 (MNL-MDPs [35]). *An MDP $\mathcal{M}(\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^{H}, r)$ is an* MNL-MDP *with a feature map $\boldsymbol{\varphi} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$, if for each $h \in [H]$, there exists $\boldsymbol{\theta}_h^* \in \mathbb{R}^d$, such that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $s' \in \mathcal{S}_{s,a} := \{s' \in \mathcal{S} : \mathbb{P}(s' \mid s, a) \neq 0\}$, the state transition kernel of s' when an action a is taken at a state s is given by,*

$$P_h(s' \mid s, a) = \frac{\exp(\boldsymbol{\varphi}(s, a, s')^\top \boldsymbol{\theta}_h^*)}{\sum_{\widetilde{s} \in \mathcal{S}_{s,a}} \exp(\boldsymbol{\varphi}(s, a, \widetilde{s})^\top \boldsymbol{\theta}_h^*)} \,. \tag{1}$$

We call each unknown vector $\boldsymbol{\theta}_h^*$ transition core. Furthermore, we denote the maximum cardinality of the set of reachable states as $\mathcal{U}$, i.e., $\mathcal{U} := \max_{s,a} |\mathcal{S}_{s,a}|$.

Remark 1. *While Hwang and Oh [35] assume a homogeneous transition kernel, we assume an inhomogeneous transition kernel, in which the probability varies depending on the current time step h even for the same state transition, which is a more general setting. Also, for notational simplicity, we denote the true transition kernel P_h as $P_{\boldsymbol{\theta}_h^*}$, and the estimated transition kernel by $\boldsymbol{\theta}$ as $P_{\boldsymbol{\theta}}$.*

2.2 Assumptions

We introduce some standard regularity assumptions.

Assumption 2 (Boundedness). *We assume $\|\boldsymbol{\varphi}(s, a, s')\|_2 \leq L_{\boldsymbol{\varphi}}$ for all $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}_{s,a}$, and $\|\boldsymbol{\theta}_h^*\|_2 \leq L_{\boldsymbol{\theta}}$ for all $h \in [H]$.*

Assumption 3 (Known reward). *We assume that the reward function r is known to the agent.*

Assumption 4 (Problem-dependent constant). *Let $\mathcal{B}_d(L_{\boldsymbol{\theta}}) := \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta}\|_2 \leq L_{\boldsymbol{\theta}}\}$. There exists $\kappa > 0$ such that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $s', \widetilde{s} \in \mathcal{S}_{s,a}$ with $s' \neq \widetilde{s}$,*

$$\inf_{\boldsymbol{\theta} \in \mathcal{B}_d(L_{\boldsymbol{\theta}})} P_{\boldsymbol{\theta}}(s' \mid s, a) P_{\boldsymbol{\theta}}(\widetilde{s} \mid s, a) \geq \kappa \,.$$

Discussion of assumptions Assumption 2 is common in the literature on RL with function approximation [43, 72, 73, 37, 35] to make the regret bounds scale-free. Assumption 3 is used to focus on the main challenge of model-based RL that learning about P of the environment is more difficult than learning r. In the model-based RL literature [71, 9, 72, 81, 35], the known reward r assumption is widely used. Assumption 4 is typical in generalized linear contextual bandit [26, 51, 23, 3, 24] and MNL contextual bandit literature [54, 8, 55, 61, 6, 76, 50] to guarantee non-singular Fisher information matrix.

3 Randomized Algorithm for MNL-MDPs having constant-time computational cost

Previous work for MNL-MDPs [35] proposed a UCB-based exploration algorithm. Constructing a UCB-based optimistic value function is not only computationally intractable but also tends to overly optimistically estimate the true optimal value function. Additionally, their algorithm incurs increasing computation costs as episodes progress, as it requires all samples from the previous episode to estimate the transition core. In this section, we present a novel model-based RL algorithm that incorporates *randomized exploration* and *online parameter estimation* for MNL-MDPs.

3.1 Algorithm: `RRL-MNL`

Online transition core estimation While Hwang and Oh [35] estimate the transition core using maximum likelihood estimation over all samples from previous episodes, we employ an efficient online parameter estimation method by exploiting the particular structure of the MNL transition model. The key insight is that the negative log-likelihood function for the MNL model in each episode is strongly convex over a bounded domain. This property allows us to utilize a variation of the online Newton step [30, 31], which inspired online algorithms for logistic bandits [74] and MNL contextual bandits [55]. Specifically, for $(k, h) \in [K] \times [H]$, we define the response variable $y_h^k =$

4

Algorithm 1 `RRL-MNL` (Randomized RL for MNL-MDPs)

1: **Inputs:** Episodic MDP $\mathcal{M}$, Feature map $\boldsymbol{\varphi} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$, Number of episodes K, Regularization parameter λ, Exploration variance $\{\sigma_k\}_{k=1}^K$, Sample size M, Problem-dependent constant κ

2: **Initialize:** $\boldsymbol{\theta}_h^1 = \mathbf{0}_d$, $\mathbf{A}_{1,h} = \lambda \mathbf{I}_d$ for $h \in [H]$

3: **for** episode $k = 1, 2, \cdots, K$ **do**

4:     Observe s_1^k and sample *i.i.d.* noise vector $\boldsymbol{\xi}_{k,h}^{(m)} \sim \mathcal{N}(\mathbf{0}_d, \sigma_k^2 \mathbf{A}_{k,h}^{-1})$ for $m \in [M]$ and $h \in [H]$

5:     Set $\left\{ Q_h^k(\cdot, \cdot) \right\}_{h \in [H]}$ as described in (4)

6:     **for** horizon $h = 1, 2, \cdots, H$ **do**

7:         Select $a_h^k = \operatorname{argmax}_{a \in \mathcal{A}} Q_h^k(s_h^k, a)$ and observe s_{h+1}^k

8:         Update $\mathbf{A}_{k+1,h} = \mathbf{A}_{k,h} + \frac{\kappa}{2} \sum_{s' \in \mathcal{S}_{k,h}} \boldsymbol{\varphi}(s_h^k, a_h^k, s') \boldsymbol{\varphi}(s_h^k, a_h^k, s')^\top$ and $\boldsymbol{\theta}_h^{k+1}$ as in (2)

9:     **end for**

10: **end for**

$\left[y_h^k(s') \right]_{s' \in \mathcal{S}_{k,h}}$ such that $y_h^k(s') = \mathbb{1}(s_{h+1}^k = s')$ for $s' \in \mathcal{S}_{k,h} := \mathcal{S}_{s_h^k, a_h^k}$. Then, y_h^k is sampled from the following multinomial distribution: $y_h^k \sim \operatorname{multinomial}(1, \left[P_{\boldsymbol{\theta}_h^*}(s' \mid s_h^k, a_h^k) \right]_{s' \in \mathcal{S}_{k,h}})$, where 1 represents that y_h^k is a single-trial sample. We define the per-episode loss $\ell_{k,h}(\boldsymbol{\theta})$ as follows:

$$\ell_{k,h}(\boldsymbol{\theta}) := - \sum_{s' \in \mathcal{S}_{k,h}} y_h^k(s') \log P_{\boldsymbol{\theta}}(s' \mid s_h^k, a_h^k) .$$

Then, the estimated transition core for $\boldsymbol{\theta}_h^*$ is given by

$$\boldsymbol{\theta}_h^k = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathcal{B}_d(L_{\boldsymbol{\theta}})} \frac{1}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_h^{k-1}\|_{\mathbf{A}_{k,h}}^2 + (\boldsymbol{\theta} - \boldsymbol{\theta}_h^{k-1})^\top \nabla \ell_{k-1,h}(\boldsymbol{\theta}_h^{k-1}) , \qquad (2)$$

where $\boldsymbol{\theta}_h^1$ can be initialized as any point in $\mathcal{B}_d(L_{\boldsymbol{\theta}})$ and $\mathbf{A}_{k,h}$ is the Gram matrix defined by

$$\mathbf{A}_{k,h} := \lambda \mathbf{I}_d + \frac{\kappa}{2} \sum_{i=1}^{k-1} \sum_{s' \in \mathcal{S}_{i,h}} \boldsymbol{\varphi}(s_h^i, a_h^i, s') \boldsymbol{\varphi}(s_h^i, a_h^i, s')^\top . \qquad (3)$$

Stochastically optimistic value function First of all, we introduce the key challenges of regret analysis for randomized algorithms, explain how previous works have overcome these challenges, and then describe why the techniques from previous works cannot be applied to MNL-MDPs. Ensuring that the estimated value function is optimistic with sufficient frequency is a crucial challenge in analyzing the frequentist regret of randomized algorithms. A common way to promote sufficient exploration in randomized algorithms is by perturbing the estimated value function or by performing posterior sampling in the transition model class. Frequentist regret analysis of randomized exploration in an RL setting has been conducted for tabular [59, 7, 62, 60, 67], linear MDPs [73, 37], and general function classes [37, 4, 5, 75]. In the case of linear MDPs [73, 37], since the property that the action-value function is linear in the feature map allows perturbing the estimated parameter directly to control the perturbation of the estimated value function. Also, even though Ishfaq et al. [37] presented a randomized algorithm for the general function class using eluder dimension, they assume stochastic optimism (anti-concentration), which is in fact one of the most challenging aspects of frequentist analysis. Other posterior sampling algorithms in RL for the general function class such as [4, 5, 75], except for very limited examples, do not discuss how to define the posterior distribution supported by the given function class and how to draw the optimistic sample from the posterior. That is why even after there exists a so-called *general function class*-based result, it is often the case that results in specific parametric models are still needed.

Note that in episodic RL, the perturbed estimated value functions are propagated back through horizontal steps, requiring careful adjustment of the perturbation scheme to maintain a sufficient probability of optimism without decaying too quickly with the horizon. For example, if the probability of the estimated value function being optimistic at horizon h is denoted as p, this would result in the probability that the estimated value function in the initial state is optimistic being on the order of p^H, implying that the regret can increase exponentially with the length of the horizon H.

5

Additionally, the non-linearity and substitution effect of the next state transition in the MNL-MDPs make applying the existing TS techniques infeasible to guarantee optimism in MNL-MDPs with sufficient frequency. Instead, we design the *stochastically optimistic value function* by exploiting the structure of the MNL transition model. In other words, the prediction error of MNL transition model (Definition 1) can be bounded by the weighted norm of the dominant feature $\hat{\varphi}$ (Lemma 4). Based on such dominant feature, we perturb the estimated value function by injecting Gaussian noise whose variance is proportional to the inverse of the Gram matrix to encourage the perturbation with higher variance in less explored directions. To guarantee the optimism with fixed probability, we adapt optimistic sampling technique [7, 54, 37, 36]. For each $m \in [M]$, sample *i.i.d.* Gaussian noise vector $\boldsymbol{\xi}_{k,h}^{(m)} \sim \mathcal{N}(\mathbf{0}_d, \sigma_k^2 \mathbf{A}_{k,h}^{-1})$ where σ_k is an exploration parameter, and add the most optimistic inner product value $\max_{m \in [M]} \hat{\boldsymbol{\varphi}}_{k,h}(s,a)^\top \boldsymbol{\xi}_{k,h}^{(m)}$ to the estimated value function. To summarize for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, set $Q_{H+1}^k(s,a) = 0$ and for $h \in [H]$,

$$Q_h^k(s,a) = \min\left\{ r(s,a) + \sum_{s' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}_h^k}(s' \mid s,a) V_{h+1}^k(s') + \max_{m \in [M]} \hat{\boldsymbol{\varphi}}_{k,h}(s,a)^\top \boldsymbol{\xi}_{k,h}^{(m)}, H \right\}, \quad (4)$$

where $V_h^k(s) = \max_{a'} Q_h^k(s,a')$ and $\hat{\boldsymbol{\varphi}}_{k,h}(s,a) := \boldsymbol{\varphi}(s,a,\hat{s})$ for $\hat{s} = \operatorname{argmax}_{s' \in \mathcal{S}_{s,a}} \|\boldsymbol{\varphi}(s,a,s')\|_{\mathbf{A}_{k,h}^{-1}}$. Based on these stochastically optimistic value function, the agent plays a greedy action $a_h^k = \operatorname{argmax}_{a'} Q_h^k(s_h^k, a')$. We layout the procedure in Algorithm 1.

Remark 2. *Note that* RRL-MNL *only requires constant-time computational cost and storage cost per episode, as it does not require storing all samples from previous episodes, and the Gram matrix $\mathbf{A}_{k,h}$ can be updated incrementally.*

3.2 Regret bound of RRL-MNL

We present the regret upper bound of RRL-MNL. The complete proof is deferred to Appendix C.

Theorem 1 (Regret Bound of RRL-MNL). *Suppose that Assumption 1-4 hold. For any $0 < \delta < \frac{\Phi(-1)}{2}$, if we set the input parameters in Algorithm 1 as $\lambda = L_\varphi^2, \sigma_k = \widetilde{\mathcal{O}}(H\sqrt{d})$ and $M = \lceil 1 - \frac{\log H}{\log \Phi(1)} \rceil$ where Φ is the normal CDF, then with probability at least $1 - \delta$, the cumulative regret of the RRL-MNL policy π is upper-bounded as follows:*

$$\boldsymbol{Regret}_\pi(K) = \widetilde{\mathcal{O}}\left(\kappa^{-1} d^{\frac{3}{2}} H^{\frac{3}{2}} \sqrt{T} \right),$$

where $T = KH$ is the total number of steps.

Discussion of Theorem 1   To our best knowledge, this is the first result to provide a frequentist regret bound for the MNL-MDPs. Among the previous RL algorithms using function approximation, the most comparable techniques to our method are *model-free* algorithms with randomized exploration [73, 37]. To guarantee stochastic optimism, Zanette et al. [73] established a lower bound on the difference between the estimated value and the optimal value by the summation of linear terms with respect to the average feature (Lemma F.1 in [73]). This property is achievable due to the linear expression of the value function in linear MDPs. Instead, we established a lower bound on the difference between value functions by the summation of the Bellman errors (Definition 1) along the sample path obtained through the optimal policy (Lemma 7). Hence, our analysis significantly differs from that of Zanette et al. [73] since the value function in MNL-MDPs is no longer linearly parametrized, and there is no closed-form expression for it.

Compared to [37], they also used an optimistic sampling technique; however, our theoretical sampling size $M = \mathcal{O}(\log H)$ is much tighter than that of [37], i.e., $\mathcal{O}(d)$ for the linear function class, $\mathcal{O}(\log(T|\mathcal{S}||\mathcal{A}|))$ for the general function class. While Ishfaq et al. [37] extend the results of the linear function class to general function class under the assumption of stochastic optimism (Assumption C in [37]), we provide the frequentist regret analysis for a *non-linear model-based* algorithm with randomized exploration *without assuming stochastic optimism*.

Compared to the optimistic exploration algorithm for MNL-MDPs [35], our randomized exploration requires a more involved proof technique to ensure that the perturbation of the estimated value function has enough variance to maintain optimism with sufficient frequency (Lemma 6). As a result,

6

the established regret of RRL-MNL differs by a factor of $\sqrt{d}$, which aligns with the difference in the existing bounds of linear bandits between a TS-based algorithm [2] and a UCB-based algorithm [1]. Additionally, we achieve statistical efficiency for the *inhomogeneous transition model*, which is a more general setting than that of Hwang and Oh [35]. Our computation cost per episode is $\mathcal{O}(1)$ while the computation cost per episode of Hwang and Oh [35] is $\mathcal{O}(K)$.

Proof Sketch of Theorem 1 We provide the proof sketch of Theorem 1. By decomposing the regret into the estimation part and the pessimism part, we have

$$\sum_{k=1}^{K}(V_1^* - V_1^{\pi_k})(s_1^k) = \sum_{k=1}^{K}\Big(\underbrace{V_1^* - V_1^k}_{\text{Pessimism}} + \underbrace{V_1^k - V_1^{\pi_k}}_{\text{Estimation}}\Big)(s_1^k)\,.$$

We bound these two parts separately. For the estimation part, for each $k \in [K], h \in [H]$, we first show that the online estimated transition core $\boldsymbol{\theta}_h^k$ (2) concentrates around the unknown transition core parameter $\boldsymbol{\theta}_h^*$ with high probability (Lemma 1). Then, we show that the prediction error induced by the estimated transition core can be bounded by the weighted norm of the dominant feature $\tilde{\boldsymbol{\varphi}}$, multiplied by the confidence radius of the estimated transition core (Lemma 4). The bounded prediction error, together with the concentration of Gaussian noise, implies the desired bound on the estimation part (Lemma 10). For the pessimism part, we first show that the stochastically optimistic value function V_1^k is optimistic than the true optimal value function V_1^* with sufficient frequency (Lemma 6). In the next step, we show that the pessimism part is upper bounded by a bound of the estimation part times the inverse probability of being optimistic (Lemma 11). Combining all the results, we can conclude the proof. Refer to Appendix C for detailed proofs.

4 Statistically Improved Algorithm for MNL-MDPs

Although RRL-MNL is provably efficient and achieves constant-time computational cost per episode, the current analysis makes its regret bound scale with κ^{-1}. Recall that the problem-dependent constant κ introduced in Assumption 4 indicates the curvature of the MNL function, i.e., how difficult it is to learn the true transition core parameter. It is required to ensure the non-singular Fisher information matrix, hence is typically used in GLM or MNL bandit algorithms that use the maximum likelihood estimator. As introduced in Faury et al. [23], κ^{-1} can be exponentially large in the worst case. The appearance of κ in existing bounds originates in the connection between the difference of estimators and the difference of gradients of negative log-likelihood, usually denoted as $\mathbf{G}$ in Filippi et al. [26]. Without considering local information at all, using a loose lower bound for $\mathbf{G}$ incurs κ^{-1} in regret bound (see Section 4.1 in Agrawal et al. [6]). Recently, improved dependence on κ has been achieved in bandit literature [23, 3, 61, 6, 76, 50] through the use of generalization of the Bernstein-like tail inequality [23] and the self-concordant-like property of the log loss [11]. However, a direct adaptation of the MNL bandit technique would result in sub-optimal dependence on the assortment size in MNL bandit, which corresponds to the size of the set of reachable states, such as $\mathcal{U}$. In this section, we introduce a new randomized algorithm for MNL-MDPs, equipped with a tight online parameter estimation and feature centralization technique that achieves a regret bound with improved dependence on κ and $\mathcal{U}$.

4.1 Algorithms: ORRL-MNL

Tight online transition core estimation Zhang and Sugiyama [76] presented a jointly efficient UCB-based MNL contextual bandit algorithm using online mirror descent algorithm. Adapting the update rule from [76], the estimated transition core run by the online mirror descent is given by

$$\widetilde{\boldsymbol{\theta}}_h^{k+1} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathcal{B}_d(L_{\boldsymbol{\theta}})} \frac{1}{2\eta}\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^k\|_{\widetilde{\mathbf{B}}_{k,h}}^2 + \boldsymbol{\theta}^\top \nabla \ell_{k,h}(\widetilde{\boldsymbol{\theta}}_h^k)\,, \tag{5}$$

where $\widetilde{\boldsymbol{\theta}}_h^1$ can be initialized as any point in $\mathcal{B}_d(L_{\boldsymbol{\theta}})$, η is a step size, and $\widetilde{\mathbf{B}}_{k,h}$ is defined as

$$\widetilde{\mathbf{B}}_{k,h} := \mathbf{B}_{k,h} + \eta \nabla^2 \ell_{k,h}(\widetilde{\boldsymbol{\theta}}_h^k)\,, \quad \mathbf{B}_{k,h} := \lambda \mathbf{I}_d + \sum_{i=1}^{k-1} \nabla^2 \ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1})\,. \tag{6}$$

7

Algorithm 2 `ORRL-MNL` (Optimistic Randomized RL for MNL-MDPs)

1: **Inputs:** Episodic MDP $\mathcal{M}$, Feature map $\boldsymbol{\varphi} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}^d$, Number of episodes K, Regularization parameter λ, Exploration variance $\{\sigma_k\}_{k=1}^K$, Confidence radius $\{\beta_k\}_{k=1}^K$, Sample size M, Step size η

2: **Initialize:** $\widetilde{\boldsymbol{\theta}}_h^1 = \mathbf{0}_d$, $\mathbf{B}_{1,h} = \lambda \mathbf{I}_d$ for all $h \in [H]$

3: **for** episode $k = 1, 2, \cdots, K$ **do**

4:      Observe s_1^k and sample *i.i.d.* noise vector $\boldsymbol{\xi}_{k,h}^{(m)} \sim \mathcal{N}(\mathbf{0}_d, \sigma_k^2 \mathbf{B}_{k,h}^{-1})$ for $m \in [M]$ and $h \in [H]$

5:      Set $\left\{ \widetilde{Q}_h^k(\cdot, \cdot) \right\}_{h \in [H]}$ as described in (7)

6:      **for** horizon $h = 1, 2, \cdots, H$ **do**

7:          Select $a_h^k = \operatorname{argmax}_{a \in \mathcal{A}} \widetilde{Q}_h^k(s_h^k, a)$ and observe s_{h+1}^k

8:          Update $\widetilde{\mathbf{B}}_{k,h} = \mathbf{B}_{k,h} + \eta \nabla^2 \ell_{k,h}(\widetilde{\boldsymbol{\theta}}_h^k)$ and $\widetilde{\boldsymbol{\theta}}_h^{k+1}$ as in (5)

9:          Update $\mathbf{B}_{k+1,h} = \mathbf{B}_{k,h} + \nabla^2 \ell_{k,h}(\widetilde{\boldsymbol{\theta}}_h^{k+1})$

10:      **end for**

11: **end for**

Note that the MNL model in Zhang and Sugiyama [76] operates in a *multiple-parameter* setting, where there are multiple unknown choice parameters and one given context feature. In contrast, our MNL model operates in a *single-parameter* setting, where there is one unknown transition core and features for up to $\mathcal{U}$ reachable states. This difference results in variations in applying the self-concordant-like property of the log-loss for the MNL model. For instance, Zhang and Sugiyama [76] utilized the fact that the log-loss for the multiple parameter MNL model is $\sqrt{6}$-self-concordant-like (Lemma 2 in Zhang and Sugiyama [76]). On the other hand, Lee and Oh [50] revisit the self-concordant-like property and demonstrate that the log-loss of the single-parameter MNL model is $3\sqrt{2}$-self-concordant-like (Proposition B.1 in Lee and Oh [50]). This results in a concentration bound that is independent of κ and $\mathcal{U}$, introduced in Lemma 12.

Remark 3. *Note that the online estimated parameters $\boldsymbol{\theta}_h^k$ (2) and $\widetilde{\boldsymbol{\theta}}_h^k$ (5) do not aim to minimize the sum of negative log-likelihoods, $\sum_{k'=1}^k \ell_{k',h}(\boldsymbol{\theta})$. Instead, we show that the online estimated parameter concentrates around the unknown transition core $\boldsymbol{\theta}_h^*$ with high probability (Lemma 1 & 12). This online update approach allows us to estimate the transition core with constant-time computational cost per episode, as the agent does not need to store all samples from previous episodes.*

Optimistic randomized value function   To achieve improved dependence on κ, a crucial point is to utilize the local gradient information of MNL transition probabilities for each reachable state when constructing the Gram matrix. In MNL bandit problems [61, 76], this can be accomplished by substituting the Hessian of the negative log-likelihood with the Gram matrix using global gradient information κ. However, there are fundamental differences between the settings in Perivier and Goyal [61], Zhang and Sugiyama [76] and ours. Perivier and Goyal [61] address the case where the reward for each product is *uniform* (i.e., all products have a reward of 1), and the reward for not selecting a product from the given assortment (also known as the outside option) is 0. On the other hand, Zhang and Sugiyama [76] deal with *non-uniform* rewards where the reward for each product may vary; however, the rewards for individual products are known a priori to the agent. In contrast, in MNL-MDPs, the value for each reachable state may vary (non-uniform) and is *not known* beforehand. Due to these differences, the analysis techniques in MNL bandits [61, 76] cannot be directly applied to our setting. Instead, we adapt the feature centralization technique [50]. Then, the Hessian of the per-round loss $\ell_{k,h}(\boldsymbol{\theta})$ is expressed in terms of the centralized feature as follows:

$$\nabla^2 \ell_{k,h}(\boldsymbol{\theta}) = \sum_{s' \in \mathcal{S}_{k,h}} P_{\boldsymbol{\theta}}(s' \mid s_h^k, a_h^k) \bar{\boldsymbol{\varphi}}(s_h^k, a_h^k, s'; \boldsymbol{\theta}) \bar{\boldsymbol{\varphi}}(s_h^k, a_h^k, s'; \boldsymbol{\theta})^\top .$$

where $\bar{\boldsymbol{\varphi}}(s, a, s'; \boldsymbol{\theta}) := \boldsymbol{\varphi}(s, a, s') - \mathbb{E}_{\widetilde{s} \sim P_{\boldsymbol{\theta}}(\cdot | s, a)}[\boldsymbol{\varphi}(s, a, \widetilde{s})]$ is the centralized feature by $\boldsymbol{\theta}$. For more details, please refer to Appendix D.2.

Now we introduce the *optimistic randomized value function* $\widetilde{Q}_h^k(\cdot, \cdot)$ for `ORRL-MNL`. The key point is that when perturbing the estimated value function, we use the centralized feature by the estimated

8

transition parameter $\widetilde{\boldsymbol{\theta}}_h^k$. For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, set $\widetilde{Q}_{H+1}^k(s,a) = 0$ and for each $h \in [H]$,

$$\widetilde{Q}_h^k(s,a) := \min\left\{ r(s,a) + \sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s,a) \widetilde{V}_{h+1}^k(s') + \nu_{k,h}^{\mathrm{rand}}(s,a), H \right\}, \tag{7}$$

where $\widetilde{V}_h^k(s) := \max_{a \in \mathcal{A}} \widetilde{Q}_h^k(s,a)$ and $\nu_{k,h}^{\mathrm{rand}}(s,a)$ is the *randomized bonus term* defined by

$$\nu_{k,h}^{\mathrm{rand}}(s,a) := \sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s,a) \bar{\boldsymbol{\varphi}}(s,a,s';\widetilde{\boldsymbol{\theta}}_h^k)^\top \boldsymbol{\xi}_{k,h}^{s'} + 3H\beta_k^2 \max_{s' \in \mathcal{S}_{s,a}} \|\boldsymbol{\varphi}(s,a,s')\|_{\mathbf{B}_{k,h}^{-1}}^2.$$

Here we sample *i.i.d.* Gaussian noise $\boldsymbol{\xi}_{k,h}^{(m)} \sim \mathcal{N}(\mathbf{0}_d, \sigma_k^2 \mathbf{B}_{k,h}^{-1})$ for each $m \in [M]$ and set $\boldsymbol{\xi}_{k,h}^{s'} := \boldsymbol{\xi}_{k,h}^{m(s')}$ where $m(s') := \mathrm{argmax}_{m \in [M]} \bar{\boldsymbol{\varphi}}(s,a,s';\boldsymbol{\theta}_h^k)^\top \boldsymbol{\xi}_{k,h}^m$ is the most optimistic sampling index for a reachable state s'. Based on these optimistic randomized value function, at each episode the agent plays a greedy action with respect to $\widetilde{Q}_h^k$ as summarized in Algorithm 2.

Remark 4. *Note that the second term in the randomized bonus always has a positive value, but it rapidly decreases as episode proceeds. While due to the randomness of $\boldsymbol{\xi}$, the randomized bonus $\nu_{k,h}^{\mathrm{rand}}$ itself cannot be guaranteed to always have a positive value. Consequently, the constructed value function $\widetilde{Q}_h^k(\cdot, \cdot)$ can be optimistic or pessimistic. However, as shown in Lemma 18, optimistic sampling technique ensures that the optimistic randomized value function $\widetilde{Q}_h^k$ has at least a constant probability of being optimistic than the true optimal value function.*

Remark 5. *As with RRL-MNL, since the transition core is estimated in an online manner and the Gram matrices with local gradient information $\mathbf{B}_{k,h}$ and $\widetilde{\mathbf{B}}_{k,h}$ are updated incrementally, ORRL-MNL also requires constant-time computational cost and storage cost per-episode. Although ORRL-MNL requires an additional $\mathcal{O}(\mathcal{U})$ computation cost for feature centralization, the computation complexity order is the same as that of RRL-MNL because it also needs to go over reachable states to calculate the dominant feature $\hat{\boldsymbol{\varphi}}$, which also incurs a $\mathcal{O}(\mathcal{U})$ computation cost. On the other hand, ORRL-MNL does not require prior knowledge of κ and achieves a regret with a better dependence on κ.*

4.2 Regret Bound of ORRL-MNL

We present the regret upper bound of ORRL-MNL. The complete proof is deferred to Appendix D.

Theorem 2 (Regret Bound of ORRL-MNL). *Suppose that Assumption 1- 4 hold. For any $0 < \delta < \frac{\Phi(-1)}{2}$, if we set the input parameters in Algorithm 2 as $\lambda = \mathcal{O}(L_{\boldsymbol{\varphi}}^2 d \log \mathcal{U}), \beta_k = \mathcal{O}(\sqrt{d} \log \mathcal{U} \log(kH)), \sigma_k = H\beta_k, M = \lceil 1 - \frac{\log(H\mathcal{U})}{\log \Phi(1)} \rceil$, and $\eta = \mathcal{O}(\log \mathcal{U})$, then with probability at least $1 - \delta$, the cumulative regret of the ORRL-MNL policy π is upper-bounded as follows:*

$$\boldsymbol{Regret}_\pi(K) = \widetilde{\mathcal{O}}\left(d^{3/2} H^{3/2} \sqrt{T} + \kappa^{-1} d^2 H^2 \right),$$

where $T = KH$ is the total number of time steps.

Dicussion of Theorem 2 Theorem 2 establishes that the leading term in the regret bound does not suffer from the problem-dependent constant κ^{-1} and the second term of the regret bound is independent of the size of set of reachable states. To the extent of our knowledge, this is the first algorithm that provides a frequentist regret guarantee with improved dependence on κ^{-1} in MNL-MDPs. Compared to RRL-MNL, the technical challenge lies in ensuring the stochastic optimism of the estimated value for ORRL-MNL. Note that the prediction error (Definition 1) for ORRL-MNL is characterized by two components: one related to the gradient information of the MNL transition model at each reachable state, and the other related to the dominant feature with respect to the Gram matrix $\mathbf{B}_{k,h}$ (Lemma 16). Hence, the probability of the Bellman error at each horizon, when following the optimal policy, being negative can depend on the size of the reachable states. This implies that the probability of stochastic optimism can be exponentially small, not only in the horizon H but also in the size of the reachable states $\mathcal{U}$. However, as shown in Lemma 18, this challenge has been overcome by using a sample size M that *logarithmically* increases with $\mathcal{U}$, effectively addressing the issue.

9

<div align="center">

(a) $S = 4, H = 12$ (b) $S = 8, H = 24$ (c) Runtime for 1,000 episodes
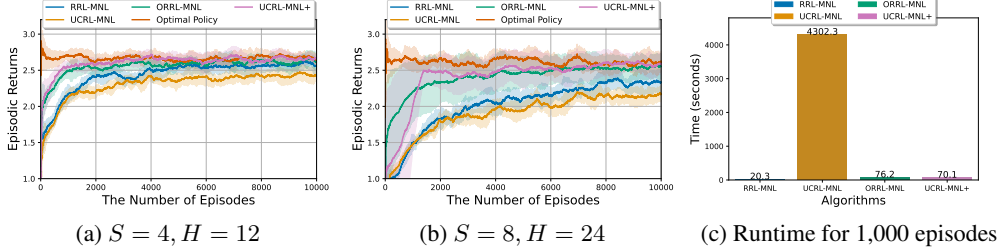
Figure 1: Riverswim experiment results

</div>

Proof Sketch of Theorem 2 The overall proof pipeline for Theorem 2 is similar to that of Theorem 1. The main differences lie in the concentration of the estimated transition core (Lemma D.2), the bound on the prediction error (Lemma D.2), and the stochastic optimism (Lemma 18). Please refer to Appendix D for detailed proofs.

Optimistic exploration extension In general, since TS-based randomized exploration requires a more rigorous proof technique than UCB-based algorithms, our technical ingredients enable the use of optimistic exploration in a straightforward manner. We introduce `UCRL-MNL+` (Algorithm 3) in the Appendix E, an optimism-based algorithm for MNL-MDPs. It is both *computationally* and *statistically* efficient compared to `UCRL-MNL` [35], achieving *the tightest regret bound* for MNL-MDPs.

Corollary 1. *`UCRL-MNL+` (Algorithm 3) has $\widetilde{\mathcal{O}}(dH^{3/2}\sqrt{T} + \kappa^{-1}d^2H^2)$ regret with high probability.*

5   Numerical Experiments

We perform a numerical evaluation on a variant of RiverSwim [58] to demonstrate practicality of our proposed algorithms. We compare our algorithms (`RRL-MNL`, `ORRL-MNL`, `UCRL-MNL+`) with the state-of-the-art `UCRL-MNL` [35] for MNL-MDPs. For each configuration, we report the averaged results over 10 independent runs. Figure 1a and 1b show the episodic return of each algorithm, which is the sum of all the rewards obtained in one episode. First, our proposed algorithms (`RRL-MNL`, `ORRL-MNL`, `UCRL-MNL+`) outperform `UCRL-MNL` [35] for both cases of $|\mathcal{S}| = 4, 8$. Second, `ORRL-MNL` and `UCRL-MNL+` reach the optimal values quickly compared to the other algorithms, demonstrating improved statistical efficiency. Figure 1c illustrates the comparison in running time of the algorithms for the first 1,000 episodes. Our proposed algorithms are at least 50 times faster than `UCRL-MNL`. These differences become more pronounced as the episodes progress because our algorithms have a constant computation cost, whereas the computation cost of `UCRL-MNL` increases over time.

6   Conclusions

We propose randomized algorithms with provable efficiency and constant-time computational cost for MNL-MDPs. For the first algorithm, `RRL-MNL`, we use an optimistic sampling technique to ensure the stochastic optimism of the estimated value functions and provide the frequentist regret analysis. This is the first frequentist regret analysis for a non-linear model-based algorithm with randomized exploration without assuming stochastic optimism. To achieve a statistically improved regret bound, we propose `ORRL-MNL` by constructing the optimistic randomized value function using the effects of the local gradient of the MNL transition model equipped with the centralized feature. As a result, we achieve a frequentist regret guarantee with improved dependence on κ in RL with the MNL transition model, which is a significant contribution. The effectiveness and practicality of our methods are supported by numerical experiments.

10

Acknowledgements

We sincerely thank the anonymous reviewers for their constructive feedback. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2022R1C1C1006859, 2022R1A4A1030579, and RS-2023-00222663) and by AI-Bio Research Grant through Seoul National University.

References

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

[2] Marc Abeille and Alessandro Lazaric. Linear Thompson Sampling Revisited. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 176–184. PMLR, PMLR, 20–22 Apr 2017.

[3] Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3691–3699. PMLR, 2021.

[4] Alekh Agarwal and Tong Zhang. Model-based rl with optimistic posterior sampling: Structural conditions and sample complexity. *Advances in Neural Information Processing Systems*, 35: 35284–35297, 2022.

[5] Alekh Agarwal and Tong Zhang. Non-linear reinforcement learning in large action spaces: Structural conditions and sample-efficiency of posterior sampling. In *Conference on Learning Theory*, pages 2776–2814. PMLR, 2022.

[6] Priyank Agrawal, Theja Tulabandhula, and Vashist Avadhanula. A tractable online learning algorithm for the multinomial logit contextual bandit. *European Journal of Operational Research*, 2023.

[7] Shipra Agrawal and Randy Jia. Posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pages 1184–1194, 2017.

[8] Sanae Amani and Christos Thrampoulidis. Ucb-based algorithms for multinomial logistic regression bandits. *Advances in Neural Information Processing Systems*, 34:2913–2924, 2021.

[9] Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

[10] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

[11] Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4(2):384 – 414, 2010.

[12] Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 2005.

[13] Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1):33–57, 1996.

[14] Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

[15] Nicolo Campolongo and Francesco Orabona. Temporal variability in implicit online learning. *Advances in neural information processing systems*, 33:12377–12387, 2020.

11

[16] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

[17] Xi Chen, Yining Wang, and Yuan Zhou. Dynamic assortment optimization with changing contextual information. *Journal of machine learning research*, 2020.

[18] Zixiang Chen, Chris Junchi Li, Huizhuo Yuan, Quanquan Gu, and Michael Jordan. A general framework for sample-efficient function approximation in reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023.

[19] Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient pac rl with rich observations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[20] Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674. PMLR, 2019.

[21] Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

[22] Simon S. Du, Sham M. Kakade, Ruosong Wang, and Lin F. Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[23] Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR, 2020.

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

[25] 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.

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

[27] Dylan J Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. In *Conference On Learning Theory*, pages 167–208. PMLR, 2018.

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

[29] David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[30] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

[31] Elad Hazan, Tomer Koren, and Kfir Y Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *Conference on Learning Theory*, pages 197–209. PMLR, 2014.

[32] Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

12

[33] Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 4171–4180. PMLR, 2021.

[34] Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for linear markov decision processes. In *International Conference on Machine Learning*, pages 12790–12822. PMLR, 2023.

[35] Taehyun Hwang and Min-hwan Oh. Model-based reinforcement learning with multinomial logistic function approximation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7971–7979, 2023.

[36] Taehyun Hwang, Kyuwook Chai, and Min-Hwan Oh. Combinatorial neural bandits. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 2023.

[37] Haque Ishfaq, Qiwen Cui, Viet Nguyen, Alex Ayoub, Zhuoran Yang, Zhaoran Wang, Doina Precup, and Lin Yang. Randomized exploration in reinforcement learning with general value function approximation. In *International Conference on Machine Learning*, volume 139, pages 4607–4616. PMLR, PMLR, 2021.

[38] Haque Ishfaq, Qingfeng Lan, Pan Xu, A. Rupam Mahmood, Doina Precup, Anima Anandkumar, and Kamyar Azizzadenesheli. Provable and practical: Efficient exploration in reinforcement learning via langevin monte carlo. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=nfIAEJFiBZ`.

[39] Haque Ishfaq, Yixin Tan, Yu Yang, Qingfeng Lan, Jianfeng Lu, A Rupam Mahmood, Doina Precup, and Pan Xu. More efficient randomized exploration for reinforcement learning via approximate sampling. *Reinforcement Learning Journal*, 3(1), 2024.

[40] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

[41] Zeyu Jia, Lin Yang, Csaba Szepesvari, and Mengdi Wang. Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*, pages 666–686. PMLR, 2020.

[42] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

[43] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

[44] Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

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

[46] Yeoneung Kim, Insoon Yang, and Kwang-Sung Jun. Improved regret analysis for variance-adaptive linear bandits and horizon-free linear mixture mdps. *Advances in Neural Information Processing Systems*, 35:1060–1072, 2022.

[47] Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

[48] Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Pac reinforcement learning with rich observations. *Advances in Neural Information Processing Systems*, 29:1840–1848, 2016.

[49] Branislav Kveton, Csaba Szepesvári, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic linear bandits. In *Uncertainty in Artificial Intelligence*, pages 530–540. PMLR, 2020.

13

[50] Joongkyu Lee and Min-hwan Oh. Nearly minimax optimal regret for multinomial logistic bandit. *arXiv preprint arXiv:2405.09831*, 2024.

[51] Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR, 2017.

[52] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[53] Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020.

[54] Min-hwan Oh and Garud Iyengar. Thompson sampling for multinomial logit contextual bandits. *Advances in Neural Information Processing Systems*, 32:3151–3161, 2019.

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

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

[57] Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*, pages 2701–2710. PMLR, 2017.

[58] Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

[59] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pages 2377–2386. PMLR, 2016.

[60] Aldo Pacchiano, Philip Ball, Jack Parker-Holder, Krzysztof Choromanski, and Stephen Roberts. Towards tractable optimism in model-based reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 1413–1423. PMLR, 2021.

[61] Noemie Perivier and Vineet Goyal. Dynamic pricing and assortment under a contextual mnl demand. *Advances in Neural Information Processing Systems*, 35:3461–3474, 2022.

[62] Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. *Advances in Neural Information Processing Systems*, 32, 2019.

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

[64] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

[65] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[66] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

[67] Daniil Tiapkin, Denis Belomestny, Daniele Calandriello, Eric Moulines, Remi Munos, Alexey Naumov, Mark Rowland, Michal Valko, and Pierre Ménard. Optimistic posterior sampling for reinforcement learning with few samples and tight guarantees. *Advances in Neural Information Processing Systems*, 35:10737–10751, 2022.

14

[68] Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33, 2020.

[69] Yining Wang, Ruosong Wang, Simon Shaolei Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[70] Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pages 1237–1264. PMLR, 2021.

[71] Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

[72] Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

[73] Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR, 2020.

[74] Lijun Zhang, Tianbao Yang, Rong Jin, Yichi Xiao, and Zhi-Hua Zhou. Online stochastic linear optimization under one-bit feedback. In *International Conference on Machine Learning*, pages 392–401. PMLR, 2016.

[75] Tong Zhang. Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022.

[76] Yu-Jie Zhang and Masashi Sugiyama. Online (multinomial) logistic bandit: Improved regret and constant computation cost. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[77] Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. In *Advances in Neural Information Processing Systems*, volume 33, pages 15198–15207, 2020.

[78] Zihan Zhang, Jiaqi Yang, Xiangyang Ji, and Simon S Du. Improved variance-aware confidence sets for linear bandits and linear mixture mdp. *Advances in Neural Information Processing Systems*, 34:4342–4355, 2021.

[79] Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. In *International Conference on Machine Learning*, pages 12653–12662. PMLR, 2021.

[80] Dongruo Zhou and Quanquan Gu. Computationally efficient horizon-free reinforcement learning for linear mixture mdps. *Advances in neural information processing systems*, 35:36337–36349, 2022.

[81] Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

[82] Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pages 12793–12802. PMLR, 2021.

15

Contents of Appendix

A Related Work 16

B Notations & Definitions 18

C Detailed Regret Analysis for RRL-MNL (Theorem 1) 22

    C.1 Concentration of Estimated Transition Core $\boldsymbol{\theta}_h^k$ 23

    C.2 Bound on Prediction Error 32

    C.3 Good Events with High Probability 33

    C.4 Stochastic Optimism 33

    C.5 Bound on Estimation Part 36

    C.6 Bound on Pessimism Part 38

    C.7 Regret Bound of RRL-MNL 41

D Detailed Regret Analysis for ORRL-MNL (Theorem 2) 42

    D.1 Concentration of Estimated Transition Core $\widetilde{\boldsymbol{\theta}}_h^k$ 42

    D.2 Bound on Prediction Error 50

    D.3 Good Events with High Probability 55

    D.4 Stochastic Optimism 55

    D.5 Bound on Estimation Part 57

    D.6 Bound on Pessimism Part 66

    D.7 Regret Bound of ORRL-MNL 67

E Optimistic Exploration Extension 67

    E.1 Optimism 69

F Experiment Details 70

G Auxiliary Lemmas 70

H Limitations 72

A Related Work

RL with linear function approximation There has been a growing interest in studies that extend beyond tabular MDPs and focus on function approximation methods with provable guarantees [42, 71, 43, 73, 53, 22, 14, 9, 68, 70, 33, 81, 82, 37, 35, 38]. In particular, for minimizing regret in linear MDPs, Jin et al. [43] propose an optimistic variant of the Least-Squares Value Iteration (LSVI) algorithm [13, 59] under the assumption that the transition model and reward function of the MDPs are linear function of a d-dimensional feature mapping and they guarantee $\widetilde{\mathcal{O}}(d^{\frac{3}{2}} H^{\frac{3}{2}} \sqrt{T})$ regret. Zanette et al. [73] propose a randomized LSVI algorithm that incorporates exploration by perturbing the least-square approximation of the action-value function, and this algorithm guarantees $\widetilde{\mathcal{O}}(d^2 H^2 \sqrt{T})$ regret. Ishfaq et al. [37] propose a variant of the randomized LSVI algorithm that combines optimism and TS by perturbing the training data with *i.i.d.* scalar noise, achieving a regret bound of $\widetilde{\mathcal{O}}(d^{\frac{3}{2}} H^{\frac{3}{2}} \sqrt{T})$. Similarly, Ishfaq et al. [38] introduce a randomized RL algorithm that employs Langevin Monte Carlo (LMC) to approximate the posterior distribution of the action-value

16

Table 1: This table compares the problem settings, online update, performance of the this paper with those of other methods in provable RL with function approximation. For computation cost, we only keep the dependence on the number of episode K.

Algorithm	Model-based	Transition model	Reward	Computation cost	Regret
LSVI-UCB [43]	✗	Linear	Linear	$\mathcal{O}(K)$	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H^{\frac{3}{2}}\sqrt{T})$
OPT-RLSVI [73]	✗	Linear	Linear	$\mathcal{O}(K)$	$\widetilde{\mathcal{O}}(d^2 H^2\sqrt{T})$
LSVI-PHE [37]	✗	Linear	Linear	$\mathcal{O}(K)$	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H^{\frac{3}{2}}\sqrt{T})$
UC-MatrixRL [72]	✓	Linear	Known	$\mathcal{O}(K)$	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H^2\sqrt{T})$
UCRL-VTR [9]	✓	Linear mixture	Known	$\mathcal{O}(K)$	$\widetilde{\mathcal{O}}(dH^{\frac{3}{2}}\sqrt{T})$
UCRL-MNL [35]	✓	MNL	Known	$\mathcal{O}(K)$	$\widetilde{\mathcal{O}}(\kappa^{-1}dH^{\frac{3}{2}}\sqrt{T})$
RRL-MNL (**this work**)	✓	MNL	Known	$\mathcal{O}(1)$	$\widetilde{\mathcal{O}}(\kappa^{-1}d^{\frac{3}{2}}H^{\frac{3}{2}}\sqrt{T})$
ORRL-MNL (**this work**)	✓	MNL	Known	$\mathcal{O}(1)$	$\widetilde{\mathcal{O}}\left(d^{\frac{3}{2}}H^{\frac{3}{2}}\sqrt{T}+\kappa^{-1}d^2H^2\right)$
UCRL-MNL+ (**this work**)	✓	MNL	Known	$\mathcal{O}(1)$	$\widetilde{\mathcal{O}}\left(dH^{\frac{3}{2}}\sqrt{T}+\kappa^{-1}d^2H^2\right)$

function, also ensuring a regret bound of $\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H^{\frac{3}{2}}\sqrt{T})$. Also, there have been studies on model-based methods with function approximation in linear MDPs, such as Yang and Wang [72], which assume that the transition probability kernel is a bilinear model parametrized by a matrix and propose a UCB-based algorithm with an upper bound of $\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H^2\sqrt{T})$ for regret. He et al. [34] propose an algorithm achieving nearly minimax optimal regret $\widetilde{\mathcal{O}}(dH\sqrt{T})$. Jia et al. [41] consider a specific type of MDPs called linear mixture MDPs in which the transition probability kernel is a linear combination of different basis kernels. This model encompasses various types of MDPs studied previously in Modi et al. [53], Yang and Wang [72]. For this model, Jia et al. [41] propose a UCB-based RL algorithm with value-targeted model parameter estimation that guarantees an upper bound of $\widetilde{\mathcal{O}}(dH^{\frac{3}{2}}\sqrt{T})$ for regret. The same linear mixture MDPs have been used in other studies such as Ayoub et al. [9], Zhou et al. [81, 82]. Specifically, in Zhou et al. [81], a variant of the method proposed by Jia et al. [41] is suggested and proved that the algorithm guarantees an upper bound of $\widetilde{\mathcal{O}}(dH\sqrt{T})$ regret with a matching lower bound of $\Omega(dH\sqrt{T})$ for linear mixture MDPs. More recently, there are also works achieving horizon-free regret bounds for linear mixture MDPs [78, 46, 80].

RL with non-linear function approximation   Studies have been conducted on extending function approximation beyond linear models. Ayoub et al. [9], Wang et al. [68], Ishfaq et al. [37] provide upper bound for regret based on eluder dimension [63]. Also, there has been an effort to develop sample-efficient methods with more "general" function approximation [48, 42, 19–21, 28, 37, 44, 4, 5, 75, 18, 39] However, these attempts may have been hindered by the difficulty of solving computationally intractable problems [48, 42, 19, 21, 28, 44, 18], the necessity of relying on stronger assumptions [20, 37], or the lack of discussion on how to define the posterior distribution supported by a given function class and how to draw the optimistic sample from the posterior [4, 5, 75]. That is why even after there exists a so-called "general function class"-based result, it is often the case that the results in specific parametric models are still needed. Despite the large number of studies on RL with linear function approximation, there is limited research on extending beyond linear models to other parametric models. Wang et al. [69] use generalized linear function approximation, where the Bellman backup of any value function is assumed to be a generalized linear function of feature mapping. Hwang and Oh [35] discuss the limitations of linear function approximation and propose a UCB-based algorithm for MNL transition model in feature space achieving $\widetilde{\mathcal{O}}(dH^{\frac{3}{2}}\sqrt{T})$. Ishfaq et al. [39] present TS-based RL algorithms that utilize approximate samplers, such as LMC or Underdamped LMC, to enhance the implementation and computational tractability of TS for RL with general function classes.

Contextual bandits   Faury et al. [23] first provide a UCB-based algorithm with κ-independent regret for binary logistic bandit and Abeille et al. [3] present UCB & TS based algorithms achieving nearly minimax optimal regret for the same setting. Faury et al. [24] propose a jointly efficient UCB-based algorithm that achieve κ-independent regret bound with $\mathcal{O}(\log t)$ computation cost. In the context of MNL model, Oh and Iyengar [54] employ TS approach, while Oh and Iyengar [55] incorporate a combination of UCB exploration and online parameter updates for MNL bandits. Both of the methods have $\mathcal{O}(\kappa^{-1}\sqrt{T})$ regret. Amani and Thrampoulidis [8] propose an optimistic algorithm with better dependence on κ. Agrawal et al. [6] design a UCB-based algorithm with

17

$\mathcal{O}(\sqrt{T})$ regret bound without κ in its leading term, and Perivier and Goyal [61] establish $\mathcal{O}(\sqrt{T/\kappa_*})$ regret for the uniform reward setting. Zhang and Sugiyama [76] develop jointly efficient UCB-based algorithm for non-uniform MNL bandit problem. Lee and Oh [50] propose nearly minimax optimal MNL bandit algorithm for both uniform and non-uniform reward structures.

B   Notations & Definitions

In this section, we formally summarize some definitions and notations used to analyze the proposed algorithm.

Inhomogeneous MNL transition model

For $h \in [H]$, the probability of state transition to $s' \in \mathcal{S}_{s,a}$ when an action a is taken at a state s is given by

$$P_h(s' \mid s,a) := P_{\boldsymbol{\theta}_h^*}(s' \mid s,a) = \frac{\exp(\boldsymbol{\varphi}(s,a,s')^\top \boldsymbol{\theta}_h^*)}{\sum_{\widetilde{s} \in \mathcal{S}_{s,a}} \exp(\boldsymbol{\varphi}(s,a,\widetilde{s})^\top \boldsymbol{\theta}_h^*)} .$$

The estimated transition probability parameterized by $\boldsymbol{\theta}$ is denoted as

$$P_{\boldsymbol{\theta}}(s' \mid s,a) := \frac{\exp(\boldsymbol{\varphi}(s,a,s')^\top \boldsymbol{\theta})}{\sum_{\widetilde{s} \in \mathcal{S}_{s,a}} \exp(\boldsymbol{\varphi}(s,a,\widetilde{s})^\top \boldsymbol{\theta})} .$$

Feature vector

We abbreviate the feature vector as follows:

$$\begin{aligned}
\boldsymbol{\varphi}_{s,a,s'} &:= \boldsymbol{\varphi}(s,a,s') \ \text{ for } (s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}_{s,a} , \\
\boldsymbol{\varphi}_{k,h,s'} &:= \boldsymbol{\varphi}(s_h^k, a_h^k, s') \ \text{ for } (k,h) \in [K] \times [H] \text{ and } s' \in \mathcal{S}_{k,h} := \mathcal{S}_{s_h^k, a_h^k} , \\
\hat{\boldsymbol{\varphi}}_{k,h}(s,a) &:= \boldsymbol{\varphi}(s,a,\hat{s}) \ \text{ for } \hat{s} := \underset{s' \in \mathcal{S}_{s,a}}{\operatorname{argmax}} \|\boldsymbol{\varphi}(s,a,s')\|_{\mathbf{A}_{k,h}^{-1}} , \\
\bar{\boldsymbol{\varphi}}_{s,a,s'}(\boldsymbol{\theta}) &:= \bar{\boldsymbol{\varphi}}(s,a,s'; \boldsymbol{\theta}) = \boldsymbol{\varphi}(s,a,s') - \mathbb{E}_{\widetilde{s} \sim P_{\boldsymbol{\theta}}(\cdot \mid s,a)}[\boldsymbol{\varphi}(s,a,\widetilde{s})] , \\
\bar{\boldsymbol{\varphi}}_{k,h,s'}(\boldsymbol{\theta}) &:= \bar{\boldsymbol{\varphi}}(s_h^k, a_h^k, s'; \boldsymbol{\theta}) .
\end{aligned}$$

Response variable & per-episode loss

The response variable y_h^k is given by

$$y_h^k := [y_h^k(s')]_{s' \in \mathcal{S}_{k,h}} \ \text{ where } y_h^k(s') := \mathbb{1}(s_{h+1}^k = s') \ \text{ for } s' \in \mathcal{S}_{k,h} .$$

The per-episode loss $\ell_{k,h}(\boldsymbol{\theta})$ is given by

$$\ell_{k,h}(\boldsymbol{\theta}) := - \sum_{s' \in \mathcal{S}_{k,h}} y_h^k(s') \log P_{\boldsymbol{\theta}}(s' \mid s_h^k, a_h^k) ,$$

$$\mathbf{G}_{k,h}(\boldsymbol{\theta}) := \nabla \ell_{k,h}(\boldsymbol{\theta}) = \sum_{s' \in \mathcal{S}_{k,h}} (P_{\boldsymbol{\theta}}(s' \mid s_h^k, a_h^k) - y_h^k(s')) \boldsymbol{\varphi}_{k,h,s'} ,$$

$$\mathbf{H}_{k,h}(\boldsymbol{\theta}) := \nabla^2 \ell_{k,h}(\boldsymbol{\theta})$$
$$= \sum_{s' \in \mathcal{S}_{k,h}} P_{\boldsymbol{\theta}}(s' \mid s_h^k, a_h^k) \boldsymbol{\varphi}_{k,h,s'} \boldsymbol{\varphi}_{k,h,s'}^\top - \sum_{s' \in \mathcal{S}_{k,h}} \sum_{\widetilde{s} \in \mathcal{S}_{k,h}} P_{\boldsymbol{\theta}}(s' \mid s_h^k, a_h^k) P_{\boldsymbol{\theta}}(\widetilde{s} \mid s_h^k, a_h^k) \boldsymbol{\varphi}_{k,h,s'} \boldsymbol{\varphi}_{k,h,\widetilde{s}}^\top .$$

18

Regularity constants

H : Horizon length

K : Episode number

$T = KH$: Total number of interactions

$L_{\boldsymbol{\varphi}}$: ℓ_2-norm upper bound of $\boldsymbol{\varphi}(s, a, s)$, i.e., $\|\boldsymbol{\varphi}(s, a, s')\|_2 \le L_{\boldsymbol{\varphi}}$,

$L_{\boldsymbol{\theta}}$: ℓ_2-norm upper bound of $\boldsymbol{\theta}_h^*$, i.e., $\|\boldsymbol{\theta}_h^*\|_2 \le L_{\boldsymbol{\theta}}$,

κ : Problem-dependent constant such that $\displaystyle\inf_{\boldsymbol{\theta} \in \mathcal{B}_d(L_{\boldsymbol{\theta}})} P_{\boldsymbol{\theta}}(s' \mid s, a) P_{\boldsymbol{\theta}}(\widetilde{s} \mid s, a) \ge \kappa$,

$\mathcal{U}$: Maximum cardinality of the set of reachable states, i.e., $\mathcal{U} := \displaystyle\max_{s,a} |\mathcal{S}_{s,a}|$.

Estimated transition core

The estimated transition core for RRL-MNL is given by

$$\boldsymbol{\theta}_h^k = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathcal{B}_d(L_{\boldsymbol{\theta}})} \frac{1}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_h^{k-1}\|_{\mathbf{A}_{k,h}}^2 + (\boldsymbol{\theta} - \boldsymbol{\theta}_h^{k-1})^\top \nabla \ell_{k-1,h}(\boldsymbol{\theta}_h^{k-1}) ,$$

and the estimated transition core for ORRL-MNL is given by

$$\widetilde{\boldsymbol{\theta}}_h^{k+1} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathcal{B}_d(L_{\boldsymbol{\theta}})} \frac{1}{2\eta} \left\| \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^k \right\|_{\widetilde{\mathbf{B}}_{k,h}}^2 + \boldsymbol{\theta}^\top \nabla \ell_{k,h}(\widetilde{\boldsymbol{\theta}}_h^k) .$$

Gram matrices

The Gram matrix with global gradient information κ is given by

$$\mathbf{A}_{k,h} := \lambda \mathbf{I}_d + \frac{\kappa}{2} \sum_{i=1}^{k-1} \sum_{s' \in \mathcal{S}_{i,h}} \boldsymbol{\varphi}(s_h^i, a_h^i, s') \boldsymbol{\varphi}(s_h^i, a_h^i, s')^\top .$$

The Gram matrices with local gradient information are given by

$$\widetilde{\mathbf{B}}_{k,h} := \mathbf{B}_{k,h} + \eta \nabla^2 \ell_{k,h}(\widetilde{\boldsymbol{\theta}}_h^k) \ \text{ and } \ \mathbf{B}_{k,h} := \lambda \mathbf{I}_d + \sum_{i=1}^{k-1} \nabla^2 \ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}) .$$

Confidence radius

For some absolute constants $C_\beta, C_{\boldsymbol{\xi}} > 0$,

$\alpha_k := \alpha_k(\delta)$

$$= \sqrt{\frac{8d}{\kappa} \log\left(1 + \frac{k\mathcal{U}L_{\boldsymbol{\varphi}}^2}{d\lambda}\right) + \left(\frac{32 L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}}{3} + \frac{16}{\kappa}\right) \log \frac{(1 + \lceil 2\log_2 k\mathcal{U}L_{\boldsymbol{\varphi}}L_{\boldsymbol{\theta}} \rceil) k^2}{\delta} + 2\sqrt{2} + 2\lambda L_{\boldsymbol{\theta}}^2}$$

$$= \widetilde{\mathcal{O}}(\kappa^{-1/2} d^{1/2}) ,$$

$$\beta_k := \beta_k(\delta) = C_\beta \sqrt{\log\mathcal{U}\left(\lambda \log(\mathcal{U}k) + \log(\mathcal{U}k)\log\left(\frac{H\sqrt{1+2k}}{\delta}\right) + d\log\left(1 + \frac{k}{d\lambda}\right)\right) + \lambda L_{\boldsymbol{\theta}}^2}$$

$$= \mathcal{O}(\sqrt{d}\log\mathcal{U}\log(kH)) ,$$

$$\gamma_k := \gamma_k(\delta) = C_{\boldsymbol{\xi}} \sigma_k \sqrt{d\log(Md/\delta)} .$$

Filtration

For an arbitrary set X, we denote the Σ-algebra generated by X as $\Sigma(X)$. Then we define the following filtrations

$$\mathcal{F}_k := \Sigma\left(\left\{s_j^i, a_j^i, r(s_j^i, a_j^i) \mid i < k, j \le H\right\} \cup \left\{\boldsymbol{\xi}_{i,j}^{(m)} \mid i < k, j \le H, 1 \le m \le M\right\}\right) ,$$

$$\mathcal{F}_{k,h} := \Sigma\left(\mathcal{F}_k \cup \left\{s_j^k, a_j^k, r(s_j^k, a_j^k) \mid j \le h\right\} \cup \left\{\boldsymbol{\xi}_{k,j}^{(m)} \mid j \ge h, 1 \le m \le M\right\}\right) .$$

19

Pseudo-noise

For `RRL-MNL`, the pseudo-noise is sampled as

$$\boldsymbol{\xi}_{k,h}^{(m)} \sim \mathcal{N}(\mathbf{0}_d, \sigma_k^2 \mathbf{A}_{k,h}^{-1}),$$

and for `ORRL-MNL`, the pseudo-noise is sampled as

$$\boldsymbol{\xi}_{k,h}^{(m)} \sim \mathcal{N}(\mathbf{0}_d, \sigma_k^2 \mathbf{B}_{k,h}^{-1}),$$

for M times independently.

Estimated value functions

The stochastically optimistic value function for `RRL-MNL` is defined as follows:

$$Q_{H+1}^k(s,a) = 0,$$

$$Q_h^k(s,a) = \min\left\{ r(s,a) + \sum_{s' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}_h^k}(s' \mid s,a) V_{h+1}^k(s') + \max_{m \in [M]} \hat{\boldsymbol{\varphi}}_{k,h}(s,a)^\top \boldsymbol{\xi}_{k,h}^{(m)}, H \right\} \text{ for } h \in [H].$$

The optimistic randomized value function for `ORRL-MNL` is defined as follows:

$$\widetilde{Q}_{H+1}^k(s,a) = 0,$$

$$\widetilde{Q}_h^k(s,a) := \min\left\{ r(s,a) + \sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s,a) \widetilde{V}_{h+1}^k(s') + \nu_{k,h}^{\mathrm{rand}}(s,a), H \right\} \text{ for } h \in [H],$$

where

$$\nu_{k,h}^{\mathrm{rand}}(s,a) := \sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s,a) \bar{\boldsymbol{\varphi}}(s,a,s'; \widetilde{\boldsymbol{\theta}}_h^k)^\top \boldsymbol{\xi}_{k,h}^{s'} + 3H\beta_k^2 \max_{s' \in \mathcal{S}_{s,a}} \|\boldsymbol{\varphi}(s,a,s')\|_{\mathbf{B}_{k,h}^{-1}}^2,$$

$$\boldsymbol{\xi}_{k,h}^{s'} := \boldsymbol{\xi}_{k,h}^{m(s')} \text{ for } m(s') := \operatorname*{argmax}_{m \in [M]} \bar{\boldsymbol{\varphi}}(s,a,s'; \widetilde{\boldsymbol{\theta}}_h^k)^\top \boldsymbol{\xi}_{k,h}^m.$$

Prediction error & Bellman error

Definition 1 (Prediction error & Bellman error). *For any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $(k,h) \in [K] \times [H]$, we define the prediction error about $\boldsymbol{\theta}_h^k$ as*

$$\Delta_h^k(s,a) := \sum_{s' \in \mathcal{S}_{s,a}} \left(P_{\boldsymbol{\theta}_h^k}(s' \mid s,a) - P_{\boldsymbol{\theta}_h^*}(s'|s,a) \right) V_{h+1}^k(s').$$

Also we define the Bellman error as follows:

$$\iota_h^k(s,a) := r(s,a) + P_h V_{h+1}^k(s,a) - Q_h^k(s,a).$$

Good events

For any $\delta \in (0,1)$, we define the following good events:

For `RRL-MNL`,

$$\mathcal{G}_{k,h}^{\Delta}(\delta) := \left\{ |\Delta_h^k(s,a)| \leq H\alpha_k(\delta) \|\hat{\boldsymbol{\varphi}}_{k,h}(s,a)\|_{\mathbf{A}_{k,h}^{-1}} \right\},$$

$$\mathcal{G}_{k,h}^{\boldsymbol{\xi}}(\delta) := \left\{ \max_{m \in [M]} \|\boldsymbol{\xi}_{k,h}^{(m)}\|_{\mathbf{A}_{k,h}} \leq \gamma_k(\delta) \right\},$$

$$\mathcal{G}_{k,h}(\delta) := \left\{ \mathcal{G}_{k,h}^{\Delta}(\delta) \cap \mathcal{G}_{k,h}^{\boldsymbol{\xi}}(\delta) \right\},$$

$$\mathcal{G}_k(\delta) := \bigcap_{h \in [H]} \mathcal{G}_{k,h}(\delta),$$

$$\mathcal{G}(K,\delta) := \bigcap_{k \leq K} \mathcal{G}_k(\delta).$$

20

For `ORRL-MNL`,

$$\mathfrak{G}_{k,h}^{\Delta}(\delta) := \left\{ |\Delta_h^k(s,a)| \leq H\beta_k(\delta) \sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s,a) \left\| \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k) \right\|_{\mathbf{B}_{k,h}^{-1}} \right.$$

$$\left. + 3H\beta_k(\delta)^2 \max_{s' \in \mathcal{S}_{s,a}} \|\boldsymbol{\varphi}_{s,a,s'}\|_{\mathbf{B}_{k,h}^{-1}}^2 \right\},$$

$$\mathfrak{G}_{k,h}^{\boldsymbol{\xi}}(\delta) := \left\{ \max_{m \in [M]} \|\boldsymbol{\xi}_{k,h}^{(m)}\|_{\mathbf{B}_{k,h}} \leq \gamma_k(\delta) \right\},$$

$$\mathfrak{G}_{k,h}(\delta) := \left\{ \mathfrak{G}_{k,h}^{\Delta}(\delta) \cap \mathfrak{G}_{k,h}^{\boldsymbol{\xi}}(\delta) \right\},$$

$$\mathfrak{G}_k(\delta) := \bigcap_{h \in [H]} \mathfrak{G}_{k,h}(\delta),$$

$$\mathfrak{G}(K,\delta) := \bigcap_{k \leq K} \mathfrak{G}_k(\delta).$$

Derivative of MNL transition model

Proposition 1 (Derivative of MNL transition model)**.** *The gradient and Hessian of $P_{\boldsymbol{\theta}}(\cdot \mid \cdot, \cdot)$ can be calculated as follows:*

$$\nabla P_{\boldsymbol{\theta}}(s' \mid s,a) = P_{\boldsymbol{\theta}}(s' \mid s,a) \left(\boldsymbol{\varphi}_{s,a,s'} - \sum_{s'' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s,a)\boldsymbol{\varphi}_{s,a,s''} \right) \tag{8}$$

$$= P_{\boldsymbol{\theta}}(s' \mid s,a)\bar{\boldsymbol{\varphi}}_{s,a,s'}(\boldsymbol{\theta}),$$

and

$$\nabla^2 P_{\boldsymbol{\theta}}(s' \mid s,a)$$
$$= P_{\boldsymbol{\theta}}(s' \mid s,a)\boldsymbol{\varphi}_{s,a,s'}\boldsymbol{\varphi}_{s,a,s'}^{\top}$$
$$- P_{\boldsymbol{\theta}}(s' \mid s,a) \sum_{s'' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s,a) \left(\boldsymbol{\varphi}_{s,a,s'}\boldsymbol{\varphi}_{s,a,s''}^{\top} + \boldsymbol{\varphi}_{s,a,s''}\boldsymbol{\varphi}_{s,a,s'}^{\top} + \boldsymbol{\varphi}_{s,a,s''}\boldsymbol{\varphi}_{s,a,s''}^{\top} \right) \tag{9}$$
$$+ 2P_{\boldsymbol{\theta}}(s' \mid s,a) \left(\sum_{s'' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s,a)\boldsymbol{\varphi}_{s,a,s''} \right) \left(\sum_{s'' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s,a)\boldsymbol{\varphi}_{s,a,s''} \right)^{\top}.$$

Proof of Proposition 1. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ and $[\boldsymbol{\varphi}_{s,a,s'}]_i$ be the i-th component of $\boldsymbol{\varphi}_{s,a,s'}$. Then, we have

$$\frac{\partial}{\partial \theta_j} P_{\boldsymbol{\theta}}(s' \mid s,a)$$

$$= \frac{\exp\left(\boldsymbol{\varphi}_{s,a,s'}^{\top}\boldsymbol{\theta}\right) [\boldsymbol{\varphi}_{s,a,s'}]_j}{\sum_{s'' \in \mathcal{S}_{s,a}} \exp\left(\boldsymbol{\varphi}_{s,a,s''}^{\top}\boldsymbol{\theta}\right)} - \frac{\exp\left(\boldsymbol{\varphi}_{s,a,s'}^{\top}\boldsymbol{\theta}\right) \sum_{s'' \in \mathcal{S}_{s,a}} \exp\left(\boldsymbol{\varphi}_{s,a,s''}^{\top}\boldsymbol{\theta}\right) [\boldsymbol{\varphi}_{s,a,s''}]_j}{\left(\sum_{s'' \in \mathcal{S}_{s,a}} \exp\left(\boldsymbol{\varphi}_{s,a,s''}^{\top}\boldsymbol{\theta}\right)\right)^2}$$

$$= P_{\boldsymbol{\theta}}(s' \mid s,a) \left([\boldsymbol{\varphi}_{s,a,s'}]_j - \sum_{s'' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s,a)[\boldsymbol{\varphi}_{s,a,s''}]_j \right).$$

Then, the gradient of $P_{\boldsymbol{\theta}}(s' \mid s,a)$ is given by

$$\nabla P_{\boldsymbol{\theta}}(s' \mid s,a) = P_{\boldsymbol{\theta}}(s' \mid s,a)\boldsymbol{\varphi}_{s,a,s'} - P_{\boldsymbol{\theta}}(s' \mid s,a) \sum_{s'' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s,a)\boldsymbol{\varphi}_{s,a,s''}$$

$$= P_{\boldsymbol{\theta}}(s' \mid s,a) \left(\boldsymbol{\varphi}_{s,a,s'} - \sum_{s'' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s,a)\boldsymbol{\varphi}_{s,a,s''} \right)$$

$$= P_{\boldsymbol{\theta}}(s' \mid s,a)\bar{\boldsymbol{\varphi}}_{s,a,s'}(\boldsymbol{\theta}).$$

21

On the other hand, the second derivative $\frac{\partial}{\partial\theta_i\partial\theta_j}P_{\boldsymbol{\theta}}(s' \mid s, a)$ can be obtained as follows:

$$\frac{\partial}{\partial\theta_i\partial\theta_j}P_{\boldsymbol{\theta}}(s' \mid s, a)$$

$$= P_{\boldsymbol{\theta}}(s' \mid s, a)\left([\boldsymbol{\varphi}_{s,a,s'}]_i - \sum_{s''\in\mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s, a)[\boldsymbol{\varphi}_{s,a,s''}]_i\right)$$

$$\cdot\left([\boldsymbol{\varphi}_{s,a,s'}]_j - \sum_{s''\in\mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s, a)[\boldsymbol{\varphi}_{s,a,s''}]_j\right)$$

$$+ P_{\boldsymbol{\theta}}(s' \mid s, a)\left(-\sum_{s''\in\mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s, a)\left([\boldsymbol{\varphi}_{s,a,s''}]_i - \sum_{\widetilde{s}\in\mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(\widetilde{s} \mid s, a)[\boldsymbol{\varphi}_{s,a,\widetilde{s}}]_i\right)[\boldsymbol{\varphi}_{s,a,s''}]_j\right)$$

$$= P_{\boldsymbol{\theta}}(s' \mid s, a)\bigg\{[\boldsymbol{\varphi}_{s,a,s'}]_i[\boldsymbol{\varphi}_{s,a,s'}]_j$$

$$- \sum_{s''\in\mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s, a)\left([\boldsymbol{\varphi}_{s,a,s''}]_i[\boldsymbol{\varphi}_{s,a,s'}]_j + [\boldsymbol{\varphi}_{s,a,s'}]_i[\boldsymbol{\varphi}_{s,a,s''}]_j\right)$$

$$+ \left(\sum_{s''\in\mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s, a)[\boldsymbol{\varphi}_{s,a,s''}]_i\right)\left(\sum_{s''\in\mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s, a)[\boldsymbol{\varphi}_{s,a,s''}]_j\right)$$

$$- \sum_{s''\in\mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s, a)[\boldsymbol{\varphi}_{s,a,s''}]_i[\boldsymbol{\varphi}_{s,a,s''}]_j$$

$$+ \left(\sum_{s''\in\mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s, a)[\boldsymbol{\varphi}_{s,a,s''}]_j\right)\left(\sum_{\widetilde{s}\in\mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(\widetilde{s} \mid s, a)[\boldsymbol{\varphi}_{s,a,\widetilde{s}}]_i\right)\bigg\}$$

$$= P_{\boldsymbol{\theta}}(s' \mid s, a)\left\{[\boldsymbol{\varphi}_{s,a,s'}]_i[\boldsymbol{\varphi}_{s,a,s'}]_j\right.$$

$$- \sum_{s''\in\mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s, a)\left([\boldsymbol{\varphi}_{s,a,s''}]_i[\boldsymbol{\varphi}_{s,a,s'}]_j + [\boldsymbol{\varphi}_{s,a,s'}]_i[\boldsymbol{\varphi}_{s,a,s''}]_j\right)$$

$$- \sum_{s''\in\mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s, a)[\boldsymbol{\varphi}_{s,a,s''}]_i[\boldsymbol{\varphi}_{s,a,s''}]_j$$

$$+ 2\left(\sum_{s''\in\mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s, a)[\boldsymbol{\varphi}_{s,a,s''}]_i\right)\left(\sum_{s''\in\mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s, a)[\boldsymbol{\varphi}_{s,a,s''}]_j\right)\right\}.$$

Thus, we get the desired result as follows:

$$\nabla^2 P_{\boldsymbol{\theta}}(s' \mid s, a)$$

$$= P_{\boldsymbol{\theta}}(s' \mid s, a)\boldsymbol{\varphi}_{s,a,s'}\boldsymbol{\varphi}_{s,a,s'}^{\top}$$

$$- P_{\boldsymbol{\theta}}(s' \mid s, a)\sum_{s''\in\mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s, a)\left(\boldsymbol{\varphi}_{s,a,s'}\boldsymbol{\varphi}_{s,a,s''}^{\top} + \boldsymbol{\varphi}_{s,a,s''}\boldsymbol{\varphi}_{s,a,s'}^{\top} + \boldsymbol{\varphi}_{s,a,s''}\boldsymbol{\varphi}_{s,a,s''}^{\top}\right)$$

$$+ 2P_{\boldsymbol{\theta}}(s' \mid s, a)\left(\sum_{s''\in\mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s, a)\boldsymbol{\varphi}_{s,a,s''}\right)\left(\sum_{s''\in\mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s, a)\boldsymbol{\varphi}_{s,a,s''}\right)^{\top}.$$

$\square$

C Detailed Regret Analysis for `RRL-MNL` (Theorem 1)

In this section, we provide the complete proof of Theorem 1. First, we introduce all the technical lemmas needed to prove Theorem 1 along with their proofs. At the end of this section, we present the proof of Theorem 1.

22

C.1 Concentration of Estimated Transition Core θ_h^k

In this section, we provide the concentration inequality for the estimated transition core run by the approximate online Newton step. The proof is similar to that given by Oh and Iyengar [55]. For completeness, we provide the detailed proof.

Lemma 1 (Concentration of online estimated transition core). *For each $h \in [H]$, if $\lambda \geq L_\varphi^2$, then we have*

$$\mathbb{P}\left(\forall k \geq 1, \|\theta_h^k - \theta_h^*\|_{\mathbf{A}_{k,h}} \leq \alpha_k(\delta)\right) \geq 1 - \delta \,.$$

where $\alpha_k(\delta)$ is given by

$\alpha_k(\delta)$

$$:= \sqrt{\frac{8d}{\kappa}\log\left(1 + \frac{k\mathcal{U}L_\varphi^2}{d\lambda}\right) + \left(\frac{32L_\varphi L_\theta}{3} + \frac{16}{\kappa}\right)\log\frac{\left(1 + \lceil 2\log_2 k\mathcal{U}L_\varphi L_\theta\rceil\right)k^2}{\delta} + 2\sqrt{2} + 2\lambda L_\theta^2} \,.$$

Proof of lemma 1. Recall that the per-round loss $\ell_{k,h}(\theta)$ and its gradient $\mathbf{G}_{k,h}(\theta)$ is defined as follows:

$$\ell_{k,h}(\theta) := -\sum_{s' \in \mathcal{S}_{k,h}} y_h^k(s')\log P_\theta(s' \mid s_h^k, a_h^k), \quad \mathbf{G}_{k,h}(\theta) := \nabla_\theta \ell_{k,h}(\theta) \,.$$

For the analysis, we define the conditional expectations of $\ell_{k,h}(\theta)$ & $\mathbf{G}_{k,h}(\theta)$ as follows:

$$\bar{\ell}_{k,h}(\theta) := \mathbb{E}_{y_h^k}\left[\ell_{k,h}(\theta) \mid \mathcal{F}_{k,h}\right], \quad \bar{\mathbf{G}}_{k,h}(\theta) := \mathbb{E}_{y_h^k}[\mathbf{G}_{k,h}(\theta) \mid \mathcal{F}_{k,h}] \,.$$

By Taylor expansion with $\bar{\theta} = \nu\theta_h^k + (1 - \nu)\theta_h^*$ for some $\nu \in (0,1)$, we have

$$\ell_{k,h}(\theta_h^*) = \ell_{k,h}(\theta_h^k) + \mathbf{G}_{k,h}(\theta_h^k)^\top(\theta_h^* - \theta_h^k) + \frac{1}{2}(\theta_h^* - \theta_h^k)^\top\mathbf{H}_{k,h}(\bar{\theta})(\theta_h^* - \theta_h^k), \quad (10)$$

where $\mathbf{H}_{k,h}(\theta)$ is the Hessian of the per-round loss evaluated at θ, i.e.,

$$\begin{aligned}
\mathbf{H}_{k,h}(\theta) &:= \nabla^2\ell_{k,h}(\theta) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (11)\\
&= \sum_{s' \in \mathcal{S}_{k,h}} P_\theta(s' \mid s_h^k, a_h^k)\varphi_{k,h,s'}\varphi_{k,h,s'}^\top \\
&\quad - \sum_{s',\tilde{s} \in \mathcal{S}_{k,h}} P_\theta(s' \mid s_h^k, a_h^k)P_\theta(\tilde{s} \mid s_h^k, a_h^k)\varphi_{k,h,s'}\varphi_{k,h,\tilde{s}}^\top \,.
\end{aligned}$$

Note that for $\bar{\theta} = \nu\theta_h^k + (1 - \nu)\theta_h^*$ with $\nu \in (0,1)$, we have

$$\begin{aligned}
\mathbf{H}_{k,h}(\bar{\theta}) &= \sum_{s' \in \mathcal{S}_{k,h}} P_{\bar{\theta}}(s' \mid s_h^k, a_h^k)\varphi_{k,h,s'}\varphi_{k,h,s'}^\top \\
&\quad - \sum_{s' \in \mathcal{S}_{k,h}}\sum_{\tilde{s} \in \mathcal{S}_{k,h}} P_{\bar{\theta}}(s' \mid s_h^k, a_h^k)P_{\bar{\theta}}(\tilde{s} \mid s_h^k, a_h^k)\varphi_{k,h,s'}\varphi_{k,h,\tilde{s}}^\top \\
&= \sum_{s' \in \mathcal{S}_{k,h}} P_{\bar{\theta}}(s' \mid s_h^k, a_h^k)\varphi_{k,h,s'}\varphi_{k,h,s'}^\top \\
&\quad - \frac{1}{2}\sum_{s' \in \mathcal{S}_{k,h}}\sum_{\tilde{s} \in \mathcal{S}_{k,h}} P_{\bar{\theta}}(s' \mid s_h^k, a_h^k)P_{\bar{\theta}}(\tilde{s} \mid s_h^k, a_h^k)(\varphi_{k,h,s'}\varphi_{k,h,\tilde{s}}^\top + \varphi_{k,h,\tilde{s}}\varphi_{i,h,s'}^\top) \\
&\succeq \sum_{s' \in \mathcal{S}_{k,h}} P_{\bar{\theta}}(s' \mid s_h^k, a_h^k)\varphi_{k,h,s'}\varphi_{k,h,s'}^\top \\
&\quad - \frac{1}{2}\sum_{s' \in \mathcal{S}_{k,h}}\sum_{\tilde{s} \in \mathcal{S}_{k,h}} P_{\bar{\theta}}(s' \mid s_h^k, a_h^k)P_{\bar{\theta}}(\tilde{s} \mid s_h^k, a_h^k)(\varphi_{k,h,s'}\varphi_{k,h,s'}^\top + \varphi_{k,h,\tilde{s}}\varphi_{k,h,\tilde{s}}^\top) \\
&= \sum_{s' \in \mathcal{S}_{k,h}} P_{\bar{\theta}}(s' \mid s_h^k, a_h^k)\varphi_{k,h,s'}\varphi_{k,h,s'}^\top \\
&\quad - \sum_{s' \in \mathcal{S}_{k,h}}\sum_{\tilde{s} \in \mathcal{S}_{k,h}} P_{\bar{\theta}}(s' \mid s_h^k, a_h^k)P_{\bar{\theta}}(\tilde{s} \mid s_h^k, a_h^k)\varphi_{k,h,s'}\varphi_{k,h,s'}^\top \,,
\end{aligned}$$

23

where the inequality utilizes the fact that $\mathbf{x}\mathbf{x}^\top + \mathbf{y}\mathbf{y}^\top \succeq \mathbf{x}\mathbf{y}^\top + \mathbf{y}\mathbf{x}^\top$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Therefore, we have

$$
\begin{aligned}
\mathbf{H}_{k,h}(\bar{\boldsymbol{\theta}}) \succeq{} & \sum_{s' \in \mathcal{S}_{k,h}} P_{\bar{\boldsymbol{\theta}}}(s' \mid s_h^k, a_h^k) \boldsymbol{\varphi}_{k,h,s'} \boldsymbol{\varphi}_{k,h,s'}^\top \\
& - \sum_{s' \in \mathcal{S}_{k,h}} \sum_{\widetilde{s} \in \mathcal{S}_{k,h}} P_{\bar{\boldsymbol{\theta}}}(s' \mid s_h^k, a_h^k) P_{\bar{\boldsymbol{\theta}}}(\widetilde{s} \mid s_h^k, a_h^k) \boldsymbol{\varphi}_{k,h,s'} \boldsymbol{\varphi}_{k,h,s'}^\top \\
={} & \sum_{s' \neq \dot{s}_{k,h}} P_{\bar{\boldsymbol{\theta}}}(s' \mid s_h^k, a_h^k) \boldsymbol{\varphi}_{k,h,s'} \boldsymbol{\varphi}_{k,h,s'}^\top \\
& - \sum_{s' \neq \dot{s}_{k,h}} \sum_{\widetilde{s} \neq \dot{s}_{k,h}} P_{\bar{\boldsymbol{\theta}}}(s' \mid s_h^k, a_h^k) P_{\bar{\boldsymbol{\theta}}}(\widetilde{s} \mid s_h^k, a_h^k) \boldsymbol{\varphi}_{k,h,s'} \boldsymbol{\varphi}_{k,h,s'}^\top \\
={} & \sum_{s' \neq \dot{s}_{k,h}} P_{\bar{\boldsymbol{\theta}}}(s' \mid s_h^k, a_h^k) \left(1 - \sum_{\widetilde{s} \neq \dot{s}_{k,h}} P_{\bar{\boldsymbol{\theta}}}(\widetilde{s} \mid s_h^k, a_h^k) \right) \boldsymbol{\varphi}_{k,h,s'} \boldsymbol{\varphi}_{k,h,s'}^\top \\
={} & \sum_{s' \neq \dot{s}_{k,h}} P_{\bar{\boldsymbol{\theta}}}(s' \mid s_h^k, a_h^k) P_{\bar{\boldsymbol{\theta}}}(\dot{s}_{k,h} \mid s_h^k, a_h^k) \boldsymbol{\varphi}_{k,h,s'} \boldsymbol{\varphi}_{k,h,s'}^\top \\
\succeq{} & \sum_{s' \neq \dot{s}_{k,h}} \kappa \boldsymbol{\varphi}_{k,h,s'} \boldsymbol{\varphi}_{k,h,s'}^\top \\
={} & \sum_{s' \in \mathcal{S}_{k,h}} \kappa \boldsymbol{\varphi}_{k,h,s'} \boldsymbol{\varphi}_{k,h,s'}^\top \,,
\end{aligned}
$$

where $\dot{s}_{k,h}$ is the state satisfying $\boldsymbol{\varphi}(s_h^k, a_h^k, \dot{s}_{k,h}) = \mathbf{0}_d$ and the last inequality comes from the Assumption 4.

Using the lower bound of the Hessian of the per-round loss evaluated at $\bar{\boldsymbol{\theta}}$, from (10) we have

$$
\ell_{k,h}(\boldsymbol{\theta}_h^*) \geq \ell_{k,h}(\boldsymbol{\theta}_h^k) + \mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k)^\top (\boldsymbol{\theta}_h^* - \boldsymbol{\theta}_h^k) + \frac{\kappa}{2}(\boldsymbol{\theta}_h^* - \boldsymbol{\theta}_h^k)^\top \left(\sum_{s' \in \mathcal{S}_{k,h}} \boldsymbol{\varphi}_{k,h,s'} \boldsymbol{\varphi}_{k,h,s'}^\top \right) (\boldsymbol{\theta}_h^* - \boldsymbol{\theta}_h^k) \,.
$$

By rearranging, we have

$$
\ell_{k,h}(\boldsymbol{\theta}_h^k) \leq \ell_{k,h}(\boldsymbol{\theta}_h^*) + \mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k)^\top (\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*) - \frac{\kappa}{2}(\boldsymbol{\theta}_h^* - \boldsymbol{\theta}_h^k)^\top \mathbf{W}_{k,h}(\boldsymbol{\theta}_h^* - \boldsymbol{\theta}_h^k) \,,
$$

where we denote $\mathbf{W}_{k,h} := \sum_{s' \in \mathcal{S}_{k,h}} \boldsymbol{\varphi}_{k,h,s'} \boldsymbol{\varphi}_{k,h,s'}^\top$. By taking expectation over y_h^k, we have

$$
\bar{\ell}_{k,h}(\boldsymbol{\theta}_h^k) \leq \bar{\ell}_{k,h}(\boldsymbol{\theta}_h^*) + \bar{\mathbf{G}}_{k,h}(\boldsymbol{\theta}_h^k)^\top (\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*) - \frac{\kappa}{2}(\boldsymbol{\theta}_h^* - \boldsymbol{\theta}_h^k)^\top \mathbf{W}_{k,h}(\boldsymbol{\theta}_h^* - \boldsymbol{\theta}_h^k) \,. \qquad (12)
$$

On the other hand, for any $\boldsymbol{\theta} \in \mathbb{R}^d$, since we have

$$
\begin{aligned}
& \bar{\ell}_{k,h}(\boldsymbol{\theta}) - \bar{\ell}_{k,h}(\boldsymbol{\theta}_h^*) \\
={} & - \sum_{s' \in \mathcal{S}_{k,h}} P_{\boldsymbol{\theta}_h^*}(s' \mid s_h^k, a_h^k) \log P_{\boldsymbol{\theta}}(s' \mid s_h^k, a_h^k) + \sum_{s' \in \mathcal{S}_{k,h}} P_{\boldsymbol{\theta}_h^*}(s' \mid s_h^k, a_h^k) \log P_{\boldsymbol{\theta}_h^*}(s' \mid s_h^k, a_h^k) \\
={} & \sum_{s' \in \mathcal{S}_{k,h}} P_{\boldsymbol{\theta}_h^*}(s' \mid s_h^k, a_h^k) \left(\log P_{\boldsymbol{\theta}_h^*}(s' \mid s_h^k, a_h^k) - \log P_{\boldsymbol{\theta}}(s' \mid s_h^k, a_h^k) \right) \\
={} & \sum_{s' \in \mathcal{S}_{k,h}} P_{\boldsymbol{\theta}_h^*}(s' \mid s_h^k, a_h^k) \log \frac{P_{\boldsymbol{\theta}_h^*}(s' \mid s_h^k, a_h^k)}{P_{\boldsymbol{\theta}}(s' \mid s_h^k, a_h^k)} \\
={} & D_{\mathrm{KL}}(P_{\boldsymbol{\theta}_h^*} \parallel P_{\boldsymbol{\theta}}) \\
\geq{} & 0 \,,
\end{aligned}
$$

where $D_{\mathrm{KL}}(P \parallel Q)$ is the Kullback-Leibler divergence of P from Q, from (12) we have

$$
\begin{aligned}
0 \leq{} & \bar{\ell}_{k,h}(\boldsymbol{\theta}_h^k) - \bar{\ell}_{k,h}(\boldsymbol{\theta}_h^*) \\
\leq{} & \bar{\mathbf{G}}_{k,h}(\boldsymbol{\theta}_h^k)^\top (\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*) - \frac{\kappa}{2} \|\boldsymbol{\theta}_h^* - \boldsymbol{\theta}_h^k\|_{\mathbf{W}_{k,h}}^2 \\
={} & \mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k)^\top (\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*) - \frac{\kappa}{2} \|\boldsymbol{\theta}_h^* - \boldsymbol{\theta}_h^k\|_{\mathbf{W}_{k,h}}^2 + \left(\bar{\mathbf{G}}_{k,h}(\boldsymbol{\theta}_h^k) - \mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k) \right)^\top (\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*) \,. \qquad (13)
\end{aligned}
$$

24

To get an upper bound of $\mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k)^\top(\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*)$, recall that the estimated transition core is given by

$$\boldsymbol{\theta}_h^{k+1} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathcal{B}_d(L_{\boldsymbol{\theta}})} \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_h^k\|_{\mathbf{A}_{k+1,h}}^2 + (\boldsymbol{\theta} - \boldsymbol{\theta}_h^k)^\top \mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k). \tag{14}$$

Since the objective function in (14) is convex, by the first-order optimality condition for any $\boldsymbol{\theta} \in \mathcal{B}_d(L_{\boldsymbol{\theta}})$, we have

$$\left(\mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k) + \mathbf{A}_{k+1,h}(\boldsymbol{\theta}_h^{k+1} - \boldsymbol{\theta}_h^k)\right)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_h^{k+1}) \geq 0,$$

which gives

$$\boldsymbol{\theta}^\top \mathbf{A}_{k+1,h}(\boldsymbol{\theta}_h^{k+1} - \boldsymbol{\theta}_h^k) \geq (\boldsymbol{\theta}_h^{k+1})^\top \mathbf{A}_{k+1,h}(\boldsymbol{\theta}_h^{k+1} - \boldsymbol{\theta}_h^k) - \mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k)^\top(\boldsymbol{\theta} - \boldsymbol{\theta}_h^{k+1}). \tag{15}$$

Then, we have

$$
\begin{aligned}
&\|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*\|_{\mathbf{A}_{k+1,h}}^2 - \|\boldsymbol{\theta}_h^{k+1} - \boldsymbol{\theta}_h^*\|_{\mathbf{A}_{k+1,h}}^2 \\
&= (\boldsymbol{\theta}_h^k)^\top \mathbf{A}_{k+1,h}\boldsymbol{\theta}_h^k - (\boldsymbol{\theta}_h^{k+1})^\top \mathbf{A}_{k+1,h}\boldsymbol{\theta}_h^{k+1} + 2(\boldsymbol{\theta}_h^*)^\top \mathbf{A}_{k+1,h}(\boldsymbol{\theta}_h^{k+1} - \boldsymbol{\theta}_h^k) \\
&\geq (\boldsymbol{\theta}_h^k)^\top \mathbf{A}_{k+1,h}\boldsymbol{\theta}_h^k - (\boldsymbol{\theta}_h^{k+1})^\top \mathbf{A}_{k+1,h}\boldsymbol{\theta}_h^{k+1} + 2(\boldsymbol{\theta}_h^{k+1})^\top \mathbf{A}_{k+1,h}(\boldsymbol{\theta}_h^{k+1} - \boldsymbol{\theta}_h^k) \\
&\quad - 2\mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k)^\top(\boldsymbol{\theta}_h^* - \boldsymbol{\theta}_h^{k+1}) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(by (15))} \\
&= (\boldsymbol{\theta}_h^k)^\top \mathbf{A}_{k+1,h}\boldsymbol{\theta}_h^k + (\boldsymbol{\theta}_h^{k+1})^\top \mathbf{A}_{k+1,h}\boldsymbol{\theta}_h^{k+1} - 2(\boldsymbol{\theta}_h^{k+1})^\top \mathbf{A}_{k+1,h}\boldsymbol{\theta}_h^k - 2\mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k)^\top(\boldsymbol{\theta}_h^* - \boldsymbol{\theta}_h^{k+1}) \\
&= \|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^{k+1}\|_{\mathbf{A}_{k+1,h}}^2 - 2\mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k)^\top(\boldsymbol{\theta}_h^* - \boldsymbol{\theta}_h^{k+1}) \\
&= \|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^{k+1}\|_{\mathbf{A}_{k+1,h}}^2 + 2\mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k)^\top(\boldsymbol{\theta}_h^{k+1} - \boldsymbol{\theta}_h^k) + 2\mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k)^\top(\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*) \\
&\geq -\|\mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k)\|_{\mathbf{A}_{k+1,h}^{-1}}^2 + 2\mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k)^\top(\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*), \tag{16}
\end{aligned}
$$

where the last inequality follows by the fact that

$$
\begin{aligned}
\|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^{k+1}\|_{\mathbf{A}_{k+1,h}}^2 + 2\mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k)^\top(\boldsymbol{\theta}_h^{k+1} - \boldsymbol{\theta}_h^k) &\geq \min_{\boldsymbol{\theta}\in\mathcal{B}_d(L_{\boldsymbol{\theta}})} \left\{\|\boldsymbol{\theta}\|_{\mathbf{A}_{k+1,h}}^2 + 2\mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k)^\top\boldsymbol{\theta}\right\} \\
&= -\|\mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k)\|_{\mathbf{A}_{k+1,h}^{-1}}^2.
\end{aligned}
$$

Therefore, from (16) we have

$$\mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k)^\top(\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*) \leq \frac{1}{2}\|\mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k)\|_{\mathbf{A}_{k+1,h}^{-1}}^2 + \frac{1}{2}\|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*\|_{\mathbf{A}_{k+1,h}}^2 - \frac{1}{2}\|\boldsymbol{\theta}_h^{k+1} - \boldsymbol{\theta}_h^*\|_{\mathbf{A}_{k+1,h}}^2. \tag{17}$$

By substituting (17) into (13), we have

$$
\begin{aligned}
0 \leq &\frac{1}{2}\|\mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k)\|_{\mathbf{A}_{k+1,h}^{-1}}^2 + \frac{1}{2}\|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*\|_{\mathbf{A}_{k+1,h}}^2 - \frac{1}{2}\|\boldsymbol{\theta}_h^{k+1} - \boldsymbol{\theta}_h^*\|_{\mathbf{A}_{k+1,h}}^2 \\
&- \frac{\kappa}{2}\|\boldsymbol{\theta}_h^* - \boldsymbol{\theta}_h^k\|_{\mathbf{W}_{k,h}} + \left(\bar{\mathbf{G}}_{k,h}(\boldsymbol{\theta}_h^k) - \mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k)\right)^\top(\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*). \tag{18}
\end{aligned}
$$

25

Note that since we have

$$\|\mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k)\|_{\mathbf{A}_{k+1,h}^{-1}}^2$$

$$= \sum_{s',\widetilde{s}\in\mathcal{S}_{k,h}} \left(P_{\boldsymbol{\theta}_h^k}(s'\mid s_h^k,a_h^k) - y_h^k(s') \right) \left(P_{\boldsymbol{\theta}_h^k}(\widetilde{s}\mid s_h^k,a_h^k) - y_h^k(\widetilde{s}) \right) \boldsymbol{\varphi}_{k,h,s'}^\top \mathbf{A}_{k+1,h}^{-1} \boldsymbol{\varphi}_{k,h,\widetilde{s}}$$

$$= \frac{1}{2} \sum_{s',\widetilde{s}\in\mathcal{S}_{k,h}} \left(P_{\boldsymbol{\theta}_h^k}(s'\mid s_h^k,a_h^k) - y_h^k(s') \right) \left(P_{\boldsymbol{\theta}_h^k}(\widetilde{s}\mid s_h^k,a_h^k) - y_h^k(\widetilde{s}) \right)$$

$$\cdot (\boldsymbol{\varphi}_{k,h,s'}^\top \mathbf{A}_{k+1,h}^{-1} \boldsymbol{\varphi}_{k,h,\widetilde{s}} + \boldsymbol{\varphi}_{k,h,\widetilde{s}}^\top \mathbf{A}_{k+1,h}^{-1} \boldsymbol{\varphi}_{k,h,s'})$$

$$\leq \frac{1}{2} \sum_{s',\widetilde{s}\in\mathcal{S}_{k,h}} \left[\left(P_{\boldsymbol{\theta}_h^k}(s'\mid s_h^k,a_h^k) - y_h^k(s') \right)^2 \boldsymbol{\varphi}_{k,h,s'}^\top \mathbf{A}_{k+1,h}^{-1} \boldsymbol{\varphi}_{k,h,s'} \right.$$

$$\left. + \left(P_{\boldsymbol{\theta}_h^k}(\widetilde{s}\mid s_h^k,a_h^k) - y_h^k(\widetilde{s}) \right)^2 \boldsymbol{\varphi}_{k,h,\widetilde{s}}^\top \mathbf{A}_{k+1,h}^{-1} \boldsymbol{\varphi}_{k,h,\widetilde{s}} \right]$$

$$= \sum_{s'\in\mathcal{S}_{k,h}} \left(P_{\boldsymbol{\theta}_h^k}(s'\mid s_h^k,a_h^k) - y_h^k(s') \right)^2 \boldsymbol{\varphi}_{k,h,s'}^\top \mathbf{A}_{k+1,h}^{-1} \boldsymbol{\varphi}_{k,h,s'}$$

$$\leq \sum_{s'\in\mathcal{S}_{k,h}} \left| P_{\boldsymbol{\theta}_h^k}(\widetilde{s}\mid s_h^k,a_h^k) - y_h^k(\widetilde{s}) \right| \boldsymbol{\varphi}_{k,h,s'}^\top \mathbf{A}_{k+1,h}^{-1} \boldsymbol{\varphi}_{k,h,s'}$$

$$\leq \sum_{s'\in\mathcal{S}_{k,h}} \left(P_{\boldsymbol{\theta}_h^k}(\widetilde{s}\mid s_h^k,a_h^k) + y_h^k(s') \right) \boldsymbol{\varphi}_{k,h,s'}^\top \mathbf{A}_{k+1,h}^{-1} \boldsymbol{\varphi}_{k,h,s'}$$

$$= \sum_{s'\in\mathcal{S}_{k,h}} P_{\boldsymbol{\theta}_h^k}(\widetilde{s}\mid s_h^k,a_h^k) \boldsymbol{\varphi}_{k,h,s'}^\top \mathbf{A}_{k+1,h}^{-1} \boldsymbol{\varphi}_{k,h,s'} + \sum_{s'\in\mathcal{S}_{k,h}} y_h^k(s') \boldsymbol{\varphi}_{k,h,s'}^\top \mathbf{A}_{k+1,h}^{-1} \boldsymbol{\varphi}_{k,h,s'}$$

$$\leq 2 \max_{s'\in\mathcal{S}_{k,h}} \|\boldsymbol{\varphi}_{k,h,s'}\|_{\mathbf{A}_{k+1,h}^{-1}}^2, \tag{19}$$

where the first inequality utilizes the inequality $\mathbf{x}^\top \mathbf{A}\mathbf{y} + \mathbf{y}^\top \mathbf{A}\mathbf{x} \leq \mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{y}^\top \mathbf{A}\mathbf{y}$ for any positive-semidefinite matrix $\mathbf{A}$, and the last inequality holds since $0 \leq P_{\boldsymbol{\theta}_h^k}(s'\mid s_h^k,a_h^k) \leq 1$ and $\sum_{s'} P_{\boldsymbol{\theta}_h^k}(s'\mid s_h^k,a_h^k) = 1$.

Combining the results of (18) and (19), we have

$$0 \leq \max_{s'\in\mathcal{S}_{k,h}} \|\boldsymbol{\varphi}_{k,h,s'}\|_{\mathbf{A}_{k+1,h}^{-1}}^2 + \frac{1}{2}\|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*\|_{\mathbf{A}_{k+1,h}}^2 - \frac{1}{2}\|\boldsymbol{\theta}_h^{k+1} - \boldsymbol{\theta}_h^*\|_{\mathbf{A}_{k+1,h}}^2$$

$$- \frac{\kappa}{2}\|\boldsymbol{\theta}_h^* - \boldsymbol{\theta}_h^k\|_{\mathbf{W}_{k,h}} + \left(\bar{\mathbf{G}}_{k,h}(\boldsymbol{\theta}_h^k) - \mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k) \right)^\top (\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*)$$

$$= \max_{s'\in\mathcal{S}_{k,h}} \|\boldsymbol{\varphi}_{k,h,s'}\|_{\mathbf{A}_{k+1,h}^{-1}}^2 + \frac{1}{2}\|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*\|_{\mathbf{A}_{k,h}}^2 + \frac{\kappa}{4}\|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*\|_{\mathbf{W}_{k,h}}^2 - \frac{1}{2}\|\boldsymbol{\theta}_h^{k+1} - \boldsymbol{\theta}_h^*\|_{\mathbf{A}_{k+1,h}}^2$$

$$- \frac{\kappa}{2}\|\boldsymbol{\theta}_h^* - \boldsymbol{\theta}_h^k\|_{\mathbf{W}_{k,h}} + \left(\bar{\mathbf{G}}_{k,h}(\boldsymbol{\theta}_h^k) - \mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k) \right)^\top (\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*)$$

$$= \max_{s'\in\mathcal{S}_{k,h}} \|\boldsymbol{\varphi}_{k,h,s'}\|_{\mathbf{A}_{k+1,h}^{-1}}^2 + \frac{1}{2}\|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*\|_{\mathbf{A}_{k,h}}^2 - \frac{\kappa}{4}\|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*\|_{\mathbf{W}_{k,h}}^2 - \frac{1}{2}\|\boldsymbol{\theta}_h^{k+1} - \boldsymbol{\theta}_h^*\|_{\mathbf{A}_{k+1,h}}^2$$

$$+ \left(\bar{\mathbf{G}}_{k,h}(\boldsymbol{\theta}_h^k) - \mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k) \right)^\top (\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*),$$

where for the first equality we use $\mathbf{A}_{k+1,h} = \mathbf{A}_{k,h} + \frac{\kappa}{2}\mathbf{W}_{k,h}$. By rearranging the terms, we have

$$\|\boldsymbol{\theta}_h^{k+1} - \boldsymbol{\theta}_h^*\|_{\mathbf{A}_{k+1,h}}^2 \leq \|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*\|_{\mathbf{A}_{k,h}}^2 + 2\max_{s'\in\mathcal{S}_{k,h}} \|\boldsymbol{\varphi}_{k,h,s'}\|_{\mathbf{A}_{k+1,h}^{-1}}^2 - \frac{\kappa}{2}\|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*\|_{\mathbf{W}_{k,h}}^2$$

$$+ 2\left(\bar{\mathbf{G}}_{k,h}(\boldsymbol{\theta}_h^k) - \mathbf{G}_{k,h}(\boldsymbol{\theta}_h^k) \right)^\top (\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*).$$

26

Then summing over k gives

$$
\|\boldsymbol{\theta}_h^{k+1} - \boldsymbol{\theta}_h^*\|_{\mathbf{A}_{k+1,h}}^2 \leq \|\boldsymbol{\theta}_{1,h} - \boldsymbol{\theta}_h^*\|_{\mathbf{A}_{1,h}}^2 + 2\sum_{i=1}^k \max_{s'\in\mathcal{S}_{i,h}} \|\boldsymbol{\varphi}_{i,h,s'}\|_{\mathbf{A}_{i+1,h}^{-1}}^2 - \frac{\kappa}{2}\sum_{i=1}^k \|\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*\|_{\mathbf{W}_{i,h}}^2
$$

$$
+ 2\sum_{i=1}^k \left(\bar{\mathbf{G}}_{i,h}(\boldsymbol{\theta}_h^i) - \mathbf{G}_{i,h}(\boldsymbol{\theta}_h^i)\right)^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*)
$$

$$
\leq 2\lambda L_\theta^2 + 2\sum_{i=1}^k \max_{s'\in\mathcal{S}_{i,h}} \|\boldsymbol{\varphi}_{i,h,s'}\|_{\mathbf{A}_{i+1,h}^{-1}}^2 - \frac{\kappa}{2}\sum_{i=1}^k \|\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*\|_{\mathbf{W}_{i,h}}^2
$$

$$
+ 2\sum_{i=1}^k \left(\bar{\mathbf{G}}_{i,h}(\boldsymbol{\theta}_h^i) - \mathbf{G}_{i,h}(\boldsymbol{\theta}_h^i)\right)^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*).
$$

For the final step, note that $\left(\bar{\mathbf{G}}_{i,h}(\boldsymbol{\theta}_h^i) - \mathbf{G}_{i,h}(\boldsymbol{\theta}_h^i)\right)^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*)$ is a martingale difference sequence. To bound this term, we invoke the following lemmas:

Lemma 2. *For $\delta \in (0,1)$ and $(k,h) \in [K] \times [H]$, with a probability at least $1 - \delta$ we have*

$$
\sum_{i=1}^k \left(\bar{\mathbf{G}}_{i,h}(\boldsymbol{\theta}_h^i) - \mathbf{G}_{i,h}(\boldsymbol{\theta}_h^i)\right)^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}^*)
$$

$$
\leq \frac{\kappa}{4}\sum_{i=1}^k \|\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*\|_{\mathbf{W}_{i,h}}^2 + \left(\frac{16 L_\varphi L_\theta}{3} + \frac{8}{\kappa}\right)\log \frac{(1 + \lceil 2\log_2 k\mathcal{U}L_\varphi L_\theta\rceil)\, k^2}{\delta} + \sqrt{2}\,.
$$

Lemma 3 (Generalized elliptical potential). *Let $S_t := \{\mathbf{x}_{t,1}, \ldots, \mathbf{x}_{t,K}\} \subset \mathbb{R}^d$. For any $1 \leq t \leq T$ and $i \in [K]$, suppose $\|\mathbf{x}_{t,i}\|_2 \leq L$. Let $\mathbf{V}_t := \lambda \mathbf{I}_d + \sum_{\tau=1}^{t-1}\sum_{i\in S_\tau} \mathbf{x}_{\tau,i}\mathbf{x}_{\tau,i}^\top$ for some $\lambda > 0$. If $\lambda \geq L^2$, then we have*

$$
\sum_{t=1}^T \max_{i\in[K]} \|\mathbf{x}_{t,i}\|_{\mathbf{V}_t^{-1}}^2 \leq 2d\log\left(1 + \frac{TKL}{d\lambda}\right).
$$

By Lemma 2, with probability at least $1 - \delta$, we have

$$
\|\boldsymbol{\theta}_h^{k+1} - \boldsymbol{\theta}_h^*\|_{\mathbf{A}_{k+1,h}}^2
$$

$$
\leq 2\lambda L_\theta^2 + 2\sum_{i=1}^k \max_{s'\in\mathcal{S}_{i,h}} \|\boldsymbol{\varphi}_{i,h,s'}\|_{\mathbf{A}_{i+1,h}^{-1}}^2
$$

$$
+ \left(\frac{32 L_\varphi L_\theta}{3} + \frac{16}{\kappa}\right)\log \frac{(1 + \lceil 2\log_2 k\mathcal{U}L_\varphi L_\theta\rceil)\, k^2}{\delta} + 2\sqrt{2}
$$

$$
\leq 2\lambda L_\theta^2 + \frac{8}{\kappa}d\log\left(1 + \frac{k\mathcal{U}L_\varphi^2}{d\lambda}\right) + \left(\frac{32 L_\varphi L_\theta}{3} + \frac{16}{\kappa}\right)\log \frac{(1 + \lceil 2\log_2 k\mathcal{U}L_\varphi L_\theta\rceil)\, k^2}{\delta} + 2\sqrt{2}\,,
$$

where the second inequality comes from Lemma 3. Note that the Gram matrix $\mathbf{A}_{k,h}$ in Algorithm 1 and the Gram matrix $\mathbf{V}$ in Lemma 3 are different by the factor of $\frac{\kappa}{2}$, which results in additional $\frac{2}{\kappa}$ factor for the bound of $\sum_{i=1}^k \max_{s'\in\mathcal{S}_{i,h}} \|\boldsymbol{\varphi}_{i,h,s'}\|_{\mathbf{A}_{i+1,h}^{-1}}^2$. $\qquad\square$

In the following, we provide all the proofs of the lemmas used to prove Lemma 1.

27

C.1.1 Proof of Lemma 2

Proof of Lemma 2. Note that $\left(\bar{\mathbf{G}}_{i,h}(\boldsymbol{\theta}_h^i) - \mathbf{G}_{i,h}(\boldsymbol{\theta}_h^i)\right)^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*)$ is a martingale difference sequence, i.e.,

$$
\mathbb{E}\left[\left(\bar{\mathbf{G}}_{i,h}(\boldsymbol{\theta}_h^i) - \mathbf{G}_{i,h}(\boldsymbol{\theta}_h^i)\right)^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*) \mid \mathcal{F}_{i,h}\right]
$$
$$
= \left(\bar{\mathbf{G}}_{i,h}(\boldsymbol{\theta}_h^i) - \mathbb{E}\left[\mathbf{G}_{i,h}(\boldsymbol{\theta}_h^i) \mid \mathcal{F}_{i,h}\right]\right)^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*)
$$
$$
= 0.
$$

On the other hand, for any $\boldsymbol{\theta} \in \mathbb{R}^d$, since we have

$$
\begin{aligned}
\|\mathbf{G}_{i,h}(\boldsymbol{\theta})\|_2 &= \left\|\sum_{s' \in \mathcal{S}_{i,h}} \left(P_{\boldsymbol{\theta}}(s' \mid s_h^i, a_h^i) - y_h^i(s')\right) \boldsymbol{\varphi}_{i,h,s'}\right\|_2 \\
&\leq \sum_{s' \in \mathcal{S}_{i,h}} \left|P_{\boldsymbol{\theta}}(s' \mid s_h^i, a_h^i) - y_h^i(s')\right| \|\boldsymbol{\varphi}_{i,h,s'}\|_2 \\
&\leq L_{\varphi} \left(\sum_{s' \in \mathcal{S}_{i,h}} P_{\boldsymbol{\theta}}(s' \mid s_h^i, a_h^i) + \sum_{s' \in \mathcal{S}_{i,h}} y_h^i(s')\right) \\
&= 2L_{\varphi},
\end{aligned}
$$

then, it follows by

$$
\begin{aligned}
&\left|\left(\bar{\mathbf{G}}_{i,h}(\boldsymbol{\theta}_h^i) - \mathbf{G}_{i,h}(\boldsymbol{\theta}_h^i)\right)^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*)\right| \\
&\leq \left|\left(\bar{\mathbf{G}}_{i,h}(\boldsymbol{\theta}_h^i)\right)^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*)\right| + \left|\left(\mathbf{G}_{i,h}(\boldsymbol{\theta}_h^i)\right)^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*)\right| \\
&\leq \|\bar{\mathbf{G}}_{i,h}(\boldsymbol{\theta}_h^i)\|_2 \|\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*\|_2 + \|\mathbf{G}_{i,h}(\boldsymbol{\theta}_h^i)\|_2 \|\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*\|_2 \\
&\leq 4L_{\varphi} \|\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*\|_2 \\
&\leq 8L_{\varphi} L_{\boldsymbol{\theta}},
\end{aligned} \tag{20}
$$

where the last inequality follows by $\|\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*\|_2 \leq \|\boldsymbol{\theta}_h^i\|_2 + \|\boldsymbol{\theta}_h^*\|_2 \leq 2L_{\boldsymbol{\theta}}$. Hence, if we denote $M_{k,h} := \sum_{i=1}^k \left(\bar{\mathbf{G}}_{i,h}(\boldsymbol{\theta}_h^i) - \mathbf{G}_{i,h}(\boldsymbol{\theta}_h^i)\right)^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*)$, then $M_{k,h}$ is a martingale. Note that we also

28

have

$$
\begin{aligned}
\Sigma_{k,h} &= \sum_{i=1}^{k} \mathbb{E}_{y_h^i} \left[\left(\left[\bar{\mathbf{G}}_{i,h}(\boldsymbol{\theta}_h^i) - \mathbf{G}_{i,h}(\boldsymbol{\theta}_h^i) \right]^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*) \right)^2 \right] \\
&= \sum_{i=1}^{k} \mathbb{E}_{y_h^i} \left[\left(\left[\mathbf{G}_{i,h}(\boldsymbol{\theta}_h^i) \right]^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*) \right)^2 \right] - \mathbb{E}_{y_h^i} \left[\left(\left[\bar{\mathbf{G}}_{i,h}(\boldsymbol{\theta}_h^i) \right]^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*) \right)^2 \right] \\
&\leq \sum_{i=1}^{k} \mathbb{E}_{y_h^i} \left[\left(\left[\mathbf{G}_{i,h}(\boldsymbol{\theta}_h^i) \right]^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*) \right)^2 \right] \\
&= \sum_{i=1}^{k} \mathbb{E}_{y_h^i} \left[\left(\sum_{s' \in \mathcal{S}_{i,h}} \left(P_{\boldsymbol{\theta}_h^i}(s' \mid s_h^i, a_h^i) - y_h^i(s') \right) \boldsymbol{\varphi}_{i,h,s'}^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*) \right)^2 \right] \\
&\leq \sum_{i=1}^{k} \mathbb{E}_{y_h^i} \left[\left(\sum_{s' \in \mathcal{S}_{i,h}} \left(P_{\boldsymbol{\theta}_h^i}(s' \mid s_h^i, a_h^i) - y_h^i(s') \right)^2 \right) \left(\sum_{s' \in \mathcal{S}_{i,h}} (\boldsymbol{\varphi}_{i,h,s'}^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*))^2 \right) \right] \\
& \hspace{11.5cm} (21) \\
&= \sum_{i=1}^{k} \mathbb{E}_{y_h^i} \left[\sum_{s' \in \mathcal{S}_{i,h}} \left(P_{\boldsymbol{\theta}_h^i}(s' \mid s_h^i, a_h^i) - y_h^i(s') \right)^2 \right] \left(\sum_{s' \in \mathcal{S}_{i,h}} (\boldsymbol{\varphi}_{i,h,s'}^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*))^2 \right) \\
&\leq 2 \sum_{i=1}^{k} \sum_{s' \in \mathcal{S}_{i,h}} \left(\boldsymbol{\varphi}_{i,h,s'}^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*) \right)^2 \hspace{5.5cm} (22) \\
&= 2 \sum_{i=1}^{k} \| \boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^* \|_{\mathbf{W}_{i,h}}^2 =: B_{k,h} \, ,
\end{aligned}
$$

where (21) holds by the Cauchy–Schwarz inequality, (22) holds because

$$
\begin{aligned}
&\sum_{s' \in \mathcal{S}_{i,h}} \left(P_{\boldsymbol{\theta}_h^i}(s' \mid s_h^i, a_h^i) - y_h^i(s') \right)^2 \\
&= \sum_{s' \in \mathcal{S}_{i,h}} \left\{ P_{\boldsymbol{\theta}_h^i}(s' \mid s_h^i, a_h^i) \right\}^2 - 2 P_{\boldsymbol{\theta}_h^i}(s' \mid s_h^i, a_h^i) y_h^i(s') + \left\{ y_h^i(s') \right\}^2 \\
&\leq 2 \, .
\end{aligned}
$$

However, if we denote $B_{k,h} := 2 \sum_{i=1}^{k} \| \boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^* \|_{\mathbf{W}_{i,h}}^2$, since $B_{k,h}$ is itself a random variable, to apply Freedman's inequality to $M_{k,h}$, we consider two cases depending on the values of $B_{k,h}$.

Case 1 : $B_{k,h} \leq \frac{4}{k\mathcal{U}}$

29

Suppose that $B_{k,h} = 2\sum_{i=1}^{k} \|\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*\|_{\mathbf{W}_{i,h}}^2 \leq \frac{4}{k\mathcal{U}}$. Then we have

$$
\begin{aligned}
M_{k,h} &= \sum_{i=1}^{k} \left(\bar{\mathbf{G}}_{i,h}(\boldsymbol{\theta}_h^i) - \mathbf{G}_{i,h}(\boldsymbol{\theta}_h^i) \right)^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*) \\
&= \sum_{i=1}^{k} \sum_{s' \in \mathcal{S}_{i,h}} \left(y_h^i(s') - \mathbb{E}[y_h^i(s')] \right) \boldsymbol{\varphi}_{i,h,s'}^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*) \\
&= \sum_{i=1}^{k} \sum_{s' \in \mathcal{S}_{i,h}} \left(y_h^i(s') - P_{\boldsymbol{\theta}_h^*}(s' \mid s_h^i, a_h^i) \right) \boldsymbol{\varphi}_{i,h,s'}^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*) \\
&\leq \sum_{i=1}^{k} \sum_{s' \in \mathcal{S}_{i,h}} |\boldsymbol{\varphi}_{i,h,s'}^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*)| \\
&\leq \sqrt{k\mathcal{U} \sum_{i=1}^{k} \sum_{s' \in \mathcal{S}_{i,h}} \left(\boldsymbol{\varphi}_{i,h,s'}^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*) \right)^2} \\
&= \sqrt{k\mathcal{U} \frac{B_{k,h}}{2}} \\
&\leq \sqrt{2}\,.
\end{aligned}
$$

Case 2 : $B_{k,h} > \frac{4}{k\mathcal{U}}$

Suppose that $B_{k,h} = 2\sum_{i=1}^{k} \|\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*\|_{\mathbf{W}_{i,h}}^2 > \frac{4}{k\mathcal{U}}$. Then, we have both a lower and upper bound for $B_{k,h}$ as follows:

$$
\frac{4}{k\mathcal{U}} < B_{k,h} \leq 2 \sum_{i=1}^{k} \sum_{s' \in \mathcal{S}_{i,h}} \|\boldsymbol{\varphi}_{i,h,s'}\|_2^2 \|\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*\|_2^2 \leq 8k\mathcal{U}L_{\boldsymbol{\varphi}}^2 L_{\boldsymbol{\theta}}^2\,.
$$

Then by the peeling process from Bartlett et al. [12], for any $\eta_k > 0$, we have

$$
\begin{aligned}
&\mathbb{P}\left(M_{k,h} \geq 2\sqrt{\eta_k B_{k,h}} + \frac{16\eta_k L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}}{3} \right) \\
&= \mathbb{P}\left(M_{k,h} \geq 2\sqrt{\eta_k B_{k,h}} + \frac{16\eta_k L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}}{3}, \frac{4}{k\mathcal{U}} < B_{k,h} \leq 8k\mathcal{U}L_{\boldsymbol{\varphi}}^2 L_{\boldsymbol{\theta}}^2 \right) \\
&= \mathbb{P}\left(M_{k,h} \geq 2\sqrt{\eta_k B_{k,h}} + \frac{16\eta_k L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}}{3}, \frac{4}{k\mathcal{U}} < B_{k,h} \leq 8k\mathcal{U}L_{\boldsymbol{\varphi}}^2 L_{\boldsymbol{\theta}}^2, \Sigma_{k,h} \leq B_{k,h} \right) \\
&\leq \sum_{j=1}^{m} \mathbb{P}\left(M_{k,h} \geq 2\sqrt{\eta_k B_{k,h}} + \frac{16\eta_k L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}}{3}, \frac{4 \cdot 2^{j-1}}{k\mathcal{U}} < B_{k,h} \leq \frac{4 \cdot 2^j}{k\mathcal{U}}, \Sigma_{k,h} \leq B_{k,h} \right) \\
&\leq \sum_{j=1}^{m} \underbrace{\mathbb{P}\left(M_{k,h} \geq \sqrt{\eta_k \frac{8 \cdot 2^j}{k\mathcal{U}}} + \frac{16\eta_k L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}}{3}, \Sigma_{k,h} \leq \frac{4 \cdot 2^j}{k\mathcal{U}} \right)}_{I_j}, \tag{23}
\end{aligned}
$$

where $m = 1 + \lceil 2\log_2 k\mathcal{U}L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}} \rceil$. For I_j, note that from (20) we have

$$
\left| \left(\bar{\mathbf{G}}_{i,h}(\boldsymbol{\theta}_h^i) - \mathbf{G}_{i,h}(\boldsymbol{\theta}_h^i) \right)^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*) \right| \leq 8L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}\,.
$$

30

By Freedman's inequality (Lemma 29), we have

$$
\mathbb{P}\left(M_{k,h} \geq \sqrt{\eta_k \frac{8 \cdot 2^j}{k\mathcal{U}}} + \frac{16\eta_k L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}}{3}, \Sigma_{k,h} \leq \frac{4 \cdot 2^j}{k\mathcal{U}} \right)
$$

$$
\leq \exp\left(\frac{-\left(\sqrt{\eta_k \frac{8\cdot 2^j}{k\mathcal{U}}} + \frac{16\eta_k L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}}{3} \right)^2}{\frac{8\cdot 2^j}{k\mathcal{U}} + \frac{2}{3} \cdot 8 L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}} \left(\sqrt{\eta_k \frac{8\cdot 2^j}{k\mathcal{U}}} + \frac{16\eta_k L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}}{3} \right)} \right)
$$

$$
= \exp\left(\frac{-\eta_k \left(\sqrt{\frac{8\cdot 2^j}{k\mathcal{U}}} + \frac{16\sqrt{\eta_k} L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}}{3} \right)^2}{\frac{8\cdot 2^j}{k\mathcal{U}} + \frac{16 L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}}{3} \sqrt{\eta_k \frac{8\cdot 2^j}{k\mathcal{U}}} + \frac{16^2 \eta_k L_{\boldsymbol{\varphi}}^2 L_{\boldsymbol{\theta}}^2}{3^2}} \right)
$$

$$
\leq \exp\left(\frac{-\eta_k \left(\sqrt{\frac{8\cdot 2^j}{k\mathcal{U}}} + \frac{16\sqrt{\eta_k} L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}}{3} \right)^2}{\frac{8\cdot 2^j}{k\mathcal{U}} + \frac{32 L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}}{3} \sqrt{\eta_k \frac{8\cdot 2^j}{k\mathcal{U}}} + \frac{16^2 \eta_k L_{\boldsymbol{\varphi}}^2 L_{\boldsymbol{\theta}}^2}{3^2}} \right)
$$

$$
= \exp(-\eta_k). \tag{24}
$$

By substituting Eq. (24) into Eq. (23), we have

$$
\mathbb{P}\left(M_{k,h} \geq 2\sqrt{\eta_k B_{k,h}} + \frac{16\eta_k L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}}{3} \right) \leq m \exp(-\eta_k).
$$

Then, combining with the result of Case 1 & 2, letting $\eta_k = \log \frac{m}{\delta/k^2} = \log \frac{(1 + \lceil 2 \log_2 k\mathcal{U} L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}} \rceil) k^2}{\delta}$ and taking union bound over k, with probability at least $1 - \delta$, we have

$$
M_{k,h} \leq 2\sqrt{2\eta_k \sum_{i=1}^k \|\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*\|_{\mathbf{W}_{i,h}}^2} + \frac{16\eta_k L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}}{3} + \sqrt{2}. \tag{25}
$$

By applying $2\sqrt{ab} \leq a + b$ to the first term on the right hand side, we have

$$
2\sqrt{2\eta_k \sum_{i=1}^k \|\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*\|_{\mathbf{W}_{i,h}}^2} \leq \frac{8\eta_k}{\kappa} + \frac{\kappa}{4} \sum_{i=1}^k \|\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*\|_{\mathbf{W}_{i,h}}^2. \tag{26}
$$

Combining the results of Eq. (25) & Eq. (26), we have

$$
M_{k,h} = \sum_{i=1}^k \left(\bar{\mathbf{G}}_{i,h}(\boldsymbol{\theta}_h^i) - \mathbf{G}_{i,h}(\boldsymbol{\theta}_h^i) \right)^\top (\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*)
$$

$$
\leq \frac{\kappa}{4} \sum_{i=1}^k \|\boldsymbol{\theta}_h^i - \boldsymbol{\theta}_h^*\|_{\mathbf{W}_{i,h}}^2 + \left(\frac{16 L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}}{3} + \frac{8}{\kappa} \right) \log \frac{(1 + \lceil 2 \log_2 k\mathcal{U} L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}} \rceil) k^2}{\delta} + \sqrt{2}.
$$

$\square$

31

C.1.2   Proof of Lemma 3

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

$$\det(\mathbf{V}_{t+1}) = \det\left(\mathbf{V}_t + \sum_{i \in S_t} \mathbf{x}_{t,i}\mathbf{x}_{t,i}^\top\right)$$

$$= \det(\mathbf{V}_t)\det\left(\mathbf{I}_d + \sum_{i \in S_t} \mathbf{V}_t^{-\frac{1}{2}}\mathbf{x}_{t,i}\mathbf{x}_{t,i}^\top\mathbf{V}_t^{-\frac{1}{2}}\right)$$

$$= \det(\mathbf{V}_t)\left(1 + \sum_{i \in S_t} \|\mathbf{x}_{t,i}\|_{\mathbf{V}_t^{-1}}^2\right)$$

$$= \det(\lambda\mathbf{I}_d)\prod_{\tau=1}^{t}\left(1 + \sum_{i \in S_\tau} \|\mathbf{x}_{\tau,i}\|_{\mathbf{V}_\tau^{-1}}^2\right)$$

$$\geq \det(\lambda\mathbf{I}_d)\prod_{\tau=1}^{t}\left(1 + \max_{i \in S_\tau} \|\mathbf{x}_{\tau,i}\|_{\mathbf{V}_t^{-1}}^2\right). \tag{27}$$

Since $\lambda \geq L^2$, we have

$$\max_{i \in S_\tau} \|\mathbf{x}_{\tau,i}\|_{\mathbf{V}_\tau^{-1}}^2 \leq \frac{L^2}{\lambda} \leq 1.$$

Since for any $z \in [0,1]$, it follows that $z \leq 2\log(1+z)$. Hence, we have

$$\sum_{t=1}^{T}\max_{i \in S_t}\|\mathbf{x}_{t,i}\|_{\mathbf{V}_t^{-1}}^2 \leq 2\sum_{t=1}^{T}\log\left(1 + \max_{i \in S_t}\|\mathbf{x}_{t,i}\|_{\mathbf{V}_t^{-1}}^2\right)$$

$$= 2\log\prod_{t=1}^{T}\left(1 + \max_{i \in S_t}\|\mathbf{x}_{t,i}\|_{\mathbf{V}_t^{-1}}^2\right)$$

$$\leq 2\log\frac{\det(\mathbf{V}_{T+1})}{\det(\lambda\mathbf{I}_d)}$$

$$\leq 2d\log\left(1 + \frac{TKL^2}{d\lambda}\right),$$

where the second inequality comes from Eq. (27) and the last inequality follows by the determinant-trace inequality (Lemma 28). $\square$

C.2   Bound on Prediction Error

In this section, we provide the bound on the prediction error induced by estimated transition core $\boldsymbol{\theta}_h^k$.

Lemma 4 (Bound on Prediction Error). *For any $\delta \in (0,1)$, suppose that Lemma 1 holds. Then for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have*

$$|\Delta_h^k(s,a)| \leq H\alpha_k(\delta)\|\hat{\boldsymbol{\varphi}}_{k,h}(s,a)\|_{\mathbf{A}_{k,h}^{-1}}.$$

Proof of Lemma 4. Recall that

$$\Delta_h^k(s,a) = \sum_{s' \in \mathcal{S}_{s,a}}\left(P_{\boldsymbol{\theta}_h^k}(s' \mid s,a) - P_{\boldsymbol{\theta}_h^*}(s' \mid s,a)\right)V_{h+1}^k(s')$$

$$= \sum_{s' \in \mathcal{S}_{s,a}}\frac{\exp(\boldsymbol{\varphi}_{s,a,s'}^\top\boldsymbol{\theta}_h^k)V_{h+1}^k(s')}{\sum_{\widetilde{s} \in \mathcal{S}_{s,a}}\exp(\boldsymbol{\varphi}_{s,a,\widetilde{s}}^\top\boldsymbol{\theta}_h^k)} - \sum_{s' \in \mathcal{S}_{s,a}}\frac{\exp(\boldsymbol{\varphi}_{s,a,s'}^\top\boldsymbol{\theta}_h^*)V_{h+1}^k(s')}{\sum_{\widetilde{s} \in \mathcal{S}_{s,a}}\exp(\boldsymbol{\varphi}_{s,a,\widetilde{s}}^\top\boldsymbol{\theta}_h^*)}.$$

32

Then by the mean value theorem, there exists $\bar{\boldsymbol{\theta}} = \rho\boldsymbol{\theta}_h^k + (1-\rho)\boldsymbol{\theta}_h^*$ for some $\rho \in [0,1]$ satisfying that

$$
\Delta_h^k(s,a) = \frac{\left(\sum_{s' \in \mathcal{S}_{s,a}} \exp(\boldsymbol{\varphi}_{s,a,s'}^\top \bar{\boldsymbol{\theta}}) V_{h+1}^k(s') \boldsymbol{\varphi}_{s,a,s'}^\top (\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*)\right) \left(\sum_{\widetilde{s} \in \mathcal{S}_{s,a}} \exp(\boldsymbol{\varphi}_{s,a,\widetilde{s}}^\top \bar{\boldsymbol{\theta}})\right)}{\left(\sum_{\widetilde{s} \in \mathcal{S}_{s,a}} \exp(\boldsymbol{\varphi}_{s,a,\widetilde{s}}^\top \bar{\boldsymbol{\theta}})\right)^2}
$$

$$
- \frac{\left(\sum_{s' \in \mathcal{S}_{s,a}} \exp(\boldsymbol{\varphi}_{s,a,s'}^\top \bar{\boldsymbol{\theta}}) V_{h+1}^k(s')\right) \left(\sum_{\widetilde{s} \in \mathcal{S}_{s,a}} \exp(\boldsymbol{\varphi}_{s,a,\widetilde{s}}^\top \bar{\boldsymbol{\theta}}) \boldsymbol{\varphi}_{s,a,\widetilde{s}}^\top (\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*)\right)}{\left(\sum_{\widetilde{s} \in \mathcal{S}_{s,a}} \exp(\boldsymbol{\varphi}_{s,a,\widetilde{s}}^\top \bar{\boldsymbol{\theta}})\right)^2}
$$

$$
= \sum_{s' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s' \mid s,a) V_{h+1}^k(s') \boldsymbol{\varphi}_{s,a,s'}^\top (\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*)
$$

$$
- \left(\frac{\sum_{s' \in \mathcal{S}_{s,a}} \exp(\boldsymbol{\varphi}_{s,a,s'}^\top \bar{\boldsymbol{\theta}}) V_{h+1}^k(s')}{\sum_{\widetilde{s} \in \mathcal{S}_{k,h}} \exp(\boldsymbol{\varphi}_{s,a,\widetilde{s}}^\top \bar{\boldsymbol{\theta}})}\right) \sum_{s' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s' \mid s,a) \boldsymbol{\varphi}_{s,a,s'}^\top (\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*)
$$

$$
= \sum_{s' \in \mathcal{S}_{s,a}} \left(V_{h+1}^k(s') - \frac{\sum_{s' \in \mathcal{S}_{s,a}} \exp(\boldsymbol{\varphi}_{s,a,s'}^\top \bar{\boldsymbol{\theta}}) V_{h+1}^k(s')}{\sum_{\widetilde{s} \in \mathcal{S}_{s,a}} \exp(\boldsymbol{\varphi}_{s,a,\widetilde{s}}^\top \bar{\boldsymbol{\theta}})}\right) P_{\bar{\boldsymbol{\theta}}}(s' \mid s,a) \boldsymbol{\varphi}_{s,a,s'}^\top (\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*) .
$$

Since $V_h^k(s') \le H$ for all $s' \in \mathcal{S}, k \in [K]$, and $h \in [H]$, we have

$$
\Delta_h^k(s,a) \le H \sum_{s' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s' \mid s,a) \boldsymbol{\varphi}_{s,a,s'}^\top (\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*)
$$

$$
\le H \max_{s' \in \mathcal{S}_{s,a}} |\boldsymbol{\varphi}_{s,a,s'}^\top (\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*)|
$$

$$
\le H \max_{s' \in \mathcal{S}_{s,a}} \|\boldsymbol{\varphi}_{s,a,s'}\|_{\mathbf{A}_{k,h}^{-1}} \|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h^*\|_{\mathbf{A}_{k,h}}
$$

$$
\le H\alpha_k(\delta) \|\hat{\boldsymbol{\varphi}}_{k,h}(s,a)\|_{\mathbf{A}_{k,h}^{-1}} ,
$$

where the second inequality comes from the fact that $P_{\bar{\boldsymbol{\theta}}}(s' \mid s,a) \le 1$ is a multinomial probability, the third inequality holds due to the Cauchy-Schwarz inequality, and the last inequality follows from Lemma 1 and the definition of $\hat{\boldsymbol{\varphi}}_{k,h}$, i.e., $\hat{\boldsymbol{\varphi}}_{k,h}(s,a) := \boldsymbol{\varphi}(s,a,\hat{s})$ for $\hat{s} = \operatorname{argmax}_{s' \in \mathcal{S}_{s,a}} \|\boldsymbol{\varphi}(s,a,s')\|_{\mathbf{A}_{k,h}^{-1}}$. $\qquad\square$

C.3 Good Events with High Probability

Lemma 5 (Good event probability). *For any $K \in \mathbb{N}$ and $\delta \in (0,1)$, the good event $\mathcal{G}(K,\delta')$ holds with probability at least $1 - \delta$ where $\delta' = \delta/(2KH)$.*

Proof of Lemma 5. For any $\delta' \in (0,1)$, we have

$$
\mathcal{G}(K,\delta') = \bigcap_{k \le K} \bigcap_{h \le H} \mathcal{G}_{k,h}(\delta') = \bigcap_{k \le K} \bigcap_{h \le H} \left\{\mathcal{G}_{k,h}^\Delta(\delta') \cap \mathcal{G}_{k,h}^{\boldsymbol{\xi}}(\delta')\right\} .
$$

On the other hand, for any $(k,h) \in [K] \times [H]$, by Lemma 30, $\mathcal{G}_{k,h}^{\boldsymbol{\xi}}(\delta')$ holds with probability at least $1 - \delta'$. Then, for $\delta' = \delta/(2KH)$ by taking union bound, we have the desired result as follows:

$$
\mathbb{P}(\mathcal{G}(K,\delta')) \ge (1-\delta')^{2KH} \ge 1 - 2KH\delta' = 1 - \delta .
$$

$\qquad\square$

C.4 Stochastic Optimism

Lemma 6 (Stochastic optimism). *For any δ with $0 < \delta < \Phi(-1)/2$, let $\sigma_k = H\alpha_k(\delta) = \widetilde{\mathcal{O}}(H\sqrt{d})$. If we take multiple sample size $M = \lceil 1 - \frac{\log H}{\log \Phi(1)} \rceil$, then for any $k \in [K]$, we have*

$$
\mathbb{P}\left((V_1^k - V_1^*)(s_1^k) \ge 0 \mid s_1^k, \mathcal{F}_k\right) \ge \Phi(-1)/2 .
$$

33

Proof of lemma 6. Before presenting the proof, we introduce the following lemmas.

Lemma 7. *For any $k \in [K]$, it holds*

$$V_1^k(s_1^k) - V_1^*(s_1^k) \geq \mathbb{E}_{\pi^*}\left[\sum_{h=1}^H -\iota_h^k(x_h, a_h) \mid x_1 = s_1^k\right],$$

where $\iota_h^k(s,a) := r(s,a) + P_h V_{h+1}^k(s,a) - Q_h^k(s,a)$.

Lemma 8. *Let $\delta \in (0,1)$ be given. For any $(k,h) \in [K] \times [H]$, let $\sigma_k = H\alpha_k(\delta)$. If we define the event $\mathcal{G}_{k,h}^\Delta(\delta)$ as*

$$\mathcal{G}_{k,h}^\Delta(\delta) := \left\{\Delta_h^k(s,a) \leq H\alpha_k(\delta)\|\hat{\varphi}_{k,h}(s,a)\|_{\mathbf{A}_{k,h}^{-1}}\right\},$$

then conditioned on $\mathcal{G}_{k,h}^\Delta(\delta)$, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\mathbb{P}\left(-\iota_h^k(s,a) \geq 0 \mid \mathcal{G}_{k,h}^\Delta(\delta)\right) \geq 1 - \Phi(1)^M.$$

Lemma 9. *Let $\delta \in (0,1)$ be given. For any $(h,k) \in [H] \times [K]$, let $\sigma_k = H\alpha_k(\delta)$. If we take multiple sample size $M = \lceil 1 - \frac{\log H}{\log \Phi(1)}\rceil$, then conditioned on the event $\mathcal{G}_k^\Delta(\delta) := \bigcap_{h \in [H]} \mathcal{G}_{k,h}^\Delta(\delta)$, we have*

$$\mathbb{P}\left(-\iota_h^k(s_h, a_h) \geq 0, \forall h \in [H] \mid \mathcal{G}_k^\Delta(\delta)\right) \geq \Phi(-1).$$

Now, we define the event of the estimated value function being optimistic at the start of the k-th episode as

$$\mathcal{X}_k := \left\{(V_1^k - V_1^*)(s_1^k) \geq 0\right\}.$$

Then for the event $\mathcal{G}_k(\delta) =: \mathcal{G}_k$, we have

$$\begin{aligned}
\mathbb{P}(\mathcal{X}_k) &= 1 - \mathbb{P}(\mathcal{X}_k^\mathsf{c}) \\
&= 1 - \mathbb{P}(\mathcal{X}_k^\mathsf{c} \cap \mathcal{G}_k) - \mathbb{P}(\mathcal{X}_k^\mathsf{c} \cap \mathcal{G}_k^\mathsf{c}) \\
&\geq 1 - \mathbb{P}(\mathcal{X}_k^\mathsf{c} \cap \mathcal{G}_k) - \mathbb{P}(\mathcal{G}_k^\mathsf{c}) \\
&\geq 1 - \mathbb{P}(\mathcal{X}_k^\mathsf{c} \cap \mathcal{G}_k) - \delta
\end{aligned}$$

where the last inequality comes from lemma 5.

On the other hand, by Lemma 7, we have

$$\begin{aligned}
V_1^k(s_1^k) - V_1^*(s_1^k) &\geq \mathbb{E}_{\pi^*}\left[\sum_{h=1}^H -\iota_h^k(x_h, a_h) \mid x_1 = s_1^k\right] \\
&= \sum_{h=1}^H \mathbb{E}_{\pi^*}\left[-\iota_h^k(x_h, a_h) \mid x_1 = s_1^k\right].
\end{aligned}$$

If we define an event

$$\mathcal{Y}_k = \left\{\sum_{h=1}^H \mathbb{E}_{\pi^*}\left[-\iota_h^k(x_h, a_h) \mid x_1 = s_1^k\right] \geq 0\right\},$$

then, by Lemma 9, we have

$$\begin{aligned}
\mathbb{P}(\mathcal{Y}_k \mid \mathcal{G}_k) \geq \Phi(-1) &\iff \mathbb{P}(\mathcal{Y}_k^\mathsf{c} \mid \mathcal{G}_k) \leq 1 - \Phi(-1) \\
&\implies \mathbb{P}(\mathcal{Y}_k^\mathsf{c} \cap \mathcal{G}_k) \leq (1 - \Phi(-1))\,\mathbb{P}(\mathcal{G}_k) \leq 1 - \Phi(-1)
\end{aligned}$$

Note that since $\mathcal{X}_k^\mathsf{c} \cap \mathcal{G}_k \subset \mathcal{Y}_k^\mathsf{c} \cap \mathcal{G}_k$, we can conclude that

$$\begin{aligned}
\mathbb{P}(\mathcal{X}_k) &\geq 1 - \mathbb{P}(\mathcal{X}_k^\mathsf{c} \cap \mathcal{G}_k) - \delta \\
&\geq 1 - \mathbb{P}(\mathcal{Y}_k^\mathsf{c} \cap \mathcal{G}_k) - \delta \\
&\geq 1 - (1 - \Phi(-1)) - \delta \\
&= \Phi(-1) - \delta \\
&\geq \Phi(-1)/2
\end{aligned}$$

where the last inequality comes from the choice of δ. $\qquad\square$

In the following, we provide all the proofs of the lemmas used to prove Lemma 6.

34

C.4.1 Proof of Lemma 7

Proof of lemma 7. In this proof, we use x_h^k as the states sampled under the π^* to distinguish with s_h^k. Since we have,

$$V_1^k(s_1^k) - V_1^*(s_1^k)$$
$$\geq Q_1^k(s_1^k, \pi^*(s_1^k)) - Q_1^*(s_1^k, \pi^*(s_1^k))$$
$$= r(s_1^k, \pi^*(s_1^k)) + P_1 V_2^k(s_1^k, \pi^*(s_1^k)) - \iota_1^k(s_1^k, \pi^*(s_1^k)) - \left(r(s_1^k, \pi^*(s_1^k)) + P_1 V_2^*(s_1^k, \pi^*(s_1^k))\right)$$
$$= P_1(V_2^k - V_2^*)(s_1^k, \pi^*(s_1^k)) - \iota_1^k(s_1^k, \pi^*(s_1^k))$$
$$= \mathbb{E}_{x|s_1^k, \pi^*(s_1^k)} \left[(V_2^k - V_2^*)(x)\right] - \iota_1^k(s_1^k, \pi^*(s_1^k))$$
$$\geq \mathbb{E}_{x_2^k|s_1^k, \pi^*(s_1^k)} \left[(Q_2^k - Q_2^*)(x_2^k, \pi^*(x_2^k))\right] - \iota_1^k(s_1^k, \pi^*(s_1^k))$$
$$= \mathbb{E}_{x_2^k \sim s_1^k, \pi^*(s_1^k)} \left[\mathbb{E}_{x|x_2^k, \pi^*(x_2^k)} \left[(V_3^k - V_3^*)(x)\right] - \iota_2^k(x_2^k, \pi^*(x_2^k))\right] - \iota_1^k(s_1^k, \pi^*(s_1^k))$$
$$= \underbrace{\mathbb{E}_{x_2^k \sim s_1^k, \pi^*(s_1^k)} \left[\mathbb{E}_{x|x_2^k, \pi^*(x_2^k)} \left[(V_3^k - V_3^*)(x)\right]\right]}_{\mathbb{E}_{x_3^k \sim \pi^* | s_1^k} \left[(V_3^k - V_3^*)(x_3^k)\right]}$$
$$- \mathbb{E}_{x_2^k \sim s_1^k, \pi^*(s_1^k)} \left[\iota_2^k(x_2^k, \pi^*(x_2^k))\right] - \iota_1^k(s_1^k, \pi^*(s_1^k))$$

then by applying this argument recursively, we finally have

$$V_1^k(s_1^k) - V_1^*(s_1^k) \geq \mathbb{E}_{\pi^*} \left[\sum_{h=1}^H -\iota_h^k(x_h, a_h) \mid x_1 = s_1^k\right].$$

$\square$

C.4.2 Proof of Lemma 8

Proof of Lemma 8. Since we have

$$-\iota_h^k(s, a) = Q_h^k(s, a) - \left(r(s, a) + P_h V_{h+1}^k(s, a)\right)$$
$$= \min \left\{r(s, a) + \sum_{s' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}_h^k}(s' \mid s, a) V_{h+1}^k(s') + \max_{m \in [M]} \hat{\boldsymbol{\varphi}}_{k,h}(s, a)^\top \xi_{k,h}^{(m)}, H\right\}$$
$$- \left(r(s, a) + P_h V_{h+1}^k(s, a)\right)$$
$$\geq \min \left\{\sum_{s' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}_h^k}(s' \mid s, a) V_{h+1}^k(s') + \max_{m \in [M]} \hat{\boldsymbol{\varphi}}_{k,h}(s, a)^\top \xi_{k,h}^{(m)} - P_h V_{h+1}^k(s, a), 0\right\},$$

it is enough to show that

$$\sum_{s' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}_h^k}(s' \mid s, a) V_{h+1}^k(s') + \max_{m \in [M]} \hat{\boldsymbol{\varphi}}_{k,h}(s, a)^\top \xi_{k,h}^{(m)} - P_h V_{h+1}^k(s, a) \geq 0$$

at least with constant probability.

On the other hand, under the event $\mathcal{G}_{k,h}(\delta)$, by Lemma 4 we have

$$\sum_{s' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}_h^k}(s' \mid s, a) V_{h+1}^k(s') + \max_{m \in [M]} \hat{\boldsymbol{\varphi}}_{k,h}(s, a)^\top \xi_{k,h}^{(m)} - P_h V_{h+1}^k(s, a)$$
$$\geq \max_{m \in [M]} \hat{\boldsymbol{\varphi}}_{k,h}(s, a)^\top \xi_{k,h}^{(m)} - H\alpha_k(\delta) \|\hat{\boldsymbol{\varphi}}_{k,h}(s, a)\|_{\mathbf{A}_{k,h}^{-1}}.$$

Now, for $\forall m \in [M]$, since $\xi_{k,h}^{(m)} \sim \mathcal{N}(\mathbf{0}_d, \sigma_k^2 \mathbf{A}_{k,h}^{-1})$, we have

$$\hat{\boldsymbol{\varphi}}_{k,h}(s, a)^\top \xi_{k,h}^{(m)} \sim \mathcal{N}(0, \sigma_k^2 \|\hat{\boldsymbol{\varphi}}_{k,h}(s, a)\|_{\mathbf{A}_{k,h}^{-1}}^2),$$

35

which means,

$$\mathbb{P}\left(\hat{\boldsymbol{\varphi}}_{k,h}(s,a)^\top \xi_{k,h}^{(m)} \geq H\alpha_k(\delta)\|\hat{\boldsymbol{\varphi}}_{k,h}(s,a)\|_{\mathbf{A}_{k,h}^{-1}}\right) \geq \Phi(-1),$$

by setting $\sigma_k = H\alpha_k(\delta)$. Then, finally we have the desired results as follows:

$$
\begin{aligned}
&\mathbb{P}\left(-\iota_h^k(s,a) \geq 0 \mid \mathcal{G}_{k,h}^\Delta(\delta)\right)\\
&\geq \mathbb{P}\left(\max_{m\in[M]} \hat{\boldsymbol{\varphi}}_{k,h}(s,a)^\top \xi_{k,h}^{(m)} \geq H\alpha_k(\delta)\|\hat{\boldsymbol{\varphi}}_{k,h}(s,a)\|_{\mathbf{A}_{k,h}^{-1}} \mid \mathcal{G}_{k,h}^\Delta(\delta)\right)\\
&= 1 - \mathbb{P}\left(\hat{\boldsymbol{\varphi}}_{k,h}(s,a)^\top \xi_{k,h}^{(m)} < H\alpha_k(\delta)\|\hat{\boldsymbol{\varphi}}_{k,h}(s,a)\|_{\mathbf{A}_{k,h}^{-1}}, \forall m \in [M] \mid \mathcal{G}_{k,h}^\Delta(\delta)\right)\\
&\geq 1 - (1 - \Phi(-1))^M\\
&= 1 - \Phi(1)^M.
\end{aligned}
$$

$\square$

C.4.3 Proof of Lemma 9

Proof of Lemma 9. For each $h \in [H]$ and $k \in [K]$, define an event $\mathcal{E}_h^k := \{-\iota_h^k(s_h, a_h) \geq 0\}$ Then it holds

$$
\begin{aligned}
\mathbb{P}\left(-\iota_h^k(s_h, a_h) \geq 0, \forall h \in [H] \mid \mathcal{G}_k^\Delta(\delta)\right) &= \mathbb{P}\left(\bigcap_{h=1}^H \mathcal{E}_h^k \mid \mathcal{G}_k^\Delta(\delta)\right)\\
&= 1 - \mathbb{P}\left(\bigcup_{h=1}^H (\mathcal{E}_h^k)^{\mathsf{c}} \mid \mathcal{G}_k^\Delta(\delta)\right)\\
&\geq 1 - \sum_{h=1}^H \mathbb{P}\left((\mathcal{E}_h^k)^{\mathsf{c}} \mid \mathcal{G}_{k,h}^\Delta(\delta)\right)\\
&\geq 1 - H\Phi(1)^M\\
&\geq \Phi(-1)
\end{aligned}
$$

where the first inequality uses the union bound, the second inequality comes from the Lemma 8 and the last inequality holds due to the choice of $M = \lceil 1 - \frac{\log H}{\log \Phi(1)} \rceil$. $\square$

C.5 Bound on Estimation Part

We decompose the regret into the estimation part and the pessimism part as follows:

$$\sum_{k=1}^K (V_1^* - V_1^{\pi^k})(s_1^k) = \sum_{k=1}^K \left(\underbrace{V_1^* - V_1^k}_{\text{Pessimism}} + \underbrace{V_1^k - V_1^{\pi^k}}_{\text{Estimation}} \right)(s_1^k),$$

and we bound these two parts in the following sections, respectively.

Lemma 10 (Bound on estimation part)**.** *For any $\delta \in (0,1)$, if $\lambda \geq L_{\boldsymbol{\varphi}}^2$, then with probability at least $1 - \delta/2$, we have*

$$\sum_{k=1}^K (V_1^k - V_1^{\pi^k})(s_1^k) = \widetilde{\mathcal{O}}\left(\kappa^{-1} d^{\frac{3}{2}} H^{\frac{3}{2}} \sqrt{T}\right).$$

Proof of lemma 10. For any given $k \in [K]$,

$$
\begin{aligned}
(V_1^k - V_1^{\pi^k})(s_1^k) &= (Q_1^k - Q_1^{\pi^k})(s_1^k, a_1^k) + \iota_1^k(s_1^k, a_1^k) - \iota_1^k(s_1^k, a_1^k)\\
&= (Q_1^k - Q_1^{\pi^k})(s_1^k, a_1^k) + P_1(V_2^k - V_2^{\pi^k})(s_1^k, a_1^k) \qquad (28)\\
&\quad + (Q_1^{\pi^k} - Q_1^k)(s_1^k, a_1^k) - \iota_1^k(s_1^k, a_1^k)\\
&= \underbrace{P_1(V_2^k - V_2^{\pi^k})(s_1^k, a_1^k) - (V_2^k - V_2^{\pi^k})(s_2^k)}_{\dot{\zeta}_1^k} + (V_2^k - V_2^{\pi^k})(s_2^k) - \iota_1^k(s_1^k, a_1^k)
\end{aligned}
$$

36

where the second equality holds due to the variant of $\iota_h^k(s_h^k, a_h^k)$ as follows:

$$
\begin{aligned}
\iota_h^k(s_h^k, a_h^k) &= r(s_h^k, a_h^k) + P_h V_{h+1}^k(s_h^k, a_h^k) - Q_h^k(s_h^k, a_h^k) + Q_h^{\pi^k}(s_h^k, a_h^k) - Q_h^{\pi^k}(s_h^k, a_h^k) \\
&= r(s_h^k, a_h^k) + P_h V_{h+1}^k(s_h^k, a_h^k) - Q_h^k(s_h^k, a_h^k) \\
&\quad + Q_h^{\pi^k}(s_h^k, a_h^k) - \left(r(s_h^k, a_h^k) + P_h V_{h+1}^{\pi^k}(s_h^k, a_h^k) \right) \\
&= P_h(V_{h+1}^k - V_{h+1}^{\pi^k})(s_h^k, a_h^k) + (Q_h^{\pi^k} - Q_h^k)(s_h^k, a_h^k).
\end{aligned}
$$

Then, by applying this argument recursively for whole horizon, we have

$$
(V_1^k - V_1^{\pi^k})(s_1^k) = \sum_{h=1}^{H} -\iota_h^k(s_h^k, a_h^k) + \sum_{h=1}^{H} \dot{\zeta}_h^k, \tag{29}
$$

where $\dot{\zeta}_h^k := P_h(V_{h+1}^k - V_{h+1}^{\pi^k})(s_h^k, a_h^k) - (V_{h+1}^k - V_{h+1}^{\pi^k})(s_{h+1}^k)$.

Let $\delta' = \delta/(8KH)$. By Lemma 5, the good event $\mathcal{G}(K, \delta')$ holds with probability at least $1 - \delta/4$. Then under the event $\mathcal{G}(K, \delta')$, for any $h \in [H]$ we have

$$
\begin{aligned}
&- \iota_h^k(s_h^k, a_h^k) \\
&= Q_h^k(s_h^k, a_h^k) - \left(r(s_h^k, a_h^k) + P_h V_{h+1}^k(s_h^k, a_h^k) \right) \\
&= \min \left\{ r(s_h^k, a_h^k) + \sum_{s' \in \mathcal{S}_{k,h}} P_{\boldsymbol{\theta}_h^k}(s' \mid s_h^k, a_h^k) V_{h+1}^k(s') + \max_{m \in [M]} \hat{\boldsymbol{\varphi}}_{k,h}(s_h^k, a_h^k)^\top \xi_{k,h}^{(m)}, H \right\} \\
&\quad - \left(r(s_h^k, a_h^k) + P_h V_{h+1}^k(s_h^k, a_h^k) \right) \\
&\leq \sum_{s' \in \mathcal{S}_{k,h}} P_{\boldsymbol{\theta}_h^k}(s' \mid s_h^k, a_h^k) V_{h+1}^k(s') + \max_{m \in [M]} \hat{\boldsymbol{\varphi}}_{k,h}(s_h^k, a_h^k)^\top \xi_{k,h}^{(m)} - P_h V_{h+1}^k(s_h^k, a_h^k) \\
&\leq \left| \sum_{s' \in \mathcal{S}_{k,h}} P_{\boldsymbol{\theta}_h^k}(s' \mid s_h^k, a_h^k) V_{h+1}^k(s') - P_h V_{h+1}^k(s_h^k, a_h^k) \right| + \max_{m \in [M]} \left| \hat{\boldsymbol{\varphi}}_{k,h}(s_h^k, a_h^k)^\top \xi_{k,h}^{(m)} \right| \\
&\leq |\Delta_h^k(s_h^k, a_h^k)| + \max_{m \in [M]} \|\hat{\boldsymbol{\varphi}}_{k,h}(s_h^k, a_h^k)\|_{\mathbf{A}_{k,h}^{-1}} \|\xi_{k,h}^{(m)}\|_{\mathbf{A}_{k,h}} \tag{30} \\
&\leq (H\alpha_k(\delta') + \gamma_k(\delta')) \|\hat{\boldsymbol{\varphi}}_{k,h}(s_h^k, a_h^k)\|_{\mathbf{A}_{k,h}^{-1}}, \tag{31}
\end{aligned}
$$

where (30) comes from the Cauchy-Schwarz inequality and (31) holds due the the Lemma 4 & 30. Then, with probability at least $1 - \delta/4$, we have

$$
\sum_{h=1}^{H} -\iota_h^k(s_h^k, a_h^k) \leq \sum_{h=1}^{H} (H\alpha_k(\delta') + \gamma_k(\delta')) \|\hat{\boldsymbol{\varphi}}_{k,h}(s_h^k, a_h^k)\|_{\mathbf{A}_{k,h}^{-1}}. \tag{32}
$$

On the other hand, for $\dot{\zeta}_h^k$, we have $|\dot{\zeta}_h^k| \leq 2H$ and $\mathbb{E}[\dot{\zeta}_h^k \mid \mathcal{F}_{k,h}] = 0$, which means $\{\dot{\zeta}_h^k \mid \mathcal{F}_{k,h}\}_{k,h}$ is a martingale difference sequence for any $k \in [K]$ and $h \in [H]$. Hence, by applying the Azuma-Hoeffding inequality with probability at least $1 - \delta/4$, we have

$$
\sum_{k=1}^{K} \sum_{h=1}^{H} \dot{\zeta}_h^k \leq 2H \sqrt{2KH \log(4/\delta)}. \tag{33}
$$

37

Combining the results of (32) and (33), with probability at least $1 - \delta/2$, we have

$$(V_1^k - V_1^{\pi^k})(s_1^k)$$

$$\leq 2H\sqrt{2T\log(4/\delta)} + \sum_{k=1}^{K}\sum_{h=1}^{H}(H\alpha_k(\delta') + \gamma_k(\delta'))\,\|\hat{\varphi}_{k,h}(s_h^k, a_h^k)\|_{\mathbf{A}_{k,h}^{-1}}$$

$$\leq 2H\sqrt{2T\log(4/\delta)} + (H\alpha_K(\delta') + \gamma_K(\delta'))\sum_{k=1}^{K}\sum_{h=1}^{H}\|\hat{\varphi}_{k,h}(s_h^k, a_h^k)\|_{\mathbf{A}_{k,h}^{-1}} \qquad (34)$$

$$\leq 2H\sqrt{2T\log(4/\delta)} + (H\alpha_K(\delta') + \gamma_K(\delta'))\sum_{h=1}^{H}\sqrt{K\sum_{k=1}^{K}\|\hat{\varphi}_{k,h}(s_h^k, a_h^k)\|_{\mathbf{A}_{k,h}^{-1}}^2} \qquad (35)$$

$$\leq 2H\sqrt{2T\log(4/\delta)} + (H\alpha_K(\delta') + \gamma_K(\delta'))\sum_{h=1}^{H}\sqrt{4\kappa^{-1}Kd\log\left(1 + \frac{KUL_{\varphi}^2}{d\lambda}\right)} \qquad (36)$$

$$= 2H\sqrt{2T\log(4/\delta)} + (H\alpha_K(\delta') + \gamma_K(\delta'))\sqrt{4\kappa^{-1}THd\log\left(1 + \frac{KUL_{\varphi}^2}{d\lambda}\right)},$$

$$= \widetilde{\mathcal{O}}\left(\kappa^{-1}d^{\frac{3}{2}}H^{\frac{3}{2}}\sqrt{T} + H\sqrt{T}\right),$$

where (34) follows from the fact that both $\alpha_k(\delta)$ and $\gamma_k(\delta)$ are increasing in k, (35) comes from Cauchy-Schwarz inequality and (36) holds by the generalized elliptical potential lemma (Lemma 3). $\qquad\square$

C.6 Bound on Pessimism Part

Lemma 11 (Bound on pessimism). *For any δ with $0 < \delta < \Phi(-1)/2$, let $\sigma_k = H\alpha_k(\delta)$. If $\lambda \geq L_{\varphi}^2$ and we take multiple sample size $M = \lceil 1 - \frac{\log H}{\log \Phi(1)} \rceil$, then with probability at least $1 - \delta/2$, we have*

$$\sum_{k=1}^{K}(V_1^* - V_1^k)(s_1^k) = \widetilde{\mathcal{O}}\left(\kappa^{-1}d^{\frac{3}{2}}H^{\frac{3}{2}}\sqrt{T}\right).$$

Proof of lemma 11. Similar to the techniques used in [73], we show that the difference between the optimal value function V_1^* and the estimated value function V_1^k can be controlled by constructing an upper bound on V_1^* and a lower bound on V_1^k. In this proof, we consider three kinds of pseudo-noises, $\xi, \bar{\xi}$ and $\underline{\xi}$ that we define later in the proof. Also, for $\delta' = \delta/10$, we denote $\mathcal{G}(K, \delta'), \bar{\mathcal{G}}(K, \delta')$ and $\underline{\mathcal{G}}(K, \delta')$ as the good events induced by $\xi, \bar{\xi}$ and $\underline{\xi}$ respectively. From now on, we denote $G(K, \delta')$ by the event $\mathcal{G}(K, \delta') \cap \bar{\mathcal{G}}(K, \delta') \cap \underline{\mathcal{G}}(K, \delta')$. Then, by Lemma 5, the event $G(K, \delta')$ holds with high probability at least $1 - 3\delta/10$.

First, we construct the lower bound of V_1^k. For any given $k \in [K]$, let $\widetilde{\xi} := \{\widetilde{\xi}_{k,h}^{(m)}\}_{m \in [M]} \subset \mathbb{R}^d$ be a set of vectors for $h \in [H]$ and $V_h^k(\cdot; \widetilde{\xi})$ be the value function obtained by the Algorithm 1 with non-random $\widetilde{\xi}_{k,h}^{(m)}$ in place of $\xi_{k,h}^{(m)}$. Then consider the following minimization problem:

$$\min_{\{\widetilde{\xi}_{k,h}^{(m)}\}_{h \in [H], m \in [M]}} \quad V_1^k(s_1^k; \widetilde{\xi})$$

$$\text{s.t.} \quad \max_{m \in [M]} \|\widetilde{\xi}_{k,h}^{(m)}\|_{\mathbf{A}_{k,h}} \leq \gamma_k(\delta), \quad \forall h \in [H]$$

And we denote $\underline{\xi} := \{\underline{\xi}_{k,h}^{(m)}\}_{h \in [H], m \in [M]}$ by a minimizer and $\underline{V}_1^k(s_1^k)$ by the minimum of the above minimization problem, i.e., $\underline{V}_h^k(\cdot) := V_h^k(\cdot; \underline{\xi})$. Then, under the event $\mathcal{G}(K, \delta')$, since $\{\xi_{k,h}^{(m)}\}_{h \in [H], m \in [M]}$ is also a feasible solution of the above optimization problem, and since $V_h^k = V_h^k(\cdot; \xi)$, thus we have

$$\underline{V}_1^k(s_1^k) \leq V_1^k(s_1^k). \qquad (37)$$

38

Second, to find an upper bound for V^*, considering i.i.d copies $\{\bar{\boldsymbol{\xi}}_{k,h}^{(m)}\}_{h\in[H],m\in[M]}$ of $\{\boldsymbol{\xi}_{k,h}^{(m)}\}_{h\in[H],m\in[M]}$ and run Algorithm 1 to get a corresponding value function $\bar{V}_h^k$ and $\bar{Q}_h^k$ for all $h \in [H]$. Define the event that $\bar{V}_1^k(s_1^k)$ is optimistic in the k-th episode as

$$\bar{\mathcal{X}}_k = \{(\bar{V}_1^k - V_1^*)(s_1^k) \geq 0\}\,.$$

Then by Lemma 6, for given δ, we have

$$\mathbb{P}(\bar{\mathcal{X}}_k \mid s_1^k, \mathcal{F}_k) \geq \Phi(-1)/2\,.$$

Then by the definition of optimism, under the event $\mathcal{G}(K, \delta')$, we have

$$(V_1^* - V_1^k)(s_1^k) \leq \mathbb{E}_{\bar{\boldsymbol{\xi}}|\bar{\mathcal{X}}_k}\left[(\bar{V}_1^k - V_1^k)(s_1^k)\right]$$
$$\leq \mathbb{E}_{\bar{\boldsymbol{\xi}}|\bar{\mathcal{X}}_k}\left[(\bar{V}_1^k - \underline{V}_1^k)(s_1^k)\right]\,, \tag{38}$$

where the expectations are over the $\bar{\boldsymbol{\xi}}$'s conditioned on the event $\bar{\mathcal{X}}_k$ and the second inequality comes from (37). On the other hand, under the event $\bar{\mathcal{G}}(K, \delta')$ by the law of the total expectation, we have

$$\mathbb{E}_{\bar{\boldsymbol{\xi}}}\left[(\bar{V}_1^k - \underline{V}_1^k)(s_1^k)\right] = \mathbb{E}_{\bar{\boldsymbol{\xi}}|\bar{\mathcal{X}}_k}\left[(\bar{V}_1^k - \underline{V}_1^k)(s_1^k)\right]\mathbb{P}(\bar{\mathcal{X}}_k) + \mathbb{E}_{\bar{\boldsymbol{\xi}}|\bar{\mathcal{X}}_k^c}\left[(\bar{V}_1^k - \underline{V}_1^k)(s_1^k)\right]\mathbb{P}(\bar{\mathcal{X}}_k^c)$$
$$\geq \mathbb{E}_{\bar{\boldsymbol{\xi}}|\bar{\mathcal{X}}_k}\left[(\bar{V}_1^k - \underline{V}_1^k)(s_1^k)\right]\mathbb{P}(\bar{\mathcal{X}}_k)\,, \tag{39}$$

where (39) comes from the fact that $\{\bar{\boldsymbol{\xi}}_{k,h}^{(m)}\}_{h\in[H],m\in[M]}$ is also a feasible solution of the above optimization problem under the event $\bar{\mathcal{G}}(K, \delta')$, i.e., $\bar{V}_1^k(s_1^k) \geq \underline{V}_1^k(s_1^k)$. Then, by combining the results of (39) and (38), under the event $G(K, \delta')$, we have

$$(V_1^* - V_1^k)(s_1^k) \leq \mathbb{E}_{\bar{\boldsymbol{\xi}}|\bar{\mathcal{X}}_k}\left[(\bar{V}_1^k - \underline{V}_1^k)(s_1^k)\right]$$
$$\leq \mathbb{E}_{\bar{\boldsymbol{\xi}}}\left[(\bar{V}_1^k - \underline{V}_1^k)(s_1^k)\right]/\mathbb{P}(\bar{\mathcal{X}}_k)$$
$$\leq \frac{2}{\Phi(-1)}\mathbb{E}_{\bar{\boldsymbol{\xi}}}\left[(\bar{V}_1^k - V_1^k + V_1^k - \underline{V}_1^k)(s_1^k)\right]$$
$$= \frac{2}{\Phi(-1)}\left((V_1^k - \underline{V}_1^k)(s_1^k)\right) + \ddot{\zeta}_k\,, \tag{40}$$

where we denote

$$\ddot{\zeta}_k := \frac{2}{\Phi(-1)}\left(\mathbb{E}_{\bar{\boldsymbol{\xi}}}\left[\bar{V}_1^k(s_1^k)\right] - V_1^k(s_1^k)\right)\,.$$

Note that since $\bar{\boldsymbol{\xi}}$ is the i.i.d copy of $\boldsymbol{\xi}$, therefore $\bar{V}_{k,1}$ and $V_{k,1}$ are independent, which means $\{\ddot{\zeta}_k \mid \mathcal{F}_{k-1}\}_{k=1}^K$ is a martingale difference sequence with $|\ddot{\zeta}_k| \leq \frac{2H}{\Phi(-1)}$. Therefore by applying Azuma-Hoeffding inequality under the event $G(K, \delta')$, with probability at least $1 - \delta'$, we have

$$\sum_{k=1}^K \ddot{\zeta}_k \leq \frac{2H}{\Phi(-1)}\sqrt{2K\log(1/\delta')}\,. \tag{41}$$

On the other hand, by dividing the first term in (40) into two terms we have

$$(V_1^k - \underline{V}_1^k)(s_1^k) = \underbrace{(V_1^k - V_1^{\pi^k})(s_1^k)}_{I_1} + \underbrace{(V_1^{\pi^k} - \underline{V}_1^k)(s_1^k)}_{I_2}\,.$$

For I_1, note that since it is related to the estimation error, under the event $G(K, \delta')$ we can bound the sum of I_1 for the total episode number using Lemma 10 as follows:

$$\sum_{k=1}^K (V_1^k - V_1^{\pi^k})(s_1^k) \leq (H\alpha_K(\delta') + \gamma_K(\delta'))\sqrt{4\kappa^{-1}THd\log\left(1 + \frac{KU L_\varphi^2}{d\lambda}\right)}$$
$$+ 2H\sqrt{2T\log(1/\delta')}\,. \tag{42}$$

39

For I_2, since we have

$$
\begin{aligned}
I_2 &= Q_1^{\pi^k}(s_1^k, a_1^k) - \underline{V}_1^k(s_1^k) \\
&\le Q_1^{\pi^k}(s_1^k, a_1^k) - \underline{Q}_1^k(s_1^k, a_1^k) \tag{43} \\
&= Q_1^{\pi^k}(s_1^k, a_1^k) - \underline{Q}_1^k(s_1^k, a_1^k) - \underline{\iota}_1^k(s_1^k, a_1^k) + \underline{\iota}_1^k(s_1^k, a_1^k) \\
&= P_1(V_2^{\pi^k} - \underline{V}_2^k)(s_1^k, a_1^k) + \underline{\iota}_1^k(s_1^k, a_1^k) \tag{44} \\
&= \underbrace{P_1(V_2^{\pi^k} - \underline{V}_2^k)(s_1^k, a_1^k) - (V_2^{\pi^k} - \underline{V}_2^k)(s_2^k)}_{\dddot{\zeta}_1^k} + (V_2^{\pi^k} - \underline{V}_2^k)(s_2^k) + \underline{\iota}_1^k(s_1^k, a_1^k)
\end{aligned}
$$

where (43) comes from $a_1^k = \mathrm{argmax}_a\, Q_1^k(s_1^k, a)$ and (44) holds by the following definition of $\underline{\iota}_h^k(s_h^k, a_h^k)$:

$$
\begin{aligned}
\underline{\iota}_h^k(s_h^k, a_h^k) &:= r(s_h^k, a_h^k) + P_h \underline{V}_{h+1}^k(s_h^k, a_h^k) - \underline{Q}_h^k(s_h^k, a_h^k) \\
&= r(s_h^k, a_h^k) + P_h \underline{V}_{h+1}^k(s_h^k, a_h^k) - \underline{Q}_h^k(s_h^k, a_h^k) + Q_h^{\pi^k}(s_h^k, a_h^k) - Q_h^{\pi^k}(s_h^k, a_h^k) \\
&= P_h(\underline{V}_{h+1}^k - V_{h+1}^{\pi^k})(s_h^k, a_h^k) + (Q_h^{\pi^k} - \underline{Q}_h^k)(s_h^k, a_h^k)\,.
\end{aligned}
$$

Then by applying the same argument recursively for the whole horizon, we have

$$
I_2 \le \sum_{h=1}^{H} \underline{\iota}_h^k(s_h^k, a_h^k) + \sum_{h=1}^{H} \dddot{\zeta}_h^k\,,
$$

where we denote

$$
\dddot{\zeta}_h^k := P_h(V_{h+1}^{\pi^k} - \underline{V}_{h+1}^k)(s_h^k, a_h^k) - (V_{h+1}^{\pi^k} - \underline{V}_{h+1}^k)(s_{h+1}^k)\,.
$$

Note that $\left\{ \dddot{\zeta}_h^k \mid \mathcal{F}_{k,h} \right\}_{k,h}$ is a martingale difference sequence with $|\dddot{\zeta}_h^k| \le 2H$. Then, under the event $G(K, \delta')$ by applying the Azuma-Hoeffding inequality with probability at least $1 - \delta'$, we have

$$
\sum_{k=1}^{K} \sum_{h=1}^{H} \dddot{\zeta}_h^k \le 2H \sqrt{2T \log(1/\delta')}\,. \tag{45}
$$

To bound $\sum_{h=1}^{H} \underline{\iota}_h^k(s_h^k, a_h^k)$, we divide the whole horizon index set into two groups as follows:

$$
\begin{aligned}
&H^+ \\
&= \left\{ j \in [H] : r(s_j^k, a_j^k) + \sum_{s' \in \mathcal{S}_{k,j}} P_{\boldsymbol{\theta}_h^k}(s' \mid s_j^k, a_j^k) \underline{V}_{j+1}^k(s') + \max_{m \in [M]} \hat{\boldsymbol{\varphi}}_{k,j}(s_j^k, a_j^k)^\top \underline{\boldsymbol{\xi}}_{k,j}^{(m)} > H \right\} \\
&H^- = [H] \backslash H^+\,.
\end{aligned}
$$

Then, for $j \in H^+$ since $\underline{Q}_j^k(s_j^k, a_j^k) = H - j + 1$, $\underline{V}_{j+1}^k \le H - j$ and $r(s_j^k, a_j^k) \le 1$, we have

$$
\underline{\iota}_j^k(s_j^k, a_j^k) = r(s_j^k, a_j^k) + P_j \underline{V}_{j+1}^k(s_j^k, a_j^k) - (H - j + 1) \le 0\,. \tag{46}
$$

On the other hand, for $j \in H^-$, under the event $G(K, \delta')$ we have

$$
\begin{aligned}
\underline{\iota}_j^k(s_j^k, a_j^k) &= P_j \underline{V}_{j+1}^k(s_j^k, a_j^k) - \sum_{s' \in \mathcal{S}_{k,j}} P_{\boldsymbol{\theta}_h^k}(s' \mid s_j^k, a_j^k) \underline{V}_{j+1}^k(s') - \max_{m \in [M]} \hat{\boldsymbol{\varphi}}_{k,j}(s_j^k, a_j^k)^\top \underline{\boldsymbol{\xi}}_{k,j}^{(m)} \\
&\le \left| P_j \underline{V}_{j+1}^k(s_j^k, a_j^k) - \sum_{s' \in \mathcal{S}_{k,j}} P_{\boldsymbol{\theta}_h^k}(s' \mid s_j^k, a_j^k) \underline{V}_{j+1}^k(s') \right| + \left| \max_{m \in [M]} \hat{\boldsymbol{\varphi}}_{k,j}(s_j^k, a_j^k)^\top \underline{\boldsymbol{\xi}}_{k,j}^{(m)} \right| \\
&\le H \alpha_k(\delta') \|\hat{\boldsymbol{\varphi}}_{k,j}(s_j^k, a_j^k)\|_{\mathbf{A}_{k,j}^{-1}} + \max_{m \in [M]} \left| \hat{\boldsymbol{\varphi}}_{k,j}(s_j^k, a_j^k)^\top \underline{\boldsymbol{\xi}}_{k,j}^{(m)} \right| \tag{47} \\
&\le H \alpha_k(\delta') \|\hat{\boldsymbol{\varphi}}_{k,j}(s_j^k, a_j^k)\|_{\mathbf{A}_{k,j}^{-1}} + \max_{m \in [M]} \|\hat{\boldsymbol{\varphi}}_{k,j}(s_j^k, a_j^k)\|_{\mathbf{A}_{k,j}^{-1}} \|\underline{\boldsymbol{\xi}}_{k,j}^{(m)}\|_{\mathbf{A}_{k,j}} \\
&\le (H \alpha_k(\delta') + \gamma_k(\delta')) \|\hat{\boldsymbol{\varphi}}_{k,j}(s_j^k, a_j^k)\|_{\mathbf{A}_{k,j}^{-1}}\,, \tag{48}
\end{aligned}
$$

40

where (47) holds by Lemma 4.

By combining the result of (46) and (48), we have

$$I_2 \le \sum_{j \in H^-} (H\alpha_k(\delta') + \gamma_k(\delta')) \, \|\hat{\boldsymbol{\varphi}}_{k,j}(s_j^k, a_j^k)\|_{\mathbf{A}_{k,j}^{-1}} + \sum_{h=1}^{H} \dddot{\zeta}_h^k$$

$$\le \sum_{h=1}^{H} (H\alpha_k(\delta') + \gamma_k(\delta')) \, \|\hat{\boldsymbol{\varphi}}_{k,h}(s_h^k, a_h^k)\|_{\mathbf{A}_{k,h}^{-1}} + \sum_{h=1}^{H} \dddot{\zeta}_h^k .$$

Then summing I_2 over the total number of episodes, under the event $G(K, \delta')$, we have

$$\sum_{k=1}^{K} (V_1^{\pi^k} - \underline{V}_1^k)(s_1^k) \le \sum_{k=1}^{K} \sum_{h=1}^{H} (H\alpha_k(\delta') + \gamma_k(\delta')) \, \|\hat{\boldsymbol{\varphi}}_{k,h}(s_h^k, a_h^k)\|_{\mathbf{A}_{k,h}^{-1}} + \sum_{k=1}^{K} \sum_{h=1}^{H} \dddot{\zeta}_h^k$$

$$\le (H\alpha_K(\delta') + \gamma_K(\delta')) \sum_{k=1}^{K} \sum_{h=1}^{H} \|\hat{\boldsymbol{\varphi}}_{k,h}(s_h^k, a_h^k)\|_{\mathbf{A}_{k,h}^{-1}} + \sum_{k=1}^{K} \sum_{h=1}^{H} \dddot{\zeta}_h^k$$

$$\le (H\alpha_K(\delta') + \gamma_K(\delta')) \sqrt{4\kappa^{-1} T H d \log\left(1 + \frac{K \mathcal{U} L_\varphi^2}{d\lambda}\right)} \qquad (49)$$

$$+ 2H \sqrt{2T \log(1/\delta')} , \qquad (50)$$

where the last inequality holds due to the Lemma 3 and (45).

Finally, by summing (40) over k and plugging the results of (42), (50) and (41) then, we have

$$\sum_{k=1}^{K} (V_1^* - V_1^k)(s_1^k)$$

$$\le \frac{4}{\Phi(-1)} \left[(H\alpha_K(\delta') + \gamma_K(\delta')) \sqrt{4\kappa^{-1} T H d \log\left(1 + \frac{K \mathcal{U} L_\varphi^2}{d\lambda}\right)} + 2H \sqrt{2T \log(1/\delta')} \right]$$

$$+ \frac{2H}{\Phi(-1)} \sqrt{2K \log(1/\delta')}$$

$$\le \tilde{\mathcal{O}} \left(\kappa^{-1} d^{3/2} H^{3/2} \sqrt{T} + H\sqrt{T} + H\sqrt{K} \right) .$$

To conclude the proof, by setting $\delta' = \delta/10$ and we take a union bound over the two applications of Azuma-Hoeffding $(\ddot{\zeta}_k, \dddot{\zeta}_h^k)$ and the event $G(K, \delta')$, we get the desired result with probability at least $1 - \delta/2$. $\qquad \square$

C.7 Regret Bound of `RRL-MNL`

Proof of Theorem 1. We can decompose the regret with estimation part and pessimism part as follows:

$$\mathbf{Regret}_\pi(K) = \sum_{k=1}^{K} (V_1^* - V_1^{\pi^k})(s_1^k)$$

$$= \sum_{k=1}^{K} (V_1^* - V_1^k)(s_1^k) + \sum_{k=1}^{K} (V_1^k - V_1^{\pi^k})(s_1^k) .$$

Since both Lemma 10 and Lemma 11 holds with probability at least $1 - \delta/2$ respectively, by taking the union bound the following holds with probability at least $1 - \delta$:

$$\mathbf{Regret}_\pi(K) = \tilde{\mathcal{O}} \left(\kappa^{-1} d^{\frac{3}{2}} H^{\frac{3}{2}} \sqrt{T} + H\sqrt{T} + H\sqrt{K} \right) + \tilde{\mathcal{O}} \left(\kappa^{-1} d^{\frac{3}{2}} H^{\frac{3}{2}} \sqrt{T} + H\sqrt{T} \right)$$

$$= \tilde{\mathcal{O}} \left(\kappa^{-1} d^{\frac{3}{2}} H^{\frac{3}{2}} \sqrt{T} \right) .$$

$\qquad \square$

41

D Detailed Regret Analysis for `ORRL-MNL` (Theorem 2)

D.1 Concentration of Estimated Transition Core $\widetilde{\boldsymbol{\theta}}_h^k$

In this section, we provide the detailed proof of Lemma 12, which demonstrates the concentration result for $\widetilde{\boldsymbol{\theta}}_h^k$ independently of κ and $\mathcal{U}$. Note that we adapt the proof provided by Zhang and Sugiyama [76] in the MNL contextual bandit setting to MNL-MDPs and improve the result, making it independent of $\mathcal{U}$. We provide the lemmas for the concentration of the online transition core for completeness, noting that there are slight differences compared to their work, which stem from the different problem setting.

Lemma 12 (Concentration of online estimated transition core). *Let* $\eta = \mathcal{O}(\log \mathcal{U})$ *and* $\lambda = \mathcal{O}(d \log \mathcal{U})$. *Then, for any* $\delta \in (0, 1]$ *and for any* $h \in [H]$, *we have*

$$\mathbb{P}\left(\forall k \geq 1, \left\|\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*\right\|_{\mathbf{B}_{k,h}} \leq \beta_k(\delta)\right) \geq 1 - \delta,$$

where $\beta_k(\delta) = \mathcal{O}(\sqrt{d} \log \mathcal{U} \log(kH))$.

Proof of Lemma 12. Recall that the transition core updated by the online mirror descent is represented by

$$\widetilde{\boldsymbol{\theta}}_h^{k+1} = \underset{\boldsymbol{\theta} \in \mathcal{B}(L_{\boldsymbol{\theta}})}{\operatorname{argmin}} \, \widetilde{\ell}_{k,h}(\boldsymbol{\theta}) + \frac{1}{2\eta} \left\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^k\right\|_{\mathbf{B}_{k,h}}^2,$$

where $\widetilde{\ell}_{k,h}(\boldsymbol{\theta}) = \ell_{k,h}(\widetilde{\boldsymbol{\theta}}_h^k) + (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^k)^\top \nabla \ell_{k,h}(\widetilde{\boldsymbol{\theta}}_h^k) + \frac{1}{2}\left\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^k\right\|_{\nabla^2 \ell_{k,h}(\widetilde{\boldsymbol{\theta}}_h^k)}$. We introduce the following lemma providing that the estimation error of the online estimator $\widetilde{\boldsymbol{\theta}}_h^k$ can be bounded by the regret.

Lemma 13 (Lemma 12 in [76]). *Let* $\alpha = \log \mathcal{U} + 2(1 + L_{\boldsymbol{\theta}} L_{\boldsymbol{\varphi}})$ *and* $\lambda > 0$. *If we set the step size* $\eta = \alpha/2$, *then we have*

$$
\begin{aligned}
\left\|\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*\right\|_{\mathbf{B}_{k,h}}^2 &\leq \alpha \sum_{i=1}^k \left(\ell_{i,h}(\boldsymbol{\theta}_h^*) - \ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1})\right) + \lambda L_{\boldsymbol{\theta}}^2 \\
&\quad + 3\sqrt{2} L_{\boldsymbol{\varphi}}^3 \alpha \sum_{i=1}^k \left\|\widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^i\right\|_2^2 - \sum_{i=1}^k \left\|\widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^i\right\|_{\mathbf{B}_{i,h}}^2.
\end{aligned}
\tag{51}
$$

Now, we bound the first term of (51). To simplify the presentation, for all $(k, h) \in [K] \times [H]$, we define the softmax function $\boldsymbol{\sigma}_{k,h} : \mathbb{R}^{|\mathcal{S}_{k,h}|} \to [0, 1]^{|\mathcal{S}_{k,h}|}$ as follows:

$$[\boldsymbol{\sigma}_{k,h}(\mathbf{z})]_{s'} = \frac{\exp([\mathbf{z}]_{s'})}{\sum_{s'' \in \mathcal{S}_{k,h}} \exp([\mathbf{z}]_{s''})},$$

where $[\cdot]_{s'}$ denote the element corresponding to $s' \in \mathcal{S}$ of the input vector. We also define the pseudo-inverse of the softmax function $\boldsymbol{\sigma}_{k,h}$ via $[\boldsymbol{\sigma}_{k,h}^+(\mathbf{p})]_{s'} = \log([\mathbf{p}]_{s'})$ which has the property that for all $\mathbf{p} \in \Delta_{|\mathcal{S}_{k,h}|}$, we have $\boldsymbol{\sigma}_{k,h}(\boldsymbol{\sigma}_{k,h}^+(\mathbf{p})) = \mathbf{p}$ and $\sum_{s \in \mathcal{S}_{k,h}} \exp\left([\boldsymbol{\sigma}_{k,h}^+(\mathbf{p})]_s\right) = 1$.

We denote $\boldsymbol{\Phi}_{k,h} = [\boldsymbol{\varphi}_{k,h,s'}]_{s' \in \mathcal{S}_{k,h}} \in \mathbb{R}^{d \times |\mathcal{S}_{k,h}|}$ for simplicity. Then, the transition model can also be written as $P_{\boldsymbol{\theta}}(s' \mid s_h^k, a_h^k) = [\boldsymbol{\sigma}_{k,h}(\boldsymbol{\Phi}_{k,h}^\top \boldsymbol{\theta}_h^*)]_{s'}$. We further define $\widetilde{\mathbf{z}}_{i,h} = \boldsymbol{\sigma}_{i,h}^+ \left(\mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{N}(\widetilde{\boldsymbol{\theta}}_h^i, c\mathbf{B}_{i,h}^{-1})}[\boldsymbol{\sigma}_{i,h}(\boldsymbol{\Phi}_{i,h}^\top \boldsymbol{\theta})]\right)$ for our analysis. Then, we have

$$\sum_{i=1}^k \left(\ell_{i,h}(\boldsymbol{\theta}_h^*) - \ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1})\right) = \sum_{i=1}^k \left(\ell_{i,h}(\boldsymbol{\theta}_h^*) - \ell(\widetilde{\mathbf{z}}_{i,h}, y_h^i)\right) + \sum_{i=1}^k \left(\ell(\widetilde{\mathbf{z}}_{i,h}, y_h^i) - \ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1})\right).$$

$$\tag{52}$$

We can bound the first term of (52) by the following lemma.

42

Lemma 14. *Let $\delta \in (0, 1]$. Then, for all $(k, h) \in [K] \times [H]$, with probability at least $1 - \delta$, we have*

$$\sum_{i=1}^{k} \left(\ell_{i,h}(\boldsymbol{\theta}_h^*) - \ell(\widetilde{\mathbf{z}}_{i,h}, y_h^i) \right) \leq \Gamma_k^A(\delta),$$

where $\Gamma_k^A(\delta) = \frac{5}{4}(3\log(\mathcal{U}k) + L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}})\lambda + 4(3\log(\mathcal{U}k) + L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}) \log \left(\frac{H\sqrt{1+2k}}{\delta} \right) + 2.$

Furthermore, we can bound the second term of (52) by the following lemma.

Lemma 15. *Let $\lambda \geq 72L_{\boldsymbol{\varphi}}^2 cd$. Then, for any $c > 0$ and all $(k, h) \in [K] \times [H]$, we have*

$$\sum_{i=1}^{k} \left(\ell(\widetilde{\mathbf{z}}_{i,h}, y_h^i) - \ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}) \right) \leq \frac{1}{2c} \sum_{i=1}^{k} \left\| \widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^{i} \right\|_{\mathbf{B}_{i,h}}^2 + \Gamma_k^B(\delta).$$

where $\Gamma_k^B(\delta) = \sqrt{6}cd \log \left(1 + \frac{2kL_{\boldsymbol{\varphi}}^2}{d\lambda} \right).$

Combining Lemma 13, Lemma 14, and Lemma 15, and by setting $\eta = \alpha/2, c = 2\alpha/3$ and $\lambda \geq \max\{12\sqrt{2}L_{\boldsymbol{\varphi}}^3\alpha, 48L_{\boldsymbol{\varphi}}^2 d\alpha\}$, we derive that

$$\left\| \widetilde{\boldsymbol{\theta}}_h^{k+1} - \boldsymbol{\theta}_h^* \right\|_{\mathbf{B}_{k,h}}^2$$

$$\leq \alpha\Gamma_k^A(\delta) + \alpha\Gamma_k^B(\delta) + \lambda L_{\boldsymbol{\theta}}^2 + 3\sqrt{2}L_{\boldsymbol{\varphi}}^3\alpha \sum_{i=1}^{k} \left\| \widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^{i} \right\|_2^2 + \left(\frac{\alpha}{2c} - 1 \right) \sum_{i=1}^{k} \left\| \widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^{i} \right\|_{\mathbf{B}_{i,h}}^2$$

$$\leq \alpha\Gamma_k^A(\delta) + \alpha\Gamma_k^B(\delta) + \lambda L_{\boldsymbol{\theta}}^2$$

$$\leq C \log \mathcal{U} \left(\lambda \log(\mathcal{U}k) + \log(\mathcal{U}k) \log \left(\frac{H\sqrt{1+2k}}{\delta} \right) + d \log \left(1 + \frac{k}{d\lambda} \right) \right) + \lambda L_{\boldsymbol{\theta}}^2$$

$$=: \beta_k(\delta)^2 \tag{53}$$

where $C > 0$ is an absolute constant. In the above, we choose $\lambda = \mathcal{O}(d \log \mathcal{U})$, $\alpha = \mathcal{O}(\log \mathcal{U})$. The second inequality of (53) is derived from the fact that

$$3\sqrt{2}L_{\boldsymbol{\varphi}}^3\alpha \sum_{i=1}^{k} \left\| \widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^{i} \right\|_2^2 + \left(\frac{\alpha}{2c} - 1 \right) \sum_{i=1}^{k} \left\| \widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^{i} \right\|_{\mathbf{B}_{i,h}}^2$$

$$= 3\sqrt{2}L_{\boldsymbol{\varphi}}^3\alpha \sum_{i=1}^{k} \left\| \widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^{i} \right\|_2^2 - \frac{1}{4} \sum_{i=1}^{k} \left\| \widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^{i} \right\|_{\mathbf{B}_{i,h}}^2$$

$$\leq 3\sqrt{2}L_{\boldsymbol{\varphi}}^3\alpha \sum_{i=1}^{k} \left\| \widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^{i} \right\|_2^2 - \frac{\lambda}{4} \sum_{i=1}^{k} \left\| \widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^{i} \right\|_2^2$$

$$\leq 0.$$

The first inequality holds from $\mathbf{B}_{i,h} \succeq \lambda \mathbf{I}_d$, and the second inequality is obvious from our setting of λ. Therefore, we can conclude that

$$\left\| \widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^* \right\|_{\mathbf{B}_{k,h}} \leq \beta_k(\delta) = \mathcal{O}(\sqrt{d} \log \mathcal{U} \log(kH)).$$

$\square$

In the following section, we provide the proofs of the lemmas used in Lemma 12.

43

D.1.1 Proof of Lemma 13

Proof of Lemma 13. Let $\widetilde{\ell}_{i,h}(\boldsymbol{\theta}) = \ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^i) + \nabla\ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^i)^\top\left(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^i\right) + \frac{1}{2}\left\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^i\right\|^2_{\nabla^2\ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^i)}$ be a second-order Taylor expansion of $\ell_{i,h}(\boldsymbol{\theta})$ at $\widetilde{\boldsymbol{\theta}}_h^i$. Since we have

$$\widetilde{\boldsymbol{\theta}}_h^{k+1} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathcal{B}_d(L_{\boldsymbol{\theta}})} \frac{1}{2\eta}\left\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^k\right\|^2_{\widetilde{\mathbf{B}}_{k,h}} + \nabla\ell_{k,h}(\widetilde{\boldsymbol{\theta}}_h^k)^\top\boldsymbol{\theta} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathcal{B}(\mathbf{0}_d,L_{\boldsymbol{\theta}})} \widetilde{\ell}_{k,h}(\boldsymbol{\theta}) + \frac{1}{2\eta}\left\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^k\right\|^2_{\mathbf{B}_{k,h}},$$

by Lemma 31, if we define $\psi(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|^2_{\mathbf{B}_{i,h}}$ we obtain

$$\nabla\widetilde{\ell}_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1})^\top(\widetilde{\boldsymbol{\theta}}_h^{i+1} - \boldsymbol{\theta}_h^*) \le \frac{1}{2\eta}\left(\left\|\widetilde{\boldsymbol{\theta}}_h^i - \boldsymbol{\theta}_h^*\right\|^2_{\mathbf{B}_{i,h}} - \left\|\widetilde{\boldsymbol{\theta}}_h^{i+1} - \boldsymbol{\theta}_h^*\right\|^2_{\mathbf{B}_{i,h}} - \left\|\widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^i\right\|_{\mathbf{B}_{i,h}}\right). \tag{54}$$

By applying Lemma 33, we have

$$\ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}) - \ell_{i,h}(\boldsymbol{\theta}_h^*) \le \left\langle\nabla\ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}), \widetilde{\boldsymbol{\theta}}_h^{i+1} - \boldsymbol{\theta}_h^*\right\rangle - \frac{1}{\alpha_{i,h}}\left\|\widetilde{\boldsymbol{\theta}}_h^{i+1} - \boldsymbol{\theta}_h^*\right\|_{\nabla^2\ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1})}, \tag{55}$$

where $\alpha_{i,h} = \log|\mathcal{S}_{i,h}| + 2(1 + L_{\boldsymbol{\varphi}}L_{\boldsymbol{\theta}})$.

By setting $\eta = \alpha_{i,h}/2$ and merging equations (54) and (55), we arrive at

$$\ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}) - \ell_{i,h}(\boldsymbol{\theta}_h^*) \le \left\langle\nabla\ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}) - \nabla\widetilde{\ell}_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}), \widetilde{\boldsymbol{\theta}}_h^{i+1} - \boldsymbol{\theta}_h^*\right\rangle$$

$$+ \frac{1}{\alpha_{i,h}}\left(\left\|\widetilde{\boldsymbol{\theta}}_h^i - \boldsymbol{\theta}_h^*\right\|^2_{\mathbf{B}_{i,h}} - \left\|\widetilde{\boldsymbol{\theta}}_h^{i+1} - \boldsymbol{\theta}_h^*\right\|^2_{\mathbf{B}_{i+1,h}} - \left\|\widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^i\right\|_{\mathbf{B}_{i,h}}\right). \tag{56}$$

Meanwhile, we obtain

$$\nabla\widetilde{\ell}_{i,h}(\boldsymbol{\theta}) = \nabla\ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^i) + \nabla^2\ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^i)(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^i) \tag{57}$$

by taking the gradient over both sides of the Taylor approximation of $\ell_{i,h}(\boldsymbol{\theta})$. Using (57), we proceed to bound the first term of (56) as follows:

$$\left\langle\nabla\ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}) - \nabla\widetilde{\ell}_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}), \widetilde{\boldsymbol{\theta}}_h^{i+1} - \boldsymbol{\theta}_h^*\right\rangle$$

$$= \left\langle\nabla\ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}) - \nabla\ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^i) - \nabla^2\ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^i)(\widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^i), \widetilde{\boldsymbol{\theta}}_h^{i+1} - \boldsymbol{\theta}^*\right\rangle$$

$$= \left\langle D^3\ell_{i,h}(\bar{\boldsymbol{\theta}}_h^{i+1})\left[\widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^i\right](\widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^i), \widetilde{\boldsymbol{\theta}}_h^{i+1} - \boldsymbol{\theta}^*\right\rangle$$

$$\le 3\sqrt{2}L_{\boldsymbol{\varphi}}\left\|\widetilde{\boldsymbol{\theta}}_h^{i+1} - \boldsymbol{\theta}_h^*\right\|_2\left\|\widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^i\right\|^2_{\nabla^2\ell_{i,h}(\bar{\boldsymbol{\theta}}_h^{i+1})}$$

$$\le 3\sqrt{2}L_{\boldsymbol{\varphi}}\left\|\widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^i\right\|^2_{\nabla^2\ell_{i,h}(\bar{\boldsymbol{\theta}}_h^{i+1})}$$

$$\le 3\sqrt{2}L_{\boldsymbol{\varphi}}^3\left\|\widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^i\right\|^2_2$$

where $\bar{\boldsymbol{\theta}}_h^{i+1}$ is a convex combination of $\widetilde{\boldsymbol{\theta}}_h^i$ and $\widetilde{\boldsymbol{\theta}}_h^{i+1}$. The second equality arises from the Taylor expansion, the first inequality is due to the self-concordant property, and the final inequality is justified by the following:

$$\nabla^2\ell_{i,h}(\bar{\boldsymbol{\theta}}_h^{i+1})$$

$$= \sum_{s'\in\mathcal{S}_{i,h}} P_{\bar{\boldsymbol{\theta}}_h^{i+1}}(s'\mid s_h^i, a_h^i)\boldsymbol{\varphi}_{i,h,s'}\boldsymbol{\varphi}_{i,h,s'}^\top$$

$$- \sum_{s'\in\mathcal{S}_{i,h}}\sum_{s''\in\mathcal{S}_{i,h}} P_{\bar{\boldsymbol{\theta}}_h^{i+1}}(s'\mid s_h^i, a_h^i)P_{\bar{\boldsymbol{\theta}}_h^{i+1}}(s''\mid s_h^i, a_h^i)\boldsymbol{\varphi}_{i,h,s'}\boldsymbol{\varphi}_{i,h,s''}^\top$$

$$\preceq \sum_{s'\in\mathcal{S}_{i,h}} P_{\bar{\boldsymbol{\theta}}_h^{i+1}}(s'\mid s_h^i, a_h^i)\boldsymbol{\varphi}_{i,h,s'}\boldsymbol{\varphi}_{i,h,s'}^\top$$

$$\preceq L_{\boldsymbol{\varphi}}^2\mathbf{I}_d.$$

44

By summing over i and reorganizing the terms, we arrive at the final result as follows:

$$
\left\| \widetilde{\boldsymbol{\theta}}_h^{k+1} - \boldsymbol{\theta}_h^* \right\|_{\mathbf{B}_{k+1,h}}^2
$$

$$
\leq \sum_{i=1}^k \alpha_{i,h} \left(\ell_{i,h}(\boldsymbol{\theta}_h^*) - \ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}) \right) + \left\| \widetilde{\boldsymbol{\theta}}_h^1 - \boldsymbol{\theta}_h^* \right\|_{\mathbf{B}_{1,h}}^2
$$

$$
+ 3\sqrt{2} L_{\boldsymbol{\varphi}}^3 \sum_{i=1}^k \alpha_{i,h} \left\| \widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^i \right\|_2^2 - \sum_{i=1}^k \left\| \widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^i \right\|_{\mathbf{B}_{i,h}}^2
$$

$$
\leq \alpha \sum_{i=1}^k \left(\ell_{i,h}(\boldsymbol{\theta}_h^*) - \ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}) \right) + \lambda L_{\boldsymbol{\theta}}^2 + 3\sqrt{2} L_{\boldsymbol{\varphi}}^3 \alpha \sum_{i=1}^k \left\| \widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^i \right\|_2^2 - \sum_{i=1}^k \left\| \widetilde{\boldsymbol{\theta}}_h^{i+1} - \widetilde{\boldsymbol{\theta}}_h^i \right\|_{\mathbf{B}_{i,h}}^2.
$$

where the first inequality holds by Assumption 2 and the last inequality holds since $\alpha = \log \mathcal{U} + 2(1 + L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}) \geq \alpha_{i,h}$ for all $i \in [k]$. □

D.1.2 Proof of Lemma 14

Proof of Lemma 14. The norm of $\widetilde{\mathbf{z}}_{i,h} = \boldsymbol{\sigma}_{i,h}^+ \left(\mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{N}(\widetilde{\boldsymbol{\theta}}_h^i, c\mathbf{B}_{i,h}^{-1})}[\boldsymbol{\sigma}_{i,h}(\boldsymbol{\Phi}_{i,h}^\top \boldsymbol{\theta})] \right)$ is generally unbounded [27]. In this proof, we utilize the smoothed version of $\widetilde{\mathbf{z}}_{i,h}$, defined as follows:

$$
\widetilde{\mathbf{z}}_{i,h}^u = \boldsymbol{\sigma}_{i,h}^+ \left(\mathrm{smooth}_{i,h}^u \, \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{N}(\widetilde{\boldsymbol{\theta}}_h^i, c\mathbf{B}_{i,h}^{-1})}[\boldsymbol{\sigma}_{i,h}(\boldsymbol{\Phi}_{i,h}^\top \boldsymbol{\theta})] \right)
$$

where the smooth function $\mathrm{smooth}_{i,h}^u(\mathbf{p}) = (1-u)\mathbf{p} + (u/\mathcal{U})\mathbf{1}$ with $u \in [0, 1/2]$, and $\mathbf{1} \in \mathbb{R}^{|\mathcal{S}_{i,h}|}$ is an all-one vector.

Exploiting the property of $\boldsymbol{\sigma}_{i,h}^+$ such that $\boldsymbol{\sigma}_{i,h}(\boldsymbol{\sigma}_{i,h}^+(\mathbf{p})) = \mathbf{p}$ for any $\mathbf{p} \in \Delta_{|\mathcal{S}_{i,h}|}$, it is straightforward to show that $\widetilde{\mathbf{z}}_{i,h}^u = \boldsymbol{\sigma}_{i,h}^+(\mathrm{smooth}_{i,h}^u(\boldsymbol{\sigma}_{i,h}(\widetilde{\mathbf{z}}_{i,h})))$. Then, by Lemma 34, we have

$$
\sum_{i=1}^k \ell(\widetilde{\mathbf{z}}_{i,h}^u, y_h^i) - \sum_{i=1}^k \ell(\widetilde{\mathbf{z}}_{i,h}, y_h^i) \leq 2uk, \quad \text{and} \quad \|\widetilde{\mathbf{z}}_{i,h}^u\|_\infty \leq \log(\mathcal{U}/u). \tag{58}
$$

Given the definition of $\ell_{i,h}$, we know that $\ell(\mathbf{z}_{i,h}^*, y_h^i) = \ell_{i,h}(\boldsymbol{\theta}_h^*)$, where $\mathbf{z}_{i,h}^* = \boldsymbol{\Phi}_{i,h}^\top \boldsymbol{\theta}_h^*$. We can bound the gap between the loss of $\boldsymbol{\theta}_h^*$ and $\widetilde{\mathbf{z}}_{i,h}^u$ as follows:

$$
\sum_{i=1}^k \left(\ell_{i,h}(\boldsymbol{\theta}_h^*) - \ell(\widetilde{\mathbf{z}}_{i,h}^u, y_h^i) \right)
$$

$$
= \sum_{i=1}^k \left(\ell(\mathbf{z}_{i,h}^*, y_h^i) - \ell(\widetilde{\mathbf{z}}_{i,h}^u, y_h^i) \right)
$$

$$
\leq \sum_{i=1}^k \langle \nabla_z \ell(\mathbf{z}_{i,h}^*, y_h^i), \mathbf{z}_{i,h}^* - \widetilde{\mathbf{z}}_{i,h}^u \rangle - \sum_{i=1}^k \frac{1}{M_{i,h}} \|\mathbf{z}_{i,h}^* - \widetilde{\mathbf{z}}_{i,h}^u\|_{\nabla_z^2 \ell(\mathbf{z}_{i,h}^*, y_h^i)}^2
$$

$$
= \sum_{i=1}^k \langle \nabla_z \ell(\mathbf{z}_{i,h}^*, y_h^i), \mathbf{z}_{i,h}^* - \widetilde{\mathbf{z}}_{i,h}^u \rangle - \sum_{i=1}^k \frac{1}{M_{i,h}} \|\mathbf{z}_{i,h}^* - \widetilde{\mathbf{z}}_{i,h}^u\|_{\nabla \boldsymbol{\sigma}_{i,h}(\mathbf{z}_{i,h}^*)}^2, \tag{59}
$$

where $M_{i,h} = \log(|\mathcal{S}_{i,h}|) + 2\log(\mathcal{U}/u)$, and the second equality holds by a direct calculation of the first order and Hessian of the logistic loss.

Now, we first bound the first term of the right-hand side. Let $\mathbf{d}_{i,h} = (\mathbf{z}_{i,h}^* - \widetilde{\mathbf{z}}_{i,h}^u)/(M + L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}})$, where $M = \log \mathcal{U} + 2\log(\mathcal{U}/u)$. Then, one can check that $\|\mathbf{d}_{i,h}\|_\infty \leq 1$ since $\|\mathbf{z}_{i,h}^*\|_\infty \leq \max_{s' \in \mathcal{S}_{i,h}} \|\boldsymbol{\varphi}_{i,h,s'}\|_2 \|\boldsymbol{\theta}_h^*\|_2 \leq L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}$ and $\|\widetilde{\mathbf{z}}_{i,h}^u\|_\infty \leq \log(\mathcal{U}/u)$. Moreover, since $\mathbf{z}_{i,h}^*$ and $\widetilde{\mathbf{z}}_{i,h}^u$ are independent of y_h^i, $\mathbf{d}_{i,h}$ is $\mathcal{F}_{i,h}$-measurable. Since $\mathbb{E}[(\boldsymbol{\sigma}_{i,h}(\mathbf{z}_{i,h}^*) - y_h^i)(\boldsymbol{\sigma}_{i,h}(\mathbf{z}_{i,h}^*) - y_h^i)^\top \mid \mathcal{F}_{i,h}] = \nabla \boldsymbol{\sigma}_{i,h}(\mathbf{z}_{i,h}^*)$ and $\|\boldsymbol{\sigma}_{i,h}(\mathbf{z}_{i,h}^*) - y_h^i\|_1 \leq 2$, we can apply Lemma 32. For any k and $\delta \in (0, 1]$,

45

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

$$\sum_{i=1}^{k} \langle \nabla_z \ell(\mathbf{z}_{i,h}^*, y_h^i), \mathbf{z}_{i,h}^* - \widetilde{\mathbf{z}}_{i,h}^u \rangle$$

$$= (M + L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}) \sum_{i=1}^{k} \langle \nabla_z \ell(\mathbf{z}_{i,h}^*, y_h^i), \mathbf{d}_{i,h} \rangle$$

$$\leq (M + L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}) \sqrt{\lambda + \sum_{i=1}^{k} \|\mathbf{d}_{i,h}\|_{\nabla \boldsymbol{\sigma}_{i,h}(\mathbf{z}_{i,h}^*)}^2}$$

$$\cdot \sqrt{\frac{\sqrt{\lambda}}{4} + \frac{4}{\sqrt{\lambda}} \log \left(\frac{H\sqrt{1 + \frac{1}{\lambda} \sum_{i=1}^{k} \|\mathbf{d}_{i,h}\|_{\nabla \boldsymbol{\sigma}_{i,h}(\mathbf{z}_{i,h}^*)}^2}}{\delta} \right)}$$

$$\leq (M + L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}) \sqrt{\lambda + \sum_{i=1}^{k} \|\mathbf{d}_{i,h}\|_{\nabla \boldsymbol{\sigma}_{i,h}(\mathbf{z}_{i,h}^*)}^2} \sqrt{\frac{\sqrt{\lambda}}{4} + 4 \log \left(\frac{H\sqrt{1 + 2k}}{\delta} \right)}, \qquad (60)$$

where the second inequality holds since $\|\mathbf{d}_{i,h}\|_{\nabla \boldsymbol{\sigma}_{i,h}(\mathbf{z}_{i,h}^*)}^2 = \mathbf{d}_{i,h}^\top \nabla \boldsymbol{\sigma}_{i,h}(\mathbf{z}_{i,h}^*) \mathbf{d}_{i,h} \leq 2$ and $\lambda \geq 1$.
Plugging (60) into (59) and rearranging the term, we get

$$\sum_{i=1}^{k} \left(\ell_{i,h}(\boldsymbol{\theta}^*) - \ell(\widetilde{\mathbf{z}}_{i,h}^u, y_h^i) \right)$$

$$\leq (M + L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}) \sqrt{\lambda + \sum_{i=1}^{k} \|\mathbf{d}_{i,h}\|_{\nabla \boldsymbol{\sigma}_{i,h}(\mathbf{z}_{i,h}^*)}^2} \sqrt{\frac{\sqrt{\lambda}}{4} + 4 \log \left(H \frac{\sqrt{1 + 2k}}{\delta} \right)}$$

$$- \sum_{i=1}^{k} \frac{1}{M_{i,h}} \|\mathbf{z}_{i,h}^* - \widetilde{\mathbf{z}}_{i,h}^u\|_{\nabla \boldsymbol{\sigma}_{i,h}(\mathbf{z}_{i,h}^*)}^2$$

$$\leq (M + L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}) \sqrt{\lambda + \sum_{i=1}^{k} \|\mathbf{d}_{i,h}\|_{\nabla \boldsymbol{\sigma}_{i,h}(\mathbf{z}_{i,h}^*)}^2} \sqrt{\frac{\sqrt{\lambda}}{4} + 4 \log \left(\frac{H\sqrt{1 + 2k}}{\delta} \right)}$$

$$- (M + L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}) \sum_{i=1}^{k} \|\mathbf{d}_{i,h}\|_{\nabla \boldsymbol{\sigma}_{i,h}(\mathbf{z}_{i,h}^*)}^2$$

$$\leq (M + L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}) \left(\lambda + \sum_{i=1}^{k} \|\mathbf{d}_{i,h}\|_{\nabla \boldsymbol{\sigma}_{i,h}(\mathbf{z}_{i,h}^*)}^2 \right) + (M + L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}) \left(\frac{\sqrt{\lambda}}{4} + 4 \log \left(\frac{H\sqrt{1 + 2k}}{\delta} \right) \right)$$

$$- (M + L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}) \sum_{i=1}^{k} \|\mathbf{d}_{i,h}\|_{\nabla \boldsymbol{\sigma}_{i,h}(\mathbf{z}_{i,h}^*)}^2$$

$$\leq \frac{5}{4}(M + L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}})\lambda + 4(M + L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}) \log \left(\frac{H\sqrt{1 + 2k}}{\delta} \right). \qquad (61)$$

Finally, combining (58) and (61), by setting $u = 1/k$, we derive that

$$\sum_{i=1}^{k} \left(\ell_{i,h}(\boldsymbol{\theta}_h^*) - \ell(\widetilde{\mathbf{z}}_{i,h}, y_h^i) \right)$$

$$\leq \frac{5}{4}(M + L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}})\lambda + 4(M + L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}) \log \left(\frac{H\sqrt{1 + 2k}}{\delta} \right) + 2uk$$

$$\leq \frac{5}{4}(3 \log(\mathcal{U}k) + L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}})\lambda + 4(3 \log(\mathcal{U}k) + L_{\boldsymbol{\varphi}} L_{\boldsymbol{\theta}}) \log \left(\frac{H\sqrt{1 + 2k}}{\delta} \right) + 2$$

where the last inequality holds by the definition of $M = \log \mathcal{U} + 2 \log(\mathcal{U}/u)$. Taking the union bound over $h \in [H]$, we conclude the proof. $\qquad \square$

46

D.1.3 Proof of Lemma 15

Proof of Lemma 15. We start the proof from the observation of Proposition 2 in Foster et al. [27], stating that $\widetilde{\mathbf{z}}_{i,h}$ represents the mixed prediction, which adheres to the following property:

$$\ell(\widetilde{\mathbf{z}}_{i,h}, y_h^i) \leq -\log\left(\mathbb{E}_{\boldsymbol{\theta}\sim\mathcal{N}(\widetilde{\boldsymbol{\theta}}_h^i, c\mathbf{B}_{i,h}^{-1})}\left[\exp\left(-\ell_{i,h}(\boldsymbol{\theta})\right)\right]\right) = -\log\left(\frac{1}{Z_{i,h}}\int_{\mathbb{R}^d}\exp\left(-L_{i,h}(\boldsymbol{\theta})\right)\mathrm{d}\boldsymbol{\theta}\right),$$
(62)

where $L_{i,h}(\boldsymbol{\theta}) := \ell_{i,h}(\boldsymbol{\theta}) + \frac{1}{2c}\left\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^i\right\|_{\mathbf{B}_{i,h}}^2$ and $Z_{i,h} := \sqrt{(2\pi)^d c|\mathbf{B}_{i,h}^{-1}|}$.

Consider the quadratic approximation

$$\widetilde{L}_{i,h}(\boldsymbol{\theta}) = L_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}) + \left\langle\nabla L_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}), \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1}\right\rangle + \frac{1}{2c}\left\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1}\right\|_{\mathbf{B}_{i,h}}^2.$$

Using the property that $\ell_{i,h}$ is $3\sqrt{2}L_{\boldsymbol{\varphi}}$-self-concordant-like function as asserted by Proposition B.1 in [50], and applying Lemma 35, we obtain

$$L_{i,h}(\boldsymbol{\theta}) \leq \widetilde{L}_{i,h}(\boldsymbol{\theta}) + \exp\left(18L_{\boldsymbol{\varphi}}^2\left\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1}\right\|_2^2\right)\left\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1}\right\|_{\nabla\ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1})}^2.$$

Also, we have

$$\frac{1}{Z_{i,h}}\int_{\mathbb{R}^d}\exp(-L_{i,h}(\boldsymbol{\theta}))\,\mathrm{d}\boldsymbol{\theta}$$

$$\geq \frac{1}{Z_{i,h}}\int_{\mathbb{R}^d}\exp\left(-\widetilde{L}_{i,h}(\boldsymbol{\theta}) - \exp\left(18L_{\boldsymbol{\varphi}}^2\left\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1}\right\|_2^2\right)\left\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1}\right\|_{\nabla\ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1})}^2\right)\mathrm{d}\boldsymbol{\theta}$$

$$= \frac{\exp\left(-L_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1})\right)}{Z_{i,h}}\int_{\mathbb{R}^d}\widetilde{f}_{i+1,h}(\boldsymbol{\theta})\cdot\exp\left(-\left\langle\nabla L_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}), \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1}\right\rangle\right)\mathrm{d}\boldsymbol{\theta},$$
(63)

where we define the function $\widetilde{f}_{i,h} : \mathcal{B}(\mathbf{0}_d, 1) \to \mathbb{R}$ as

$$\widetilde{f}_{i+1,h}(\boldsymbol{\theta}) = \exp\left(-\frac{1}{2c}\left\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1}\right\|_{\mathbf{B}_{i,h}}^2 - \exp\left(18L_{\boldsymbol{\varphi}}^2\left\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1}\right\|_2^2\right)\left\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1}\right\|_{\nabla^2\ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1})}^2\right).$$

We denote $\widetilde{Z}_{i+1,h} = \int_{\mathbb{R}^d}\widetilde{f}_{i+1,h}(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \leq +\infty$ and define $\widetilde{\Theta}_{i+1,h}$ as the distribution whose density function is $\widetilde{f}_{i+1,h}(\boldsymbol{\theta})/\widetilde{Z}_{i+1,h}$. Then, we can rewrite (63) as follows:

$$\frac{1}{Z_{i,h}}\int_{\mathbb{R}^d}\exp(-L_{i,h}(\boldsymbol{\theta}))\,\mathrm{d}\boldsymbol{\theta}$$

$$\geq \frac{\exp\left(-L_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1})\right)\widetilde{Z}_{i+1,h}}{Z_{i,h}}\mathbb{E}_{\boldsymbol{\theta}\sim\widetilde{\Theta}_{i+1,h}}\left[\exp\left(-\left\langle\nabla L_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}), \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1}\right\rangle\right)\right]$$

$$\geq \frac{\exp\left(-L_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1})\right)\widetilde{Z}_{i+1,h}}{Z_{i,h}}\exp\left(-\mathbb{E}_{\boldsymbol{\theta}\sim\widetilde{\Theta}_{i+1,h}}\left[\left\langle\nabla L_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}), \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1}\right\rangle\right]\right)$$

$$= \frac{\exp\left(-L_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1})\right)\widetilde{Z}_{i+1,h}}{Z_{i,h}},$$
(64)

where the second inequality is by Jensen's inequality and the last inequality holds because $\widetilde{\Theta}_{i+1,h}$ is symmetric around $\widetilde{\boldsymbol{\theta}}_h^{i+1}$ and thus $\mathbb{E}_{\boldsymbol{\theta}\sim\widetilde{\Theta}_{i+1,h}}\left[\left\langle\nabla L_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}), \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1}\right\rangle\right] = 0$.

Combining (62) and (64), we get

$$\ell_{i,h}(\widetilde{\mathbf{z}}) \leq L_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}) + \log Z_{i,h} - \log\widetilde{Z}_{i+1,h}.$$
(65)

47

Moreover, we have

$$-\log \widetilde{Z}_{i+1,h}$$

$$= -\log \left(\int_{\mathbb{R}^d} \exp \left(-\frac{1}{2c} \left\| \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1} \right\|_{\mathbf{B}_{i,h}}^2 \right. \right.$$

$$\left. \left. - \exp \left(18 L_{\boldsymbol{\varphi}}^2 \left\| \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1} \right\|_2^2 \right) \left\| \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1} \right\|_{\nabla^2 \ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1})}^2 \right) d\boldsymbol{\theta} \right)$$

$$= -\log \left(\widehat{Z}_{i+1,h} \cdot \mathbb{E}_{\boldsymbol{\theta} \sim \widehat{\Theta}_{i+1,h}} \left[\exp \left(-\exp \left(18 L_{\boldsymbol{\varphi}}^2 \left\| \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1} \right\|_2^2 \right) \left\| \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1} \right\|_{\nabla^2 \ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1})}^2 \right) \right] \right)$$

$$\leq -\log \widehat{Z}_{i+1,h} + \mathbb{E}_{\boldsymbol{\theta} \sim \widehat{\Theta}_{i+1,h}} \left[\exp \left(18 L_{\boldsymbol{\varphi}}^2 \left\| \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1} \right\|_2^2 \right) \left\| \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1} \right\|_{\nabla^2 \ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1})}^2 \right]$$

$$= -\log Z_{i,h} + \mathbb{E}_{\boldsymbol{\theta} \sim \widehat{\Theta}_{i+1,h}} \left[\exp \left(18 L_{\boldsymbol{\varphi}}^2 \left\| \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1} \right\|_2^2 \right) \left\| \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1} \right\|_{\nabla^2 \ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1})}^2 \right], \tag{66}$$

where $\widehat{\Theta}_{i+1,h} = \mathcal{N}(\widetilde{\boldsymbol{\theta}}_h^{i+1}, c\mathbf{B}_{i,h}^{-1})$ and $\widehat{Z}_{i+1,h} = \int_{\mathbb{R}^d} \exp \left(-\frac{1}{2c} \left\| \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1} \right\|_{\mathbf{B}_{i,h}}^2 \right) d\boldsymbol{\theta}$, and the last inequality holds because $\widehat{Z}_{i+1,h}$ and $Z_{i,h}$ are identical normalizing factors. Integrating (65) and (66) and summing over k, yields

$$\sum_{i=1}^k \ell(\widetilde{\mathbf{z}}_{i,h}, y_h^i)$$

$$= \sum_{i=1}^k L_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}) + \sum_{i=1}^k \mathbb{E}_{\boldsymbol{\theta} \sim \widehat{\Theta}_{i+1,h}} \left[\exp \left(18 L_{\boldsymbol{\varphi}}^2 \left\| \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1} \right\|_2^2 \right) \left\| \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1} \right\|_{\nabla^2 \ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1})}^2 \right].$$

Moreover, we can further bound the second term on the right-hand side of (66). By Cauchy-Schwarz inequality, we get

$$\mathbb{E}_{\boldsymbol{\theta} \sim \widehat{\Theta}_{i+1,h}} \left[\exp \left(18 L_{\boldsymbol{\varphi}}^2 \left\| \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1} \right\|_2^2 \right) \left\| \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1} \right\|_{\nabla^2 \ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1})}^2 \right]$$

$$\leq \underbrace{\sqrt{\mathbb{E}_{\boldsymbol{\theta} \sim \widehat{\Theta}_{i+1,h}} \left[\exp \left(36 L_{\boldsymbol{\varphi}}^2 \left\| \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1} \right\|_2^2 \right) \right]}}_{(\mathrm{I})} \underbrace{\sqrt{\mathbb{E}_{\boldsymbol{\theta} \sim \widehat{\Theta}_{i+1,h}} \left[\left\| \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1} \right\|_{\nabla^2 \ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1})}^4 \right]}}_{(\mathrm{II})}.$$

Since $\widehat{\Theta}_{i+1,h} = \mathcal{N} \left(\widetilde{\boldsymbol{\theta}}_h^{i+1}, c\mathbf{B}_{i,h}^{-1} \right)$, $\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1}$ follows the same distribution as

$$\sum_{j=1}^d \sqrt{c\lambda_j \left(\mathbf{B}_{i,h}^{-1} \right)} X_j \mathbf{e}_j, \quad \text{where } X_j \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \forall j \in [d], \tag{67}$$

where $\lambda_j \left(\mathbf{B}_{i,h}^{-1} \right)$ denotes the j-th largest eigenvalue of $\mathbf{B}_{i,h}^{-1}$ and $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$ are orthogonal basis of $\mathbb{R}^d$. Furthermore, since we know that $\mathbf{B}_{i,h}^{-1} \leq \lambda^{-1} \mathbf{I}_d$, we can bound the term (I) by

$$(\mathrm{I}) \leq \sqrt{\mathbb{E}_{X_j} \left[\prod_{j=1}^d \exp \left(36 L_{\boldsymbol{\varphi}}^2 c\lambda^{-1} X_j^2 \right) \right]} = \sqrt{\prod_{j=1}^d \mathbb{E}_{X_j} \left[\exp \left(36 L_{\boldsymbol{\varphi}}^2 c\lambda^{-1} X_j^2 \right) \right]}$$

$$\leq \left(\mathbb{E}_{W \sim \chi^2} \left[\exp \left(36 L_{\boldsymbol{\varphi}}^2 c\lambda^{-1} W \right) \right] \right)^{\frac{d}{2}} \leq \mathbb{E}_{W \sim \chi^2} \left[\exp \left(18 L_{\boldsymbol{\varphi}}^2 c\lambda^{-1} W d \right) \right]$$

where χ^2 is the chi-square distribution and the last inequality holds due to Jensen's inequality. By choosing $\lambda \geq 72 L_{\boldsymbol{\varphi}}^2 cd$, we arrive that

$$(\mathrm{I}) \leq \mathbb{E}_{W \sim \chi^2} \left[\exp \left(\frac{W}{4} \right) \right] \leq \sqrt{2}, \tag{68}$$

48

where the last inequality holds because the moment-generating function for χ^2-distribution is bounded by $\mathbb{E}_{W\sim\chi^2}[\exp(tW)]\le 1/\sqrt{1-2t}$ for all $t\le 1/2$. Now, we bound the term (II).

$$\text{(II)}=\sqrt{\mathbb{E}_{\boldsymbol\theta\sim\widehat\Theta_{i+1,h}}\left[\left\|\boldsymbol\theta-\widetilde{\boldsymbol\theta}_h^{i+1}\right\|_{\nabla^2\ell_{i,h}(\widetilde{\boldsymbol\theta}_h^{i+1})}^4\right]}=\sqrt{\mathbb{E}_{\boldsymbol\theta\sim\mathcal{N}\left(0,c\mathbf{B}_{i,h}^{-1}\right)}\left[\|\boldsymbol\theta\|_{\nabla^2\ell_{i,h}(\widetilde{\boldsymbol\theta}_h^{i+1})}^4\right]}$$

$$=\sqrt{\mathbb{E}_{\boldsymbol\theta\sim\mathcal{N}\left(0,c\bar{\mathbf{B}}_{i,h}^{-1}\right)}[\|\boldsymbol\theta\|_2^4]},$$

where $\bar{\mathbf{B}}_{i,h}=\left(\nabla^2\ell_{i,h}(\widetilde{\boldsymbol\theta}_h^{i+1})\right)^{-1/2}\mathbf{B}_{i,h}\left(\nabla^2\ell_{i,h}(\widetilde{\boldsymbol\theta}_h^{i+1})\right)^{-1/2}$. Let $\bar\lambda_j:=\lambda_j\left(c\bar{\mathbf{B}}_{i,h}^{-1}\right)$ be the j-th largest eigenvalue of the matrix. Then, a similar analysis as (67) gives that

$$\text{(II)}=\sqrt{\mathbb{E}_{X_j\sim\mathcal{N}(0,1)}\left[\left\|\sum_{j=1}^d\sqrt{\bar\lambda_j}X_j\mathbf{e}_j\right\|_2^4\right]}=\sqrt{\mathbb{E}_{X_j\sim\mathcal{N}(0,1)}\left[\left(\sum_{j=1}^d\bar\lambda_j X_j^2\right)^2\right]}$$

$$=\sqrt{\sum_{j=1}^d\sum_{j'=1}^d\bar\lambda_j\bar\lambda_{j'}\mathbb{E}_{X_j,X_{j'}\sim\mathcal{N}(0,1)}[X_j^2 X_{j'}^2]}\le\sqrt{3\sum_{j=1}^d\sum_{j'=1}^d\bar\lambda_j\bar\lambda_{j'}}=\sqrt3 c\,\mathrm{tr}\left(\bar{\mathbf{B}}_{i,h}^{-1}\right),$$

where the last inequality holds due to $\mathbb{E}_{X_j,X_{j'}\sim\mathcal{N}(0,1)}[X_j^2 X_{j'}^2]\le 3$ when considering the case where $j=j'$ and the last equality is derived from the fact that $\left(\sum_{j=1}^d\bar\lambda_j\right)^2=\mathrm{tr}\left(c\bar{\mathbf{B}}_{i,h}^{-1}\right)$. Here, we denote $\mathrm{tr}(A)$ as the trace of the matrix A.

We define matrix $\mathbf{R}_{i+1,h}:=\lambda\mathbf{I}_d/2+\sum_{\tau=1}^i\nabla^2\ell_{\tau,h}(\boldsymbol\theta_{\tau+1,h})$. Under the condition $\lambda\ge 2L_\varphi^2$, we have $\nabla^2\ell_{i,h}(\boldsymbol\theta_{i+1,h})\preceq L_\varphi^2\mathbf{I}_d\le\frac{\lambda}{2}\mathbf{I}_d$. Then, we have $\mathbf{B}_{i,h}\succeq\mathbf{R}_{i+1,h}$. Therefore, we can bound the trace by

$$\mathrm{tr}\left(\bar{\mathbf{B}}_{i,h}^{-1}\right)=\mathrm{tr}\left(\mathbf{B}_{i,h}^{-1}\nabla^2\ell_{i,h}(\boldsymbol\theta_{i+1,h})\right)\le\mathrm{tr}\left(\mathbf{R}_{i+1,h}^{-1}\nabla^2\ell_{i,h}(\boldsymbol\theta_{i+1,h})\right)$$

$$=\mathrm{tr}\left(\mathbf{R}_{i+1,h}^{-1}(\mathbf{R}_{i+1,h}-\mathbf{R}_{i,h})\right)\le\log\frac{\det(\mathbf{R}_{i+1,h})}{\det(\mathbf{R}_{i,h})},$$

where the last inequality holds due to Lemma 4.7 of Hazan et al. [32]. Therefore we can bound the term (II) as

$$\text{(II)}\le\sqrt3 c\log\frac{\det(\mathbf{R}_{i+1,h})}{\det(\mathbf{R}_{i,h})}.\tag{69}$$

Combining (68) and (69), we get

$$\mathbb{E}_{\boldsymbol\theta\sim\widehat\Theta_{i+1,h}}\left[\exp\left(6\left\|\boldsymbol\theta-\widetilde{\boldsymbol\theta}_h^{i+1}\right\|_2^2\right)\left\|\boldsymbol\theta-\widetilde{\boldsymbol\theta}_h^{i+1}\right\|_{\nabla^2\ell_{i,h}(\widetilde{\boldsymbol\theta}_h^{i+1})}^2\right]\le\sqrt6 c\log\frac{\det(\mathbf{R}_{i+1,h})}{\det(\mathbf{R}_{i,h})}.\tag{70}$$

Plugging (66) and (70) into (65), and taking summation over k, we derive that

$$\sum_{i=1}^k\ell(\widetilde{\mathbf{z}}_{i,h},y_h^i)\le\sum_{i=1}^k L_{i,h}(\widetilde{\boldsymbol\theta}_h^{i+1})+\sqrt6 c\sum_{i=1}^k\log\frac{\det(\mathbf{R}_{i+1,h})}{\det(\mathbf{R}_{i,h})}$$

$$=\sum_{i=1}^k\left(\ell_{i,h}(\widetilde{\boldsymbol\theta}_h^{i+1})+\frac{1}{2c}\left\|\widetilde{\boldsymbol\theta}_h^{i+1}-\widetilde{\boldsymbol\theta}_h^i\right\|_{\mathbf{B}_{i,h}}^2\right)+\sqrt6 c\sum_{i=1}^k\log\frac{\det(\mathbf{R}_{i+1,h})}{\det(\mathbf{R}_{i,h})}$$

$$\le\sum_{i=1}^k\left(\ell_{i,h}(\widetilde{\boldsymbol\theta}_h^{i+1})+\frac{1}{2c}\left\|\widetilde{\boldsymbol\theta}_h^{i+1}-\widetilde{\boldsymbol\theta}_h^i\right\|_{\mathbf{B}_{i,h}}^2\right)+\sqrt6 cd\log\left(1+\frac{2kL_\varphi^2}{d\lambda}\right),$$

where the last inequality holds because $\sum_{i=1}^k\log\frac{\det(\mathbf{R}_{i+1,h})}{\det(\mathbf{R}_{i,h})}=\log(\det(\mathbf{R}_{k+1,h})/\det(\lambda/2\mathbf{I}_d))\le d\log\left(1+\frac{2kL_\varphi^2}{d\lambda}\right)$. By rearranging the terms, we conclude the proof. $\qquad\square$

49

D.2 Bound on Prediction Error

In this section, we present the bound on the prediction error of parameters updated by `ORRL-MNL`. First, we compare the problem setting of MNL contextual bandits with ours and introduce the challenges of applying their analysis to our setting.

MNL dynamic assortment optimization (single-parameter & uniform reward) [61] Perivier and Goyal [61] consider an assortment selection problem where the user choice is given by a MNL choice model with the single-parameter. At each time t, the agent observes context features $\{\mathbf{x}_{t,i}\}_{i=1}^{M} \subset \mathbb{R}^d$. Then the agent decides on the set $S_t \subset [M]$ to offer to a user, with $|S_t| \leq N$. Without loss of generality, we may assume $|S_t| = N$. Then the user purchases one single product $j \in S_t \cup \{0\}$ and the probability of each product j is purchased by a user follows the MNL model parametrized by a unknown fixed parameter $\boldsymbol{\theta}^* \in \mathbb{R}^d$,

$$
q_{t,j}(S_t, \boldsymbol{\theta}^*) := \begin{cases} \frac{\exp(\mathbf{x}_{t,j}^\top \boldsymbol{\theta}^*)}{1+\sum_{k \in S_t} \exp(\mathbf{x}_k^\top \boldsymbol{\theta}^*)} & \text{if } j \in S_t \\ \frac{1}{1+\sum_{k \in S_t} \exp(\mathbf{x}_k^\top \boldsymbol{\theta}^*)} & \text{if } j = 0 \,. \end{cases}
$$

Then the difference between the revenue induced by $\boldsymbol{\theta}^*$ and that by an estimator $\boldsymbol{\theta}$ in Perivier and Goyal [61] is expressed as follows:

$$
\sum_{j \in S_t} q_{t,j}(S_t, \boldsymbol{\theta}^*) - \sum_{j \in S_t} q_{t,j}(S_t, \boldsymbol{\theta}) \,. \tag{71}
$$

If we define $Q : \mathbb{R}^N \to \mathbb{R}$, such that for all $\mathbf{u} = (u_1, \ldots, u_N) \in \mathbb{R}^N, Q(\mathbf{u}) := \sum_{i=1}^{N} \frac{\exp(u_i)}{1+\sum_{j=1}^{N}\exp(u_j)}$ and let $\mathbf{v}^* = (\mathbf{x}_{t,i_1}^\top \boldsymbol{\theta}^*, \ldots, \mathbf{x}_{t,i_N}^\top \boldsymbol{\theta}^*)$ and $\mathbf{v} = (\mathbf{x}_{t,i_1}^\top \boldsymbol{\theta}, \ldots, \mathbf{x}_{t,i_N}^\top \boldsymbol{\theta})$, then Eq. (71) can be expressed as follows:

$$
\sum_{j \in S_t} q_{t,j}(S_t, \boldsymbol{\theta}^*) - \sum_{j \in S_t} q_{t,j}(S_t, \boldsymbol{\theta}) = Q(\mathbf{v}^*) - Q(\mathbf{v})
$$

$$
= \nabla Q(\mathbf{v}^*)^\top (\mathbf{v}^* - \mathbf{v}) + \frac{1}{2}(\mathbf{v}^* - \mathbf{v})^\top \nabla^2 Q(\bar{\mathbf{v}})(\mathbf{v}^* - \mathbf{v}) \,, \tag{72}
$$

where $\bar{\mathbf{v}}$ is a convex combination of $\mathbf{v}^*$ and $\mathbf{v}$. For the first term in Eq. (72), we have

$$
\nabla Q(\mathbf{v}^*)^\top (\mathbf{v}^* - \mathbf{v})
$$
$$
= \frac{\sum_{i \in S_t} \exp(\mathbf{x}_{t,j}^\top \boldsymbol{\theta}^*)(v_j - v_j^*)}{1 + \sum_{j \in S_t} \exp(\mathbf{x}_{t,j}^\top \boldsymbol{\theta}^*)} - \frac{\sum_{i \in S_t} \exp(\mathbf{x}_{t,j}^\top \boldsymbol{\theta}^*) \sum_{i \in S_t} \exp(\mathbf{x}_{t,i}^\top \boldsymbol{\theta}^*)(v_j - v_j^*)}{\left(1 + \sum_{j \in S_t} \exp(\mathbf{x}_{t,j}^\top \boldsymbol{\theta}^*)\right)^2}
$$
$$
= \sum_{j \in S_t} q_{t,j}(S_t, \boldsymbol{\theta}^*)\mathbf{x}_{t,j}^\top(\boldsymbol{\theta}^* - \boldsymbol{\theta}) - \sum_{j \in S_t}\sum_{i \in S_t} q_{t,j}(S_t, \boldsymbol{\theta}^*)q_{t,j}(S_t, \boldsymbol{\theta}^*)\mathbf{x}_{t,i}^\top(\boldsymbol{\theta}^* - \boldsymbol{\theta})
$$
$$
= \sum_{j \in S_t} q_{t,j}(S_t, \boldsymbol{\theta}^*)\left(1 - \sum_{i \in S_t} q_{t,i}(S_t, \boldsymbol{\theta}^*)\right)\mathbf{x}_{t,j}^\top(\boldsymbol{\theta}^* - \boldsymbol{\theta})
$$
$$
= \sum_{j \in S_t} q_{t,j}(S_t, \boldsymbol{\theta}^*)q_{t,0}(S_t, \boldsymbol{\theta}^*)\mathbf{x}_{t,j}^\top(\boldsymbol{\theta}^* - \boldsymbol{\theta})
$$
$$
\leq \sum_{j \in S_t} q_{t,j}(S_t, \boldsymbol{\theta}^*)q_{t,0}(S_t, \boldsymbol{\theta}^*)\|\mathbf{x}_{t,j}\|_{\mathbf{H}_t^{-1}(\boldsymbol{\theta}^*)}\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_{\mathbf{H}_t(\boldsymbol{\theta}^*)} \,, \tag{73}
$$

where $\mathbf{H}_t(\boldsymbol{\theta})$ is the Gram matrix used in [61] defined by

$$
\mathbf{H}_t(\boldsymbol{\theta}^*) := \sum_{\tau=1}^{t-1} \sum_{j \in S_\tau} q_{\tau,j}(S_\tau, \boldsymbol{\theta}^*)\mathbf{x}_{\tau,j}\mathbf{x}_{\tau,j}^\top - \sum_{j \in S_\tau}\sum_{i \in S_\tau} q_{\tau,j}(S_\tau, \boldsymbol{\theta}^*)q_{\tau,i}(S_\tau, \boldsymbol{\theta}^*)\mathbf{x}_{\tau,j}\mathbf{x}_{\tau,i}^\top \,.
$$

Note that the term $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_{\mathbf{H}_t(\boldsymbol{\theta}^*)}$ can be bounded by the concentration result of the estimated parameter. On the other hand, to apply the elliptical potential lemma to the term

50

$\sum_{j \in S_t} q_{t,j}(S_t, \boldsymbol{\theta}^*) q_{t,0}(S_t, \boldsymbol{\theta}^*) \|\mathbf{x}_{t,j}\|_{\mathbf{H}_t^{-1}(\boldsymbol{\theta}^*)}$, note that $\mathbf{H}_t(\boldsymbol{\theta}^*)$ can be bounded as follows:

$$
\begin{aligned}
&\mathbf{H}_t(\boldsymbol{\theta}^*) \\
&= \mathbf{H}_{t-1}(\boldsymbol{\theta}^*) + \sum_{j \in S_t} q_{t,j}(S_t, \boldsymbol{\theta}^*) \mathbf{x}_{t,j} \mathbf{x}_{t,j}^\top - \frac{1}{2} \sum_{i,j \in S_t} q_{t,j}(S_t, \boldsymbol{\theta}^*) q_{t,i}(S_t, \boldsymbol{\theta}^*) \left(\mathbf{x}_{t,j} \mathbf{x}_{t,i}^\top + \mathbf{x}_{t,i} \mathbf{x}_{t,j}^\top \right) \\
&\succeq \mathbf{H}_{t-1}(\boldsymbol{\theta}^*) + \sum_{j \in S_t} q_{t,j}(S_t, \boldsymbol{\theta}^*) \mathbf{x}_{t,j} \mathbf{x}_{t,j}^\top - \frac{1}{2} \sum_{i,j \in S_t} q_{t,j}(S_t, \boldsymbol{\theta}^*) q_{t,i}(S_t, \boldsymbol{\theta}^*) \left(\mathbf{x}_{t,j} \mathbf{x}_{t,j}^\top + \mathbf{x}_{t,i} \mathbf{x}_{t,i}^\top \right) \\
&= \mathbf{H}_{t-1}(\boldsymbol{\theta}^*) + \sum_{j \in S_t} q_{t,j}(S_t, \boldsymbol{\theta}^*) \left(1 - \sum_{i \in S_t} q_{t,i}(S_t, \boldsymbol{\theta}^*) \right) \mathbf{x}_{t,j} \mathbf{x}_{t,j}^\top \\
&= \mathbf{H}_{t-1}(\boldsymbol{\theta}^*) + \sum_{j \in S_t} q_{t,j}(S_t, \boldsymbol{\theta}^*) q_{t,0}(S_t, \boldsymbol{\theta}^*) \mathbf{x}_{t,j} \mathbf{x}_{t,j}^\top \,.
\end{aligned}
\tag{74}
$$

Now since the coefficient $q_{t,j}(S_t, \boldsymbol{\theta}^*) q_{t,0}(S_t, \boldsymbol{\theta}^*)$ of $\|\mathbf{x}\|_{\mathbf{H}_t^{-1}(\boldsymbol{\theta}^*)}$ in Eq. (73) aligns with the coefficients of the lower bound of $\mathbf{H}_t(\boldsymbol{\theta}^*)$ in Eq. (74), the elliptical potential lemma can be applied. Note that such a lower bound in Eq. (74) holds since Perivier and Goyal [61] deals with the uniform reward, i.e., $1 - \sum_{i \in S_t} q_{t,i}(S_t, \boldsymbol{\theta}^*) = q_{t,0}(S_t, \boldsymbol{\theta}^*)$.

Mulitinomial logistic bandit problem [76] Zhang and Sugiyama [76] address the multiple-parameter MNL contextual bandit problem where at each time step t the agent selects an action $\mathbf{x}_t \in \mathbb{R}^d$ and receives response feedback $y_t \in \{0\} \cup [N]$ with $N + 1$ possible outcomes. Each outcome $i \in [N]$ is associated with a ground-truth parameter $\boldsymbol{\theta}_i^* \in \mathbb{R}^d$, and the probability of the outcome $\mathbb{P}(y_t = i \mid \mathbf{x}_t)$ follows the MNL model,

$$
\mathbb{P}(y_t = i \mid \mathbf{x}_t) = \frac{\exp(\mathbf{x}_t^\top \boldsymbol{\theta}_i^*)}{1 + \sum_{j=1}^N \exp(\mathbf{x}_t^\top \boldsymbol{\theta}_j^*)}, \quad \mathbb{P}(y_t = 0 \mid \mathbf{x}_t) = 1 - \sum_{j=1}^N \mathbb{P}(y_t = j \mid \mathbf{x}_t) \,.
$$

In this model, there are N unknown choice parameter $\boldsymbol{\Theta}^* := [\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_N^*] \in \mathbb{R}^{d \times N}$ and the agent chooses one context feature $\mathbf{x}_t$, that is why we call multiple-parameter MNL model. Then, the expected revenue of an action $\mathbf{x}_t$ in [76] is given by

$$
\sum_{i=1}^N \frac{\exp(\mathbf{x}_t^\top \boldsymbol{\theta}_i^*) \rho_i}{1 + \sum_{j=1}^N \exp(\mathbf{x}_t^\top \boldsymbol{\theta}_j^*)} := \boldsymbol{\rho}^\top \boldsymbol{\sigma}(\boldsymbol{\Theta}^* \mathbf{x}_t) \,,
$$

where we define the softmax function $\boldsymbol{\sigma} : \mathbb{R}^N \to [0,1]^N$ by

$$
[\boldsymbol{\sigma}(\mathbf{z})]_k = \frac{\exp([\mathbf{z}]_i)}{1 + \sum_{j=1}^N \exp([\mathbf{z}]_j)} \quad \forall k \in [N] \quad \text{and} \quad [\boldsymbol{\sigma}(\mathbf{z})]_0 = \frac{1}{1 + \sum_{j=1}^N \exp([\mathbf{z}]_j)} \quad \forall k \in [N] \,,
$$

and $\boldsymbol{\rho} := [\rho_1, \ldots, \rho_N] \in \mathbb{R}_+^{N+1}$ represents the reward for each outcome $j \in [N]$ with $\rho_0 = 0$. Then, the difference between the revenue induced by $\boldsymbol{\Theta}^*$ and that by an estimator $\hat{\boldsymbol{\Theta}}$ in [76] is expressed by

$$
\begin{aligned}
&\boldsymbol{\rho}^\top \left(\boldsymbol{\sigma}(\boldsymbol{\Theta}^* \mathbf{x}_t) - \boldsymbol{\sigma}(\hat{\boldsymbol{\Theta}} \mathbf{x}_t) \right) \\
&= \sum_{k=1}^N \rho_k \left([\boldsymbol{\sigma}(\boldsymbol{\Theta}^* \mathbf{x}_t)]_k - [\boldsymbol{\sigma}(\hat{\boldsymbol{\Theta}} \mathbf{x}_t)]_k \right) \\
&= \sum_{k=1}^N \rho_k \left(\nabla [\boldsymbol{\sigma}(\hat{\boldsymbol{\Theta}} \mathbf{x}_t)]_k \right)^\top (\boldsymbol{\Theta}^* - \hat{\boldsymbol{\Theta}}) \mathbf{x}_t + \sum_{k=1}^N \rho_k \|(\boldsymbol{\Theta}^* - \boldsymbol{\Theta}) \mathbf{x}_t\|_{\boldsymbol{\Xi}_k} \,,
\end{aligned}
\tag{75}
$$

51

where $\Xi_k = \int_0^1 (1-\nu)\nabla^2[\boldsymbol{\sigma}(\hat{\boldsymbol{\Theta}}\mathbf{x}_t + \nu(\boldsymbol{\Theta}^* - \hat{\boldsymbol{\Theta}})\mathbf{x}_t)]_k d\nu$. Then for the first term in Eq. (75), we have

$$
\begin{aligned}
\sum_{k=1}^{N} &\rho_k \left(\nabla[\boldsymbol{\sigma}(\hat{\boldsymbol{\Theta}}\mathbf{x}_t)]_k\right)^\top (\boldsymbol{\Theta}^* - \hat{\boldsymbol{\Theta}})\mathbf{x}_t \\
&\leq \left|\boldsymbol{\rho}^\top \nabla\boldsymbol{\sigma}(\hat{\boldsymbol{\Theta}}\mathbf{x}_t)(\boldsymbol{\Theta}^* - \hat{\boldsymbol{\Theta}})\mathbf{x}_t\right| \\
&= \left|\boldsymbol{\rho}^\top \nabla\boldsymbol{\sigma}(\hat{\boldsymbol{\Theta}}\mathbf{x}_t)(\mathbf{I}_N \otimes \mathbf{x}_t^\top)(\mathrm{vec}(\boldsymbol{\Theta}^*) - \mathrm{vec}(\hat{\boldsymbol{\Theta}}))\right| \\
&\leq \|\mathrm{vec}(\boldsymbol{\Theta}^*) - \mathrm{vec}(\hat{\boldsymbol{\Theta}})\|_{\mathbf{H}_t} \|\mathbf{H}_t^{-\frac{1}{2}}(\mathbf{I}_N \otimes \mathbf{x}_t^\top)\nabla\boldsymbol{\sigma}(\hat{\boldsymbol{\Theta}}\mathbf{x}_t)\boldsymbol{\rho}\|_2 \qquad (76)
\end{aligned}
$$

where $\mathbf{H}_t$ is the Gram matrix used in [76] defined by

$$
\mathbf{H}_t := \lambda\mathbf{I}_N + \sum_{s=1}^{t-1} \nabla\boldsymbol{\sigma}(\hat{\boldsymbol{\Theta}}_{s+1}\mathbf{x}_s) \otimes \mathbf{x}_s\mathbf{x}_s^\top .
$$

Note that the term $\|\mathrm{vec}(\boldsymbol{\Theta}^*) - \mathrm{vec}(\hat{\boldsymbol{\Theta}})\|_{\mathbf{H}_t}$ in Eq. (76) can be bounded by the concentration result of the estimated parameter, and the term $\|\mathbf{H}_t^{-\frac{1}{2}}(\mathbf{I}_N \otimes \mathbf{x}_t^\top)\nabla\boldsymbol{\sigma}(\hat{\boldsymbol{\Theta}}\mathbf{x}_t)\boldsymbol{\rho}\|_2$ also can be bounded as follows:

$$
\|\mathbf{H}_t^{-\frac{1}{2}}(\mathbf{I}_N \otimes \mathbf{x}_t^\top)\nabla\boldsymbol{\sigma}(\hat{\boldsymbol{\Theta}}\mathbf{x}_t)\boldsymbol{\rho}\|_2 \leq \|\boldsymbol{\rho}\|_2 \|\mathbf{H}_t^{-\frac{1}{2}}(\mathbf{I}_N \otimes \mathbf{x}_t^\top)\nabla\boldsymbol{\sigma}(\hat{\boldsymbol{\Theta}}\mathbf{x}_t)\|_2 .
$$

Here Zhang and Sugiyama [76] bound the term $\|\mathbf{H}_t^{-\frac{1}{2}}(\mathbf{I}_N \otimes \mathbf{x}_t^\top)\nabla\boldsymbol{\sigma}(\hat{\boldsymbol{\Theta}}\mathbf{x}_t)\|_2$ using a matrix version of elliptical lemma. However, they assume $\|\boldsymbol{\rho}\|_2 \leq R$ (Assumption 2 in [76]).

Now, regarding the prediction error in our setting, the estimated values $(\widetilde{V}_{h+1}^k(\cdot))$ for each reachable state are typically distinct, and we do not assume a constant upper bound on the ℓ_2-norm of the estimated value vector for all reachable states. Instead, we can bound the ℓ_2-norm of the estimated value vector for all reachable states as follows:

$$
\|\widetilde{\mathbf{V}}_{h+1}^k(s,a)\|_2 \leq \max_{s' \in \mathcal{S}_{s,a}} \left|\widetilde{V}_{h+1}^k(s')\right| \sqrt{|\mathcal{S}_{s,a}|} \leq H\sqrt{\mathcal{U}} ,
$$

where $\widetilde{\mathbf{V}}_{h+1}^k(s,a) := \left[\widetilde{V}_{h+1}^k(s')\right]_{s' \in \mathcal{S}_{s,a}} \in \mathbb{R}^{|\mathcal{S}_{s,a}|}$. However, such a bound leads to a looser regret by a factor of $\sqrt{\mathcal{U}}$. To address, we adapt the *feature centralization technique* [50] to bound the prediction error independently of $\mathcal{U}$, without making any additional assumptions. The key point is that the Hessian of per-round loss $\ell_{k,h}(\boldsymbol{\theta})$ is expressed in terms of the centralized feature as follows:

$$
\nabla^2 \ell_{k,h}(\boldsymbol{\theta}) = \sum_{s' \in \mathcal{S}_{k,h}} P_{\boldsymbol{\theta}}(s' \mid s_h^k, a_h^k)\bar{\boldsymbol{\varphi}}(s_h^k, a_h^k, s'; \boldsymbol{\theta})\bar{\boldsymbol{\varphi}}(s_h^k, a_h^k, s'; \boldsymbol{\theta})^\top .
$$

where $\bar{\boldsymbol{\varphi}}(s, a, s'; \boldsymbol{\theta}) := \boldsymbol{\varphi}(s, a, s') - \mathbb{E}_{\widetilde{s} \sim P_{\boldsymbol{\theta}}(\cdot|s,a)}[\boldsymbol{\varphi}(s, a, \widetilde{s})]$ is the centralized feature by $\boldsymbol{\theta}$. Now, we provide the bound on prediction error of the estimated parameter updated by ORRL-MNL.

Lemma 16 (Bound on the prediction error). *For any $\delta \in (0, 1)$, suppose that Lemma 12 holds. Let us denote the prediction error about $\widetilde{\boldsymbol{\theta}}_h^k$ by*

$$
\Delta_h^k(s,a) := \sum_{s' \in \mathcal{S}_{s,a}} \left(P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s, a) - P_{\boldsymbol{\theta}_h^*}(s' \mid s, a)\right) \widetilde{V}_{h+1}^k(s') .
$$

Then, for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$
|\Delta_h^k(s,a)| \leq H\beta_k(\delta) \sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s, a) \left\|\bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)\right\|_{\mathbf{B}_{k,h}^{-1}} + 3H\beta_k(\delta)^2 \max_{s' \in \mathcal{S}_{s,a}} \|\boldsymbol{\varphi}_{s,a,s'}\|_{\mathbf{B}_{k,h}^{-1}}^2 .
$$

Proof of Lemma 16. Let us define $F(\boldsymbol{\theta}) := \sum_{s' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s' \mid s, a)\widetilde{V}_{h+1}^k(s')$. Then, by Taylor expansion we have

$$
F(\boldsymbol{\theta}_h^*) = F(\widetilde{\boldsymbol{\theta}}_h^k) + \nabla F(\widetilde{\boldsymbol{\theta}}_h^k)^\top (\boldsymbol{\theta}_h^* - \widetilde{\boldsymbol{\theta}}_h^k) + \frac{1}{2}(\boldsymbol{\theta}_h^* - \widetilde{\boldsymbol{\theta}}_h^k)^\top \nabla^2 F(\bar{\boldsymbol{\theta}})(\boldsymbol{\theta}_h^* - \widetilde{\boldsymbol{\theta}}_h^k) ,
$$

52

where $\bar{\boldsymbol{\theta}} = (1-v)\boldsymbol{\theta}_h^* + v\widetilde{\boldsymbol{\theta}}_h^k$ for some $v \in (0,1)$. By Proposition 1, we have

$$
\begin{aligned}
\nabla F(\boldsymbol{\theta}) &= \sum_{s' \in \mathcal{S}_{s,a}} \nabla P_{\boldsymbol{\theta}}(s' \mid s,a) \widetilde{V}_{h+1}^k(s') \\
&= \sum_{s' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s' \mid s,a) \left(\boldsymbol{\varphi}_{s,a,s'} - \sum_{\widetilde{s} \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(\widetilde{s} \mid s,a) \boldsymbol{\varphi}_{s,a,\widetilde{s}} \right) \widetilde{V}_{h+1}^k(s') \\
&= \sum_{s' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s' \mid s,a) \bar{\boldsymbol{\varphi}}_{s,a,s'}(\boldsymbol{\theta}) \widetilde{V}_{h+1}^k(s'),
\end{aligned}
$$

and

$$
\begin{aligned}
\nabla^2 F(\boldsymbol{\theta}) &= \sum_{s' \in \mathcal{S}_{s,a}} \nabla^2 P_{\boldsymbol{\theta}}(s' \mid s,a) \widetilde{V}_{h+1}^k(s') \\
&= \sum_{s' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s' \mid s,a) \widetilde{V}_{h+1}^k(s') \boldsymbol{\varphi}_{s,a,s'} \boldsymbol{\varphi}_{s,a,s'}^\top \\
&\quad - \sum_{s' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s' \mid s,a) \widetilde{V}_{h+1}^k(s') \\
&\qquad \cdot \sum_{s'' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s,a) \left(\boldsymbol{\varphi}_{s,a,s'} \boldsymbol{\varphi}_{s,a,s''}^\top + \boldsymbol{\varphi}_{s,a,s''} \boldsymbol{\varphi}_{s,a,s'}^\top + \boldsymbol{\varphi}_{s,a,s''} \boldsymbol{\varphi}_{s,a,s''}^\top \right) \\
&\quad + 2 \sum_{s' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s' \mid s,a) \widetilde{V}_{h+1}^k(s') \\
&\qquad \cdot \left(\sum_{s'' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s,a) \boldsymbol{\varphi}_{s,a,s''} \right) \left(\sum_{s'' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}}(s'' \mid s,a) \boldsymbol{\varphi}_{s,a,s''} \right)^\top.
\end{aligned}
$$

Then, the prediction error can be bounded as follows:

$$
\begin{aligned}
|\Delta_h^k(s,a)| &= |F(\boldsymbol{\theta}_h^*) - F(\widetilde{\boldsymbol{\theta}}_h^k)| \\
&\le \left| \nabla F(\widetilde{\boldsymbol{\theta}}_h^k)^\top (\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*) \right| + \frac{1}{2} \left| (\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*)^\top \nabla^2 F(\bar{\boldsymbol{\theta}})(\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*) \right|.
\end{aligned}
\tag{77}
$$

For the first term in Eq. (77),

$$
\begin{aligned}
\left| \nabla F(\widetilde{\boldsymbol{\theta}}_h^k)^\top (\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*) \right| &= \left| \sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s,a) \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)^\top (\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*) \widetilde{V}_{h+1}^k(s') \right| \\
&\le H \sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s,a) \left\| \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k) \right\|_{\mathbf{B}_{k,h}^{-1}} \left\| \widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^* \right\|_{\mathbf{B}_{k,h}} \\
&\le H\beta_k(\delta) \sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s,a) \left\| \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k) \right\|_{\mathbf{B}_{k,h}^{-1}},
\end{aligned}
\tag{78}
$$

where in the first inequality we use $\widetilde{V}_{h+1}^k(s') \le H$ and Cauchy-Scharwz inequality, and the second inequality follows by the concentration result of Lemma 12.

53

For the second term in Eq. (77), since $0 \leq \widetilde{V}_{h+1}^k(s') \leq H$,

$$
\begin{aligned}
&\left| (\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*)^\top \nabla^2 F(\bar{\boldsymbol{\theta}})(\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*) \right| \\
&\leq H \sum_{s' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s' \mid s, a) \left((\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*)^\top \boldsymbol{\varphi}_{s,a,s'} \right)^2 \\
&\quad + H \sum_{s' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s' \mid s, a) \\
&\qquad \cdot \sum_{s'' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s'' \mid s, a) \left| (\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*)^\top \left(\boldsymbol{\varphi}_{s,a,s'} \boldsymbol{\varphi}_{s,a,s''}^\top + \boldsymbol{\varphi}_{s,a,s''} \boldsymbol{\varphi}_{s,a,s'}^\top \right) (\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*) \right| \\
&\quad + H \sum_{s' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s' \mid s, a) \sum_{s'' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s'' \mid s, a) \left((\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*)^\top \boldsymbol{\varphi}_{s,a,s''} \right)^2 \\
&\quad + 2H \left((\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*)^\top \left(\sum_{s'' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s'' \mid s, a) \boldsymbol{\varphi}_{s,a,s''} \right) \right)^2 \\
&\leq H \sum_{s' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s' \mid s, a) \|\boldsymbol{\varphi}_{s,a,s'}\|_{\mathbf{B}_{k,h}^{-1}}^2 \left\| \widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^* \right\|_{\mathbf{B}_{k,h}}^2 \\
&\quad + H \sum_{s' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s' \mid s, a) \\
&\qquad \cdot \sum_{s'' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s'' \mid s, a) \left| (\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*)^\top \left(\boldsymbol{\varphi}_{s,a,s'} \boldsymbol{\varphi}_{s,a,s'}^\top + \boldsymbol{\varphi}_{s,a,s''} \boldsymbol{\varphi}_{s,a,s''}^\top \right) (\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*) \right| \\
&\quad + H \sum_{s'' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s'' \mid s, a) \|\boldsymbol{\varphi}_{s,a,s''}\|_{\mathbf{B}_{k,h}^{-1}}^2 \left\| \widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^* \right\|_{\mathbf{B}_{k,h}}^2 \\
&\quad + 2H \left(\sum_{s'' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s'' \mid s, a) \|\boldsymbol{\varphi}_{s,a,s''}\|_{\mathbf{B}_{k,h}^{-1}} \left\| \widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^* \right\|_{\mathbf{B}_{k,h}} \right)^2,
\end{aligned}
\tag{79}
$$

where for the second inequality we use Cauchy-Schwarz inequality, $\mathbf{x}\mathbf{x}^\top + \mathbf{y}\mathbf{y}^\top \succeq \mathbf{x}\mathbf{y}^\top + \mathbf{y}\mathbf{x}^\top$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, and triangle inequality. Note that

$$
\begin{aligned}
&H \sum_{s' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s' \mid s, a) \\
&\quad \cdot \sum_{s'' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s'' \mid s, a) \left| (\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*)^\top \left(\boldsymbol{\varphi}_{s,a,s'} \boldsymbol{\varphi}_{s,a,s'}^\top + \boldsymbol{\varphi}_{s,a,s''} \boldsymbol{\varphi}_{s,a,s''}^\top \right) (\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*) \right| \\
&= H \sum_{s' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s' \mid s, a) \left((\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*)^\top \boldsymbol{\varphi}_{s,a,s'} \right)^2 \\
&\quad + H \sum_{s'' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s'' \mid s, a) \left((\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*)^\top \boldsymbol{\varphi}_{s,a,s''} \right)^2 \\
&\leq 2H \sum_{s' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s' \mid s, a) \|\boldsymbol{\varphi}_{s,a,s'}\|_{\mathbf{B}_{k,h}^{-1}}^2 \left\| \widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^* \right\|_{\mathbf{B}_{k,h}}^2 .
\end{aligned}
\tag{80}
$$

54

By substituting Eq. (80) into Eq. (79) we have

$$
\left| (\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*)^\top \nabla^2 F(\bar{\boldsymbol{\theta}}) (\widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^*) \right|
$$

$$
\leq 4H \sum_{s' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s' \mid s, a) \| \boldsymbol{\varphi}_{s,a,s'} \|_{\mathbf{B}_{k,h}^{-1}}^2 \left\| \widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^* \right\|_{\mathbf{B}_{k,h}}^2
$$

$$
+ 2H \left(\sum_{s'' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s'' \mid s, a) \| \boldsymbol{\varphi}_{s,a,s''} \|_{\mathbf{B}_{k,h}^{-1}} \left\| \widetilde{\boldsymbol{\theta}}_h^k - \boldsymbol{\theta}_h^* \right\|_{\mathbf{B}_{k,h}} \right)^2
$$

$$
\leq 4H \beta_k^2 \max_{s' \in \mathcal{S}_{s,a}} \| \boldsymbol{\varphi}_{s,a,s'} \|_{\mathbf{B}_{k,h}^{-1}}^2 + 2H \left(\beta_k \max_{s' \in \mathcal{S}_{s,a}} \| \boldsymbol{\varphi}_{s,a,s'} \|_{\mathbf{B}_{k,h}^{-1}} \right)^2
$$

$$
\leq 6H \beta_k^2 \max_{s' \in \mathcal{S}_{s,a}} \| \boldsymbol{\varphi}_{s,a,s'} \|_{\mathbf{B}_{k,h}^{-1}}^2 , \tag{81}
$$

where for the second inequality follows by Lemma 12 and $\sum_{s' \in \mathcal{S}_{s,a}} P_{\bar{\boldsymbol{\theta}}}(s' \mid s, a) = 1$. Combining the results of Eq. (78) and Eq. (81) and , we conclude the proof. $\qquad\square$

D.3 Good Events with High Probability

In this section, we introduce the good events used to prove Theorem 2 and show that the good events happen with high probability.

Lemma 17 (Good event probability). *For any $K \in \mathbb{N}$ and $\delta \in (0, 1)$, the good event $\mathfrak{G}(K, \delta')$ holds with probability at least $1 - \delta$ where $\delta' = \delta/(2KH)$.*

Proof of Lemma 17. For any $\delta' \in (0, 1)$, we have

$$
\mathfrak{G}(K, \delta') = \bigcap_{k \leq K} \bigcap_{h \leq H} \mathfrak{G}_{k,h}(\delta') = \bigcap_{k \leq K} \bigcap_{h \leq H} \left\{ \mathfrak{G}_{k,h}^{\Delta}(\delta') \cap \mathfrak{G}_{k,h}^{\boldsymbol{\xi}}(\delta') \right\} .
$$

On the other hand, for any $(k, h) \in [K] \times [H]$, by Lemma 30 $\mathfrak{G}_{k,h}^{\boldsymbol{\xi}}(\delta')$ holds with probability at least $1 - \delta'$. Then, for $\delta' = \delta/(2KH)$ by taking union bound, we have the desired result as follows:

$$
\mathbb{P}(\mathfrak{G}(K, \delta')) \geq (1 - \delta')^{2KH} \geq 1 - 2KH\delta' = 1 - \delta .
$$

$\qquad\square$

D.4 Stochastic Optimism

Lemma 18 (Stochastic optimism). *For any δ with $0 < \delta < \Phi(-1)/2$, let $\sigma_k = H\beta_k(\delta)$. If we take multiple sample size $M = \lceil 1 - \frac{\log(H\mathcal{U})}{\log \Phi(1)} \rceil$, then for any $k \in [K]$, we have*

$$
\mathbb{P} \left((\widetilde{V}_1^k - V_1^*)(s_1^k) \geq 0 \mid s_1^k, \mathcal{F}_k \right) \geq \Phi(-1)/2 .
$$

Proof of Lemma 18. First, we introduce the following lemmas.

Lemma 19. *Let $\delta \in (0, 1)$ be given. For any $(k, h) \in [K] \times [H]$, let $\sigma_k = H\beta_k(\delta)$. If we define the event $\mathfrak{G}_{k,h}^{\Delta}(\delta)$ as*

$$
\mathfrak{G}_{k,h}^{\Delta}(\delta) := \left\{ |\Delta_h^k(s,a)| \leq H\beta_k(\delta) \sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s, a) \left\| \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k) \right\|_{\mathbf{B}_{k,h}^{-1}} \right.
$$

$$
\left. + 3H\beta_k(\delta)^2 \max_{s' \in \mathcal{S}_{s,a}} \| \boldsymbol{\varphi}_{s,a,s'} \|_{\mathbf{B}_{k,h}^{-1}}^2 \right\} ,
$$

then conditioned on $\mathfrak{G}_{k,h}^{\Delta}(\delta)$, for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$
\mathbb{P} \left(-\iota_h^k(s, a) \geq 0 \mid \mathfrak{G}_{k,h}^{\Delta}(\delta) \right) \geq 1 - \Phi(1)^M .
$$

55

Lemma 20. *Let $\delta \in (0,1)$ be given. For any $(h,k) \in [H] \times [K]$, let $\sigma_k = H\beta_k(\delta)$. If we take multiple sample size $M = \lceil 1 - \frac{\log(H\mathcal{U})}{\log \Phi(1)} \rceil$, then conditioned on the event $\mathfrak{G}_k^{\Delta}(\delta) := \cap_{h \in [H]} \mathfrak{G}_{k,h}^{\Delta}(\delta)$, we have*

$$\mathbb{P}\left(-\iota_h^k(s_h, a_h) \geq 0, \forall h \in [H] \mid \mathfrak{G}_k^{\Delta}(\delta)\right) \geq \Phi(-1).$$

Based on the result of Lemma 20, using the same argument as in Lemma 6 we obtain the desired result. $\qquad\square$

D.4.1 Proof of Lemma 19

Proof of Lemma 19. Recall the definition of Bellman error (Definition 1), we have

$$
\begin{aligned}
&- \iota_h^k(s,a)\\
&= \widetilde{Q}_h^k(s,a) - \left(r(s,a) + P_h \widetilde{V}_{h+1}^k(s,a)\right)\\
&= \min\left\{ r(s,a) + \sum_{s' \in \mathcal{S}_{s,a}} P_{\widehat{\boldsymbol{\theta}}_h^k}(s' \mid s,a)\widetilde{V}_{h+1}^k(s') + \nu_{k,h}^{\mathrm{rand}}(s,a) \right\} - \left(r(s,a) + P_h \widetilde{V}_{h+1}^k(s,a)\right)\\
&\geq \min\left\{ \sum_{s' \in \mathcal{S}_{s,a}} P_{\widehat{\boldsymbol{\theta}}_h^k}(s' \mid s,a)\widetilde{V}_{h+1}^k(s') - P_h \widetilde{V}_{h+1}^k(s,a) + \nu_{k,h}^{\mathrm{rand}}(s,a), 0 \right\}.
\end{aligned}
$$

Then, it is enough to show that

$$\sum_{s' \in \mathcal{S}_{s,a}} P_{\widehat{\boldsymbol{\theta}}_h^k}(s' \mid s,a)\widetilde{V}_{h+1}^k(s') - P_h \widetilde{V}_{h+1}^k(s,a) + \nu_{k,h}^{\mathrm{rand}}(s,a) \geq 0$$

at least with constant probability. On the other hand, under the event $\mathfrak{G}_{k,h}^{\Delta}(\delta)$, by Lemma 16 we have

$$
\begin{aligned}
&\sum_{s' \in \mathcal{S}_{s,a}} P_{\widehat{\boldsymbol{\theta}}_h^k}(s' \mid s,a)\widetilde{V}_{h+1}^k(s') - P_h \widetilde{V}_{h+1}^k(s,a) + \nu_{k,h}^{\mathrm{rand}}(s,a)\\
&= \Delta_h^k(s,a) + \nu_{k,h}^{\mathrm{rand}}(s,a)\\
&\geq -H\beta_k(\delta) \sum_{s' \in \mathcal{S}_{s,a}} P_{\widehat{\boldsymbol{\theta}}_h^k}(s' \mid s,a)\left\|\bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)\right\|_{\mathbf{B}_{k,h}^{-1}}\\
&\quad - 3H\beta_k(\delta)^2 \max_{s' \in \mathcal{S}_{s,a}} \|\boldsymbol{\varphi}_{s,a,s'}\|_{\mathbf{B}_{k,h}^{-1}}^2 + \nu_{k,h}^{\mathrm{rand}}(s,a)\\
&= \sum_{s' \in \mathcal{S}_{s,a}} P_{\widehat{\boldsymbol{\theta}}_h^k}(s' \mid s,a)\bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)^\top \boldsymbol{\xi}_{k,h}^{s'} - H\beta_k(\delta) \sum_{s' \in \mathcal{S}_{s,a}} P_{\widehat{\boldsymbol{\theta}}_h^k}(s' \mid s,a)\left\|\bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)\right\|_{\mathbf{B}_{k,h}^{-1}}.
\end{aligned}
$$

Note that since $\boldsymbol{\xi}_{k,h}^{(m)} \sim \mathcal{N}(\mathbf{0}, \sigma_k^2 \mathbf{B}_{k,h}^{-1})$, it follows that

$$\bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)^\top \boldsymbol{\xi}_{k,h}^{(m)} \sim \mathcal{N}\left(0, \sigma_k^2 \left\|\bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)\right\|_{\mathbf{B}_{k,h}^{-1}}^2\right), \quad \forall m \in [M].$$

Therefore, by setting $\sigma_k = H\beta_k(\delta)$, we have for $m \in [M]$ and $s' \in \mathcal{S}_{s,a}$,

$$\mathbb{P}\left(\bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)^\top \boldsymbol{\xi}_{k,h}^{(m)} \geq H\beta_k(\delta)\left\|\bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)\right\|_{\mathbf{B}_{k,h}^{-1}}\right) = \Phi(-1).$$

56

Recall that $\boldsymbol{\xi}_{k,h}^{s'} := \boldsymbol{\xi}_{k,h}^{m(s')}$ where $m(s') := \operatorname{argmax}_{m \in [M]} \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)^\top \boldsymbol{\xi}_{k,h}^{(m)}$. Then, we can deduce

$$\mathbb{P}\left(\bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)^\top \boldsymbol{\xi}_{k,h}^{s'} \geq H\beta_k(\delta) \left\|\bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)\right\|_{\mathbf{B}_{k,h}^{-1}}\right)$$

$$= \mathbb{P}\left(\max_{m \in [M]} \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)^\top \boldsymbol{\xi}_{k,h}^{(m)} \geq H\beta_k(\delta) \left\|\bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)\right\|_{\mathbf{B}_{k,h}^{-1}}\right)$$

$$= 1 - \mathbb{P}\left(\bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)^\top \boldsymbol{\xi}_{k,h}^{(m)} < H\beta_k(\delta) \left\|\bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)\right\|_{\mathbf{B}_{k,h}^{-1}}, \forall m \in [M]\right)$$

$$\geq 1 - (1 - \Phi(-1))^M$$

$$= 1 - \Phi(1)^M. \tag{82}$$

Consequently, we arrive at the conclusion as follows:

$$\mathbb{P}(-\iota_h^k(s,a) \geq 0 \mid \mathfrak{G}_{k,h}^\Delta(\delta))$$

$$\geq \mathbb{P}\left(\sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s, a)\widetilde{V}_{h+1}^k(s') - P_h\widetilde{V}_{h+1}^k(s,a) + \nu_{k,h}^{\text{rand}}(s,a) \geq 0 \mid \mathfrak{G}_{k,h}^\Delta(\delta)\right)$$

$$\geq \mathbb{P}\left(\sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s, a)\bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)^\top \boldsymbol{\xi}_{k,h}^{s'} \right. \tag{83}$$

$$\left. \geq H\beta_k(\delta) \sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s, a) \left\|\bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)\right\|_{\mathbf{B}_{k,h}^{-1}} \mid \mathfrak{G}_{k,h}^\Delta(\delta)\right)$$

$$\geq \mathbb{P}\left(\bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)^\top \boldsymbol{\xi}_{k,h}^{s'} \geq H\beta_k(\delta) \left\|\bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)\right\|_{\mathbf{B}_{k,h}^{-1}}, \forall s' \in \mathcal{S}_{s,a} \mid \mathfrak{G}_{k,h}^\Delta(\delta)\right)$$

$$= 1 - \mathbb{P}\left(\exists s' \in \mathcal{S}_{s,a} \text{ s.t. } \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)^\top \boldsymbol{\xi}_{k,h}^{s'} < H\beta_k(\delta) \left\|\bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)\right\|_{\mathbf{B}_{k,h}^{-1}} \mid \mathfrak{G}_{k,h}^\Delta(\delta)\right)$$

$$\geq 1 - \mathcal{U}\mathbb{P}\left(\bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)^\top \boldsymbol{\xi}_{k,h}^{s'} < H\beta_k(\delta) \left\|\bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)\right\|_{\mathbf{B}_{k,h}^{-1}} \mid \mathfrak{G}_{k,h}^\Delta(\delta)\right) \tag{84}$$

$$\geq 1 - \mathcal{U}\Phi(1)^M, \tag{85}$$

where (84) comes from the fact that $\max_{s,a} |\mathcal{S}_{s,a}| = \mathcal{U}$ and the union bound, and (85) follows by (82). □

D.4.2 Proof of Lemma 20

Proof of Lemma 20. It holds

$$\mathbb{P}\left(-\iota_h^k(s_h, a_h) \geq 0, \forall h \in [H]\right) = 1 - \mathbb{P}\left(\exists h \in [H] \text{ s.t. } -\iota_h^k(s_h, a_h) < 0\right)$$

$$\geq 1 - H\mathbb{P}\left(-\iota_h^k(s_h, a_h) < 0\right)$$

$$\geq 1 - H\mathcal{U}\Phi(1)^M$$

$$\geq \Phi(-1)$$

where the first inequality uses the Bernoulli's inequality, the second inequality follows by Lemma 19, and the last inequality holds due to the choice of $M = \lceil 1 - \frac{\log(\mathcal{U}H)}{\log \Phi(1)} \rceil$. □

D.5 Bound on Estimation Part

In this section, we provide the upper bound on the estimation part of the regret: $\sum_{k=1}^K (\widetilde{V}_1^k - V_1^*)(s_1^k)$.

Lemma 21 (Bound on estimation). *For any $\delta \in (0,1)$, if $\lambda = \mathcal{O}(L_\varphi^2 d \log \mathcal{U})$, then with probability at least $1 - \delta/2$, we have*

$$\sum_{k=1}^K (\widetilde{V}_1^k - V_1^{\pi^k})(s_1^k) = \widetilde{\mathcal{O}}\left(d^{3/2}H^{3/2}\sqrt{T} + \kappa^{-1}d^2H^2\right).$$

57

Proof of Lemma 21. With the same argument in Lemma 10, we have

$$(\widetilde{V}_1^k - V_1^{\pi^k})(s_1^k) = \sum_{h=1}^{H} -\iota_h^k(s_h^k, a_h^k) + \sum_{h=1}^{H} \dot{\zeta}_h^k, \tag{86}$$

where $\dot{\zeta}_h^k := P_h(\widetilde{V}_{h+1}^k - V_{h+1}^{\pi^k})(s_h^k, a_h^k) - (\widetilde{V}_{h+1}^k - V_{h+1}^{\pi^k})(s_{h+1}^k)$. Note that

$$
\begin{aligned}
-\iota_h^k(s_h^k, a_h^k) &= \widetilde{Q}_h^k(s_h^k, a_h^k) - \left(r(s_h^k, a_h^k) + P_h \widetilde{V}_{h+1}^k(s_h^k, a_h^k) \right) \\
&\leq \sum_{s' \in \mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s_h^k, a_h^k) \widetilde{V}_{h+1}^k(s') - P_h \widetilde{V}_{h+1}^k(s_h^k, a_h^k) + \nu_{k,h}^{\mathrm{rand}}(s_h^k, a_h^k) \\
&\leq \left| \Delta_h^k(s_h^k, a_h^k) \right| + \nu_{k,h}^{\mathrm{rand}}(s_h^k, a_h^k) \\
&\leq H\beta_k \sum_{s' \in \mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s_h^k, a_h^k) \left\| \bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^k) \right\|_{\mathbf{B}_{k,h}^{-1}} + 3H\beta_k^2 \max_{s' \in \mathcal{S}_{k,h}} \| \boldsymbol{\varphi}_{k,h,s'} \|_{\mathbf{B}_{k,h}^{-1}}^2 \\
&\quad + \nu_{k,h}^{\mathrm{rand}}(s_h^k, a_h^k), \tag{87}
\end{aligned}
$$

where the last inequality follows by Lemma 16. Now we introduce the following lemma.

Lemma 22. *For any $(k,h) \in [K] \times [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, it holds*

$$
\begin{aligned}
&\sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s, a) \left\| \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k) \right\|_{\mathbf{B}_{k,h}^{-1}} \\
&\leq \sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s, a) \left\| \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}} + \frac{16\eta L_{\varphi}}{\sqrt{\lambda}} \max_{s' \in \mathcal{S}_{s,a}} \left\| \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2.
\end{aligned}
$$

By plugging the result of Lemma 22 into Eq. (87), we have

$$
\begin{aligned}
&-\iota_h^k(s_h^k, a_h^k) \\
&\leq H\beta_k \sum_{s' \in \mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s_h^k, a_h^k) \left\| \bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}} \\
&\quad + H\beta_k \frac{16\eta L_{\varphi}}{\sqrt{\lambda}} \max_{s' \in \mathcal{S}_{k,h}} \left\| \bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2 + 3H\beta_k^2 \max_{s' \in \mathcal{S}_{k,h}} \| \boldsymbol{\varphi}_{k,h,s'} \|_{\mathbf{B}_{k,h}^{-1}}^2 + \nu_{k,h}^{\mathrm{rand}}(s_h^k, a_h^k) \\
&\leq H\beta_k \sum_{s' \in \mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s_h^k, a_h^k) \left\| \bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}} \\
&\quad + \sum_{s' \in \mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s_h^k, a_h^k) \bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^k)^{\top} \boldsymbol{\xi}_{k,h}^{s'} \\
&\quad + H\beta_k \frac{16\eta L_{\varphi}}{\sqrt{\lambda}} \max_{s' \in \mathcal{S}_{k,h}} \left\| \bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2 + 6H\beta_k^2 \max_{s' \in \mathcal{S}_{k,h}} \| \boldsymbol{\varphi}_{k,h,s'} \|_{\mathbf{B}_{k,h}^{-1}}^2.
\end{aligned}
$$

By letting us denote

$$\Upsilon_h^k(s,a) := H\beta_k \frac{16\eta L_{\varphi}}{\sqrt{\lambda}} \max_{s' \in \mathcal{S}_{s,a}} \left\| \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2 + 6H\beta_k^2 \max_{s' \in \mathcal{S}_{s,a}} \| \boldsymbol{\varphi}_{s,a,s'} \|_{\mathbf{B}_{k,h}^{-1}}^2, \tag{88}$$

58

and summing over all episodes, we have

$$\sum_{k=1}^{K}(\widetilde{V}_1^k - V_1^{\pi^k})(s_1^k) = \sum_{k=1}^{K}\sum_{h=1}^{H} -\iota_h^k(s_h^k, a_h^k) + \sum_{k=1}^{K}\sum_{h=1}^{H} \dot{\zeta}_h^k$$

$$\leq \underbrace{H\beta_K \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s'\in\mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s_h^k, a_h^k) \left\|\bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1})\right\|_{\mathbf{B}_{k,h}^{-1}}}_{\text{(i)}}$$

$$+ \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s'\in\mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s_h^k, a_h^k)\bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^k)^\top \boldsymbol{\xi}_{k,h}^{s'}}_{\text{(ii)}}$$

$$+ \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} \Upsilon_h^k(s_h^k, a_h^k)}_{\text{(iii)}} + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} \dot{\zeta}_h^k}_{\text{(iv)}}. \tag{89}$$

Note that $\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s'\in\mathcal{S}_{k,h}}$ is hereafter abbreviated as $\sum_{k,h,s'}$.

For term (i), we have

$$\sum_{k,h,s'} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s_h^k, a_h^k) \left\|\bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1})\right\|_{\mathbf{B}_{k,h}^{-1}}$$

$$\leq \sqrt{\sum_{k,h,s'} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s_h^k, a_h^k)}\sqrt{\sum_{k,h,s'} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s_h^k, a_h^k) \left\|\bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1})\right\|_{\mathbf{B}_{k,h}^{-1}}^2}$$

$$= \sqrt{T}\sqrt{\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{s'\in\mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s_h^k, a_h^k) \left\|\bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1})\right\|_{\mathbf{B}_{k,h}^{-1}}^2}$$

$$\leq \sqrt{T}\sqrt{2Hd\log\left(1 + \frac{KUL_\varphi^2}{d\lambda}\right)}, \tag{90}$$

where the last inequality follows by the following lemma:

Lemma 23. *For each $h \in [H]$, if $\lambda \geq L_\varphi^2$, then we have*

$$\sum_{k=1}^{K}\sum_{s'\in\mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s_h^k, a_h^k) \left\|\bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1})\right\|_{\mathbf{B}_{k,h}^{-1}}^2 \leq 2d\log\left(1 + \frac{KUL_\varphi^2}{d\lambda}\right).$$

Then, term (i) can be bounded as follows:

$$\text{(i)} = H\beta_K \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s'\in\mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s_h^k, a_h^k) \left\|\bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1})\right\|_{\mathbf{B}_{k,h}^{-1}}$$

$$\leq H\beta_K\sqrt{T}\sqrt{2Hd\log\left(1 + \frac{KUL_\varphi^2}{d\lambda}\right)}$$

$$= \widetilde{\mathcal{O}}(dH^{3/2}\sqrt{T}). \tag{91}$$

For term (ii), we introduce the following lemma:

59

Lemma 24. *Let $\delta \in (0,1)$ be given. For any $(k,h) \in [K] \times [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, with probability at least $1 - \delta$, it holds*

$$
\sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s,a) \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k)^\top \boldsymbol{\xi}_{k,h}^{s'}
$$

$$
\leq \gamma_k(\delta) \bigg(\sum_{s' \in \mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s_h^k, a_h^k) \left\| \bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}
$$

$$
+ \frac{16\eta L_\varphi}{\sqrt{\lambda}} \max_{s' \in \mathcal{S}_{k,h}} \left\| \bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2 \bigg),
$$

where $\gamma_k(\delta) := C_{\boldsymbol{\xi}} \sigma_k \sqrt{d \log(Md/\delta)}$ for an absolute constant $C_{\boldsymbol{\xi}} > 0$.

By Lemma 24, we have

$$
\sum_{k,h,s'} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s_h^k, a_h^k) \bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^k)^\top \boldsymbol{\xi}_{k,h}^{s'}
$$

$$
\leq \gamma_K(\delta) \bigg(\sum_{k,h,s'} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s_h^k, a_h^k) \left\| \bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}
$$

$$
+ \frac{16\eta L_\varphi}{\sqrt{\lambda}} \sum_{k=1}^{K} \sum_{h=1}^{H} \max_{s' \in \mathcal{S}_{k,h}} \left\| \bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2 \bigg)
$$

$$
\leq \gamma_K(\delta) \bigg(\sqrt{T} \sqrt{2Hd \log \left(1 + \frac{K\mathcal{U}L_\varphi^2}{d\lambda} \right)} + \frac{16\eta L_\varphi}{\sqrt{\lambda}} \sum_{k=1}^{K} \sum_{h=1}^{H} \max_{s' \in \mathcal{S}_{k,h}} \left\| \bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2 \bigg),
\tag{92}
$$

where the last inequality follows by Eq. (90). Note that

$$
\sum_{k=1}^{K} \sum_{h=1}^{H} \max_{s' \in \mathcal{S}_{k,h}} \left\| \bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2
$$

$$
\leq \sum_{k=1}^{K} \sum_{h=1}^{H} \max_{s' \in \mathcal{S}_{k,h}} \left\| \bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{A}_{k,h}^{-1}}^2
$$

$$
= \sum_{k=1}^{K} \sum_{h=1}^{H} \max_{s' \in \mathcal{S}_{k,h}} \left\| \boldsymbol{\varphi}_{k,h,s'} - \sum_{\widetilde{s} \in \mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(\widetilde{s} \mid s_h^k, a_h^k) \boldsymbol{\varphi}_{k,h,\widetilde{s}} \right\|_{\mathbf{A}_{k,h}^{-1}}^2
$$

$$
\leq \sum_{k=1}^{K} \sum_{h=1}^{H} \max_{s' \in \mathcal{S}_{k,h}} \left(2\|\boldsymbol{\varphi}_{k,h,s'}\|_{\mathbf{A}_{k,h}^{-1}}^2 + 2 \left\| \sum_{\widetilde{s} \in \mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(\widetilde{s} \mid s_h^k, a_h^k) \boldsymbol{\varphi}_{k,h,\widetilde{s}} \right\|_{\mathbf{A}_{k,h}^{-1}}^2 \right)
$$

$$
\leq 2 \sum_{k=1}^{K} \sum_{h=1}^{H} \max_{s' \in \mathcal{S}_{k,h}} \|\boldsymbol{\varphi}_{k,h,s'}\|_{\mathbf{A}_{k,h}^{-1}}^2 + 2 \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{\widetilde{s} \in \mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(\widetilde{s} \mid s_h^k, a_h^k) \|\boldsymbol{\varphi}_{k,h,\widetilde{s}}\|_{\mathbf{A}_{k,h}^{-1}}^2
$$

$$
\leq 4 \sum_{k=1}^{K} \sum_{h=1}^{H} \max_{s' \in \mathcal{S}_{k,h}} \|\boldsymbol{\varphi}_{k,h,s'}\|_{\mathbf{A}_{k,h}^{-1}}^2
$$

$$
\leq 16\kappa^{-1} dH \log \left(1 + \frac{K\mathcal{U}L_\varphi^2}{d\lambda} \right),
\tag{93}
$$

where the first inequality holds since $\mathbf{B}_{k,h}^{-1} \preceq \mathbf{A}_{k,h}^{-1}$, the second inequality follows from $(x + y)^2 \leq 2x^2 + 2y^2$, and the third inequality uses the triangle inequality, and the fourth inequality uses $\sum_{\widetilde{s} \in \mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(\widetilde{s} \mid s_h^k, a_h^k) = 1$, and the last inequality follows by Lemma 3. By substituting

60

Eq. (93) into Eq. (92), we have

$$\text{(ii)} \le \gamma_K(\delta)\left(\sqrt{T}\sqrt{2Hd\log\left(1 + K\mathcal{U}L_\varphi^2/(d\lambda)\right)} + \frac{256\eta L_\varphi}{\sqrt{\lambda}}\kappa^{-1}dH\log\left(1 + K\mathcal{U}L_\varphi^2/(d\lambda)\right) \right)$$

$$= \widetilde{\mathcal{O}}(d^{3/2}H^{3/2}\sqrt{T} + \kappa^{-1}d^{3/2}H^2). \tag{94}$$

For term (iii),

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\Upsilon_h^k(s_h^k, a_h^k)$$

$$= \sum_{k=1}^{K}\sum_{h=1}^{H}\left(H\beta_k\frac{16\eta L_\varphi}{\sqrt{\lambda}}\max_{s'\in\mathcal{S}_{k,h}}\left\|\bar\varphi_{k,h,s'}(\widetilde{\boldsymbol\theta}_h^{k+1})\right\|_{\mathbf{B}_{k,h}^{-1}}^2 + 6H\beta_k^2\max_{s'\in\mathcal{S}_{k,h}}\|\varphi_{k,h,s'}\|_{\mathbf{B}_{k,h}^{-1}}^2 \right)$$

$$\le H\beta_K\frac{16\eta L_\varphi}{\sqrt{\lambda}}\sum_{k=1}^{K}\sum_{h=1}^{H}\max_{s'\in\mathcal{S}_{k,h}}\left\|\bar\varphi_{k,h,s'}(\widetilde{\boldsymbol\theta}_h^{k+1})\right\|_{\mathbf{B}_{k,h}^{-1}}^2 + 6H\beta_K^2\sum_{k=1}^{K}\sum_{h=1}^{H}\max_{s'\in\mathcal{S}_{k,h}}\|\varphi_{k,h,s'}\|_{\mathbf{A}_{k,h}^{-1}}^2$$

$$\le \beta_K\frac{256\eta L_\varphi}{\sqrt{\lambda}}\kappa^{-1}dH^2\log\left(1 + K\mathcal{U}L_\varphi^2/(d\lambda)\right) + 24\kappa^{-1}dH^2\beta_K^2\log\left(1 + K\mathcal{U}L_\varphi^2/(d\lambda)\right)$$

$$= \widetilde{\mathcal{O}}(\kappa^{-1}d^2H^2), \tag{95}$$

where for the second inequality we use the same argument used to derive Eq. (93) and Lemma 3.

For term (iv), since we have $|\dot\zeta_h^k| \le 2H$ and $\mathbb{E}[\dot\zeta_h^k \mid \mathcal{F}_{k,h}] = 0$, which means $\{\dot\zeta_h^k \mid \mathcal{F}_{k,h}\}_{k,h}$ is a martingale difference sequence for any $k \in [K]$ and $h \in [H]$. Hence, by applying the Azuma-Hoeffding inequality with probability at least $1 - \delta/4$, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\dot\zeta_h^k \le 2H\sqrt{2KH\log(4/\delta)}. \tag{96}$$

Combining all results of Eq. (91), (94), (95), and (96), we have the desired result.

$$\sum_{k=1}^{K}(\widetilde{V}_1^k - V_1^{\pi^k})(s_1^k) = \widetilde{\mathcal{O}}(dH^{3/2}\sqrt{T} + d^{3/2}H^{3/2}\sqrt{T} + \kappa^{-1}d^{3/2}H^2 + \kappa^{-1}d^2H^2 + H\sqrt{T})$$

$$= \widetilde{\mathcal{O}}(d^{3/2}H^{3/2}\sqrt{T} + \kappa^{-1}d^2H^2).$$

$\square$

In the following, we provide the proof of the lemmas used in Lemma 21.

61

D.5.1 Proof of Lemma 22

Proof of Lemma 22. Note that

$$
\sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s, a) \left\| \bar{\boldsymbol{\varphi}}_{s,a,s'}(\boldsymbol{\theta}_h^k) \right\|_{\mathbf{B}_{k,h}^{-1}}
$$

$$
\leq \sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s, a) \left\| \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}
$$

$$
+ \sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s, a) \left\| \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k) - \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}
$$

$$
\leq \sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s, a) \left\| \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}
$$

$$
+ \underbrace{\sum_{s' \in \mathcal{S}_{s,a}} \left(P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s, a) - P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s, a) \right) \left\| \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}}_{\text{(i)}}
$$

$$
+ \underbrace{\sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s, a) \left\| \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k) - \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}}_{\text{(ii)}},
$$

where the first inequality holds by triangle inequality.

For (i), we have

$$
\text{(i)} = \sum_{s' \in \mathcal{S}_{s,a}} \nabla P_{\boldsymbol{\vartheta}_h^k}(s' \mid s, a)^\top (\widetilde{\boldsymbol{\theta}}_h^k - \widetilde{\boldsymbol{\theta}}_h^{k+1}) \left\| \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}
$$

$$
\leq \sum_{s' \in \mathcal{S}_{s,a}} \left\| \nabla P_{\boldsymbol{\vartheta}_h^k}(s' \mid s, a) \right\|_{\mathbf{B}_{k,h}^{-1}} \left\| \widetilde{\boldsymbol{\theta}}_h^k - \widetilde{\boldsymbol{\theta}}_h^{k+1} \right\|_{\mathbf{B}_{k,h}} \left\| \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}} \tag{97}
$$

where in the equality we apply the mean value theorem with $\boldsymbol{\vartheta}_h^k = v \widetilde{\boldsymbol{\theta}}_h^k + (1-v) \widetilde{\boldsymbol{\theta}}_h^{k+1}$ for some $v \in [0,1]$, and the inequality follows by Cauchy-Schwarz inequality. Meanwhile, since we have

$$
P_{\boldsymbol{\vartheta}_h^k}(s' \mid s, a) \left(\bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) - \sum_{s'' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\vartheta}_h^k}(s'' \mid s, a) \bar{\boldsymbol{\varphi}}_{s,a,s''}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right) \tag{98}
$$

$$
= P_{\boldsymbol{\vartheta}_h^k}(s' \mid s, a) \left(\boldsymbol{\varphi}_{s,a,s'} - \sum_{\widetilde{s} \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(\widetilde{s} \mid s, a) \boldsymbol{\varphi}_{s,a,\widetilde{s}} \right.
$$

$$
\left. - \sum_{s'' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\vartheta}_h^k}(s'' \mid s, a) \left[\boldsymbol{\varphi}_{s,a,s''} - \sum_{\widetilde{s}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(\widetilde{s} \mid s, a) \boldsymbol{\varphi}_{s,a,\widetilde{s}} \right] \right)
$$

$$
= P_{\boldsymbol{\vartheta}_h^k}(s' \mid s, a) \boldsymbol{\varphi}_{s,a,s'} - P_{\boldsymbol{\vartheta}_h^k}(s' \mid s, a) \sum_{\widetilde{s} \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(\widetilde{s} \mid s, a) \boldsymbol{\varphi}_{s,a,\widetilde{s}}
$$

$$
- P_{\boldsymbol{\vartheta}_h^k}(s' \mid s, a) \sum_{s'' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\vartheta}_h^k}(s'' \mid s, a) \boldsymbol{\varphi}_{s,a,s''}
$$

$$
+ P_{\boldsymbol{\vartheta}_h^k}(s' \mid s, a) \underbrace{\left(\sum_{s'' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\vartheta}_h^k}(s'' \mid s, a) \right)}_{1} \sum_{\widetilde{s}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(\widetilde{s} \mid s, a) \boldsymbol{\varphi}_{s,a,\widetilde{s}}
$$

$$
= P_{\boldsymbol{\vartheta}_h^k}(s' \mid s, a) \boldsymbol{\varphi}_{s,a,s'} - P_{\boldsymbol{\vartheta}_h^k}(s' \mid s, a) \sum_{s'' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\vartheta}_h^k}(s'' \mid s, a) \boldsymbol{\varphi}_{s,a,s''}
$$

$$
= \nabla P_{\boldsymbol{\vartheta}_h^k}(s' \mid s, a),
$$

62

by substituting (98) into (97) we have

$$
\begin{aligned}
(i) \\
\leq \sum_{s' \in \mathcal{S}_{s,a}} &\Big\{ \Big\| P_{\vartheta_h^k}(s' \mid s,a) \bar{\varphi}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \\
&- P_{\vartheta_h^k}(s' \mid s,a) \sum_{s'' \in \mathcal{S}_{s,a}} P_{\vartheta_h^k}(s'' \mid s,a) \bar{\varphi}_{s,a,s''}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \Big\|_{\mathbf{B}_{k,h}^{-1}} \\
&\cdot \Big\| \widetilde{\boldsymbol{\theta}}_h^k - \widetilde{\boldsymbol{\theta}}_h^{k+1} \Big\|_{\mathbf{B}_{k,h}} \Big\| \bar{\varphi}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \Big\|_{\mathbf{B}_{k,h}^{-1}} \Big\} \\
\leq \sum_{s' \in \mathcal{S}_{s,a}} & P_{\vartheta_h^k}(s' \mid s,a) \Big\| \bar{\varphi}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \Big\|^2_{\mathbf{B}_{k,h}^{-1}} \Big\| \widetilde{\boldsymbol{\theta}}_h^k - \widetilde{\boldsymbol{\theta}}_h^{k+1} \Big\|_{\mathbf{B}_{k,h}} \\
&+ \Big(\sum_{s' \in \mathcal{S}_{s,a}} P_{\vartheta_h^k}(s' \mid s,a) \Big\| \bar{\varphi}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \Big\|_{\mathbf{B}_{k,h}^{-1}} \Big)^2 \Big\| \widetilde{\boldsymbol{\theta}}_h^k - \widetilde{\boldsymbol{\theta}}_h^{k+1} \Big\|_{\mathbf{B}_{k,h}} . \quad (99)
\end{aligned}
$$

Note that by Jensen's inequality, we have

$$
\begin{aligned}
\Big(\sum_{s' \in \mathcal{S}_{s,a}} P_{\vartheta_h^k}(s' \mid s,a) \Big\| \bar{\varphi}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \Big\|_{\mathbf{B}_{k,h}^{-1}} \Big)^2 &= \Big(\mathbb{E}_{s' \sim P_{\vartheta_h^k}(\cdot \mid s,a)} \Big[\Big\| \bar{\varphi}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \Big\|_{\mathbf{B}_{k,h}^{-1}} \Big] \Big)^2 \\
&\leq \mathbb{E}_{s' \sim P_{\vartheta_h^k}(\cdot \mid s,a)} \Big[\Big\| \bar{\varphi}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \Big\|^2_{\mathbf{B}_{k,h}^{-1}} \Big] \\
&= \sum_{s' \in \mathcal{S}_{s,a}} P_{\vartheta_h^k}(s' \mid s,a) \Big\| \bar{\varphi}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \Big\|^2_{\mathbf{B}_{k,h}^{-1}} . \\
& \hspace{3cm} (100)
\end{aligned}
$$

Also, we introduce the following lemma:

Lemma 25. *For any $k \in [K]$ and $h \in [H]$, the following holds:*

$$
\Big\| \widetilde{\boldsymbol{\theta}}_h^{k+1} - \widetilde{\boldsymbol{\theta}}_h^k \Big\|_{\mathbf{B}_{k,h}} \leq \frac{4\eta L_{\varphi}}{\sqrt{\lambda}} .
$$

Then, substituting (100) into (99), we have

$$
\begin{aligned}
(i) &\leq 2 \sum_{s' \in \mathcal{S}_{s,a}} P_{\vartheta_h^k}(s' \mid s,a) \Big\| \bar{\varphi}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \Big\|^2_{\mathbf{B}_{k,h}^{-1}} \Big\| \widetilde{\boldsymbol{\theta}}_h^k - \widetilde{\boldsymbol{\theta}}_h^{k+1} \Big\|_{\mathbf{B}_{k,h}} \\
&\leq \frac{8\eta L_{\varphi}}{\sqrt{\lambda}} \sum_{s' \in \mathcal{S}_{s,a}} P_{\vartheta_h^k}(s' \mid s,a) \Big\| \bar{\varphi}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \Big\|^2_{\mathbf{B}_{k,h}^{-1}} \\
&\leq \frac{8\eta L_{\varphi}}{\sqrt{\lambda}} \max_{s' \in \mathcal{S}_{s,a}} \Big\| \bar{\varphi}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \Big\|^2_{\mathbf{B}_{k,h}^{-1}} , \quad (101)
\end{aligned}
$$

where the second inequality comes from Lemma 25, and the last inequality holds due to $\sum_{s' \in \mathcal{S}_{s,a}} P_{\vartheta_h^k}(s' \mid s,a) = 1$.

63

For (ii), we have

$$
\begin{aligned}
\text{(ii)} &= \sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s, a) \left\| \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k) - \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}} \\
&= \sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s, a) \left\| \mathbb{E}_{\widetilde{s} \sim P_{\widetilde{\boldsymbol{\theta}}_h^k}(\cdot \mid s,a)} \left[\boldsymbol{\varphi}_{s,a,\widetilde{s}} \right] - \mathbb{E}_{\widetilde{s} \sim P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(\cdot \mid s,a)} \left[\boldsymbol{\varphi}_{s,a,\widetilde{s}} \right] \right\|_{\mathbf{B}_{k,h}^{-1}} \\
&= \left\| \sum_{\widetilde{s} \in \mathcal{S}_{s,a}} \left(P_{\widetilde{\boldsymbol{\theta}}_h^k}(\widetilde{s} \mid s, a) - P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(\widetilde{s} \mid s, a) \right) \boldsymbol{\varphi}_{s,a,\widetilde{s}} \right\|_{\mathbf{B}_{k,h}^{-1}} \\
&= \left\| \sum_{\widetilde{s} \in \mathcal{S}_{s,a}} \left(P_{\widetilde{\boldsymbol{\theta}}_h^k}(\widetilde{s} \mid s, a) - P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(\widetilde{s} \mid s, a) \right) \left(\boldsymbol{\varphi}_{s,a,\widetilde{s}} - \mathbb{E}_{s' \sim P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(\cdot \mid s,a)} \left[\boldsymbol{\varphi}_{s,a,s'} \right] \right) \right\|_{\mathbf{B}_{k,h}^{-1}} \\
&= \left\| \sum_{\widetilde{s} \in \mathcal{S}_{s,a}} \left(P_{\widetilde{\boldsymbol{\theta}}_h^k}(\widetilde{s} \mid s, a) - P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(\widetilde{s} \mid s, a) \right) \bar{\boldsymbol{\varphi}}_{s,a,\widetilde{s}}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}} \\
&\leq \frac{8 \eta L_{\boldsymbol{\varphi}}}{\sqrt{\lambda}} \max_{s' \in \mathcal{S}_{s,a}} \left\| \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2 , \tag{102}
\end{aligned}
$$

where the last inequality is obtained through the same argument as used to bound (i). Combining the results of Eq. (101) and Eq. (102), we have

$$
\begin{aligned}
&\sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s, a) \left\| \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^k) \right\|_{\mathbf{B}_{k,h}^{-1}} \\
&\leq \sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s, a) \left\| \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}} + \frac{16 \eta L_{\boldsymbol{\varphi}}}{\sqrt{\lambda}} \max_{s' \in \mathcal{S}_{s,a}} \left\| \bar{\boldsymbol{\varphi}}_{s,a,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2
\end{aligned}
$$

$\square$

D.5.2 Proof of Lemma 23

Proof of Lemma 23. Note that

$$
\begin{aligned}
\mathbf{B}_{k+1,h} &= \mathbf{B}_{k,h} + \sum_{s' \in \mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s_h^k, a_h^k) \bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1})^\top \\
&= \mathbf{B}_{k,h} + \sum_{s' \in \mathcal{S}_{k,h}} \widetilde{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \widetilde{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1})^\top ,
\end{aligned}
$$

where we define $\widetilde{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) := \sqrt{P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s_h^k, a_h^k)} \, \bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1})$. Then, we have

$$
\begin{aligned}
\det(\mathbf{B}_{k+1,h}) &= \det(\mathbf{B}_{k,h}) \det \left(\mathbf{I}_d + \mathbf{B}_{k,h}^{-\frac{1}{2}} \sum_{s' \in \mathcal{S}_{k,h}} \widetilde{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \widetilde{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1})^\top \mathbf{B}_{k,h}^{-\frac{1}{2}} \right) \\
&= \det(\mathbf{B}_{k,h}) \left(1 + \sum_{s' \in \mathcal{S}_{k,h}} \left\| \widetilde{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2 \right) \\
&= \det(\lambda \mathbf{I}_d) \prod_{k=1}^{K} \left(1 + \sum_{s' \in \mathcal{S}_{k,h}} \left\| \widetilde{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2 \right) .
\end{aligned}
$$

Taking the logarithm on both sides yields

$$
\log \frac{\det(\mathbf{B}_{k+1,h})}{\det(\lambda \mathbf{I}_d)} = \sum_{k=1}^{K} \log \left(1 + \sum_{s' \in \mathcal{S}_{k,h}} \left\| \widetilde{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2 \right) .
$$

64

On the other hand, since $\lambda \geq L_\varphi^2$,

$$\sum_{s' \in \mathcal{S}_{k,h}} \left\| \widetilde{\varphi}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2 \leq \sum_{s' \in \mathcal{S}_{k,h}} \frac{1}{\lambda} \left\| \widetilde{\varphi}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_2^2$$

$$= \sum_{s' \in \mathcal{S}_{k,h}} \frac{1}{\lambda} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s_h^k, a_h^k) \left\| \bar{\varphi}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_2^2$$

$$\leq \frac{L_\varphi^2}{\lambda} \sum_{s' \in \mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s_h^k, a_h^k)$$

$$\leq 1,$$

where the last inequality uses $\sum_{s' \in \mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s_h^k, a_h^k) = 1$. From the fact that $z \leq 2 \log(1 + z)$ for any $z \in [0, 1]$, it follows that

$$\sum_{k=1}^K \log \left(1 + \sum_{s' \in \mathcal{S}_{k,h}} \left\| \widetilde{\varphi}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2 \right) \geq \sum_{k=1}^K \frac{1}{2} \sum_{s' \in \mathcal{S}_{k,h}} \left\| \widetilde{\varphi}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2 .$$

Finally, we obtain

$$\sum_{k=1}^K \sum_{s' \in \mathcal{S}_{k,h}} \left\| \widetilde{\varphi}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2 \leq 2 \sum_{k=1}^K \log \left(1 + \sum_{s' \in \mathcal{S}_{k,h}} \left\| \widetilde{\varphi}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2 \right)$$

$$= 2 \log \frac{\det(\mathbf{B}_{K+1,h})}{\det(\lambda \mathbf{I}_d)}$$

$$\leq 2d \log \left(1 + \frac{K \mathcal{U} L_\varphi^2}{d\lambda} \right) ,$$

where the last inequality follows by the determinant-trace inequality (Lemma 28). □

D.5.3 Proof of Lemma 24

Proof of Lemma 24. Since $\boldsymbol{\xi}_{k,h}^{(m)} \sim \mathcal{N}(\mathbf{0}, \sigma_k^2 \mathbf{B}_{k,h}^{-1})$, by Lemma 30 for each $m \in [M]$, we have

$$\| \boldsymbol{\xi}_{k,h}^{(m)} \|_{\mathbf{B}_{k,h}} \leq C_{\boldsymbol{\xi}} \sigma_k \sqrt{d \log(Md/\delta)} .$$

Following the result of Lemma 22, we have

$$\sum_{s' \in \mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s_h^k, a_h^k) \left\| \bar{\varphi}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^k) \right\|_{\mathbf{B}_{k,h}^{-1}} \leq \sum_{s' \in \mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s_h^k, a_h^k) \left\| \bar{\varphi}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}$$

$$+ \frac{16 \eta L_\varphi}{\sqrt{\lambda}} \max_{s' \in \mathcal{S}_{k,h}} \left\| \bar{\varphi}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2 .$$

Then, we obtain

$$\sum_{s' \in \mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s_h^k, a_h^k) \bar{\varphi}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^k)^\top \boldsymbol{\xi}_{k,h}^{s'}$$

$$\leq \sum_{s' \in \mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s_h^k, a_h^k) \left\| \bar{\varphi}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^k) \right\|_{\mathbf{B}_{k,h}^{-1}} \| \boldsymbol{\xi}_{k,h}^{s'} \|_{\mathbf{B}_{k,h}}$$

$$\leq C_{\boldsymbol{\xi}} \sigma_k \sqrt{d \log(Md/\delta)} \sum_{s' \in \mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s_h^k, a_h^k) \| \bar{\varphi}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^k) \|_{\mathbf{B}_{k,h}^{-1}}$$

$$\leq \gamma_k(\delta) \left(\sum_{s' \in \mathcal{S}_{k,h}} P_{\widetilde{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s_h^k, a_h^k) \left\| \bar{\varphi}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}} \right.$$

$$\left. + \frac{16 \eta L_\varphi}{\sqrt{\lambda}} \max_{s' \in \mathcal{S}_{k,h}} \left\| \bar{\varphi}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2 \right) .$$

□

65

D.5.4 Proof of Lemma 25

Proof of Lemma 25. We provide a proof for Lemma 25 since it is slight modification of Lemma 20 of [76]. From the definition, we know that

$$\left(\widetilde{\boldsymbol{\theta}}_h^{k+1}\right)^\top \nabla \ell_{k,h}(\widetilde{\boldsymbol{\theta}}_h^k) + \frac{1}{2\eta} \left\|\widetilde{\boldsymbol{\theta}}_h^{k+1} - \widetilde{\boldsymbol{\theta}}_h^k\right\|_{\widetilde{\mathbf{B}}_{k,h}}^2 \leq \left(\widetilde{\boldsymbol{\theta}}_h^k\right)^\top \nabla \ell_{k,h}(\widetilde{\boldsymbol{\theta}}_h^k).$$

By rearranging the terms, the following holds:

$$\frac{1}{2\eta} \left\|\widetilde{\boldsymbol{\theta}}_h^{k+1} - \widetilde{\boldsymbol{\theta}}_h^k\right\|_{\widetilde{\mathbf{B}}_{k,h}}^2 \leq \left(\widetilde{\boldsymbol{\theta}}_h^k - \widetilde{\boldsymbol{\theta}}_h^{k+1}\right)^\top \nabla \ell_{k,h}(\widetilde{\boldsymbol{\theta}}_h^k)$$

$$\leq \left\|\widetilde{\boldsymbol{\theta}}_h^k - \widetilde{\boldsymbol{\theta}}_h^{k+1}\right\|_{\widetilde{\mathbf{B}}_{k,h}} \left\|\nabla \ell_{k,h}(\widetilde{\boldsymbol{\theta}}_h^k)\right\|_{\widetilde{\mathbf{B}}_{k,h}^{-1}}$$

Thus, we get

$$\left\|\widetilde{\boldsymbol{\theta}}_h^{k+1} - \widetilde{\boldsymbol{\theta}}_h^k\right\|_{\widetilde{\mathbf{B}}_{k,h}} \leq 2\eta \left\|\nabla \ell_{k,h}(\widetilde{\boldsymbol{\theta}}_h^k)\right\|_{\widetilde{\mathbf{B}}_{k,h}^{-1}}.$$

Since $\mathbf{B}_{k,h} \preceq \widetilde{\mathbf{B}}_{k,h}$ and $\widetilde{\mathbf{B}}_{k,h}^{-1} \preceq \lambda^{-1}\mathbf{I}_d$, we obtain

$$\left\|\widetilde{\boldsymbol{\theta}}_h^{k+1} - \widetilde{\boldsymbol{\theta}}_h^k\right\|_{\mathbf{B}_{k,h}} \leq \left\|\widetilde{\boldsymbol{\theta}}_h^{k+1} - \widetilde{\boldsymbol{\theta}}_h^k\right\|_{\widetilde{\mathbf{B}}_{k,h}} \leq 2\eta \left\|\nabla \ell_{k,h}(\widetilde{\boldsymbol{\theta}}_h^k)\right\|_{\widetilde{\mathbf{B}}_{k,h}^{-1}} \leq \frac{2\eta}{\sqrt{\lambda}} \left\|\nabla \ell_{k,h}(\widetilde{\boldsymbol{\theta}}_h^k)\right\|_2 \leq \frac{4\eta L_\varphi}{\sqrt{\lambda}}.$$
$$(103)$$

For the last inequality of (103), we provide the upper bound of l_2-norm of $\nabla \ell_{k,h}(\boldsymbol{\theta})$. Since

$$\ell_{k,h}(\boldsymbol{\theta}) = - \sum_{s' \in \mathcal{S}_{k,h}} y_h^k(s') \log P_{\boldsymbol{\theta}}(s' \mid s_h^k, a_h^k),$$

the gradient of the loss function is given by

$$\nabla \ell_{k,h}(\boldsymbol{\theta}) = - \sum_{s' \in \mathcal{S}_{k,h}} y_h^k(s') \left(\boldsymbol{\varphi}_{s,a,s'} - \sum_{s'' \in \mathcal{S}_{k,h}} P_{\boldsymbol{\theta}}(s'' \mid s_h^k, a_h^k) \boldsymbol{\varphi}_{s,a,s''} \right)$$

$$= \sum_{s' \in \mathcal{S}_{k,h}} y_h^k(s') \sum_{s'' \in \mathcal{S}_{k,h}} P_{\boldsymbol{\theta}}(s'' \mid s_h^k, a_h^k) \boldsymbol{\varphi}_{s,a,s''} - \sum_{s' \in \mathcal{S}_{k,h}} y_h^k(s') \boldsymbol{\varphi}_{s,a,s'}$$

$$= \sum_{s'' \in \mathcal{S}_{k,h}} P_{\boldsymbol{\theta}}(s'' \mid s_h^k, a_h^k) \boldsymbol{\varphi}_{s,a,s''} - \sum_{s' \in \mathcal{S}_{k,h}} y_h^k(s') \boldsymbol{\varphi}_{s,a,s'}$$

$$= \sum_{s' \in \mathcal{S}_{k,h}} \left(P_{\boldsymbol{\theta}}(s' \mid s_h^k, a_h^k) - y_h^k(s') \right) \boldsymbol{\varphi}_{s,a,s'}.$$

Therefore, we have

$$\|\nabla \ell_{k,h}(\boldsymbol{\theta})\|_2 = \left\| \sum_{s' \in \mathcal{S}_{k,h}} \left(P_{\boldsymbol{\theta}}(s' \mid s_h^k, a_h^k) - y_h^k(s') \right) \boldsymbol{\varphi}_{s,a,s'} \right\|_2$$

$$\leq \sum_{s' \in \mathcal{S}_{k,h}} \left| P_{\boldsymbol{\theta}}(s' \mid s_h^k, a_h^k) - y_h^k(s') \right| \|\boldsymbol{\varphi}_{s,a,s'}\|_2$$

$$\leq 2L_\varphi$$

and this concludes the proof. $\qquad\square$

D.6 Bound on Pessimism Part

In this section, we provide the upper bound on the pessimism part of the regret: $\sum_{k=1}^K (V_1^* - \widetilde{V}_1^k)(s_1^k)$.

Lemma 26 (Bound on pessimism). *For any δ with $0 < \delta < \Phi(-1)/2$, let $\sigma_k = H\beta_k$. If $\lambda = \mathcal{O}(L_\varphi^2 d \log \mathcal{U})$ and we take multiple sample size $M = \lceil 1 - \frac{\log(H\mathcal{U})}{\log \Phi(1)} \rceil$, then with probability at least $1 - \delta/2$, we have*

$$\sum_{k=1}^K (V_1^* - V_1^k)(s_1^k) = \widetilde{\mathcal{O}}\left(d^{3/2} H^{3/2} \sqrt{T} + \kappa^{-1} d^2 H^2 \right).$$

66

Proof of Lemma 26. As seen in Lemma 18, by using multiple sampling technique we show that the optimistic randomized value function $\widetilde{V}$ of `ORRL-MNL` is optimistic than the true optimal value with constant probability Hence, with the same argument used in Lemma 11, we can show that the pessimism term of `ORRL-MNL` is upper bounded by a bound of the estimation term times the inverse probability of being optimistic, i.e.,

$$\sum_{k=1}^{K} \left(V_1^* - V_1^k\right)(s_1^k) \leq \widetilde{\mathcal{O}}\left(\frac{1}{\Phi(-1)}\sum_{k=1}^{K}\left(V_1^k - V_1^{\pi^k}\right)(s_1^k)\right).$$

$\square$

D.7   Regret Bound of `ORRL-MNL`

Proof of Theorem 2. Since both Lemma 21 and Lemma 26 holds with probability at least $1 - \delta/2$ respectively, by taking the union bound we conclude the proof. $\square$

E   Optimistic Exploration Extension

In this section, we introduce `UCRL-MNL+` (Algorithm 3), which is both *computationally* and *statistically* efficient for MNL-MDPs with UCB-based exploration. The main difference compared to `ORRL-MNL` is that `UCRL-MNL+` constructs an *optimistic value function* that is greater than the optimal value function with high probability. At each episode $k \in [K]$, with the estimated transition core parameter $\widetilde{\boldsymbol{\theta}}_h^k$ (5), for $(s, a) \in \mathcal{S} \times \mathcal{A}$, set $\hat{Q}_{H+1}^k(s, a) = 0$. For each $h \in [H]$,

$$\hat{Q}_h^k(s,a) := r(s,a) + \sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s,a)\hat{V}_{h+1}^k(s') + \nu_{k,h}^{\text{opt}}(s,a), \tag{104}$$

where $\hat{V}_h^k(s) := \min\{\max_{a \in \mathcal{A}} \hat{Q}_h^k(s,a), H\}$ and $\nu_{k,h}^{\text{opt}}(s,a)$ is the *optimistic bonus term* defined by

$$\nu_{k,h}^{\text{opt}}(s,a) := H\beta_k \sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s,a)\|\bar{\boldsymbol{\varphi}}(s,a,s';\widetilde{\boldsymbol{\theta}}_h^k)\|_{\mathbf{B}_{k,h}^{-1}} + 3H\beta_k^2 \max_{s' \in \mathcal{S}_{s,a}} \|\boldsymbol{\varphi}(s,a,s')\|_{\mathbf{B}_{k,h}^{-1}}^2.$$

Based on these *optimistic value function* $\hat{Q}_h^k$, at each episode the agent plays a greedy action with respect to $\hat{Q}_h^k$ as summarized in Algorithm 3.

Algorithm 3 `UCRL-MNL+` (Upper Confidence RL for MNL-MDPs)

1: **Inputs:** Episodic MDP $\mathcal{M}$, Feature map $\boldsymbol{\varphi}: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$, Number of episodes K, Regularization parameter λ, Confidence radius $\{\beta_k\}_{k=1}^K$, Step size η
2: **Initialize:** $\widetilde{\boldsymbol{\theta}}_h^1 = \mathbf{0}_d$, $\mathbf{B}_{1,h} = \lambda\mathbf{I}_d$ for all $h \in [H]$
3: **for** episode $k = 1, 2, \cdots, K$ **do**
4:     Observe s_1^k and set $\left\{\hat{Q}_h^k(\cdot, \cdot)\right\}_{h \in [H]}$ as described in (104)
5:     **for** horizon $h = 1, 2, \cdots, H$ **do**
6:         Select $a_h^k = \text{argmax}_{a \in \mathcal{A}} \hat{Q}_h^k(s_h^k, a)$ and observe s_{h+1}^k
7:         Update $\widetilde{\mathbf{B}}_{k,h} = \mathbf{B}_{k,h} + \eta\nabla^2\ell_{k,h}(\widetilde{\boldsymbol{\theta}}_h^k)$ and $\widetilde{\boldsymbol{\theta}}_h^{k+1}$ as in (5)
8:         Update $\mathbf{B}_{k+1,h} = \mathbf{B}_{k,h} + \nabla^2\ell_{k,h}(\widetilde{\boldsymbol{\theta}}_h^{k+1})$
9:     **end for**
10: **end for**

The main difference in regret analysis lies in ensuring the optimism of the estimated value function $\hat{Q}_h^k$ (Lemma 27). In the following statement (formal statement of Corollary 1), we provide a regret guarantee for `UCRL-MNL+`, which enjoys the tightest regret bound for MNL-MDPs.

Theorem 3 (Regret Bound of `UCRL-MNL+`). *Suppose that Assumption 1- 4 hold. For any $\delta \in (0, 1)$, if we set the input parameters in Algorithm 3 as $\lambda = \mathcal{O}(L_{\boldsymbol{\varphi}}^2 d\log\mathcal{U})$, $\beta_k = \mathcal{O}(\sqrt{d}\log\mathcal{U}\log(kH))$*

67

$\eta = \mathcal{O}(\log \mathcal{U})$, then with probability at least $1 - \delta$, the cumulative regret of the UCRL-MNL+ policy π is upper-bounded by

$$\boldsymbol{Regret}_\pi(K) = \widetilde{\mathcal{O}}\left(dH^{3/2}\sqrt{T} + \kappa^{-1}d^2H^2\right),$$

where $T = KH$ is the total number of time steps.

Proof of Theorem 3. By Lemma 17, suppose that the good event $\mathfrak{G}(K, \delta')$ holds with probability at least $1 - \delta$. Then, we show that the optimistic value function $\hat{Q}_h^k$ is deterministically greater than the true optimal value function as follows:

Lemma 27 (Optimism). *Suppose that the event $\mathfrak{G}_{k,h}^\Delta(\delta)$ holds for all $k \in [K]$ and $h \in [H]$. Then for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have*

$$Q_h^*(s, a) \leq \hat{Q}_h^k(s, a).$$

Conditioned on $\mathfrak{G}(K, \delta')$, by Lemma 27 we have

$$
\begin{aligned}
(V_1^* - V_1^{\pi^k})(s_1^k) &= Q_1^*(s_1^k, \pi^*(s_1^k)) - Q_1^{\pi^k}(s_1^k, a_1^k) \\
&\leq \hat{Q}_1^k(s_1^k, \pi^*(s_1^k)) - Q_1^{\pi^k}(s_1^k, a_1^k) \\
&\leq \hat{Q}_1^k(s_1^k, a_1^k) - Q_1^{\pi^k}(s_1^k, a_1^k) = \nu_{k,1}^{\mathrm{opt}}(s_1^k, a_1^k) + P_1(\hat{V}_2^k - V_2^{\pi^k})(s_1^k, a_1^k).
\end{aligned}
$$

Note that

$$P_1(\hat{V}_2^k - V_2^{\pi^k})(s_1^k, a_1^k) = \mathbb{E}_{\widetilde{s}|s_1^k, a_1^k}\left[(\hat{V}_2^k - V_2^{\pi^k})(\widetilde{s})\right] = (\hat{V}_2^k - V_2^{\pi^k})(s_2^k) + \dot{\zeta}_1^k,$$

where we denote $\dot{\zeta}_h^k := (\hat{V}_{h+1}^k - V_{h+1}^{\pi^k})(s_{h+1}^k) - \mathbb{E}_{\widetilde{s}|s_h^k, a_h^k}\left[(\hat{V}_{h+1}^k - V_{h+1}^{\pi^k})(\widetilde{s})\right]$. Then, with the same argument, we have

$$(V_1^* - V_1^{\pi^k})(s_1^k) \leq \sum_{h=1}^H \nu_{k,h}^{\mathrm{opt}}(s_h^k, a_h^k) + \sum_{h=1}^H \dot{\zeta}_h^k.$$

By summing over all episodes, we have

$$\boldsymbol{Regret}_\pi(K) \leq \sum_{k=1}^K \sum_{h=1}^H \nu_{k,h}^{\mathrm{opt}}(s_h^k, a_h^k) + \sum_{k=1}^K \sum_{h=1}^H \dot{\zeta}_h^k. \tag{105}$$

On the other hand, note that

$$
\begin{aligned}
&\sum_{k=1}^K \sum_{h=1}^H \nu_{k,h}^{\mathrm{opt}}(s_h^k, a_h^k) \\
&= \sum_{k=1}^K \sum_{h=1}^H H\beta_k \sum_{s' \in \mathcal{S}_{k,h}} P_{\widetilde{\theta}_h^k}(s' \mid s_h^k, a_h^k) \|\bar{\varphi}_{k,h,s'}(\widetilde{\theta}_h^k)\|_{\mathbf{B}_{k,h}^{-1}} + \sum_{k=1}^K \sum_{h=1}^H 3H\beta_k^2 \max_{s' \in \mathcal{S}_{k,h}} \|\varphi_{k,h,s'}\|_{\mathbf{B}_{k,h}^{-1}}^2 \\
&\leq H\beta_K \sum_{k=1}^K \sum_{h=1}^H \sum_{s' \in \mathcal{S}_{k,h}} P_{\widetilde{\theta}_h^k}(s' \mid s_h^k, a_h^k) \|\bar{\varphi}_{k,h,s'}(\widetilde{\theta}_h^k)\|_{\mathbf{B}_{k,h}^{-1}} \\
&\quad + 3H\beta_K^2 \sum_{k=1}^K \sum_{h=1}^H \max_{s' \in \mathcal{S}_{k,h}} \|\varphi_{k,h,s'}\|_{\mathbf{B}_{k,h}^{-1}}^2 \\
&\leq \underbrace{H\beta_K \sum_{k=1}^K \sum_{h=1}^H \sum_{s' \in \mathcal{S}_{k,h}} P_{\widetilde{\theta}_h^{k+1}}(s' \mid s_h^k, a_h^k) \left\|\bar{\varphi}_{k,h,s'}(\widetilde{\theta}_h^{k+1})\right\|_{\mathbf{B}_{k,h}^{-1}}}_{(\mathrm{i})} \\
&\quad + \underbrace{\frac{16\eta L_\varphi}{\sqrt{\lambda}} H\beta_K \sum_{k=1}^K \sum_{h=1}^H \max_{s' \in \mathcal{S}_{k,h}} \left\|\bar{\varphi}_{k,h,s'}(\widetilde{\theta}_h^{k+1})\right\|_{\mathbf{B}_{k,h}^{-1}}^2}_{(\mathrm{ii})} + \underbrace{3H\beta_K^2 \sum_{k=1}^K \sum_{h=1}^H \max_{s' \in \mathcal{S}_{k,h}} \|\varphi_{k,h,s'}\|_{\mathbf{B}_{k,h}^{-1}}^2}_{(\mathrm{iii})},
\end{aligned}
$$

68

where the last inequality follows by Lemma 22.

Term (i) can be bounded as in Eq. (91):

$$H\beta_K \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s' \in \mathcal{S}_{k,h}} P_{\widehat{\boldsymbol{\theta}}_h^{k+1}}(s' \mid s_h^k, a_h^k) \left\| \bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}} = \widetilde{\mathcal{O}}(dH^{3/2}\sqrt{T}).$$ (106)

For term (ii), recall that as in Eq. (93) we have

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \max_{s' \in \mathcal{S}_{k,h}} \left\| \bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2 \le 16\kappa^{-1}dH \log\left(1 + \frac{K\mathcal{U}L_{\boldsymbol{\varphi}}^2}{d\lambda} \right).$$

Then, we have

$$\frac{16\eta L_{\boldsymbol{\varphi}}}{\sqrt{\lambda}} H\beta_K \sum_{k=1}^{K} \sum_{h=1}^{H} \max_{s' \in \mathcal{S}_{k,h}} \left\| \bar{\boldsymbol{\varphi}}_{k,h,s'}(\widetilde{\boldsymbol{\theta}}_h^{k+1}) \right\|_{\mathbf{B}_{k,h}^{-1}}^2 = \widetilde{\mathcal{O}}(\kappa^{-1}dH^2).$$ (107)

For term (iii), since we have

$$\begin{aligned}
3H\beta_K^2 \sum_{k=1}^{K} \sum_{h=1}^{H} \max_{s' \in \mathcal{S}_{k,h}} \|\boldsymbol{\varphi}_{k,h,s'}\|_{\mathbf{B}_{k,h}^{-1}}^2 &\le 3H\beta_K^2 \sum_{k=1}^{K} \sum_{h=1}^{H} \max_{s' \in \mathcal{S}_{k,h}} \|\boldsymbol{\varphi}_{k,h,s'}\|_{\mathbf{A}_{k,h}^{-1}}^2 \\
&\le 12\kappa^{-1}dH^2\beta_K^2 \log\left(1 + K\mathcal{U}L_{\boldsymbol{\varphi}}^2/(d\lambda)\right) \\
&= \widetilde{\mathcal{O}}(\kappa^{-1}d^2H^2).
\end{aligned}$$ (108)

Combining the results of Eq. (106), (107), and (108), we have

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \nu_{k,h}^{\mathrm{opt}}(s_h^k, a_h^k) = \widetilde{\mathcal{O}}(dH^{3/2}\sqrt{T} + \kappa^{-1}d^2H^2).$$

Finally, by Azuma-Hoeffiding inequality as in Eq. (96) we have

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \dot{\zeta}_h^k = \widetilde{\mathcal{O}}(H\sqrt{T}).$$

This concludes the proof. □

In the following, we provide the proof of Lemma 27.

E.1 Optimism

Proof of Lemma 27. We prove this by backwards induction on h. For the base case $h = H$, since $V_{H+1}^*(s) = \hat{V}_{H+1}^k(s) = 0$ for all $s \in \mathcal{S}$, we have

$$\hat{Q}_H^k(s, a) = r(s, a) = Q_H^*(s, a).$$

69

Suppose that the statement holds for $h+1$ where $h \in [H-1]$. Then, for h and for any $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$
\begin{aligned}
\hat{Q}_h^k(s,a) \\
&= r(s,a) + \sum_{s' \in \mathcal{S}_{s,a}} P_{\widehat{\boldsymbol{\theta}}_h^k}(s' \mid s,a)\hat{V}_{h+1}^k(s') + \nu_{k,h}^{\mathrm{opt}}(s,a) \\
&\geq r(s,a) + \sum_{s' \in \mathcal{S}_{s,a}} P_{\widehat{\boldsymbol{\theta}}_h^k}(s' \mid s,a)V_{h+1}^*(s') + \nu_{k,h}^{\mathrm{opt}}(s,a) \\
&= r(s,a) + \sum_{s' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}_h^*}(s' \mid s,a)V_{h+1}^*(s') \\
&\quad + \sum_{s' \in \mathcal{S}_{s,a}} \left(P_{\widehat{\boldsymbol{\theta}}_h^k}(s' \mid s,a) - P_{\boldsymbol{\theta}_h^*}(s' \mid s,a) \right) V_{h+1}^*(s') + \nu_{k,h}^{\mathrm{opt}}(s,a) \\
&\geq r(s,a) + \sum_{s' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}_h^*}(s' \mid s,a)V_{h+1}^*(s') \\
&= Q_h^*(s,a),
\end{aligned}
$$

where the first inequality follows from the induction hypothesis and the second inequality holds by Lemma 16. $\qquad\square$

F   Experiment Details
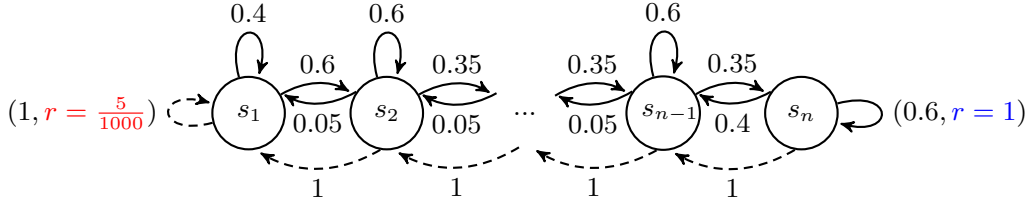
Figure 2: The "RiverSwim" environment with n states [58]

The RiverSwim environment (Figure 2) consists of n states that are arranged in a chain. The agent starts in the leftmost state with a relatively small reward of 0.005 and aims to reach the rightmost state, which has a relatively large reward of 1. Choosing to swim to the left moves the agent deterministically to the left, while swimming to the right has a probability of transitioning the agent toward the right state, but also a high chance of remaining in the current state or even moving left due to the strong current of river. Therefore, efficient exploration is crucial in order to learn the optimal policy for this environment.

We fine-tuned the hyperparameters for each algorithm within specific ranges. Figures 1a and 1b show the episodic returns in the RiverSwim environment over 10 independent runs with $|\mathcal{S}| = 4, H = 12$, and $K = 10{,}000$ and $|\mathcal{S}| = 8, H = 24$, and $K = 10{,}000$, respectively. The shaded areas represent the standard deviations (1-sigma error). Figure 1c compares the running time of the algorithms over the first 1,000 episodes. All experiments were conducted on a Xeon(R) Gold 6226R CPU @ 2.90GHz (16 cores).

G   Auxiliary Lemmas

Lemma 28 (Determinant-trace inequality [1]). *Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_t \in \mathbb{R}^d$ and for any $1 \leq \tau \leq t$, $\|\mathbf{x}_\tau\|_2 \leq L$. Let $\mathbf{V}_t = \lambda\mathbf{I}_d + \sum_{\tau=1}^t \mathbf{x}_\tau\mathbf{x}_\tau^\top$ for some $\lambda > 0$. Then,*

$$
\det(\mathbf{V}_t) \leq (\lambda + tL^2/d)^d.
$$

Lemma 29 (Freedman's inequality [29]). *Consider a real-valued martingale $\{Y_k : k = 0, 1, 2, \ldots\}$ with difference sequence $\{X_k : k = 0, 1, 2, 3, \ldots\}$. Assume that the difference sequence is uniformly*

70

bounded, $X_k \leq R$ almost surely for $k = 1, 2, 3, \ldots$. Define the predictable quadratic variation process of the martingale:

$$W_k := \sum_{j=1}^{k} \mathbb{E}_{j-1}[X_j^2] \quad \text{for } k = 1, 2, 3, \ldots.$$

Then, for all $t \geq 0$ and $\sigma^2 > 0$,

$$\mathbb{P}\left(\exists k \geq 0 : Y_k \geq t \text{ and } W_k \leq \sigma^2\right) \leq \exp\left(-\frac{-t^2/2}{\sigma^2 + Rt/3}\right).$$

Lemma 30 (Gaussian noise concentration (Lemma D.2 in [37])). *Let $\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(M)}$ be M independent d-dimensional multivariate normal distributed vector with mean $\mathbf{0}_d$ and covariance $\sigma^2 \mathbf{A}^{-1}$ for some $\sigma > 0$ and a positive definite matrix $\mathbf{A}^{-1}$, i.e., $\boldsymbol{\xi}^{(m)} \sim \mathcal{N}(\mathbf{0}_d, \sigma^2 \mathbf{A}^{-1})$ for $m \in [M]$. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have*

$$\max_{m \in [M]} \|\boldsymbol{\xi}^{(m)}\|_{\mathbf{A}} \leq C_{\boldsymbol{\xi}} \sigma \sqrt{d \log(Md/\delta)} := \gamma(\delta),$$

where $C_{\boldsymbol{\xi}}$ is an absolute constant.

Lemma 31 (Proposition 4.1 of 15). *Let the w_{t+1} be the solution of the update rule*

$$w_{t+1} = \arg\min_{w \in \mathcal{V}} \eta \ell_t(w) + D_\psi(w, w_t),$$

where $\mathcal{V} \subseteq \mathcal{W} \subseteq \mathbb{R}^d$ is a non-empty convex set and $D_\psi(w_1, w_2) = \psi(w_1) - \psi(w_2) - \langle \nabla \psi(w_2), w_1 - w_2 \rangle$ is the Bregman Divergence w.r.t. a strictly convex and continuously differentiable function $\psi : \mathcal{W} \to \mathbb{R}$. Further supposing $\psi(w)$ is 1-strongly convex w.r.t. a certain norm $\|\cdot\|$ in $\mathcal{W}$, then there exists a $g_t \in \partial \ell_t(w_{t+1})$ such that

$$\langle \eta_t g_t', w_{t+1} - u \rangle \leq \langle \nabla \psi(w_t) - \nabla \psi(w_{t+1}), w_{t+1} - u \rangle$$

for any $u \in \mathcal{W}$.

Lemma 32. *Let $\{\mathcal{F}_t\}_{t=1}^{\infty}$ be a filtration. Let $\{\mathbf{z}_t\}_{t=1}^{\infty}$ be a stochastic process in $\mathcal{B}_2(\mathcal{U}) = \{\mathbf{z} \in \mathbb{R}^{\mathcal{U}} \mid \|\mathbf{z}\|_\infty \leq 1\}$ such that $\mathbf{z}_t$ is $\mathcal{F}_t$ measurable. Let $\{\boldsymbol{\varepsilon}_t\}_{t=1}^{\infty}$ be a martingale difference sequence such that $\boldsymbol{\varepsilon}_t \in \mathbb{R}^{\mathcal{U}}$ is $\mathcal{F}_{t+1}$ measurable. Furthermore, assume that conditional on $\mathcal{F}_t$, we have $\|\boldsymbol{\varepsilon}_t\|_1 \leq 2$ almost surely, and denote by $\Sigma_t = \mathbb{E}[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^\top | \mathcal{F}_t]$. Let $\lambda > 0$ and for any $t \geq 1$ define*

$$U_t = \sum_{i=1}^{t-1} \langle \boldsymbol{\varepsilon}_i, \mathbf{z}_i \rangle \quad \text{and} \quad \mathbf{B}_t = \lambda + \sum_{i=1}^{t-1} \|\mathbf{z}_i\|_{\Sigma_i}^2,$$

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

$$\Pr\left[\exists t \geq 1, U_t \geq \sqrt{\mathbf{B}_t}\left(\frac{\sqrt{\lambda}}{4} + \frac{4}{\sqrt{\lambda}} \log\left(\sqrt{\frac{\mathbf{B}_t}{\lambda}}\right) + \frac{4}{\sqrt{\lambda}} \log\left(\frac{2}{\delta}\right)\right)\right] \leq \delta.$$

Lemma 33 (Lemma 1 of 76). *Let $\ell(\mathbf{z}, y) = \sum_{k=0}^{K} \mathbf{1}\{y = k\} \cdot \log\left(\frac{1}{[\sigma(\mathbf{z})]_k}\right)$, $\mathbf{a} \in [-C, C]^K$, $y \in \{0\} \cup [K]$ and $\mathbf{b} \in \mathbb{R}^K$ where $C > 0$. Then, we have*

$$\ell(\mathbf{a}, y) \geq \ell(\mathbf{b}, y) + \nabla \ell(\mathbf{b}, y)^\top (\mathbf{a} - \mathbf{b}) + \frac{1}{\log(K+1) + 2(C+1)} (\mathbf{a} - \mathbf{b})^\top \nabla^2 \ell(\mathbf{b}, y)(\mathbf{a} - \mathbf{b}).$$

Lemma 34 (Lemma 17 of 76). *Let $\ell(\mathbf{z}, y) = \sum_{k=0}^{K} \mathbf{1}\{y = k\} \cdot \log\left(\frac{1}{[\sigma(\mathbf{z})]_k}\right)$ and $\mathbf{z} \in \mathbb{R}^K$ be a K-dimensional vector. Define $\mathbf{z}^\mu \triangleq \sigma^+ (\text{smooth}_\mu(\sigma(\mathbf{z})))$, where $\text{smooth}_\mu(\mathbf{p}) = (1 - \mu)\mathbf{p} + \mu \mathbf{1}/(K+1)$. Then, for $\mu \in [0, 1/2]$, we have*

$$\ell(\mathbf{z}^\mu, y) - \ell(\mathbf{z}, y) \leq 2\mu$$

for any $y \in \{0\} \cup [K]$. We also have $\|\mathbf{z}^\mu\|_\infty \leq \log(K/\mu)$.

Lemma 35 (Lemma 18 of 76). *Let $L_{i,h}(\boldsymbol{\theta}) := \ell_{i,h}(\boldsymbol{\theta}) + \frac{1}{2c}\|\boldsymbol{\theta} - \boldsymbol{\theta}_h^i\|_{\mathbf{B}_{i,h}}^2$. Assume that $\ell_{i,h}$ is a $\sqrt{N}$-self-concordant-like function. Then, for any $\boldsymbol{\theta}, \boldsymbol{\theta}_h^i \in \mathcal{B}(\mathbf{0}_d, 1)$, the quadratic approximation $\widetilde{L}_{i,h}(\boldsymbol{\theta}) = L_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}) + \langle \nabla L_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1}), \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1} \rangle + \frac{1}{2c}\left\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1}\right\|_{\mathbf{B}_{i,h}}^2$ satisfies*

$$L_{i,h}(\boldsymbol{\theta}) \leq \widetilde{L}_{i,h}(\boldsymbol{\theta}) + \exp\left(N \left\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1}\right\|_2^2\right) \left\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h^{i+1}\right\|_{\nabla \ell_{i,h}(\widetilde{\boldsymbol{\theta}}_h^{i+1})}^2.$$

71

H Limitations

We make an assumption about the transition model of MDPs by using the MNL model, which is a specific parametric model. This assumption implies that we assume the realizability of the MNL model. It's worth noting that the realizability assumption has also been commonly made in previous literature on provable reinforcement learning with function approximation, including works such as [72, 43, 73, 53, 22, 14, 9, 68, 70, 33, 81, 82, 37, 35]. However, we hope that this condition can be relaxed in the future work.

72

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 is to propose provably efficient randomized algorithms for MNL-MDPs. In Section 1 (Introduction), we provide the motivation and main contributions of this paper.

    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 limitation of this work in Appendix H.

    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?

73

Answer: [Yes]

Justification: We provide the full set of assumptions in Section 2.2 and a complete proof of main results in Appendix C and D.

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: [Yes]

Justification: We provide numerical experiments that support our main claims in Section 5 and the detailed information of experiments in Appendix F.

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**

74

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: [Yes]

Justification: We have attached the data and code with sufficient instructions to reproduce the main experimental results in the supplementary material.

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: [Yes]

Justification: We provide the detailed explanation for the experimental setting in Appendix F.

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: [Yes]

Justification: We report error bars (standard deviation) in our numerical experiment results shown in Section 5.

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).

75

- 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: [Yes]

Justification: We provide the detailed information on the computer resources used to conduct numerical experiments in Appendix F.

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: The research conducted in this paper adheres to the NeurIPS Code of Ethics in all aspects.

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: There is no negative societal impacts of the work performed because this research focuses on theoretical aspects.

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.

76

- 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: The research conducted in this paper does not pose any such risks.

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 paper does not use any external assets such as code, data, or models.

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.

77

- 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 paper 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.

78