Infinite-Horizon Reinforcement Learning with Multinomial Logit Function Approximation

Jaehyun Park
KAIST

Junyeop Kwon
KAIST

Dabeen Lee
KAIST

Abstract

We study model-based reinforcement learning with non-linear function approximation where the transition function of the underlying Markov decision process (MDP) is given by a multinomial logit (MNL) model. We develop a provably efficient discounted value iteration-based algorithm that works for both infinite-horizon average-reward and discounted-reward settings. For average-reward communicating MDPs, the algorithm guarantees a regret upper bound of $\tilde{\mathcal{O}}(dD\sqrt{T})$ where d is the dimension of feature mapping, D is the diameter of the underlying MDP, and T is the horizon. For discounted-reward MDPs, our algorithm achieves $\tilde{\mathcal{O}}(d(1-\gamma)^{-2}\sqrt{T})$ regret where γ is the discount factor. Then we complement these upper bounds by providing several regret lower bounds. We prove a lower bound of $\Omega(d\sqrt{DT})$ for learning communicating MDPs of diameter D and a lower bound of $\Omega(d(1-\gamma)^{-3/2}\sqrt{T})$ for learning discounted-reward MDPs with discount factor γ. Lastly, we show a regret lower bound of $\Omega(dH^{3/2}\sqrt{K})$ for learning H-horizon episodic MDPs with MNL function approximation where K is the number of episodes, which improves upon the best-known lower bound for the finite-horizon setting.

1 INTRODUCTION

Function approximation schemes have been successful in modern reinforcement learning (RL) under the presence of large state and action spaces. Applications and domains where function approximation approaches have been deployed include video games (Mnih et al., 2015), Go (Silver et al., 2017), robotics (Kober et al., 2013), and autonomous driving (Yurtsever et al., 2020). Such empirical success has motivated a plethora of theoretical studies that establish provable guarantees for RL with function approximation. The first line of theoretical work considers linear function approximation, such as linear Markov Decision Processes (MDPs) (Yang and Wang, 2019; Jin et al., 2020) and linear mixture MDPs (Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021b) where the reward and transition functions are linear in some feature mappings. While (nearly) minimax optimal algorithms have been developed for linear MDPs (He et al., 2023; Agarwal et al., 2023; Hu et al., 2022) and for linear mixture MDPs (Zhou et al., 2021a), the linearity assumption is restrictive and rarely holds in practice. In particular, when a linear model is misspecified, these algorithms may suffer from linear regret (Jin et al., 2020).

RL with general function approximation has recently emerged as an alternative to the linear function approximation framework. The term general here means that it makes minimal structural assumptions about the family of functions taken for approximation. Some concepts that lead to conditions ensuring sample-efficient learning are the Bellman rank (Jiang et al., 2017), the eluder dimension (Wang et al., 2020), the Bellman eluder dimension (Jin et al., 2021), the bilinear class (Du et al., 2021), the decision-estimation coefficient (Foster et al., 2023), and the generalized eluder coefficient (Zhong et al., 2023). Recently, He et al. (2024) considered infinite-horizon average-reward MDPs with general function approximation. However, algorithms for these frameworks require an oracle to query from some abstract function class. In practice, the oracle would correspond to solving an abstract non-convex optimization/regression problem. Furthermore, no regret lower bound has been identified for a general function approximation framework.

More concrete non-linear function approximation

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

models have been proposed recently. Yang et al. (2020); Xu and Gu (2020); Fan et al. (2020) considered representing the Q function by an overparametrized neural network based on the neural tangent kernel. Wang et al. (2021) studied generalized linear models for approximating the Q function. Liu et al. (2022); Zhang et al. (2023) focused on the case where the Q function is smooth and lies in the Besov space or the Barron space, and they used a two-layer neural network to approximate the Q function. Hwang and Oh (2023) proposed a framework to represent the transition function by a multinomial logit (MNL) model. Indeed, the multinomial logit model can naturally represent state transition probabilities, providing a practical alternative to linear function approximation. The model is widely used for modeling multiple outcomes, such as multiclass classification (Bishop, 2006), news recommendations (Li et al., 2010, 2012), and assortment optimization (Caro and Gallien, 2007).

For RL with MNL approximation, Hwang and Oh (2023) developed an efficient model-based algorithm that achieves an $\tilde{\mathcal{O}}(\kappa^{-1}dH^2\sqrt{K})$ regret where d is the dimension of the transition core, H is the horizon, K is the number of episodes, and $\kappa \in (0,1)$ is a problem-dependent quantity. Recently, Cho et al. (2024) and Li et al. (2024) developed algorithms that both achieve a regret bound of $\tilde{\mathcal{O}}(dH^2\sqrt{K} + \kappa^{-1}d^2H^2)$, avoiding a dependence on κ in the leading term. Their algorithms are based on recently proposed online Newton-based parameter estimation schemes for logistic bandits (Zhang and Sugiyama, 2023) and multinomial logit bandits (Lee and Oh, 2024b). Moreover, Li et al. (2024) presented the first lower bound for this setting, given by $\Omega(dH\sqrt{\kappa^* K})$ where $\kappa^* \in (0,1)$ is another problem-dependent constant similar to κ.

Our Contributions This paper contributes to the RL with MNL approximation literature with the following new theoretical results. Our results are summarized in Table 1.

- We prove that there is a family of H-horizon episodic MDPs with MNL transitions for which any algorithm incurs a regret of $\Omega(dH^{3/2}\sqrt{K})$. This improves upon the lower bound due to Li et al. (2024) by a factor of $O(\sqrt{H/\kappa^*})$.
- We develop UCMNLK, a discounted extended value iteration-based algorithm that works for infinite-horizon average-reward and discounted-reward MDPs with MNL function approximation. For learning average-reward MDPs with diameter at most D, UCMNLK guarantees $\tilde{\mathcal{O}}(dD\sqrt{T} + \kappa^{-1}d^2D)$ regret. For learning discounted-reward MDPs with discount factor γ, UCMNLK provides a regret upper bound of $\tilde{\mathcal{O}}((d\sqrt{T} + \kappa^{-1}d^2)/(1-\gamma)^2)$.
- We prove a lower bound of $\Omega(d\sqrt{DT})$ for learning infinite-horizon average-reward communicating MDPs with MNL transitions and diameter D.
- We show a lower bound of $\Omega(d\sqrt{T}/(1-\gamma)^{3/2})$ for learning infinite-horizon discounted-reward MDPs with discount factor γ.

While UCMNLK is inspired by UCLK of Zhou et al. (2021b) developed for discounted-reward linear mixture MDPs, it has several novel components and thus works for the average-reward setting as well. First, we develop an efficient extended value iteration scheme for MNL function approximation. As the multinomial logit probability function is non-convex in the transition parameter vector θ, optimization over θ is not tractable. Instead, we construct and optimize over confidence polytopes for the true transition probability, thereby achieving computational efficiency. Second, for the average-reward setting, we approximate a given average-reward MDP by a discounted-reward MDP with an appropriate discount factor. We show that the discounted value function returned by extended value iteration has a bounded span. This leads to an analysis based on a novel regret decomposition.

We derive the lower bounds by approximating a multinomial logit function to a linear function, based on the mean value theorem. This approximation technique allows us to bridge MDPs with a multinomial logit transition model and linear mixture MDPs. Then we deduce our results from the known regret lower bounds for linear mixture MDPs by Zhou et al. (2021a,b); Wu et al. (2022).

2 PRELIMINARIES

Notations For a vector $x \in \mathbb{R}^d$ and a positive semidefinite matrix $A \in \mathbb{R}^{d\times d}$, $\|x\|_2$ denotes the ℓ_2-norm of x, and we denote by $\|x\|_A = \sqrt{x^\top Ax}$ the weighted ℓ_2-norm of x. Given a matrix A, $\|A\|_2$ denotes its spectral norm. For a symmetric matrix A, let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote its minimum and maximum eigenvalues, respectively. Let $\mathbf{1}\{\mathcal{E}\}$ be the indicator function of event $\mathcal{E}$. A random variable $Y \in \mathbb{R}$ is R-sub-Gaussian if $\mathbb{E}[Y] = 0$ and $\mathbb{E}[\exp(sY)] \leq \exp(R^2s^2/2)$ for any $s \in \mathbb{R}$. Let $\Delta(\mathcal{X})$ denote the family of probability measures on $\mathcal{X}$. For any positive integers m, n with $m < n$, $[n]$ and $[m : n]$ denote $\{1, \ldots, n\}$ and $\{m, \ldots, n\}$, respectively.

2.1 Infinite-Horizon Average-Reward MDP

We consider an infinite-horizon MDP specified by $M = (\mathcal{S}, \mathcal{A}, p, r)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $p(s' \mid s, a)$ denotes the unknown transition prob-

Table 1: Summary of Our Results on Regret Upper and Lower Bounds for RL with MNL Approximation

Setting	Regret Upper Bound	Regret Lower Bound
Finite-Horizon	$\tilde{\mathcal{O}}\left(dH^2\sqrt{K}+\kappa^{-1}d^2H^2\right)$ (Cho et al., 2024; Li et al., 2024)	$\Omega\left(dH^{3/2}\sqrt{K}\right)$ (Theorem 3)
Average-Reward	$\tilde{\mathcal{O}}\left(dD\sqrt{T}+\kappa^{-1}d^2D\right)$ (Theorem 1)	$\Omega\left(d\sqrt{DT}\right)$ (Theorem 4)
Discounted-Reward	$\tilde{\mathcal{O}}\left(d(1-\gamma)^{-2}\sqrt{T}+\kappa^{-1}d^2(1-\gamma)^{-2}\right)$ (Theorem 2)	$\Omega\left(d(1-\gamma)^{-3/2}\sqrt{T}\right)$ (Theorem 5)

ability of transitioning to state s' from state s after taking action a, and $r:\mathcal{S}\times\mathcal{A}\rightarrow[0,1]$ is the known reward function. Throughout this paper, we assume that both $\mathcal{S}$ and $\mathcal{A}$ are finite. A stationary policy $\pi:\mathcal{S}\rightarrow\Delta(\mathcal{A})$ is given by $\pi(a\mid s)$ specifying the probability of taking action a at state s. When π is deterministic, i.e., for each $s\in\mathcal{S}$ there exists $a\in\mathcal{A}$ with $\pi(a\mid s)=1$, we write that $a=\pi(s)$ with abuse of notation. Starting from an initial state $s_1=s$, for each time step t, an algorithm $\mathfrak{A}$ selects action a_t based on state s_t, and then s_{t+1} is drawn according to the transition function $p(\cdot\mid s_t,a_t)$. Then the cumulative reward of $\mathfrak{A}$ incurred over T times steps is given by $R(\mathfrak{A},s,T)=\sum_{t=1}^{T}r(s_t,a_t)$, and the average reward of $\mathfrak{A}$ is defined as $J(\mathfrak{A},s)=\liminf_{T\rightarrow\infty}\mathbb{E}\left[R(\mathfrak{A},s,T)\mid s_1=s\right]/T$. It is known that the average reward can be maximized by a deterministic stationary policy (See Puterman, 2014). Given a stationary policy π starting from state s, the average reward is given by $J^{\pi}(s)=\liminf_{T\rightarrow\infty}\mathbb{E}\left[\sum_{t=1}^{T}r(s_t,a_t)\mid s_1=s\right]/T$.

In this paper, following Jaksch et al. (2010), we focus on the class of communicating MDPs that have a finite diameter. Here, the diameter is defined as follows. Given an MDP M and a policy π, let $T(s'\mid M,\pi,s)$ denote the number of steps after which state s' is reached from state s for the first time. Then the diameter of M is defined as $D(M)=\max_{s\neq s'\in\mathcal{S}}\min_{\pi:\mathcal{S}\rightarrow\mathcal{A}}\mathbb{E}\left[T(s'\mid M,\pi,s)\right]$. For a communicating MDP M, it is known that the optimal average reward does not depend on the initial state s (See Puterman, 2014), and therefore, there exists J^* such that $J^*=J^*(s):=\max_{\pi}J^{\pi}(s)$. Based on this, we consider $\mathrm{Regret}(T)=T\cdot J^*-\sum_{t=1}^{T}r(s_t,a_t)$ as our notion of regret to assess the performance of any algorithm $\mathfrak{A}$ for infinite-horizon average-reward MDPs.

2.2 Discounted-Reward MDP

Given an infinite-horizon MDP $M=(\mathcal{S},\mathcal{A},p,r)$, consider a non-stationary policy π given by $\{\pi_t\}_{t=1}^{\infty}$ where $\pi_t:\{\mathcal{S}\times\mathcal{A}\}^{t-1}\times\mathcal{S}\rightarrow\Delta(A)$ samples an action from $\mathcal{A}$ based on history $(s_1,a_1,\ldots,s_{t-1},a_{t-1},s_t)$. Given a discount factor $\gamma\in[0,1)$, we consider the value function and the action-value function defined as $V_t^{\pi}(s)=\mathbb{E}\left[\sum_{i=0}^{\infty}\gamma^i r(s_{t+i},a_{t+i})\mid s_t=s\right]$ and $Q_t^{\pi}(s,a)=\mathbb{E}\left[\sum_{i=0}^{\infty}\gamma^i r(s_{t+i},a_{t+i})\mid s_t=s,a_t=a\right]$ for $(s,a)\in\mathcal{S}\times\mathcal{A}$. Note that $V_t^{\pi}(s)$ and $Q_t^{\pi}(s,a)$ capture the infinite-horizon discounted reward under policy π from time step t. The functions, however, are well-defined only when the probability of the event that $s_t=s$ and $a_t=a$ is positive. Nevertheless, (Zhou et al., 2021a, Appendix A) provides a slightly more technical definition that avoids the issue and is consistent with the above definition. Furthermore, we define the optimal value function V^* and the optimal action-value function Q^* as $V^*(s)=\max_{\pi}V_1^{\pi}(s)$ and $Q^*(s,a)=\max_{\pi}Q_1^{\pi}(s,a)$ for $(s,a)\in\mathcal{S}\times\mathcal{A}$. It is known that there exists a deterministic stationary policy π^* such that $V_1^{\pi^*}(s)=V^*(s)$ and $Q_1^{\pi^*}(s,a)=Q^*(s,a)$ for $(s,a)\in\mathcal{S}\times\mathcal{A}$ (See Puterman, 2014; Agarwal et al., 2021). Moreover, V^* and Q^* satisfy the following Bellman optimality equation.

$$\begin{aligned}Q^*(s,a)&=r(s,a)+\gamma\sum_{s'\in\mathcal{S}}p(s'\mid s,a)V^*(s')\\V^*(s)&=\max_{a\in\mathcal{A}}Q^*(s,a).\end{aligned}\tag{1}$$

For discounted-reward MDPs, following (Liu and Su, 2021; Zhou et al., 2021b), we consider $\mathrm{Regret}(\pi,T)=\sum_{t=1}^{T}V^*(s_t)-\sum_{t=1}^{T}V_t^{\pi}(s_t)$ as our notion of regret of a non-stationary policy π for discounted-reward MDPs.

2.3 Multinomial Logit Model

Despite being finite, the state space $\mathcal{S}$ and the action space $\mathcal{A}$ can be intractably large, in which case tabular model-based reinforcement learning algorithms suffer from a large regret. To remedy this, linear and linear mixture MDPs take some structural assumptions on the underlying MDP which lead to efficient learning. However, imposing linearity structures is indeed restrictive and limits the scope of practical applications. Inspired by this issue, we consider the recent framework of MNL function approximation proposed by Hwang and Oh (2023), assuming that the transition function is given by a feature-based multinomial logit model as follows. For each $(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}$, its associated feature vector $\varphi(s,a,s') \in \mathbb{R}^d$ is known, and the transition probability is given by

$$p(s' \mid s,a) := p(s' \mid s,a,\theta^*) \tag{2}$$

where

$$p(s' \mid s,a,\theta) := \frac{\exp\left(\varphi(s,a,s')^\top \theta\right)}{\sum_{s'' \in \mathcal{S}_{s,a}} \exp\left(\varphi(s_t,a_t,s'')^\top \theta\right)}.$$

Here, $\theta^* \in \mathbb{R}^d$ is an unknown parameter, which we refer to as the transition core, and $\mathcal{S}_{s,a} := \{s' \in \mathcal{S} : \mathbb{P}(s' \mid s,a) > 0\}$ is the set of reachable states from s in one step after taking action a. Let $\mathcal{U} := \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |\mathcal{S}_{s,a}|$. The general intuition is that the ambient dimension d of the feature vectors and the parameter vector is small compared to the size of $\mathcal{S}$ and that of $\mathcal{A}$. Moreover, it is often the case that $\mathcal{S}_{s,a}$ is small in comparison with $\mathcal{S}$.

Throughout this paper, we assume the following.

Assumption 1. *There exist some L_φ, L_θ such that $\|\varphi(s,a,s')\|_2 \le L_\varphi$ for all $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$ and $\|\theta^*\|_2 \le L_\theta$.*

Assumption 1 is standard in contextual bandits and RL with function approximation. Let $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le L_\theta\}$.

Assumption 2. *There exists $\kappa \in (0,1)$ such that we have $\inf_{\theta\in\Theta} p_{t,s'}(\theta)p_{t,s''}(\theta) \ge \kappa$ for all $t \in [T]$ and $s', s'' \in \mathcal{S}_{s_t,a_t}$.*

Assumption 2 is also common in the generalized linear contextual bandit literature (Filippi et al., 2010a; Li et al., 2017; Oh and Iyengar, 2019; Kveton et al., 2020; Russac et al., 2020) and is taken for RL with MNL function approximation (Hwang and Oh, 2023; Li et al., 2024; Cho et al., 2024). It guarantees that the associated Fisher information matrix of the log-likelihood function in our setting is non-singular.

Assumption 3. *For every $(s,a) \in \mathcal{S}\times\mathcal{A}$, there exists $s' \in \mathcal{S}_{s,a}$ such that $\varphi(s,a,s') = 0$.*

In fact, we may impose Assumption 3 without loss of generality, by the following procedure. For a given pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, we take an arbitrary $s' \in \mathcal{S}_{s,a}$ and replace $\varphi(s,a,s'')$ by $\varphi(s,a,s'') - \varphi(s,a,s')$ for all $s'' \in \mathcal{S}_{s,a}$. Note that $\|\varphi(s,a,s'') - \varphi(s,a,s')\|_2 \le 2L_\varphi$ and the probability term $p(s' \mid s,a,\theta)$ remains the same. Therefore, up to doubling the parameter L_φ, Assumptions 1 and 2 remain valid even after the procedure to enforce Assumption 3.

3 ALGORITHM AND REGRET BOUNDS

In this section, we present our algorithm, upper-confidence multinomial logit kernel reinforcement learning (`UCMNLK` described by Algorithm 1). `UCMNLK` runs extended value iteration on a discounted-reward MDP. For the average-reward MDP, we approximate it by a discounted-reward MDP. `UCMNLK` is inspired by `UCLK` by Zhou et al. (2021b) for infinite-horizon discounted-reward linear mixture MDPs. In contrast to `UCLK`, however, `UCMNLK` optimizes over the transition probability p, not the parameter vector θ. This is because for our MNL function approximation framework, optimization over θ is a non-convex problem while optimizing over p is a linear program.

We construct certain confidence polytopes for the true transition probability function based on the recent online Newton method-based technique for estimating the transition probability vector due to (Zhang and Sugiyama, 2023; Lee and Oh, 2024b; Cho et al., 2024; Li et al., 2024), explained in Section 3.1.

3.1 Confidence Polytope for the True Transition Function

For simplicity, we use shorthand notation $\mathcal{S}_t$ for $\mathcal{S}_{s_t,a_t}$ for $t \in [T]$. We define the transition response variable $y_{t,s'} := \mathbf{1}\{s_{t+1} = s'\}$ for $t \in [T]$ and $s' \in \mathcal{S}_t$. Here, $y_{t,s'}$ basically corresponds to a sample from the multinomial distribution over $\mathcal{S}_t$ with probability $p(s'|s_t,a_t)$. Next, we introduce notation $p_{t,s'}(\theta)$ to denote $p_{t,s'}(\theta) = p(s' \mid s_t,a_t,\theta)$. Then we have $p(s' \mid s,a,\theta^*) = p(s' \mid s,a)$ and $p_{t,s'}(\theta^*) = p(s' \mid s_t,a_t)$. For each time step $t \in [T]$, we consider a per-time loss function, its gradient, and its Hessian given by

$$\begin{aligned}
\ell_t(\theta) &= -\sum_{s'\in\mathcal{S}_t} y_{t,s'} \log p_{t,s'}(\theta),\\
\nabla_\theta\left(\ell_t(\theta)\right) &= -\sum_{s'\in\mathcal{S}_t} \left(y_{t,s'} - p_{t,s'}(\theta)\right)\varphi_{t,s'},\\
\nabla^2_\theta(\ell_t(\theta)) &= \sum_{s'\in\mathcal{S}_t} p_{t,s'}(\theta)\varphi_{t,s'}\varphi_{t,s'}^\top\\
&\quad - \sum_{s'\in\mathcal{S}_t}\sum_{s''\in\mathcal{S}_t} p_{t,s'}(\theta)p_{t,s''}(\theta)\varphi_{t,s'}\varphi_{t,s''}^\top,
\end{aligned} \tag{3}$$

respectively. We show in Appendix B.1 that the Hessian $\nabla^2_\theta(\ell_t(\theta))$ is positive semidefinite for any $\theta \in \Theta$ under Assumption 2. Motivated by recent progress on online learning frameworks for multinomial logistic bandit (Zhang and Sugiyama, 2023), multinomial logit contextual bandit (Lee and Oh, 2024b), and RL with MNL approximation (Li et al., 2024; Cho et al., 2024), we apply the following online algorithm to estimate the true transition core θ^*. We start with $\widehat{\theta}_1 = 0$. At time step $t \in [T]$, given $\widehat{\theta}_1, \ldots, \widehat{\theta}_t$, we prepare

$$\Sigma_t = \lambda I_d + \sum_{i=1}^{t-1} \nabla^2_\theta(\ell_i(\widehat{\theta}_{i+1})),\ \widehat{\Sigma}_t = \Sigma_t + \eta\nabla^2_\theta(\ell_t(\widehat{\theta}_t)) \tag{4}$$

where η is a step size and I_d is the $d \times d$ identity matrix. Note that $\widehat{\Sigma}_t$ is positive definite. Then we set $\widehat{\theta}_{t+1}$ to

$$\operatorname*{argmin}_{\theta\in\Theta} \left\{ \nabla_\theta(\ell_t(\widehat{\theta}_t))^\top(\theta - \widehat{\theta}_t) + \frac{1}{2\eta}\|\theta - \widehat{\theta}_t\|^2_{\widehat{\Sigma}_t} \right\}. \quad (5)$$

As (5) can be viewed as an online mirror descent step with the associated Bregman divergence given by $\|\theta - \vartheta\|^2_{\widehat{\Sigma}_t}/2$, we may compute $\widehat{\theta}_{t+1}$ as follows.

$$\widehat{\theta}_{t+1} = \operatorname*{argmin}_{\theta\in\Theta} \|\theta - (\,\widehat{\theta}_t - \eta\widehat{\Sigma}_t^{-1}\nabla_\theta(\ell_t(\widehat{\theta}_t))\,)\|_{\widehat{\Sigma}_t}$$

The following lemma provides confidence ellipsoids for estimating the transition core θ^*.

Lemma 1. *Suppose that Assumptions 1–3 hold. Let $\delta \in (0,1)$, $\eta = (1/2)\log\mathcal{U} + (L_\theta L_\varphi + 1)$, and $\lambda \geq 84\sqrt{2}(L_\theta L_\varphi^3 + dL_\varphi^2)\eta$. With probability at least $1-\delta$, θ^* is contained in*

$$\mathcal{C}_t := \left\{\theta \in \Theta : \|\widehat{\theta}_t - \theta^*\|_{\Sigma_t} \leq \beta_t\right\} \quad (6)$$

where $\beta_t = f(L_\theta, L_\varphi)\sqrt{d}(\log(\mathcal{U}t/\delta))^2$ for every $t \in [T]$ and f is a polynomial in L_θ, L_φ.

Based on Lemma 1, we may construct confidence sets for the true transition function. Let $p^* \in \mathbb{R}^{S\times A\times S}$ denote a vector representation of the true transition function. That is, the coordinate $p^*_{s,a,s'}$ of p^* corresponding to $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$ equals $p(s' \mid s,a,\theta^*)$.

Lemma 2. *Suppose that $\theta^* \in \mathcal{C}_t$ where $\mathcal{C}_t$ is defined as in (6). Let $p^* \in \mathbb{R}^{S\times A\times S}$ be the vector representation of the true transition function. Then for $t \in [T]$,*

$$p^* \in \mathcal{P}_t := \left\{p \in [0,1]^{S\times A\times S} : p \text{ satisfies } (8), (9)\right\} \quad (7)$$

where

$$\sum_{s'\in\mathcal{S}_{s,a}} p_{s,a,s'} = 1, \quad (8)$$

$$\sum_{s'\in\mathcal{S}_{s,a}} \left|p_{s,a,s'} - p(s' \mid s,a,\widehat{\theta}_t)\right| \leq B^{1,t}_{s,a} + B^{2,t}_{s,a} \quad (9)$$

with $B^{1,t}_{s,a} = \beta_t \sum_{s'\in\mathcal{S}_{s,a}} p(s' \mid s,a,\widehat{\theta}_t)\|\varphi(s,a,s') - \sum_{s''\in\mathcal{S}_{s,a}} p(s' \mid s,a,\widehat{\theta}_t)\varphi(s,a,s'')\|_{\Sigma_t^{-1}}$ and $B^{2,t}_{s,a} = 3\beta_t^2 \max_{s'\in\mathcal{S}_{s,a}} \|\varphi(s,a,s')\|^2_{\Sigma_t^{-1}}$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$.

Note that $\mathcal{P}_t$ defined in Lemma 2 is a polytope. Hence, one can efficiently optimize a linear function over $\mathcal{P}_t$.

Remark 1. *UCLK (Zhou et al., 2021b) and UCRL2-VTR (Wu et al., 2022) optimize directly over the parameter vector θ for extended value iteration. Although optimization over θ for linear mixture MDPs can be done by a linear program, the MNL case requires maximizing logistic functions, which is essentially a non-convex problem. Here, Lemma 2 lets us avoid the challenge of solving a non-convex optimization for value iteration. Let us elaborate on this as follows. Extended value iteration in UCMNLK optimizes over distributions p, not the parameter vector θ. For this, we need a confidence set for the true transition probability distribution p directly. Hence, we first prove Lemma 1 which provides a confidence set for θ. Using this, we prove Lemma 2 that characterizes a confidence set for p. The important part is that the confidence set for distribution p can be expressed as a polytope, optimizing which can be done by linear programming. With this technique, we may keep the extended value iteration procedure within UCMNLK computationally tractable.*

Algorithm 1 Upper-Confidence Multinomial Logit Kernel Reinforcement Learning (`UCMNLK`)

```
Input: feature map $\varphi : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}^d$, confidence level $\delta \in (0,1)$, discount factor $\gamma \in [0,1)$, number of rounds $N$, and parameters $\lambda, L_\varphi, L_\theta, \kappa, \mathcal{U}$
Initialize: $t = 1$, $\widehat{\theta}_1 = 0$, $\Sigma_1 = \lambda I_d$, and observe the initial state $s_1 \in \mathcal{S}$
for episodes $k = 1, 2, \ldots,$ do
  Set $t_k = t$
  Set $Q_k$ as the output of DEVI$(\gamma, \mathcal{P}_{t_k}, N)$ where $\mathcal{P}_{t_k}$ is given as in (7)
  Take a deterministic policy $\pi_k$ by taking $\pi_k(s) \in \operatorname{argmax}_{a\in\mathcal{A}} Q_k(s,a)$ for $s \in \mathcal{S}$
  while $\det(\Sigma_t) \leq 2\det(\Sigma_{t_k})$ do
    Take action $a_t = \pi_k(s_t)$
    Observe $s_{t+1}$ sampled from $p(\cdot \mid s_t, a_t)$
    Compute $\widehat{\theta}_{t+1}$ as in (5)
    Set $\Sigma_{t+1} = \Sigma_t + \nabla^2_\theta(\ell_t(\widehat{\theta}_{t+1}))$ as in (4)
    Update $t \leftarrow t + 1$
  end while
end for
```

3.2 Algorithm Description of UCMNLK

Algorithm 1 describes UCMNLK. As UCRL2-VTR (Wu et al., 2022) and UCLK (Zhou et al., 2021b), UCMNLK proceeds with multiple episodes. Each episode consists of the planning phase and the execution phase.

In the planning phase, UCMNLK computes an optimistic policy by running discounted extended value iteration (DEVI) as follows. For the kth episode, we denote by t_k the first time step of episode k. Before episode k begins, we construct $\widehat{\theta}_{t_k}$ and $\mathcal{C}_{t_k}$ for estimating θ^* based on (5) and (6). Then we prepare the confidence polytope $\mathcal{P}_{t_k}$ according to (7), over which we run DEVI. Lastly, based on the action-value function Q_k returned by DEVI, we deduce a greedy policy π_k.

In the execution phase, UCMNLK applies the policy π_k

Discounted Extended Value Iteration ($\texttt{DEVI}(\gamma, \mathcal{P}, N)$)

```
Inputs: discount factor $\gamma$, number of rounds $N$, confidence polytope $\mathcal{P}$
Initialize: $Q^{(0)}(s,a) = (1-\gamma)^{-1}$ for $(s,a) \in \mathcal{S} \times \mathcal{A}$
for rounds $n = 1, 2, \ldots, N$ do
    Set $V^{(n-1)}(s) = \max_{a \in \mathcal{A}} Q^{(n-1)}(s,a)$ for $s \in \mathcal{S}$
    For $(s,a) \in \mathcal{S} \times \mathcal{A}$, set
    $Q^{(n)}(s,a) = r(s,a) + \gamma \max_{p \in \mathcal{P}} \left\{ \sum_{s' \in \mathcal{S}_{s,a}} p_{s,a,s'} V^{(n-1)}(s') \right\}$
end for
Return $Q^{(N)}(s,a)$ for $(s,a) \in \mathcal{S} \times \mathcal{A}$
```

and receives a trajectory with corresponding rewards. Note that UCMNLK switches to the next episode when the determinant of the matrix Σ_t doubles compared to the beginning of episode k.

Note that in one round of extended value iteration, we optimize over the probability distributions p in the confidence polytope $\mathcal{P}_{t_k}$. In our case, optimizing over θ requires maximizing the sum of multinomial logit functions, which is a non-convex optimization problem. Instead, we maximize over probability distribution p, which boils down to solving a linear program.

Remark 2. *UCMNLK is a single unified framework for both discounted-reward and average-reward settings. For the average-reward setting, we approximate the given average-reward MDP by a discounted-reward MDP and show a novel regret decomposition lemma for UCMNLK. Hong et al. (2025) also propose an idea of approximating an average-reward MDP by a discounted-reward MDP for the linear MDP case. However, value iteration for the linear MDP setting does not converge. In contrast, we argue that value iteration within our algorithm converges, so we may use a fixed value function V_k for each episode k. As a result, our regret decomposition result differs from theirs.*

Computational Efficiency The total number of iterations of extended value iteration is $N \times K$ where $N = \mathcal{O}(\sqrt{T}\log(T))$ and $K = (d\log(T))$, while one iteration requires solving a linear program whose size is polynomial in S and A. We also remark that one iteration of the online Newton method has a time complexity polynomial in d, S, and A. Hence, the computational complexity of UCMNLK is linear in T and polynomial in d, S, A. Although the complexity of our algorithm indeed depends on the maximum number $\mathcal{U}$ of reachable states, the existing model-based reinforcement learning approaches cannot avoid a dependency on the number S of states in their computational complexity. This is inherent in model-based algorithms, which also include linear function approximation frameworks. At the same time, $\mathcal{U}$ can be much smaller than S, as shown in many examples by Lee and Oh (2024a).

3.3 Regret Analysis of UCMNLK

The following results state our regret upper bounds of UCMNLK for the average-reward and the discounted-reward settings.

Theorem 1 (Average-Reward). *Let M be an average-reward MDP governed by the model (2) with diameter at most D. Let $\delta \in (0,1)$, $\eta = (1/2)\log\mathcal{U} + (L_\theta L_\varphi + 1)$, $\lambda \geq 84\sqrt{2}(L_\theta L_\varphi^3 + dL_\varphi^2)\eta$, $\gamma = 1 - \sqrt{d/DT}$, and $N \geq \sqrt{DT/d}\log(\sqrt{T}/dD)$. Then UCMNLK guarantees that with probability at least $1-2\delta$,*

$$\mathrm{Regret}(T) = \tilde{\mathcal{O}}\left(f(L_\theta, L_\varphi)\left(dD\sqrt{T} + \kappa^{-1}d^2D\right)\right)$$

where f is a polynomial in (L_θ, L_φ) and $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic factors of T, $\mathcal{U}$, and $1/\delta$.

Theorem 2 (Discounted-Reward). *Let M be a discounted-reward MDP governed by (2). Let $\delta \in (0,1)$, $\eta = (1/2)\log\mathcal{U} + (L_\theta L_\varphi + 1)$, $\lambda \geq 84\sqrt{2}(L_\theta L_\varphi^3 + dL_\varphi^2)\eta$, and $N \geq \log(\sqrt{T}/d)/(1-\gamma)$. Then UCMNLK guarantees that with probability at least $1-2\delta$,*

$$\begin{aligned}&\mathrm{Regret}(\pi, T)\\ &= \tilde{\mathcal{O}}\left(f(L_\theta, L_\varphi)\left(d(1-\gamma)^{-2}\sqrt{T} + \kappa^{-1}d^2(1-\gamma)^{-2}\right)\right)\end{aligned}$$

where f is a polynomial in (L_θ, L_φ) and $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic factors of T, $\mathcal{U}$, and $1/\delta$.

The remainder of this subsection provides a brief overview of our proofs for the regret upper bounds.

We denote by V_k the value function for the kth episode of Algorithm 1 given by $V_k(s) = \max_{a\in\mathcal{A}} Q_k(s,a)$ for $s \in \mathcal{S}$. We prove the following lemma establishing convergence of DEVI.

Lemma 3. *Suppose that $\theta^* \in \mathcal{C}_t$ for $t \in [T]$ where $\mathcal{C}_t$ is defined as in (6). Then for each episode k and $t_k \leq t < t_{k+1} - 1$, it holds that*

$$\begin{aligned}&Q_k(s_t, a_t)\\ &\leq r(s_t, a_t) + \gamma \max_{p \in \mathcal{P}_{t_k}} \left\{ \sum_{s' \in \mathcal{S}_t} p_{s_t, a_t, s'} V_k(s') \right\} + \gamma^N.\end{aligned}$$

Let K_T denote the total number of distinct episodes over the horizon of T time steps. For simplicity, we assume that the last time step of the last episode and that time step $T+1$ is the beginning of the $(K_T + 1)$th episode, i.e., $t_{K_T+1} = T + 1$. Then it follows from Lemma 3 that the regret function for the average-reward case satisfies the following.

$$\begin{aligned}&\text{Regret}(T) = T \cdot J^* - \sum_{t=1}^{T} r(s_t, a_t) \leq \\ &T\gamma^N + \underbrace{\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left(J^* - (1-\gamma)V_k(s_{t+1})\right)}_{(a)} + \\ &\underbrace{\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left(V_k(s_{t+1}) - Q_k(s_t, a_t)\right)}_{(b)} + \\ &\underbrace{\gamma \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left(\sum_{s' \in \mathcal{S}_t} p^*_{s_t,a_t,s'} V_k(s') - V_k(s_{t+1})\right)}_{(c)} + \\ &\underbrace{\gamma \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \max_{p \in \mathcal{P}_{t_k}} \left\{\sum_{s' \in \mathcal{S}_t} \left(p_{s_t,a_t,s'} - p^*_{s_t,a_t,s'}\right) V_k(s')\right\}}_{(d)}.\end{aligned}$$

For regret term (a), recall that V^* and Q^* are the optimal value function and the optimal action-value function for the discounted-reward setting with discount factor γ. The following lemma proves that V_k and Q_k are optimistic estimators of V^* and Q^*, respectively.

Lemma 4. *Suppose that $\theta^* \in \mathcal{C}_t$ for $t \in [T]$ where $\mathcal{C}_t$ is defined as in (6). Then for each episode k, $1/(1-\gamma) \geq V_k(s) \geq V^*(s)$ and $1/(1-\gamma) \geq Q_k(s,a) \geq Q^*(s,a)$.*

Lemma 4 implies that $J^* - (1-\gamma)V_k(s_{t+1}) \leq J^* - (1-\gamma)V^*(s_{t+1})$. Then we apply (Lemma 2, Wei et al., 2020) to argue that $J^* - (1-\gamma)V^*(s_{t+1}) \leq (1-\gamma)D$. For regret term (b), note that $V_k(s_{t+1}) = Q_k(s_{t+1}, a_{t+1})$ for $t \in [t_k : t_{k+1} - 2]$, which leads to a telescoping structure. For regret term (c), we first observe that $\sum_{s' \in \mathcal{S}_t} p^*_{s_t,a_t,s'} V_k(s') - V_k(s_{t+1})$ equals $\sum_{s' \in \mathcal{S}_t} p^*_{s_t,a_t,s'} W_k(s') - W_k(s_{t+1})$ where $W_k = V_k - \min_{s' \in \mathcal{S}} V_k(s')$. Then the following lemma implies that $W_k(s) \in [0, D]$ for any $s \in \mathcal{S}$.

Lemma 5. *Suppose that $\theta^* \in \mathcal{C}_t$ for $t \in [T]$ where $\mathcal{C}_t$ is defined as in (6). If the underlying MDP has diameter at most D, $\max_{s \in \mathcal{S}} V_k(s) - \min_{s \in \mathcal{S}} V_k(s) \leq D$ for each episode k.*

Remark 3. *Lemma 5 shows that the span of the "discounted" value function V_k for each episode is bounded above by the diameter D for the communicating case. We need this result to apply the Azuma-Hoeffding inequality for the third regret term, for which a global upper bound on the term $|V_k(s') - V_k(s_{t+1}|$ is necessary. Wu et al. (2022) shows that the span of their "average-reward"value functions is bounded above by the diameter D, but we need to control the span of our discounted value functions.*

Based on this, regret term (c) is the sum of a martingale difference sequence where each element has an absolute value at most D. Regret term (d) is the cumulative estimation error. Based on Lemmas 2 and 5, we can show that (d) is bounded above by $D \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k_1}-1} (B^{1,t_k}_{s_t,a_t} + B^{2,t_k}_{s_t,a_t}) = \tilde{\mathcal{O}}(dD\sqrt{T})$ which corresponds to the leading term in the regret upper bound of Theorem 1.

Remark 4. *For term (d), the major step is to analyze the sum $\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k_1}-1} B^{1,t_k}_{s_t,a_t}$ indeed. In contrast to Cho et al. (2024), our algorithm for infinite-horizon MNL MDPs runs with episodes, and we use a fixed policy for an episode k spanning time steps $t_k, \ldots, t_{k+1} - 1$. At the beginning of episode k, at $t = t_k$, we obtain a policy π_k based on the estimator $\widehat{\theta}_{t_k}$ and the confidence polytope $\mathcal{P}_{t_k}$. Then we use π_k computed at $t = t_k$ for multiple time steps $t_k, \ldots, t_{k+1} - 1$. This raises a challenge in analyzing the resulting estimation error terms. To be specific, we need to control*

$$\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k_1}-1} \beta_{t_k} \sum_{s' \in \mathcal{S}_{s_t,a_t}} p_{t,s'}(\widehat{\theta}_{t_k}) \times \left\| \varphi(s_t, a_t, s') - \sum_{s'' \in \mathcal{S}_{s_t,a_t}} p_{t,s''}(\widehat{\theta}_{t_k}) \varphi(s_t, a_t, s'') \right\|_{\Sigma^{-1}_{t_k}}$$

Here, what makes the analysis nontrivial is that $\widehat{\theta}_{t_{k+1}}$ for episode $k+1$ and $\widehat{\theta}_{t_k}$ for episode k can be arbitrarily far from each other if the gap $t_{k+1} - t_k$ is large. For the same reason, the distance between $\widehat{\theta}_{t+1}$ and $\widehat{\theta}_{t_k}$ can be large if $t - t_k$ is large. Nevertheless, we develop a remedy to get around this issue and provide an upper bound on the error term.

The discounted-reward case is similar to the average-reward case, and its analysis follows the analysis of UCLK due to Zhou et al. (2021b).

4 REGRET LOWER BOUNDS

In this section, we provide regret lower bounds for learning MDPs with MNL function approximation. Section 4.1 provides a lower bound for learning H-horizon episodic MDPs with distinct transition cores over the horizon. Section 4.2 presents lower bounds for learning infinite-horizon average-reward MDPs with diameter at most D and discounted-reward MDPs with discount factor γ.

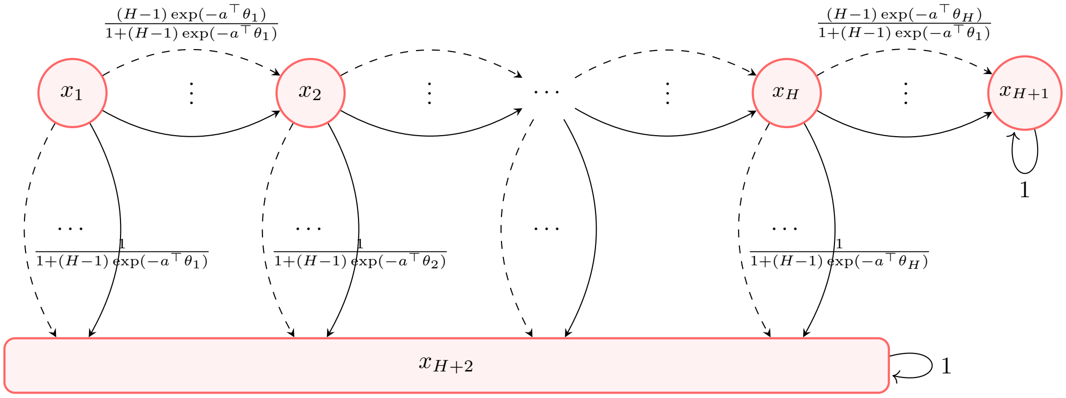
$\frac{(H-1)\exp(-a^\top\theta_1)}{1+(H-1)\exp(-a^\top\theta_1)}$
$\frac{(H-1)\exp(-a^\top\theta_H)}{1+(H-1)\exp(-a^\top\theta_1)}$
x_1
x_2
x_H
x_{H+1}
1
$\frac{1}{1+(H-1)\exp(-a^\top\theta_1)}$
$\frac{1}{1+(H-1)\exp(-a^\top\theta_2)}$
$\frac{1}{1+(H-1)\exp(-a^\top\theta_H)}$
x_{H+2}
1

Figure 1: Illustration of the Hard Finite-Horizon MDP Instance

4.1 Lower Bound for Learning Finite-Horizon Episodic MDPs

To provide a regret lower bound on learning finite-horizon MDPs with MNL approximation, we consider an instance inspired by Zhou et al. (2021a) illustrated as in Figure 1. There are $H+2$ states $x_1, \dots, x_{H+2}$ where x_{H+1} and x_{H+2} are absorbing states. We have action space $\mathcal{A} = \{-1, 1\}^{d-1}$. For any action $a \in \mathcal{A}^{d-1}$, the reward function is given by $f(x_i, a) = 1$ if $i = H+2$ and $f(x_i, a) = 0$ if $i \neq H+2$. The transition core $\bar{\theta}_h$ for each step $h \in [H]$ is given by

$$\bar{\theta}_h = \left(\frac{\theta_h}{\alpha}, \frac{1}{\beta}\right) \quad \text{where} \quad \theta_h \in \left\{-\bar{\Delta}, \bar{\Delta}\right\}^{d-1},$$

$$\bar{\Delta} = \frac{1}{d-1} \log\left(\frac{(1-\delta)(\delta + (d-1)\Delta)}{\delta(1-\delta-(d-1)\Delta)}\right),$$

with $\delta = 1/H$, $\Delta = 1/(4\sqrt{2HK})$, $\alpha = \sqrt{\bar{\Delta}/(1+(d-1)\bar{\Delta})}$, and $\beta = \sqrt{1/(1+(d-1)\bar{\Delta})}$. Moreover, the feature vector is given by $\varphi(x_h, a, x_{H+2}) = (0, 0)$ and $\varphi(x_h, a, x_{h+1}) = (-\alpha a, \beta \log(H-1))$ for $h \in [H]$. Here, we denote this MDP by M_θ to indicate that it is parameterized by $\theta = \{\theta_h\}_{h=1}^H$.

Theorem 3. *Suppose that $d \geq 2$, $H \geq 3$, $K \geq \{(d-1)^2 H/2, H^3(d-1)^2/32\}$. Then for any algorithm $\mathfrak{A}$, there exists an MDP M_θ described as in Figure 1 such that $L_\theta \leq 3/2$ and $L_\varphi \leq 1 + \log(H-1)$,*

$$\mathbb{E}\left[\mathrm{Regret}(M_\theta, \mathfrak{A}, K)\right] \geq \frac{(d-1)H^{3/2}\sqrt{K}}{480\sqrt{2}}$$

where the expectation is taken over the randomness generated by M_θ and $\mathfrak{A}$.

Recall that the lower bound provided by Li et al. (2024) is $\Omega(dH\sqrt{K\kappa^*})$ where κ^* is a constant satisfying $p_t(s', \theta^*)p_t(x'', \theta^*) \geq \kappa^*$ for all $t \in [T]$ and $s', s'' \in \mathcal{S}_{s_t, a_t}$. Hence, our lower bound from Theorem 3 improves the previous lower bound by a factor of $O(\sqrt{H/\kappa^*})$.

Notice that the instance M_θ has $L_\varphi \leq 1 + \log(H-1)$. Nonetheless, the regret upper bounds by Hwang and Oh (2023); Cho et al. (2024); Li et al. (2024) grow polynomially in L_φ, so the upper bounds remain the same up to logarithmic factors in $\log H$.

Let us briefly explain how the lower bound is derived. We consider a multinomial logit function given by $f : \mathbb{R} \to \mathbb{R}$ as

$$f(x) = \frac{1}{1 + (\delta^{-1} - 1)\exp(-x)}. \tag{10}$$

Then it follows that

$$p(x_i \mid x_h, a, \bar{\theta}_h) = \begin{cases} f(a^\top \theta_h), & \text{if } i = H+2 \\ 1 - f(a^\top \theta_h), & \text{if } i = h+1 \end{cases}$$

with $-(d-1)\bar{\Delta} \leq a^\top \theta_h \leq (d-1)\bar{\Delta}$ for any $a \in \mathcal{A}$. One of the main steps to derive the lower bound is to construct an upper bound on the gap between $p(x_i \mid x_h, a, \bar{\theta})$ and $p(x_i \mid x_h, a, \bar{\theta}')$ for $\bar{\theta} \neq \bar{\theta}'$. We use the mean value theorem to argue that the gap is bounded above by $c^\top(\theta - \theta')$ for some $c \in \mathbb{R}^{d-1}$. To be more precise, we can show that for any $x, y \in [-(d-1)\bar{\Delta}, (d-1)\bar{\Delta}]$ with $x \geq y$, we have

$$0 \leq f(x) - f(y) \leq (\delta + (d-1)\Delta)(x - y).$$

This bridges the multinomial logit function to a linear function. Then we may reduce our analysis to the linear case, and therefore, we may follow some arguments of Zhou et al. (2021a).

4.2 Lower Bounds for Learning Infinite-Horizon MDPs

In this section, we prove regret lower bounds for learning communicating MDPs of diameter at most D and

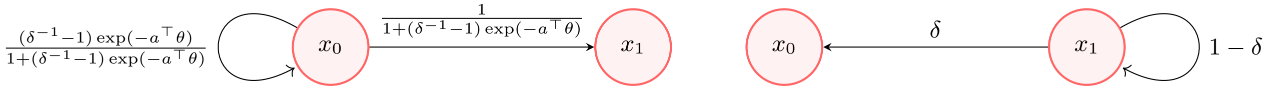
$\frac{(\delta^{-1}-1)\exp(-a^\top\theta)}{1+(\delta^{-1}-1)\exp(-a^\top\theta)}$
x_0
$\frac{1}{1+(\delta^{-1}-1)\exp(-a^\top\theta)}$
x_1
x_0
δ
x_1
$1-\delta$

Figure 2: Illustration of the Hard-to-Learn Infinite-Horizon MDP Instance

discounted-reward MDPs with discount factor γ. Our construction of the following hard-to-learn MDP is motivated by the instance proposed by Wu et al. (2022) for the linear mixture MDP case. There are two states x_0 and x_1 as in Figure 2. The action space is given by $\mathcal{A} = \{-1, 1\}^{d-1}$. Let the reward function be given by $r(x_0, a) = 0$ and $r(x_1, a) = 1$ for any $a \in \mathcal{A}$. Then a higher stationary probability at state x_1 means a larger average reward. We set the transition core $\bar{\theta}$ as

$$\bar{\theta} = \left(\frac{\theta}{\alpha}, \frac{1}{\beta}\right) \quad \text{where} \quad \theta \in \left\{-\frac{\bar{\Delta}}{d-1}, \frac{\bar{\Delta}}{d-1}\right\}^{d-1},$$

$$\bar{\Delta} = \log\left(\frac{(1-\delta)(\delta+\Delta)}{\delta(1-\delta-\Delta)}\right)$$

with $\Delta = (d-1)/(45\sqrt{(2/5)(T/\delta)\log 2})$,

$$\delta = \begin{cases} 1/D & \text{for the average-reward case,} \\ 1-\gamma & \text{for the discounted-reward case,} \end{cases}$$

$\alpha = \sqrt{\bar{\Delta}/((d-1)(1+\bar{\Delta}))}$, and $\beta = \sqrt{1/(1+\bar{\Delta})}$. The feature vector is given by $\varphi(x_0, a, x_0) = (-\alpha a, \beta\log(\delta^{-1}-1))$, $\varphi(x_0, a, x_1) = \varphi(x_1, a, x_0) = (0, 0)$, and $\varphi(x_1, a, x_1) = (0, \beta\log(\delta^{-1}-1))$. We denote this MDP by M_θ to show its dependence on θ.

Theorem 4. *Suppose that $d \geq 2$, $D \geq 101$, $T \geq 45(d-1)^2 D$. Then for any algorithm $\mathfrak{A}$, there exists an MDP M_θ described as in Figure 2 such that $L_\theta \leq 100/99$ and $L_\varphi \leq 1 + \log(D-1)$,*

$$\mathbb{E}\left[\text{Regret}(M_\theta, \mathfrak{A}, x_0, T)\right] \geq \frac{1}{2025} d\sqrt{DT}$$

where the expectation is taken over the randomness generated by M_θ and $\mathfrak{A}$.

Theorem 5. *Suppose that $d \geq 2$, $\gamma \geq 100/101$, $T \geq 45(d-1)^2/(1-\gamma)$. Then for any policy π, there exists an MDP M_θ described as in Figure 2 such that $L_\theta \leq 100/99$ and $L_\varphi \leq 1 + \log(\gamma/(1-\gamma))$,*

$$\mathbb{E}\left[\text{Regret}(M_\theta, \mathfrak{A}, x_0, T)\right] \geq \frac{\gamma}{3375(1-\gamma)^{3/2}} d\sqrt{T} - \frac{\gamma}{(1-\gamma)^2}$$

where the expectation is taken over the randomness generated by M_θ and π.

Note that L_φ can grow logarithmically in δ^{-1}, which equals D for the average-reward case and $(1-\gamma)^{-1}$ for the discounted-reward setting. Nevertheless, the upper bounds by Theorems 1 and 2 grow polynomially in L_φ. This means that The regret upper bounds on the hard-to-learn MDP remain the same up to additional logarithmic factors in δ^{-1}.

As for the finite-horizon case, our main technique is to connect the MNL function approximation model and linear mixture MDPs. With the multinomial logit function given in (10), we know that $p(x_1 \mid x_0, a, \bar{\theta}) = f(a^\top\theta)$ and $p(x_0 \mid x_1, a, \bar{\theta}) = f(0)$. Moreover, for any $a \in \mathcal{A}$, we have $a^\top\theta \in [-\bar{\Delta}, \bar{\Delta}]$. Then we can also argue that for any $x, y \in [-\bar{\Delta}, \bar{\Delta}]$ with $x \geq y$, we have

$$0 \leq f(x) - f(y) \leq (\delta + \Delta)(x - y).$$

This lays down a bridge between our instance in Figure 2 and the lower bound instance of Wu et al. (2022).

5 CONCLUSION

This paper studies infinite-horizon reinforcement learning with multinomial logit function approximation. We develop an algorithm, `UCMNLK`, that works for both the average-reward and discounted-reward settings. We provide regret lower bounds for the two settings as well as the finite-horizon setting, which demonstrates that the algorithm achieves tight regret upper bounds. We provide a more comprehensive literature review on online learning of MDPs in the appendix.

Acknowledgements

We would like to thank the review team for their careful review and valuable feedback. This work was supported by the National Research Foundation of Korea (NRF) grant (No. RS-2024-00350703) and the Institute of Information & communications Technology Planning & evaluation (IITP) grant (No. RS-2024-00437268) funded by the Korea government (MSIT).

References

Abbasi-Yadkori, Y., Bartlett, P., Bhatia, K., Lazic, N., Szepesvari, C., and Weisz, G. (2019). POLITEX: Regret bounds for policy iteration using expert prediction. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3692–3702. PMLR.

Abbasi-yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. (2021). Reinforcement learning: Theory and algorithms.

Agarwal, A., Jin, Y., and Zhang, T. (2023). Voql: Towards optimal regret in model-free rl with nonlinear function approximation. In Neu, G. and Rosasco, L., editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 987–1063. PMLR.

Agrawal, S. and Jia, R. (2017). Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77.

Auer, P., Jaksch, T., and Ortner, R. (2008). Near-optimal regret bounds for reinforcement learning. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. (2020). Model-based reinforcement learning with value-targeted regression. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 463–474. PMLR.

Bartlett, P. L. and Tewari, A. (2009). Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 35–42, Arlington, Virginia, USA. AUAI Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag, Berlin, Heidelberg.

Boone, V. and Zhang, Z. (2024). Achieving tractable minimax optimal regret in average reward mdps.

Bourel, H., Maillard, O., and Talebi, M. S. (2020). Tightening exploration in upper confidence reinforcement learning. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1056–1066. PMLR.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. (2020). Provably efficient exploration in policy optimization. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1283–1294. PMLR.

Caro, F. and Gallien, J. (2007). Dynamic assortment with demand learning for seasonal consumer goods. *Management Science*, 53(2):276–292.

Cho, W., Hwang, T., Lee, J., and hwan Oh, M. (2024). Randomized exploration for reinforcement learning with multinomial logistic function approximation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems.*

Dann, C., Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. (2018). On oracle-efficient pac rl with rich observations. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Du, S., Kakade, S., Lee, J., Lovett, S., Mahajan, G., Sun, W., and Wang, R. (2021). Bilinear classes: A structural framework for provable generalization in rl. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2826–2836. PMLR.

Fan, J., Wang, Z., Xie, Y., and Yang, Z. (2020). A theoretical analysis of deep q-learning. In Bayen, A. M., Jadbabaie, A., Pappas, G., Parrilo, P. A., Recht, B., Tomlin, C., and Zeilinger, M., editors, *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 486–489. PMLR.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. (2010a). Parametric bandits: The generalized linear case. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

Filippi, S., Cappé, O., and Garivier, A. (2010b). Optimism in reinforcement learning and kullback-leibler divergence. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 115–122.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. (2023). The statistical complexity of interactive decision making.

Fruit, R., Pirotta, M., and Lazaric, A. (2020). Improved analysis of ucrl2 with empirical bernstein inequality.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. (2018). Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1578–1586. PMLR.

He, J., Zhao, H., Zhou, D., and Gu, Q. (2023). Nearly minimax optimal reinforcement learning for linear Markov decision processes. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 12790–12822. PMLR.

He, J., Zhong, H., and Yang, Z. (2024). Sample-efficient learning of infinite-horizon average-reward MDPs with general function approximation. In *The Twelfth International Conference on Learning Representations*.

He, J., Zhou, D., and Gu, Q. (2021). Logarithmic regret for reinforcement learning with linear function approximation. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4171–4180. PMLR.

Hong, K., Chae, W., Zhang, Y., Lee, D., and Tewari, A. (2025). Reinforcement learning for infinite-horizon average-reward linear MDPs via approximation by discounted-reward MDPs. In *The 28th International Conference on Artificial Intelligence and Statistics*.

Hu, P., Chen, Y., and Huang, L. (2022). Nearly minimax optimal reinforcement learning with linear function approximation. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8971–9019. PMLR.

Hwang, T. and Oh, M.-h. (2023). Model-based reinforcement learning with multinomial logistic function approximation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):7971–7979.

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600.

Jia, Z., Yang, L., Szepesvari, C., and Wang, M. (2020). Model-based reinforcement learning with value-targeted regression. In Bayen, A. M., Jadbabaie, A., Pappas, G., Parrilo, P. A., Recht, B., Tomlin, C., and Zeilinger, M., editors, *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 666–686. PMLR.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. (2017). Contextual decision processes with low Bellman rank are PAC-learnable. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713. PMLR.

Jin, C., Liu, Q., and Miryoosefi, S. (2021). Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In Abernethy, J. and Agarwal, S., editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR.

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

Kveton, B., Szepesvári, C., Ghavamzadeh, M., and Boutilier, C. (2020). Perturbed-history exploration in stochastic linear bandits. In Adams, R. P. and Gogate, V., editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 530–540. PMLR.

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

Lee, J. and Oh, M.-h. (2024b). Nearly minimax optimal regret for multinomial logistic bandit. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Li, L., Chu, W., Langford, J., Moon, T., and Wang, X. (2012). An unbiased offline evaluation of contextual bandit algorithms with generalized linear models. In Glowacka, D., Dorard, L., and Shawe-Taylor, J., editors, *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*, volume 26 of *Proceedings of Machine Learning Research*, pages 19–36, Bellevue, Washington, USA. PMLR.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 661–670, New York, NY, USA. Association for Computing Machinery.

Li, L., Lu, Y., and Zhou, D. (2017). Provably optimal algorithms for generalized linear contextual bandits. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2071–2080. PMLR.

Li, L.-F., Zhang, Y.-J., Zhao, P., and Zhou, Z.-H. (2024). Provably efficient reinforcement learning with multinomial logit function approximation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Liu, F., Viano, L., and Cevher, V. (2022). Understanding deep neural function approximation in reinforcement learning via ϵ-greedy exploration. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 5093–5108. Curran Associates, Inc.

Liu, S. and Su, H. (2021). Regret bounds for discounted mdps.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Modi, A., Jiang, N., Tewari, A., and Singh, S. (2020). Sample complexity of reinforcement learning using linearly combined model ensembles. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2010–2020. PMLR.

Oh, M.-h. and Iyengar, G. (2019). Thompson sampling for multinomial logit contextual bandits. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R. (2017). Learning unknown markov decision processes: A thompson sampling approach. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

Russac, Y., Cappé, O., and Garivier, A. (2020). Algorithms for non-stationary generalized linear bandits.

Russo, D. and Van Roy, B. (2013). Eluder dimension and the sample complexity of optimistic exploration. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. (2019). Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In Beygelzimer, A. and Hsu, D., editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2898–2933. PMLR.

Talebi, M. S. and Maillard, O.-A. (2018). Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In Janoos, F., Mohri, M., and Sridharan, K., editors, *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 770–805. PMLR.

Wang, R., Salakhutdinov, R. R., and Yang, L. (2020). Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6123–6135. Curran Associates, Inc.

Wang, Y., Wang, R., Du, S. S., and Krishnamurthy, A. (2021). Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*.

Wei, C.-Y., Jafarnia Jahromi, M., Luo, H., and Jain, R. (2021). Learning infinite-horizon average-reward mdps with linear function approximation. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings* *of Machine Learning Research*, pages 3007–3015. PMLR.

Wei, C.-Y., Jahromi, M. J., Luo, H., Sharma, H., and Jain, R. (2020). Model-free reinforcement learning in infinite-horizon average-reward Markov decision processes. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10170–10180. PMLR.

Weisz, G., Amortila, P., and Szepesvári, C. (2021). Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In Feldman, V., Ligett, K., and Sabato, S., editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 1237–1264. PMLR.

Wu, Y., Zhou, D., and Gu, Q. (2022). Nearly minimax optimal regret for learning infinite-horizon average-reward mdps with linear function approximation. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3883–3913. PMLR.

Xu, P. and Gu, Q. (2020). A finite-time analysis of q-learning with neural network function approximation. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10555–10565. PMLR.

Yang, L. and Wang, M. (2019). Sample-optimal parametric q-learning using linearly additive features. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6995–7004. PMLR.

Yang, L. and Wang, M. (2020). Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10746–10756. PMLR.

Yang, Z., Jin, C., Wang, Z., Wang, M., and Jordan, M. (2020). On function approximation in reinforcement learning: Optimism in the face of large state spaces. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13903–13916. Curran Associates, Inc.

Yurtsever, E., Lambert, J., Carballo, A., and Takeda, K. (2020). A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8:58443–58469.

Zanette, A., Brandfonbrener, D., Brunskill, E., Pirotta, M., and Lazaric, A. (2020a). Frequentist regret bounds for randomized least-squares value iteration. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1954–1964. PMLR.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. (2020b). Learning near optimal policies with low inherent Bellman error. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10978–10989. PMLR.

Zhang, S., Li, H., Wang, M., Liu, M., Chen, P.-Y., Lu, S., Liu, S., Murugesan, K., and Chaudhury, S. (2023). On the convergence and sample complexity analysis of deep q-networks with ϵ-greedy exploration. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhang, Y.-J. and Sugiyama, M. (2023). Online (multinomial) logistic bandit: Improved regret and constant computation cost. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 29741–29782. Curran Associates, Inc.

Zhang, Z. and Ji, X. (2019). Regret minimization for reinforcement learning by evaluating the optimal bias function. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zhang, Z. and Xie, Q. (2023). Sharper model-free reinforcement learning for average-reward markov decision processes. In Neu, G. and Rosasco, L., editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5476–5477. PMLR.

Zhong, H., Xiong, W., Zheng, S., Wang, L., Wang, Z., Yang, Z., and Zhang, T. (2023). Gec: A unified framework for interactive decision making in mdp, pomdp, and beyond.

Zhou, D. and Gu, Q. (2022). Computationally efficient horizon-free reinforcement learning for linear mixture mdps. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 36337–36349. Curran Associates, Inc.

Zhou, D., Gu, Q., and Szepesvari, C. (2021a). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In Belkin, M. and Kpotufe, S., editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4532–4576. PMLR.

Zhou, D., He, J., and Gu, Q. (2021b). Provably efficient reinforcement learning for discounted mdps with feature mapping. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12793–12802. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes, in the appendix]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A RELATED WORK

In this section, we provide a more detailed discussion of related work for learning infinite-horizon average-reward Markov decision processes (MDPs).

Infinite-Horizon Average-Reward Tabular MDPs Auer et al. (2008) initiated the study of online learning of MDPs. They developed an algorithm, UCRL2, that guarantees a regret upper bound of $\widetilde{\mathcal{O}}(DS\sqrt{AT})$ over T time steps where D is the diameter of the underlying MDP, S is the number of states and A is the number of actions. They also provided a regret lower bound of $\Omega(\sqrt{DSAT})$, which shows that UCRL2 is nearly optimal. Then Bartlett and Tewari (2009) considered the class of weakly communicating MDPs. For the class, they developed an algorithm, REGAL, that guarantees a regret upper bound of $\widetilde{\mathcal{O}}(\mathrm{sp}(v^*)S\sqrt{AT})$ where $\mathrm{sp}(v^*)$ is the span of the optimal associated bias function. After these works, there has been a flurry of activities for closing the gap between regret upper and lower bounds (Filippi et al., 2010b; Talebi and Maillard, 2018; Fruit et al., 2018, 2020; Bourel et al., 2020; Zhang and Ji, 2019; Agrawal and Jia, 2017; Ouyang et al., 2017; Abbasi-Yadkori et al., 2019; Wei et al., 2021; Zhang and Xie, 2023; Boone and Zhang, 2024). In particular, Zhang and Ji (2019) developed a nearly minimax optimal algorithm with a regret upper bound of $\widetilde{\mathcal{O}}(\sqrt{\mathrm{sp}(v^*)SAT})$ based on the framework of Fruit et al. (2020), but their algorithm is not computationally efficient. Recently, Boone and Zhang (2024) proposed a provably efficient algorithm with a nearly minimax optimal regret upper bound of $\widetilde{\mathcal{O}}(\sqrt{\mathrm{sp}(v^*)SAT})$.

Reinforcement Learning with Linear Function Approximation There has been notable progress recently in reinforcement learning frameworks that leverage linear function approximations (Jiang et al., 2017; Yang and Wang, 2019, 2020; Jin et al., 2020; Wang et al., 2021; Modi et al., 2020; Dann et al., 2018; Du et al., 2021; Sun et al., 2019; Zanette et al., 2020a,b; Cai et al., 2020; Jia et al., 2020; Ayoub et al., 2020; Weisz et al., 2021; Zhou et al., 2021b,a; He et al., 2021; Zhou and Gu, 2022; Hu et al., 2022; He et al., 2023; Agarwal et al., 2023; Hong et al., 2025). Among these works, the most relevant to our paper are infinite-horizon average-reward linear MDPs and linear mixture MDPs. Wei et al. (2021) proposed a couple of algorithms for learning infinite-horizon average-reward linear MDPs. First, `FOPO` is a fixed-point iteration-based algorithm that guarantees a regret upper bound of $\widetilde{\mathcal{O}}(d^{1.5}\mathrm{sp}(v^*)\sqrt{T})$ where d is the dimension of the underlying feature mapping. Although the regret upper bound is currently the best-known upper bound, the algorithm is impractical. They also proposed `OLSVI.FH` that divides the time horizon into pieces and runs a finite-horizon algorithm for each piece, thereby achieving computational efficiency. However, the regret upper bound of `OLSVI.FH` is suboptimal. Lastly, they came up with another efficient algorithm `MDP-EXP2` that guarantees a regret upper bound of order $\widetilde{\mathcal{O}}(\sqrt{T})$, but it applies to only ergodic MDPs. He et al. (2024) studied a general function approximation framework that can be applied to the linear MDP setting. Their algorithm, `LOOP`, achieves a regret upper bound of $\widetilde{\mathcal{O}}(d^{1.5}\mathrm{sp}(v^*)^{1.5}\sqrt{T})$, but as `FOPO`, `LOOP` is not computationally tractable. Recently, Hong et al. (2025) developed an algorithm that is provably efficient and, at the same time, achieves the best-known upper bound $\widetilde{\mathcal{O}}(d^{1.5}\mathrm{sp}(v^*)\sqrt{T})$. For infinite-horizon average-reward linear mixture MDPs, Wu et al. (2022) designed an algorithm that is nearly minimax optimal for the class of communicating MDPs.

Reinforcement Learning with General Function Approximation Reinforcement learning with general function approximation is a recent framework to capture possibly non-linear structures in MDPs, thereby providing an alternative to the linear function approximation framework. Jiang et al. (2017) studied the Bellman rank as a way of extending the linear class to non-linear function classes. Wang et al. (2020) adopted the notion of eluder dimension due to Russo and Van Roy (2013) for RL with general function approximation. Jin et al. (2021) proposed the concept of the Bellman eluder (BE) dimension, which combines the eluder dimension and the Bellman error. Du et al. (2021) extended the witness ranking on low-rank structures due to Sun et al. (2019) and considered the bilinear class. Foster et al. (2023) developed a general framework based on the notion of the decision-estimation coefficient. Zhong et al. (2023) proposed a unified framework with the generalized eluder coefficient (GEC). Recently, He et al. (2024) extended the notion of GEC to the infinite-horizon average-reward setting and introduced the notion of the average-reward generalized eluder coefficient (AGEC).

Reinforcement Learning with Multinomial Logit Function Approximation Hwang and Oh (2023) initiated the study of the multinomial logit function approximation framework for learning MDPs. They proposed `UCRL-MNL` for the finite-horizon setting that achieves a regret upper bound of $\widetilde{\mathcal{O}}(\kappa^{-1}dH^2\sqrt{K})$ regret where d is the dimension of the transition core, H is the horizon, K is the number of episodes, and $\kappa \in (0,1)$ is a lowed bound of the product of the probability of transitioning to one state and the probability of transitioning to another state. Later, Cho et al. (2024) and Li et al. (2024) developed algorithms that both guarantee a regret bound of $\tilde{\mathcal{O}}(dH^2\sqrt{K} + \kappa^{-1}d^2H^2)$, removing the dependence on κ from the previous result. Their algorithms use recently developed online Newton-based parameter estimation methods for logistic bandits due to Zhang and Sugiyama (2023); Lee and Oh (2024b). For the finite-horizon setting, Li et al. (2024) found a regret lower bound of $\Omega(dH\sqrt{\kappa^* K})$ where $\kappa^* \in (0,1)$ is a quantity similar to κ.

B PROOFS FOR SECTION 3.1

B.1 Properties of Multinomial Logit Function Approximation

Recall that $\mathcal{S}_t$ denotes $\mathcal{S}_{s_t,a_t}$, the set of reachable states from state a_t in one step after taking action a_t. Moreover, the per-time loss function for time step $t \in [T]$, its gradient, and its Hessian are given by

$$\ell_t(\theta) = -\sum_{s'\in\mathcal{S}_t} y_{t,s'} \log p_{t,s'}(\theta), \quad \nabla_\theta\left(\ell_t(\theta)\right) = -\sum_{s'\in\mathcal{S}_t}\left(y_{t,s'} - p_{t,s'}(\theta)\right)\varphi_{t,s'},$$
$$\nabla^2_\theta(\ell_t(\theta)) = \sum_{s'\in\mathcal{S}_t} p_{t,s'}(\theta)\varphi_{t,s'}\varphi_{t,s'}^\top - \sum_{s'\in\mathcal{S}_t}\sum_{s''\in\mathcal{S}_t} p_{t,s'}(\theta)p_{t,s''}(\theta)\varphi_{t,s'}\varphi_{t,s''}^\top,$$

respectively. The following lemma is an immediate consequence of Taylor's theorem.

Lemma 6. *For any $\theta_1, \theta_2 \in \mathbb{R}^d$ and $t \in [T]$, there exists some $\alpha \in [0,1]$ such that $\vartheta := \alpha\theta_1 + (1-\alpha)\theta_2$ satisfies*

$$\ell_t(\theta_2) = \ell_t(\theta_1) + \nabla_\theta\left(\ell_t(\theta_1)\right)^\top(\theta_2 - \theta_1) + \frac{1}{2}\|\theta_2 - \theta_1\|^2_{\nabla^2_\theta(\ell_t(\vartheta))}.$$

Assumption 3 implies that that for each $t \in [T]$, there exists a state ς_t such that $\varphi_{t,\varsigma_t} = 0$. This implies that for any $s' \in \mathcal{S}_t$,

$$p_{t,s'}(\theta) = \frac{\exp\left(\varphi_{t,s'}^\top\theta\right)}{\sum_{s''\in\mathcal{S}_t}\exp\left(\varphi_{t,s''}^\top\theta\right)} = \frac{\exp\left(\varphi_{t,s'}^\top\theta\right)}{1 + \sum_{s''\in\mathcal{S}_t\setminus\{\varsigma_t\}}\exp\left(\varphi_{t,s''}^\top\theta\right)}.$$

Lemma 7. *For any $\theta \in \mathbb{R}^d$, we have*

$$\nabla^2_\theta(\ell_t(\theta)) \succeq \sum_{s'\in\mathcal{S}_t\setminus\{\varsigma_t\}} p_{t,\varsigma_t}(\theta)p_{t,s'}(\theta)\varphi_{t,s'}\varphi_{t,s'}^\top \succeq \sum_{s'\in\mathcal{S}_t\setminus\{\varsigma_t\}} \kappa\varphi_{t,s'}\varphi_{t,s'}^\top.$$

Proof. Note that $(x-y)(x-y)^\top = xx^\top + yy^\top - xy^\top - yx^\top \succeq 0$ where $A \succeq B$ means that $A - B$ is positive semidefinite. This implies that $xx^\top + yy^\top \succeq xy^\top + yx^\top$. This implies that

$$\begin{aligned}
&\nabla^2_\theta(\ell_t(\theta))\\
&= \sum_{s'\in\mathcal{S}_t} p_{t,s'}(\theta)\varphi_{t,s'}\varphi_{t,s'}^\top - \frac{1}{2}\sum_{s'\in\mathcal{S}_t}\sum_{s''\in\mathcal{S}_t} p_{t,s'}(\theta)p_{t,s''}(\theta)\left(\varphi_{t,s'}\varphi_{t,s''}^\top + \varphi_{t,s''}\varphi_{t,s'}^\top\right)\\
&= \sum_{s'\in\mathcal{S}_t\setminus\{\varsigma_t\}} p_{t,s'}(\theta)\varphi_{t,s'}\varphi_{t,s'}^\top - \frac{1}{2}\sum_{s'\in\mathcal{S}_t\setminus\{\varsigma_t\}}\sum_{s''\in\mathcal{S}_t\setminus\{\varsigma_t\}} p_{t,s'}(\theta)p_{t,s''}(\theta)\left(\varphi_{t,s'}\varphi_{t,s''}^\top + \varphi_{t,s''}\varphi_{t,s'}^\top\right)\\
&\succeq \sum_{s'\in\mathcal{S}_t\setminus\{\varsigma_t\}} p_{t,s'}(\theta)\varphi_{t,s'}\varphi_{t,s'}^\top - \frac{1}{2}\sum_{s'\in\mathcal{S}_t\setminus\{\varsigma_t\}}\sum_{s''\in\mathcal{S}_t\setminus\{\varsigma_t\}} p_{t,s'}(\theta)p_{t,s''}(\theta)\left(\varphi_{t,s'}\varphi_{t,s'}^\top + \varphi_{t,s''}\varphi_{t,s''}^\top\right)\\
&= \sum_{s'\in\mathcal{S}_t\setminus\{\varsigma_t\}} p_{t,s'}(\theta)\varphi_{t,s'}\varphi_{t,s'}^\top - \sum_{s'\in\mathcal{S}_t\setminus\{\varsigma_t\}}\sum_{s''\in\mathcal{S}_t\setminus\{\varsigma_t\}} p_{t,s'}(\theta)p_{t,s''}(\theta)\varphi_{t,s'}\varphi_{t,s'}^\top
\end{aligned}$$

where the first equality holds because $\varphi_{t,\varsigma_t} = 0$. Then it follows that

$$\begin{aligned}
\nabla_\theta^2(\ell_t(\theta)) &\succeq \sum_{s' \in \mathcal{S}_t \setminus \{\varsigma_t\}} \left\{ 1 - \sum_{s'' \in \mathcal{S}_t \setminus \{\varsigma_t\}} p_{t,s''}(\theta) \right\} p_{t,s'}(\theta) \varphi_{t,s'} \varphi_{t,s'}^\top \\
&= \sum_{s' \in \mathcal{S}_t \setminus \{\varsigma_t\}} p_{t,\varsigma_t}(\theta) p_{t,s'}(\theta) \varphi_{t,s'} \varphi_{t,s'}^\top \\
&\succeq \sum_{s' \in \mathcal{S}_t \setminus \{\varsigma_t\}} \kappa \varphi_{t,s'} \varphi_{t,s'}^\top
\end{aligned} \tag{11}$$

and the last inequality is from Assumption 2. □

Next we consider

$$p(s' \mid s, a, \theta) = \frac{\exp\left(\varphi(s,a,s')^\top \theta\right)}{\sum_{s'' \in \mathcal{S}_{s,a}} \exp\left(\varphi(s_t,a_t,s'')^\top \theta\right)} = \frac{\exp\left(\varphi(s,a,s')^\top \theta\right)}{1 + \sum_{s'' \in \mathcal{S}_{s,a} \setminus \{\varsigma_{s,a}\}} \exp\left(\varphi(s_t,a_t,s'')^\top \theta\right)}$$

where $\varsigma_{s,a}$ is the state where $\varphi(s,a,s') = 0$. By (Proposition 1, Cho et al., 2024), we deduce that

$$\begin{aligned}
&\nabla_\theta(p(s' \mid s,a,\theta)) \\
&= p(s' \mid s,a,\theta)\varphi(s,a,s') - p(s' \mid s,a,\theta) \sum_{s'' \in \mathcal{S}_{s,a} \setminus \{\varsigma_{s,a}\}} p(s'' \mid s,a,\theta)\varphi(s,a,s'') \\
&= p(s' \mid s,a,\theta)\varphi(s,a,s') - p(s' \mid s,a,\theta) \sum_{s'' \in \mathcal{S}_{s,a}} p(s'' \mid s,a,\theta)\varphi(s,a,s'')
\end{aligned} \tag{12}$$

where the second equality holds because $\varphi_{i,\varsigma_i} = 0$. Moreover,

$$\begin{aligned}
&\nabla_\theta^2(p(s' \mid s,a,\theta)) \\
&= p(s' \mid s,a,\theta)\varphi(s,a,s')\varphi(s,a,s')^\top \\
&\quad - p(s' \mid s,a,\theta) \sum_{s'' \in \mathcal{S}_{s,a}} p(s'' \mid s,a,\theta) \left(\varphi(s,a,s')\varphi(s,a,s'')^\top + \varphi(s,a,s'')\varphi(s,a,s')^\top\right) \\
&\quad - p(s' \mid s,a,\theta) \sum_{s'' \in \mathcal{S}_{s,a}} p(s'' \mid s,a,\theta)\varphi(s,a,s'')\varphi(s,a,s'')^\top \\
&\quad + 2p(s' \mid s,a,\theta) \left(\sum_{s'' \in \mathcal{S}_{s,a}} p(s'' \mid s,a,\theta)\varphi(s,a,s'') \right) \left(\sum_{s'' \in \mathcal{S}_{s,a}} p(s'' \mid s,a,\theta)\varphi(s,a,s'') \right)^\top .
\end{aligned} \tag{13}$$

B.2 Proof of Lemma 1: Concentration of the Transition Core

In this section, we prove Lemma 1 by adapting the proof of (Theorem 3, Zhang and Sugiyama, 2023), (Lemma 1, Lee and Oh, 2024b), (Lemma 3, Li et al., 2024), and (Lemma 12, Cho et al., 2024) to our setting of infinite-horizon MDPs.

Let us consider the second-order Taylor approximation of the per-time loss function ℓ_t at $\widehat{\theta}_t$, given by

$$\widehat{\ell}_t(\theta) = \ell_t(\widehat{\theta}_t) + \nabla_\theta(\ell_t(\widehat{\theta}_t))^\top(\theta - \widehat{\theta}_t) + \frac{1}{2}\|\theta - \widehat{\theta}_t\|^2_{\nabla_\theta^2(\ell_t(\widehat{\theta}_t))}.$$

Since $\widehat{\Sigma}_t = \eta\nabla_\theta^2(\ell_t(\widehat{\theta}_t)) + \Sigma_t$, the update rule in (5) is equivalent to

$$\widehat{\theta}_{t+1} = \operatorname*{argmin}_{\theta \in \Theta} \left\{ \widehat{\ell}_t(\theta) + \frac{1}{2\eta}\|\theta - \widehat{\theta}_t\|^2_{\Sigma_t} \right\}.$$

Lemma 8. (Lee and Oh, 2024b, Lemma F.1). *Let $\eta = (1/2)\log\mathcal{U} + (L_\theta L_\varphi + 1)$. Then*

$$\begin{aligned}
&\|\widehat{\theta}_{t+1} - \theta^*\|^2_{\Sigma_{t+1}} \\
&\leq 2\eta \sum_{i=1}^t \left(\ell_i(\theta^*) - \ell_i(\widehat{\theta}_{i+1})\right) + 4\lambda L_\theta^2 + 12\sqrt{2}L_\theta L_\varphi^3 \eta \sum_{i=1}^t \|\widehat{\theta}_{i+1} - \widehat{\theta}_i\|_2^2 - \sum_{i=1}^t \|\widehat{\theta}_{i+1} - \widehat{\theta}_i\|^2_{\Sigma_i}.
\end{aligned}$$

To prove Lemma 1, we bound the right-hand side of the inequality in Lemma 8. Let $t \in [T]$. For a vector $z \in \mathbb{R}^{|\mathcal{S}_t \setminus \{\varsigma_t\}|}$, we denote by $[z]_{s'}$ the s'th coordinate of z for $s' \in \mathcal{S}_t \setminus \{\varsigma_t\}$. Then we define the softmax function $\sigma_t : \mathbb{R}^{|\mathcal{S}_t \setminus \{\varsigma_t\}|} \to \mathbb{R}^{|\mathcal{S}_t \setminus \{\varsigma_t\}|}$ as follows. For $s' \in \mathcal{S}_t \setminus \{\varsigma_t\}$,

$$[\sigma_t(z)]_{s'} = \frac{\exp([z]_{s'})}{1 + \sum_{s'' \in \mathcal{S}_t \setminus \{\varsigma_t\}} \exp([z]_{s''})}.$$

For simplicity, we define

$$[\sigma_t(z)]_{\varsigma_t} = 1 - \sum_{s' \in \mathcal{S}_t \setminus \{\varsigma_t\}} [\sigma_t(z)]_{s'} = \frac{1}{1 + \sum_{s' \in \mathcal{S}_t \setminus \{\varsigma_t\}} \exp([z]_{s'})},$$

although it is not part of the output of the softmax function σ_t. We denote by $\Phi_t \in \mathbb{R}^{d \times |\mathcal{S}_t \setminus \{\varsigma_t\}|}$ the matrix whose columns are $\varphi_{t,s'} \in \mathbb{R}^d$ for $s' \in \mathcal{S}_t \setminus \{\varsigma_t\}$. Then $p_{t,s'}(\theta) = [\sigma_t(\Phi_t^\top \theta)]_{s'}$ for $s' \in \mathcal{S}_t \setminus \{\varsigma_t\}$. Moreover, given $y_t = \{y_{t,s'} : s' \in \mathcal{S}_t\}$, we define

$$\ell(z, y_t) := \sum_{s' \in \mathcal{S}_t} \mathbf{1}\{y_{t,s'} = 1\} \cdot \log\left(\frac{1}{[\sigma_t(z)]_{s'}}\right) \quad \text{for } z \in \mathbb{R}^{|\mathcal{S}_t \setminus \{\varsigma_t\}|}.$$

Then the per-time loss function can be rewritten as $\ell_t(\theta) = \ell(\Phi_t^\top \theta, y_t)$. Next, we define a pseudo-inverse function $\sigma_t^+ : \mathbb{R}^{|\mathcal{S}_t \setminus \{\varsigma_t\}|} \to \mathbb{R}^{|\mathcal{S}_t \setminus \{\varsigma_t\}|}$ of σ_t as

$$[\sigma_t^+(q)]_{s'} = \log\left(\frac{q_{s'}}{1 - \|q\|_1}\right)$$

for any $q \in \{p \in [0,1]^{|\mathcal{S}_t \setminus \{\varsigma_t\}|} : \|p\|_1 < 1\}$. Let z_t be defined as

$$z_t = \sigma_t^+\left(\mathbb{E}_{\theta \sim \mathcal{N}_t}\left[\sigma_t(\Phi_t^\top \theta)\right]\right) \quad \text{where} \quad \mathcal{N}_t = \mathcal{N}(\widehat{\theta}_t, c\Sigma_t^{-1}).$$

Here, $\mathcal{N}_t$ is the Guassian distribution with mean $\widehat{\theta}_t$ and covariance matrix $c\Sigma_t^{-1}$ where coefficient c is specified later. Having defined z_t for $t \in [T]$, we deduce that

$$\sum_{i=1}^{t} \left(\ell_i(\theta^*) - \ell_i(\widehat{\theta}_{i+1})\right) = \underbrace{\sum_{i=1}^{t} (\ell_i(\theta^*) - \ell(z_i, y_i))}_{(a)} + \underbrace{\sum_{i=1}^{t} \left(\ell(z_i, y_i) - \ell_i(\widehat{\theta}_{i+1})\right)}_{(b)}.$$

Term (a) can be bounded based on the following lemma.

Lemma 9. (Lee and Oh, 2024b, Lemma F.2). *Let $\delta \in (0,1]$ and $\lambda \geq 1$. With probability at least $1 - \delta$, for all $t \in [T]$, we have*

$$\begin{aligned}&\sum_{i=1}^{t} (\ell_i(\theta^*) - \ell(z_i, y_i)) \\ &\leq (3\log(1 + \mathcal{U}t) + 2 + L_\theta L_\varphi)\left(\frac{17}{16}\lambda + 2\sqrt{\lambda}\log\left(\frac{2\sqrt{1+2t}}{\delta}\right) + 16\left(\log\left(\frac{2\sqrt{1+2t}}{\delta}\right)\right)^2\right) + 2.\end{aligned}$$

For term (b), we consider the following lemma.

Lemma 10. (Lee and Oh, 2024b, Lemma F.3). *For any $c > 0$, let $\lambda \geq \max\{2L_\varphi^2, 72cdL_\varphi^2\}$. Then, for all $t \in [T]$, we have*

$$\sum_{i=1}^{t} \left(\ell(z_i, y_i) - \ell_i(\widehat{\theta}_{i+1})\right) \leq \frac{1}{2c}\sum_{i=1}^{t} \|\widehat{\theta}_i - \widehat{\theta}_{i+1}\|_{\Sigma_i}^2 + \sqrt{6}cd\log\left(1 + \frac{2tL_\varphi^2}{d\lambda}\right).$$

With Lemmas 8–10, we are ready to complete the proof of Lemma 1. It follows from Lemmas 8–10 that for $\lambda \geq \max\{2L_\varphi^2, 72cdL_\varphi^2\}$,

$$\begin{aligned}
&\|\widehat{\theta}_{t+1} - \theta^*\|_{\Sigma_{t+1}}^2 \\
&\leq 2\eta(3\log(1+\mathcal{U}t) + 2 + L_\theta L_\varphi)\left(\frac{17}{16}\lambda + 2\sqrt{\lambda}\log\left(\frac{2\sqrt{1+2t}}{\delta}\right) + 16\left(\log\left(\frac{2\sqrt{1+2t}}{\delta}\right)\right)^2\right) \\
&\quad + 4\eta + 2\eta\sqrt{6}cd\log\left(1 + \frac{2tL_\varphi^2}{d\lambda}\right) + 4\lambda L_\theta^2 \\
&\quad + 12\sqrt{2}L_\theta L_\varphi^3\eta\sum_{i=1}^t \|\widehat{\theta}_{i+1} - \widehat{\theta}_i\|_2^2 + \left(\frac{\eta}{c} - 1\right)\sum_{i=1}^t \|\widehat{\theta}_{i+1} - \widehat{\theta}_i\|_{\Sigma_i}^2.
\end{aligned}$$

Setting $c = 7\eta/6$ and $\lambda \geq 84\sqrt{2}L_\theta L_\varphi^3\eta$, we have

$$\begin{aligned}
&12\sqrt{2}L_\theta L_\varphi^3\eta\sum_{i=1}^t \|\widehat{\theta}_{i+1} - \widehat{\theta}_i\|_2^2 + \left(\frac{\eta}{c} - 1\right)\sum_{i=1}^t \|\widehat{\theta}_{i+1} - \widehat{\theta}_i\|_{\Sigma_i}^2 \\
&\leq 12\sqrt{2}L_\theta L_\varphi^3\eta\sum_{i=1}^t \|\widehat{\theta}_{i+1} - \widehat{\theta}_i\|_2^2 - \frac{\lambda}{7}\sum_{i=1}^t \|\widehat{\theta}_{i+1} - \widehat{\theta}_i\|_2^2 \\
&\leq 0.
\end{aligned}$$

Note that $84\sqrt{2}(L_\theta L_\varphi^3 + dL_\varphi^2)\eta \geq \max\{2L_\varphi^2, 72cdL_\varphi^2, 84\sqrt{2}L_\theta L_\varphi^3\eta\}$, so we set $\lambda \geq 84\sqrt{2}(L_\theta L_\varphi^3 + dL_\varphi^2)\eta$. As we have $\eta = (1/2)\log\mathcal{U} + (L_\theta L_\varphi + 1)$, we deduce that

$$\|\widehat{\theta}_{t+1} - \theta^*\|_{\Sigma_{t+1}} \leq C\sqrt{d}(\log(Ut/\delta))^2$$

for some constant C that depends only on L_θ, L_φ, as required.

B.3 Proof of Lemma 2: Confidence Polytope for the True Transition Function

Since $\left|p(s' \mid s, a, \theta^*) - p(s' \mid s, a, \widehat{\theta}_t)\right| \leq 1$ for any $s' \in \mathcal{S}_{s,a}$, we may show that

$$\sum_{s' \in \mathcal{S}_{s,a}} \left|p(s' \mid s, a, \theta^*) - p(s' \mid s, a, \widehat{\theta}_t)\right| \leq B_{s,a}^{1,t} + B_{s,a}^{2,t}$$

for any $(s, a) \in \mathcal{S} \times \mathcal{A}$. By Taylor's theorem, there exists some $\alpha \in [0, 1]$ such that $\vartheta = \alpha\widehat{\theta}_t + (1-\alpha)\theta^*$ satisfies

$$\begin{aligned}
&\sum_{s' \in \mathcal{S}_{s,a}} \left|p(s' \mid s, a, \theta^*) - p(s' \mid s, a, \widehat{\theta}_t)\right| \\
&= \sum_{s' \in \mathcal{S}_{s,a}} \left|\nabla_\theta(p(s' \mid s, a, \widehat{\theta}_t))^\top(\theta^* - \widehat{\theta}_t) + \frac{1}{2}\|\theta^* - \widehat{\theta}_t\|_{\nabla_\theta^2(p(s'|s,a,\vartheta))}^2\right| \\
&\leq \underbrace{\sum_{s' \in \mathcal{S}_{s,a}} \left|\nabla_\theta(p(s' \mid s, a, \widehat{\theta}_t))^\top(\theta^* - \widehat{\theta}_t)\right|}_{(a)} + \underbrace{\frac{1}{2}\sum_{s' \in \mathcal{S}_{s,a}} \|\theta^* - \widehat{\theta}_t\|_{\nabla_\theta^2(p(s'|s,a,\vartheta))}^2}_{(b)}.
\end{aligned}$$

Term (a) can be bounded as follows.

$$\begin{aligned}
&\sum_{s'\in\mathcal{S}_{s,a}} \left|\nabla_\theta(p(s' \mid s,a,\widehat{\theta}_t))^\top(\theta^*-\widehat{\theta}_t)\right| \\
&= \sum_{s'\in\mathcal{S}_{s,a}} \left| p(s' \mid s,a,\widehat{\theta}_t)\left(\varphi(s,a,s') - \sum_{s''\in\mathcal{S}_{s,a}} p(s'' \mid s,a,\widehat{\theta}_t)\varphi(s,a,s'')\right)^\top (\theta^*-\widehat{\theta}_t)\right| \\
&\le \sum_{s'\in\mathcal{S}_{s,a}} p(s' \mid s,a,\widehat{\theta}_t)\left\|\varphi(s,a,s') - \sum_{s''\in\mathcal{S}_{s,a}} p(s'' \mid s,a,\widehat{\theta}_t)\varphi(s,a,s'')\right\|_{\Sigma_t^{-1}} \left\|\theta^*-\widehat{\theta}_t\right\|_{\Sigma_t} \\
&\le \beta_t \sum_{s'\in\mathcal{S}_{s,a}} p(s' \mid s,a,\widehat{\theta}_t)\left\|\varphi(s,a,s') - \sum_{s''\in\mathcal{S}_{s,a}} p(s'' \mid s,a,\widehat{\theta}_t)\varphi(s,a,s'')\right\|_{\Sigma_t^{-1}}
\end{aligned}$$

where the equality is due to (12), the first inequality is by the Cauchy-Schwarz inequality, and the second inequality holds because we assumed that $\theta^* \in \mathcal{C}_t$. Hence, term (a) is bounded above by $B^{1,t}_{s,a}$. For term (b), note that

$$\begin{aligned}
&\sum_{s'\in\mathcal{S}_{s,a}} \|\theta^*-\widehat{\theta}_t\|^2_{\nabla^2_\theta(p(s'|s,a,\vartheta))} \\
&= \sum_{s'\in\mathcal{S}_{s,a}} \left|(\widehat{\theta}_t-\theta^*)^\top \nabla^2_\theta(p(s' \mid s,a,\vartheta))(\widehat{\theta}_t-\theta^*)\right| \\
&\le \sum_{s'\in\mathcal{S}_{s,a}} p(s' \mid s,a,\vartheta)\Bigg[\left((\widehat{\theta}_t-\theta^*)^\top\varphi(s,a,s')\right)^2 \\
&\quad + \sum_{s''\in\mathcal{S}_{s,a}} p(s'' \mid s,a,\vartheta)\left|2\left((\widehat{\theta}_t-\theta^*)^\top\varphi(s,a,s')\right)\left((\widehat{\theta}_t-\theta^*)^\top\varphi(s,a,s'')\right)\right| \\
&\quad + \sum_{s''\in\mathcal{S}_{s,a}} p(s'' \mid s,a,\vartheta)\left((\widehat{\theta}_t-\theta^*)^\top\varphi(s,a,s'')\right)^2 \\
&\quad +2\left((\widehat{\theta}_t-\theta^*)^\top\left(\sum_{s''\in\mathcal{S}_{s,a}} p(s'' \mid s,a,\vartheta)\varphi(s,a,s'')\right)\right)^2\Bigg]
\end{aligned}$$

where the inequality follows from (13). Note that for any $s' \in \mathcal{S}_{s,a}$, the Cauchy-Schwarz inequality implies

$$(\widehat{\theta}_t-\theta^*)^\top\varphi(s,a,s') \le \|\widehat{\theta}_t-\theta^*\|_{\Sigma_t}\|\varphi(s,a,s')\|_{\Sigma_t^{-1}} \le \beta_t\|\varphi(s,a,s')\|_{\Sigma_t^{-1}}$$

where the second inequality holds because $\theta^* \in \mathcal{C}_t$. Then we deduce that

$$\begin{aligned}
&\sum_{s'\in\mathcal{S}_{s,a}} \|\theta^*-\widehat{\theta}_t\|^2_{\nabla^2_\theta(p(s'|s,a,\vartheta))} \\
&\le 4\beta_t^2 \sum_{s'\in\mathcal{S}_{s,a}} p(s' \mid s,a,\vartheta)\|\varphi(s,a,s')\|^2_{\Sigma_t^{-1}} + 2\beta_t^2\left(\sum_{s''\in\mathcal{S}_{s,a}} p(s'' \mid s,a,\vartheta)\|\varphi(s,a,s'')\|_{\Sigma_t^{-1}}\right)^2 \\
&\le 6\beta_t^2 \max_{s'\in\mathcal{S}_{s,a}} \|\varphi(s,a,s')\|^2_{\Sigma_t^{-1}}.
\end{aligned}$$

Therefore, term (b) is bounded above by $B^{2,t}_{s,a}$, as required.

We also prove the following lemma which will be useful for our analysis.

Lemma 11. *Suppose that $\theta^* \in \mathcal{C}_t$ where $\mathcal{C}_t$ is defined as in (6). Let $p^* \in \mathbb{R}^{S\times A\times S}$ be the vector representation of the true transition function. Then for $t \in [T]$,*

$$p^* \in \mathcal{P}'_t := \left\{p \in [0,1]^{S\times A\times S} : p \text{ satisfies } (15), (16)\right\} \tag{14}$$

where

$$\sum_{s' \in \mathcal{S}_{s,a}} p_{s,a,s'} = 1, \tag{15}$$

$$\sum_{s' \in \mathcal{S}_{s,a}} \left| p_{s,a,s'} - p(s' \mid s,a,\widehat{\theta}_t) \right| \leq R_{t,s,a} \tag{16}$$

with $R_{t,s,a} = 2\beta_t \max_{s' \in \mathcal{S}_{s,a}} \|\varphi(s,a,s')\|_{\Sigma_t^{-1}}$ *for all* $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Proof. Since $\left|p(s' \mid s,a,\theta^*) - p(s' \mid s,a,\widehat{\theta}_t)\right| \leq 1$ for any $s' \in \mathcal{S}_{s,a}$, it is sufficient to show that

$$\sum_{s' \in \mathcal{S}_{s,a}} \left|p(s' \mid s,a,\theta^*) - p(s' \mid s,a,\widehat{\theta}_t)\right| \leq 2\beta_t \max_{s' \in \mathcal{S}_{s,a}} \|\varphi(s,a,s')\|_{\Sigma_t^{-1}}$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. By Taylor's theorem, for any $s' \in \mathcal{S}_{s,a}$, there exists some $\alpha_{s'} \in [0,1]$ such that $\vartheta_{s'} = \alpha_{s'}\widehat{\theta}_t + (1-\alpha_{s'})\theta^*$ satisfies

$$\begin{aligned}
&\sum_{s' \in \mathcal{S}_{s,a}} \left|p(s' \mid s,a,\theta^*) - p(s' \mid s,a,\widehat{\theta}_t)\right| \\
&= \sum_{s' \in \mathcal{S}_{s,a}} \left|\nabla_\theta (p(s' \mid s,a,\vartheta_{s'}))^\top (\theta^* - \widehat{\theta}_t)\right| \\
&= \sum_{s' \in \mathcal{S}_{s,a}} \left| p(s' \mid s,a,\vartheta_{s'}) \left(\varphi(s,a,s') - \sum_{s'' \in \mathcal{S}_{s,a}} p(s'' \mid s,a,\vartheta_{s'})\varphi(s,a,s'') \right)^\top (\theta^* - \widehat{\theta}_t) \right| \\
&\leq \sum_{s' \in \mathcal{S}_{s,a}} p(s' \mid s,a,\vartheta_{s'}) \left\| \varphi(s,a,s') - \sum_{s'' \in \mathcal{S}_{s,a}} p(s'' \mid s,a,\vartheta_{s'})\varphi(s,a,s'') \right\|_{\Sigma_t^{-1}} \left\|\theta^* - \widehat{\theta}_t\right\|_{\Sigma_t} \\
&\leq \beta_t \sum_{s' \in \mathcal{S}_{s,a}} p(s' \mid s,a,\vartheta_{s'}) \left(\|\varphi(s,a,s')\|_{\Sigma_t^{-1}} + \sum_{s'' \in \mathcal{S}_{s,a}} p(s'' \mid s,a,\vartheta_{s'}) \|\varphi(s,a,s'')\|_{\Sigma_t^{-1}} \right) \\
&\leq 2\beta_t \max_{s' \in \mathcal{S}_{s,a}} \|\varphi(s,a,s')\|_{\Sigma_t^{-1}}.
\end{aligned}$$

where the second equality is due to (12), the first inequality is by the Cauchy-Schwarz inequality, and the second inequality holds because we assumed that $\theta^* \in \mathcal{C}_t$. □

C PROOF OF THEOREM 1: PERFORMANCE ANALYSIS OF UCMNLK FOR THE AVERAGE-REWARD SETTING

Let K_T denote the total number of distinct episodes over the horizon of T time steps. For simplicity, we assume that the last time step of the last episode and that time step $T+1$ is the beginning of the (K_T+1)th episode, i.e., $t_{K_T+1} = T+1$. Then we have

$$\begin{aligned}
\mathrm{Regret}(T) &= \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} (J^* - r(s_t,a_t)) \\
&\leq \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left(J^* - Q_k(s_t,a_t) + \gamma \max_{p \in \mathcal{P}_{t_k}} \left\{ \sum_{s' \in \mathcal{S}_t} p_{s_t,a_t,s'} V_k(s') \right\} + \gamma^N \right) \\
&= \underbrace{\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} (J^* - (1-\gamma) V_k(s_{t+1}))}_{(a)} + \underbrace{\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} (V_k(s_{t+1}) - Q_k(s_t,a_t))}_{(b)} \\
&\quad + \underbrace{\gamma \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left(\sum_{s' \in \mathcal{S}_t} p^*_{s_t,a_t,s'} V_k(s') - V_k(s_{t+1}) \right)}_{(c)} \\
&\quad + \underbrace{\gamma \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \max_{p \in \mathcal{P}_{t_k}} \left\{ \sum_{s' \in \mathcal{S}_t} \left(p_{s_t,a_t,s'} - p^*_{s_t,a_t,s'} \right) V_k(s') \right\}}_{(d)} + T\gamma^N .
\end{aligned} \tag{17}$$

where the inequality comes from Lemma 3. We provide upper bounds on terms (a)–(d) in Appendices C.1 to C.4. Based on them, we prove a regret upper bound on `UCMNLK` for the average-reward setting in Appendix C.5.

C.1 Term (a)

Recall that J^* is the optimal average reward of the MDP. It is known that any communicating MDP satisfies the following Bellman optimality condition (See Puterman, 2014). The condition states that there exist $v^* : \mathcal{S} \to \mathbb{R}$ and $q^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$J^* + q^*(s,a) = r(s,a) + \sum_{s' \in \mathcal{S}} p(s' \mid s,a) v^*(s') \quad \text{and} \quad v^*(s) = \max_{a \in \mathcal{A}} q^*(s,a). \tag{18}$$

Here, $v^* : \mathcal{S} \to \mathbb{R}$ is referred to as the optimal bias function. For any function $h : \mathcal{S} \to \mathbb{R}$, we define its span as $\mathrm{sp}(h) := \max_{s \in \mathcal{S}} h(s) - \min_{s \in \mathcal{S}} h(s)$. In particular, it is known that for a communicating MDP with diameter at most D, we have $\mathrm{sp}(v^*) \leq D$ (Jaksch et al., 2010; Puterman, 2014). Furthermore, we also have the following lemma that compares the span of the optimal discounted value function V^* and that of the optimal bias function v^*.

Lemma 12. (Wei et al., 2020, Lemma 2) *For any $\gamma \in [0,1)$, $\mathrm{sp}(V^*) \leq 2 \cdot \mathrm{sp}(v^*)$ and $|(1-\gamma)V^*(s) - J^*| \leq (1-\gamma)\mathrm{sp}(v^*)$ for all $s \in \mathcal{S}$.*

We can provide an upper bound on term (a) using Lemma 12. Note that for any t,

$$J^* - (1-\gamma)V_k(s_{t+1}) \leq J^* - (1-\gamma)V^*(s_{t+1}) \leq (1-\gamma)\mathrm{sp}(v^*) \leq (1-\gamma)D$$

where the first inequality follows from Lemma 4 and the second inequality is due to Lemma 12. Therefore, it follows that

$$\text{Term } (a) = \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} (J^* - (1-\gamma)V_k(s_{t+1})) \leq T(1-\gamma)D. \tag{19}$$

C.2 Term (b)

Note that term (b) can be rewritten as follows.

$$\begin{aligned}
&\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\left(V_k(s_{t+1})-Q_k(s_t,a_t)\right)\\
&=\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-2}\left(Q_k(s_{t+1},a_{t+1})-Q_k(s_t,a_t)\right)+\sum_{k=1}^{K_T}\left(V_k(s_{t_{k+1}})-Q_k(s_{t_{k+1}-1},a_{t_{k+1}-1})\right)\\
&=\sum_{k=1}^{K_T}\left(V_k(s_{t_{k+1}})-Q_k(s_{t_k},a_{t_k})\right).
\end{aligned}$$

Here, for any $k\geq 1$, we know from Lemma 4 that

$$V_k(s_{t_{k+1}})-Q_k(s_{t_k},a_{t_k})\leq V_k(s_{t_{k+1}})\leq\frac{1}{1-\gamma}.$$

Therefore, term (b) can be bounded above as

$$\text{Term }(b)=\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\left(V_k(s_{t+1})-Q_k(s_t,a_t)\right)\leq\frac{K_T}{1-\gamma}. \tag{20}$$

C.3 Term (c)

For $t\in[T]$, recall that $p^*_{s_t,a_t,s'}$ denotes the true transition probability of transitioning to state s' given that the state-action pair for t is (s_t,a_t). Take Y_t as

$$Y_t=\sum_{s'\in\mathcal{S}_t}p^*_{s_t,a_t,s'}V_k(s')-V_k(s_{t+1})$$

for $t_k\leq t\leq t_{k+1}-1$ and $k\in[K_T]$. For $t\in[T]$, let $\mathcal{F}_t$ be the σ-algebra generated by the randomness up to time step t. Then we have $\mathbb{E}\left[Y_t\mid\mathcal{F}_t\right]=0$, which means that $Y_1,\ldots,Y_T$ gives rise to a martingale difference sequence. Then Term (c), which is essentially the summation of $Y_1,\ldots,Y_T$, can be bounded by Azuma's inequality given as follows.

Lemma 13 (Azuma's inequality)**.** *Let $Y_1,\ldots,Y_T$ be a martingale difference sequence with respect to a filtration $\mathcal{F}_1,\ldots,\mathcal{F}_T$. Assume that $|Y_t|\leq B$ for $t\in[T]$. Then with probability at least $1-\delta$, we have $\sum_{t=1}^T Y_t\leq B\sqrt{2T\log\left(1/\delta\right)}$.*

To apply Azuma's inequality, we need a global bound on Y_t terms. Note that

$$|Y_t|\leq\sum_{s'\in\mathcal{S}_t}p^*_{s_t,a_t,s'}\left|V_k(s')-V_k(s_{t+1})\right|\leq D$$

where the last inequality follows from Lemma 5. Then it follows from Lemma 13 that with probability at least $1-\delta$,

$$\text{Term }(c)=\gamma\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\left(\sum_{s'\in\mathcal{S}_t}p^*_{s_t,a_t,s'}V_k(s')-V_k(s_{t+1})\right)\leq\sum_{t=1}^{T}Y_T\leq D\sqrt{2T\log(1/\delta)}. \tag{21}$$

C.4 Term (d)

Let $p \in \mathcal{P}_{t_k}$. Then we have

$$\begin{aligned}
\sum_{s' \in \mathcal{S}_t} \left(p_{s_t,a_t,s'} - p^*_{s_t,a_t,s'}\right) V_k(s') &= \sum_{s' \in \mathcal{S}_t} \left(p_{s_t,a_t,s'} - p^*_{s_t,a_t,s'}\right) \left(V_k(s') - \min_{s'' \in \mathcal{S}} V_k(s'')\right) \\
&\leq \sum_{s' \in \mathcal{S}_t} \left|p_{s_t,a_t,s'} - p^*_{s_t,a_t,s'}\right| \mathrm{sp}(V_k) \\
&\leq D \sum_{s' \in \mathcal{S}_t} \left|p_{s_t,a_t,s'} - p^*_{s_t,a_t,s'}\right| \\
&\leq D \sum_{s' \in \mathcal{S}_t} \left(\left|p_{s_t,a_t,s'} - p_t(s', \widehat{\theta}_{t_k})\right| + \left|p_t(s', \widehat{\theta}_{t_k}) - p^*_{s_t,a_t,s'}\right|\right) \\
&\leq 2D \left(B^{1,t_k}_{s_t,a_t} + B^{2,t_k}_{s_t,a_t}\right)
\end{aligned}$$

where the equality holds because $\sum_{s' \in \mathcal{S}_t} (p_{s_t,a_t,s'} - p^*_{s_t,a_t,s'}) = 0$, the second inequality follows from Lemma 5, and the last inequality is implied by Lemma 2 as $p, p^* \in \mathcal{P}_{t_k}$. Therefore, it follows that

$$\text{Term } (d) \leq 2D \underbrace{\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} B^{1,t_k}_{s_t,a_t}}_{(\star)} + 2D \underbrace{\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} B^{2,t_k}_{s_t,a_t}}_{(\star\star)}.$$

We may provide an upper bound on the right-most side of this inequality based on the following lemma.

Lemma 14. *Suppose that $\theta^* \in \mathcal{C}_t$ for all $t \in [T]$. Let $\lambda \geq L_\varphi^2$. Then*

$$\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} B^{1,t_k}_{s_t,a_t} \leq \beta_T \left(\left(\frac{32\beta_T}{\kappa} + \frac{128\sqrt{2}L_\varphi \eta}{\kappa\sqrt{\lambda}}\right) d \log\left(1 + \frac{T\mathcal{U}L_\varphi^2}{\lambda d}\right) + 2\sqrt{dT \log\left(1 + \frac{T\mathcal{U}L_\varphi^2}{\lambda d}\right)}\right).$$

Proof. See Appendix F. □

Lemma 15. *Let $\lambda \geq L_\varphi^2$. Then*

$$\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} B^{2,t_k}_{s_t,a_t} \leq \frac{12d}{\kappa} \beta_T^2 \log\left(1 + \frac{T\mathcal{U}L_\varphi^2}{\lambda d}\right).$$

Proof. See Appendix F. □

Hence, we deduce that

$$\text{Term } (d) \leq \left(\frac{88dD}{\kappa}\beta_T^2 + \frac{256\sqrt{2}L_\varphi \eta dD}{\kappa\sqrt{\lambda}}\beta_T\right) \log\left(1 + \frac{T\mathcal{U}L_\varphi^2}{\lambda d}\right) + 4D\beta_T \sqrt{dT \log\left(1 + \frac{T\mathcal{U}L_\varphi^2}{\lambda d}\right)}. \tag{22}$$

C.5 Completing the Regret Bound for the Average-Reward Case

Combining (19), (20), (21), and (22), we deduce that

$$\begin{aligned}
&\mathrm{Regret}(T) \\
&\leq T(1-\gamma)D + \frac{K_T}{1-\gamma} + D\sqrt{2T\log(1/\delta))} + T\gamma^N \\
&\quad + \left(\frac{88dD}{\kappa}\beta_T^2 + \frac{256\sqrt{2}L_\varphi \eta dD}{\kappa\sqrt{\lambda}}\beta_T\right) \log\left(1 + \frac{T\mathcal{U}L_\varphi^2}{\lambda d}\right) + 4D\beta_T \sqrt{dT \log\left(1 + \frac{T\mathcal{U}L_\varphi^2}{\lambda d}\right)}
\end{aligned}$$

Here, K_T can be bounded by the following lemma.

Lemma 16. $K_T \leq 1 + d\log_2\left(1 + 2TL_\varphi^2/\lambda\right)$.

Proof. Since $\Sigma_1 = \lambda I_d$, we have $\det(\Sigma_1) = \lambda^d$. Note that for any θ and t,

$$\begin{aligned}\left\|\nabla_\theta^2(\ell_t(\theta))\right\|_2 &\leq \sum_{s'\in\mathcal{S}_t} p_{t,s'}(\theta)\left\|\varphi_{t,s'}\varphi_{t,s'}^\top\right\|_2 + \sum_{s'\in\mathcal{S}_t}\sum_{s''\in\mathcal{S}_t} p_{t,s'}(\theta)p_{t,s''}(\theta)\left\|\varphi_{t,s'}\varphi_{t,s''}^\top\right\|_2\\ &\leq \sum_{s'\in\mathcal{S}_t} p_{t,s'}(\theta)\left\|\varphi_{t,s'}\right\|_2^2 + \sum_{s'\in\mathcal{S}_t}\sum_{s''\in\mathcal{S}_t} p_{t,s'}(\theta)p_{t,s''}(\theta)\left\|\varphi_{t,s'}\right\|_2\left\|\varphi_{t,s''}\right\|_2\\ &\leq 2L_\varphi^2\end{aligned}$$

where the first two inequalities are by the triangle inequality and the last is due to Assumption 1. Then it follows that

$$\left\|\Sigma_T\right\|_2 = \left\|\lambda I_d + \sum_{t=1}^{T-1}\nabla_\theta^2(\ell_t(\widehat{\theta}_{t+1}))\right\|_2 \leq \lambda + 2TL_\varphi^2.$$

This implies that $\det(\Sigma_T) \leq (\lambda + 2TL_\varphi^2)^d$. Therefore, we have

$$(\lambda + 2TL_\varphi^2)^d \geq \det(\Sigma_T) \geq \det(\Sigma_{t_{K_T}}) \geq 2^{K_T-1}\det(\Sigma_{t_1}) = 2^{K_T-1}\lambda^d, \tag{23}$$

where the second inequality holds because $\Sigma_T \succeq \Sigma_{t_{K_T}}$ and the last holds due to $\det(\Sigma_{t_{k+1}}) \geq 2\det(\Sigma_{t_k})$. Then it follows from (23) that $K_T \leq 1 + d\log_2\left(1 + 2TL_\varphi^2/\lambda\right)$, as required. □

Setting γ and N as

$$\gamma = 1 - \sqrt{\frac{d}{DT}} \quad \text{and} \quad N = \frac{1}{1-\gamma}\log\left(\frac{\sqrt{T}}{dD}\right) = \sqrt{\frac{DT}{d}}\log\left(\frac{\sqrt{T}}{dD}\right),$$

we have

$$N \geq \frac{\log\left(\sqrt{T}/dD\right)}{\log(1/\gamma)},$$

in which case we get $T\gamma^N \leq dD\sqrt{T}$. Therefore,

$$\begin{aligned}\mathrm{Regret}(T) \leq{}& 3\sqrt{dDT}\log_2\left(1 + \frac{2TL_\varphi^2}{\lambda}\right) + D\sqrt{2T\log(1/\delta)} + T\gamma^N\\ &+ 8f(L_\theta, L_\varphi)\left(\frac{88dD}{\kappa}\beta_T^2 + \frac{256\sqrt{2}L_\varphi\eta dD}{\kappa\sqrt{\lambda}}\beta_T\right)\log\left(1 + \frac{T\mathcal{U}L_\varphi^2}{\lambda d}\right)(\log(\mathcal{U}t/\delta))^2\\ &+ 32f(L_\theta, L_\varphi)D\beta_T\sqrt{dT\log\left(1 + \frac{T\mathcal{U}L_\varphi^2}{\lambda d}\right)}(\log(\mathcal{U}t/\delta))^2\\ ={}& \widetilde{\mathcal{O}}\left(dD\sqrt{T} + \kappa^{-1}d^2D\right),\end{aligned}$$

as required.

D PROOF OF THEOREM 2: PERFORMANCE ANALYSIS OF UCMNLK FOR THE DISCOUNTED-REWARD SETTING

Recall that K_T denotes the total number of distinct episodes over T time steps and that t_{K_T+1} is defined as $T+1$ for simplicity. Let π denote the non-stationary policy taken by UCMNLK. Then by Lemma 4, we have

$$\mathrm{Regret}(\pi, T) = \sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\left(V^*(s_t) - V_t^\pi(s_t)\right) \leq \underbrace{\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\left(V_k(s_t) - V_t^\pi(s_t)\right)}_{\mathrm{Regret}'(\pi,T)}. \tag{24}$$

Note that for $t_k \leq t \leq t_{k+1} - 1$,

$$V_k(s_t) = Q_k(s_t, a_t) \leq r(s_t, a_t) + \gamma \max_{p \in \mathcal{P}_{t_k}} \left\{ \sum_{s' \in \mathcal{S}_t} p_{s_t, a_t, s'} V_k(s') \right\} + \gamma^N \tag{25}$$

where the inequality follows from Lemma 3. Moreover, by the Bellman equation,

$$V_t^\pi(s_t) = r(s_t, a_t) + \gamma \sum_{s' \in \mathcal{S}_t} p^*_{s_t, a_t, s'} V_{t+1}^\pi(s'). \tag{26}$$

Combining (24), (25), and (26), we deduce that

$$\begin{aligned}
&\text{Regret}(\pi, T) - T\gamma^N \\
&\leq \text{Regret}'(\pi, T) - T\gamma^N \\
&\leq \underbrace{\gamma \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \max_{p \in \mathcal{P}_{t_k}} \left\{ \sum_{s' \in \mathcal{S}_t} \left(p_{s_t, a_t, s'} - p^*_{s_t, a_t, s'} \right) V_k(s') \right\}}_{(i)} \\
&\quad + \underbrace{\gamma \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left(\sum_{s' \in \mathcal{S}_t} p^*_{s_t, a_t, s'} \left(V_k(s') - V_{t+1}^\pi(s') \right) - \left(V_k(s_{t+1}) - V_{t+1}^\pi(s_{t+1}) \right) \right)}_{(ii)} \\
&\quad + \underbrace{\gamma \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left(V_k(s_{t+1}) - V_{t+1}^\pi(s_{t+1}) \right)}_{(iii)}.
\end{aligned} \tag{27}$$

Note that term (i) is the same as term (d) in (17). Following the same argument in Appendix C.4 and using the fact that $V_k(s') \leq 1/(1-\gamma)$ for any $s' \in \mathcal{S}_t$ due to Lemma 4, we deduce that

$$\text{Term } (i) \leq \left(\frac{88d}{\kappa(1-\gamma)} \beta_T^2 + \frac{256\sqrt{2} L_\varphi \eta d}{\kappa \sqrt{\lambda}(1-\gamma)} \beta_T \right) \log\left(1 + \frac{T\mathcal{U}L_\varphi^2}{\lambda d}\right) + \frac{4\beta_T}{(1-\gamma)} \sqrt{dT \log\left(1 + \frac{T\mathcal{U}L_\varphi^2}{\lambda d}\right)}. \tag{28}$$

For term (ii), we observe that taking Y_t for $t_k \leq t \leq t_{k+1} - 1$ and $k \in [K_T]$ as

$$Y_t = \sum_{s' \in \mathcal{S}_t} p^*_{s_t, a_t, s'} \left(V_k(s') - V_{t+1}^\pi(s') \right) - \left(V_k(s_{t+1}) - V_{t+1}^\pi(s_{t+1}) \right)$$

gives rise to a martingale difference sequence. Moreover, we have $|Y_t| \leq 2/(1-\gamma)$. Then applying Azuma's inequality (Lemma 13), we deduce that

$$\text{Term } (ii) \leq \frac{2}{1-\gamma} \sqrt{2T \log(1/\delta)}. \tag{29}$$

For term (iii), observe that

$$\begin{aligned}
\text{term } (iii) &= \gamma \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left(V_k(s_t) - V_{t+1}^\pi(s_t) \right) \\
&\quad + \gamma \sum_{k=1}^{K_T} \left(-(V_k(s_{t_k}) - V_{t_k}^\pi(s_{t_k})) + (V_k(s_{t_{k+1}}) - V_{t_{k+1}}^\pi(s_{t_{k+1}})) \right) \\
&\leq \gamma \cdot \text{Regret}'(\pi, T) + \frac{\gamma}{1-\gamma} K_T.
\end{aligned} \tag{30}$$

Combining (27), (28), (29), and (30), it follows that

$$\begin{aligned}
&\mathrm{Regret}(\pi, T)\\
&\leq \frac{T\gamma^N}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}K_T + \frac{2}{(1-\gamma)^2}\sqrt{2T\log(1/\delta)}\\
&\quad + \left(\frac{88d}{\kappa(1-\gamma)^2}\beta_T^2 + \frac{256\sqrt{2}L_\varphi \eta d}{\kappa\sqrt{\lambda}(1-\gamma)^2}\beta_T\right)\log\left(1+\frac{TUL_\varphi^2}{\lambda d}\right) + \frac{4\beta_T}{(1-\gamma)^2}\sqrt{dT\log\left(1+\frac{TUL_\varphi^2}{\lambda d}\right)}.
\end{aligned}$$

Setting N as

$$N \geq \frac{1}{1-\gamma}\log\left(\frac{\sqrt{T}}{d}\right),$$

we obtain

$$\mathrm{Regret}(\pi, T) = \widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^2}d\sqrt{T} + \frac{1}{\kappa(1-\gamma)^2}d^2\right),$$

as required.

E PROOFS FOR SECTION 3.3

In this section, we prove Lemmas 4, 3, and 5 given in Section 3.3.

E.1 Proof of Lemma 4: Optimistic Value Functions for UCMNLK

Let $V^{(0)}, \ldots, V^{(N-1)}$ denote the sequence of value functions generated by DEVI for episode k, and let $Q^{(0)}, \ldots, Q^{(N)}$ be the sequence of action-value functions generated by DEVI for episode k. Then Q_k equals $Q^{(N)}$, and $V_k(s)$ is given by $\max_{a\in\mathcal{A}} Q^{(N)}(s,a)$ for $s \in \mathcal{S}$. For simplicity, we define $V^{(N)}$ as V_k.

To prove that $1/(1-\gamma) \geq V_k(s) \geq V^*(s)$ and $1/(1-\gamma) \geq Q_k(s,a) \geq Q^*(s,a)$ for $(s,a) \in \mathcal{S}\times\mathcal{A}$, we will argue by induction on n that for each $n \in \{0,\ldots,N\}$, $1/(1-\gamma) \geq V^{(n)}(s) \geq V^*(s)$ and $1/(1-\gamma) \geq Q^{(n)}(s,a) \geq Q^*(s,a)$ for $(s,a) \in \mathcal{S}\times\mathcal{A}$. For $n = 0$, $1/(1-\gamma) = Q^{(0)}(s,a) \geq Q^*(s,a)$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Moreover, note that $V^{(0)}(s) = \max_{a\in\mathcal{A}} Q^{(0)}(s,a) = 1/(1-\gamma) \geq V^*(s)$ for all $s \in \mathcal{S}$.

Assume that $1/(1-\gamma) \geq V^{(n)}(s) \geq V^*(s)$ and $1/(1-\gamma) \geq Q^{(n)}(s,a) \geq Q^*(s,a)$ for $(s,a) \in \mathcal{S}\times\mathcal{A}$ for some $n \in \{0,\ldots,N-1\}$. Note that

$$\begin{aligned}
Q^{(n+1)}(s,a) &= r(s,a) + \gamma \max_{p\in\mathcal{P}_{t_k}}\left\{\sum_{s'\in\mathcal{S}_{s,a}} p_{s,a,s'}V^{(n)}(s')\right\}\\
&\geq r(s,a) + \gamma \max_{p\in\mathcal{P}_{t_k}}\left\{\sum_{s'\in\mathcal{S}_{s,a}} p_{s,a,s'}V^{*}(s')\right\}\\
&\geq r(s,a) + \gamma \sum_{s'\in\mathcal{S}_{s,a}} p(s' \mid s,a,\theta^*)V^*(s')\\
&= Q^*(s,a)
\end{aligned}$$

where the first inequality follows from the induction hypothesis that $V^{(n)}(s) \geq V^*(s)$, the second inequality is by Lemma 2, and the last equality is due to the Bellman optimality equation. Moreover,

$$Q^{(n+1)}(s,a) = r(s,a) + \gamma \max_{p\in\mathcal{P}_{t_k}}\left\{\sum_{s'\in\mathcal{S}_{s,a}} p_{s,a,s'}V^{(n)}(s')\right\} \leq r(s,a) + \frac{\gamma}{1-\gamma} \leq \frac{1}{1-\gamma}$$

holds because the induction hypothesis implies that $V^{(n)}(s') \leq 1/(1-\gamma)$ for $s' \in \mathcal{S}_{s,a}$ and $\sum_{s'\in\mathcal{S}_{s,a}} p_{s,a,s'} = 1$ for any $p \in \mathcal{P}_{t_k}$.

Next, we argue that $1/(1-\gamma) \geq V^{(n+1)}(s) \geq V^*(s)$ for $s \in \mathcal{S}$. Since $V^{(n+1)}(s) = \max_{a\in\mathcal{A}} Q^{(n+1)}(s,a)$ and $Q^{(n+1)}(s,a) \leq 1/(1-\gamma)$ for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, it follows that $V^{(n+1)}(s) \leq 1/(1-\gamma)$ for any $s \in \mathcal{S}$. Furthermore, we have

$$V^{(n+1)}(s) = \max_{a\in\mathcal{A}} Q^{(n+1)}(s,a) \geq \max_{a\in\mathcal{A}} Q^*(s,a) = V^*(s).$$

By the induction argument, we have just proved that

$$\frac{1}{1-\gamma} \geq V^{(n)}(s) \geq V^*(s), \quad \frac{1}{1-\gamma} \geq Q^{(n)}(s,a) \geq Q^*(s,a)$$

for any $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $n \in \{0,\ldots,N\}$, as required.

E.2 Proof of Lemma 3: Convergence of Discounted Extended Value Iteration

We will first show the following lemma.

Lemma 17. *Let N be the number of rounds for discounted extended value iteration (`DEVI`). Then $Q^{(N-1)}(s,a) - Q^{(N)}(s,a) \leq \gamma^{N-1}$ for any $(s,a) \in \mathcal{S}\times\mathcal{A}$.*

Proof. Note that for $n \geq 2$, we have

$$\begin{aligned} Q^{(n)}(s,a) &= r(s,a) + \gamma \max_{p\in\mathcal{P}} \left\{ \sum_{s'\in\mathcal{S}_{s,a}} p_{s,a,s'} V^{(n-1)}(s') \right\}, \\ Q^{(n-1)}(s,a) &= r(s,a) + \gamma \max_{p\in\mathcal{P}} \left\{ \sum_{s'\in\mathcal{S}_{s,a}} p_{s,a,s'} V^{(n-2)}(s') \right\}. \end{aligned}$$

This implies that for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, there exists some $\tilde{p} \in \mathcal{P}$ such that

$$\begin{aligned} &Q^{(n-1)}(s,a) - Q^{(n)}(s,a) \\ &= \gamma \left(\max_{p\in\mathcal{P}} \left\{ \sum_{s'\in\mathcal{S}_{s,a}} p_{s,a,s'} V^{(n-2)}(s') \right\} - \max_{p\in\mathcal{P}} \left\{ \sum_{s'\in\mathcal{S}_{s,a}} p_{s,a,s'} V^{(n-1)}(s') \right\} \right) \\ &\leq \gamma \max_{p\in\mathcal{P}} \sum_{s'\in\mathcal{S}_{s,a}} p_{s,a,s'} \left(V^{(n-2)}(s') - V^{(n-1)}(s') \right) \\ &= \gamma \sum_{s'\in\mathcal{S}_{s,a}} \tilde{p}_{s,a,s'} \left(V^{(n-2)}(s') - V^{(n-1)}(s') \right) \end{aligned}$$

where the inequality holds because $\max_p\{f(p)+g(p)\} \leq \max_p\{f(p)\} + \max_p\{g(p)\}$. The right-most side can be further bounded as follows.

$$\begin{aligned} &\gamma \sum_{s'\in\mathcal{S}_{s,a}} \tilde{p}_{s,a,s'} \left(V^{(n-2)}(s') - V^{(n-1)}(s') \right) \\ &\leq \gamma \max_{s'\in\mathcal{S}} \left(V^{(n-2)}(s') - V^{(n-1)}(s') \right) \\ &= \gamma \max_{s'\in\mathcal{S}} \left(\max_{a'\in\mathcal{A}} Q^{(n-2)}(s',a') - \max_{a'\in\mathcal{A}} Q^{(n-1)}(s',a') \right) \\ &\leq \gamma \max_{(s',a')\in\mathcal{S}\times\mathcal{A}} \left(Q^{(n-2)}(s',a') - Q^{(n-1)}(s',a') \right) \end{aligned} \tag{31}$$

where the first inequality holds because the left-most side is a convex combination of $V^{(n-2)}(s') - V^{(n-1)}(s')$ for $s' \in \mathcal{S}_{s,a}$ and the second inequality is due to $\max_{a'}\{f(a')+g(a')\} \leq \max_{a'}\{f(a')\} + \max_{a'}\{g(a')\}$ as before. Therefore, it follows that for any $n \geq 2$,

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(Q^{(n-1)}(s,a) - Q^{(n)}(s,a) \right) \leq \gamma \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(Q^{(n-2)}(s,a) - Q^{(n-1)}(s,a) \right).$$

In particular, this implies that

$$\begin{aligned}\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left(Q^{(N-1)}(s,a)-Q^{(N)}(s,a)\right) &\le \gamma^{N-1}\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left(Q^{(0)}(s,a)-Q^{(1)}(s,a)\right)\\ &= \gamma^{N-1}\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left(\frac{1}{1-\gamma}-r(s,a)-\frac{\gamma}{1-\gamma}\right)\\ &\le \gamma^{N-1}\end{aligned}$$

where the last inequality holds because $r(s,a)\le 1$. □

Based on Lemma 17, we complete the proof of Lemma 3. Note that

$$\begin{aligned}&Q^{(N)}(s_t,a_t)\\ &= r(s_t,a_t)+\gamma\max_{p\in\mathcal{P}_{t_k}}\left\{\sum_{s'\in\mathcal{S}_t} p_{s_t,a_t,s'}V^{(N-1)}(s')\right\}\\ &= r(s_t,a_t)+\gamma\max_{p\in\mathcal{P}_{t_k}}\left\{\sum_{s'\in\mathcal{S}_t} p_{s_t,a_t,s'}V^{(N)}(s')\right\}\\ &\quad+\gamma\max_{p\in\mathcal{P}_{t_k}}\left\{\sum_{s'\in\mathcal{S}_t} p_{s_t,a_t,s'}\left(V^{(N-1)}(s')-V^{(N)}(s')\right)\right\}\\ &\le r(s_t,a_t)+\gamma\max_{p\in\mathcal{P}_{t_k}}\left\{\sum_{s'\in\mathcal{S}_t} p_{s_t,a_t,s'}V^{(N)}(s')\right\}+\gamma\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left(Q^{(N-1)}(s,a)-Q^{(N)}(s,a)\right)\\ &\le r(s_t,a_t)+\gamma\max_{p\in\mathcal{P}_{t_k}}\left\{\sum_{s'\in\mathcal{S}_t} p_{s_t,a_t,s'}V^{(N)}(s')\right\}+\gamma^N\end{aligned}$$

where $V^{(N)}$ is given by V_k, the first inequality can be established following the same argument as in (31), the second inequality is implied by Lemma 17. Since $Q^{(N)}$ equals Q_k and $V^{(N)}$ equals V_k, we have

$$Q_k(s_t,a_t)\le r(s_t,a_t)+\gamma\max_{p\in\mathcal{P}_{t_k}}\left\{\sum_{s'\in\mathcal{S}_t} p_{s_t,a_t,s'}V_k(s')\right\}+\gamma^N,$$

as required.

E.3 Proof of Lemma 5: Bound on the Span of the Optimistic Value Function

First, we prove the following lemma.

Lemma 18. *For any $n\ge 1$ and $s\in\mathcal{S}$, $V^{(n)}(s)\le V^{(n-1)}(s)$.*

Proof. We argue by induction on n. Since $V^{(0)}(s)=1/(1-\gamma)$ and $V^{(1)}(s)\le 1/(1-\gamma)$ by Lemma 4, it is clear that $V^{(1)}(s)\le V^{(0)}(s)$. We assume that for some $n\ge 1$, $V^{(n)}(s)\le V^{(n-1)}(s)$ for any $s\in\mathcal{S}$. Then it follows from the induction hypothesis that

$$\begin{aligned}Q^{(n+1)}(s,a) &= r(s,a)+\gamma\max_{p\in\mathcal{P}_{t_k}}\left\{\sum_{s'\in\mathcal{S}_{s,a}} p_{s,a,s'}V^{(n)}(s')\right\}\\ &\le r(s,a)+\gamma\max_{p\in\mathcal{P}_{t_k}}\left\{\sum_{s'\in\mathcal{S}_{s,a}} p_{s,a,s'}V^{(n-1)}(s')\right\}\\ &= Q^{(n)}(s,a)\end{aligned}$$

for any $(s,a)\in\mathcal{S}\times\mathcal{A}$. This implies that for any $s\in\mathcal{S}$,

$$V^{(n+1)}(s)=\max_{a\in\mathcal{A}}Q^{(n+1)}(s,a)\le\max_{a\in\mathcal{A}}Q^{(n)}(s,a)=V^{(n)}(s),$$

as required. □

Using Lemma 18, we complete the proof of Lemma 5. Let $\tau(\pi)$ denote the number of steps after which state s is reached from state s' for the first time under some policy π. As the diameter of the MDP is at most D, there exists a policy $\widetilde{\pi}$ such that $\mathbb{E}\left[\tau(\widetilde{\pi})\right] \leq D$. For discounted extended value iteration, we may think of the following non-stationary policy. First, starting from the initial state s', we run the policy $\widetilde{\pi}$ under the true transition p^* until we reach state s. If we reach state s within n steps, i.e., $n \geq \tau(\widetilde{\pi})$, then we take the non-stationary policy and the non-stationary transition function that give rise to $V^{(n-\tau(\widetilde{\pi}))}(s)$. Let $V(s')$ denote the total expected discounted reward under this procedure. Note that

$$V(s') \geq \gamma^n \cdot \frac{1}{1-\gamma}$$

since $V^{(0)}(s'') = 1/(1-\gamma)$ for any $s'' \in \mathcal{S}$. Then it follows that

$$V(s') \geq \mathbb{P}\left[n < \tau(\widetilde{\pi})\right] \cdot \frac{\gamma^n}{1-\gamma} + \mathbb{P}\left[n \geq \tau(\widetilde{\pi})\right] \cdot \mathbb{E}\left[\gamma^{\tau(\widetilde{\pi})} V^{(n-\tau(\widetilde{\pi}))}(s) \mid n \geq \tau(\widetilde{\pi})\right].$$

Here, as $V^{(n)}(s) \leq 1/(1-\gamma)$ by Lemma 4, we have

$$\gamma^n \cdot \frac{1}{1-\gamma} \geq \mathbb{E}\left[\gamma^{\tau(\widetilde{\pi})} V^{(n)}(s) \mid n < \tau(\widetilde{\pi})\right].$$

Moreover,

$$\mathbb{E}\left[\gamma^{\tau(\widetilde{\pi})} V^{(n-\tau(\widetilde{\pi}))}(s) \mid n \geq \tau(\widetilde{\pi})\right] \geq \mathbb{E}\left[\gamma^{\tau(\widetilde{\pi})} V^{(n)}(s) \mid n \geq \tau(\widetilde{\pi})\right].$$

Therefore, it follows that

$$\begin{aligned} V(s') &\geq \mathbb{P}\left[n < \tau(\widetilde{\pi})\right] \cdot \mathbb{E}\left[\gamma^{\tau(\widetilde{\pi})} V^{(n)}(s) \mid n < \tau(\widetilde{\pi})\right] \\ &\quad + \mathbb{P}\left[n \geq \tau(\widetilde{\pi})\right] \cdot \mathbb{E}\left[\gamma^{\tau(\widetilde{\pi})} V^{(n)}(s) \mid n \geq \tau(\widetilde{\pi})\right] \\ &= \mathbb{E}\left[\gamma^{\tau(\widetilde{\pi})} V^{(n)}(s)\right] \\ &\geq \gamma^{\mathbb{E}[\tau(\widetilde{\pi})]} V^{(n)}(s) \\ &\geq \gamma^D V^{(n)}(s) \end{aligned}$$

where the second inequality is by Jensen's inequality and the third inequality holds because $\gamma < 1$ and $\mathbb{E}\left[\tau(\widetilde{\pi})\right] \leq D$. Furthermore, it is clear that $V(s') \leq V^{(n)}(s')$. This is because $V^{(n)}(s')$ is the largest possible total expected discounted reward achievable by a policy that maximizes the n-step discounted reward for the discounted extended value iteration procedure. Consequently, we have just proved that

$$\gamma^D V^{(n)}(s) \leq V(s') \leq V^{(n)}(s').$$

This implies that

$$V^{(n)}(s) - V^{(n)}(s') \leq (1-\gamma^D) V^{(n)}(s) \leq \frac{1-\gamma^D}{1-\gamma} \leq D$$

where the second inequality comes from $V^{(n)}(s) \leq 1/(1-\gamma)$ by Lemma 4 and the second inequality holds because $(1-\gamma^D)/(1-\gamma) = 1 + \gamma + \cdots + \gamma^{D-1}$.

F CUMULATIVE ERROR BOUNDS

In this section, we prove Lemmas 14 and 15, providing a tight upper bound on the following.

$$\underbrace{\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} B^{1,t_k}_{s_t,a_t}}_{(\star)} + \underbrace{\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} B^{2,t_k}_{s_t,a_t}}_{(\star\star)}.$$

The following lemmas are useful for our analysis.

Lemma 19. (Abbasi-yadkori et al., 2011, Lemma 12) *Let $A, B \in \mathbb{R}^{d\times d}$ be positive semidefinite matrices such that $A \succeq B$. Then for any $x \in \mathbb{R}^d$, we have $\|x\|_A \le \|x\|_B \sqrt{\det(A)/\det(B)}$.*

Lemma 20. (Oh and Iyengar, 2019, Lemma 7) *Let $x_1, \ldots, x_n \in \mathbb{R}^d$. Then*

$$\det\left(I_d + \sum_{i=1}^n x_i x_i^\top\right) \ge 1 + \sum_{i=1}^n \|x_i\|_2^2.$$

Lemma 21. (Abbasi-yadkori et al., 2011, Lemma 10). *Suppose $x_1, \ldots, x_t \in \mathbb{R}^d$ and $\|x_s\|_2 \le L$ for any $1 \le s \le t$. Let $V_t = \lambda I_d + \sum_{i=1}^t x_i x_i^\top$ for some $\lambda > 0$. Then $\det(V_t)$ is increasing with respect to t and*

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

Based on the lemmas, we prove the following technical lemma that analyzes several summation terms.

Lemma 22. *Let $\lambda \ge L_\varphi^2$. For $t \ge 1$, let $\widehat{\varphi}_{t,s'}$ be defined as*

$$\widehat{\varphi}_{t,s'} = \varphi_{t,s'} - \sum_{s'' \in \mathcal{S}_t} p_{t,s''}(\widehat{\theta}_{t+1})\varphi_{t,s''}.$$

Then the following statements hold.

$$(1) \quad \sum_{t=1}^T \max_{s' \in \mathcal{S}_t} \|\varphi_{t,s'}\|_{\Sigma_t^{-1}}^2 \le \frac{2d}{\kappa} \log\left(1 + \frac{T\mathcal{U}L_\varphi^2}{\lambda d}\right).$$

$$(2) \quad \sum_{t=1}^T \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \|\widehat{\varphi}_{t,s'}\|_{\Sigma_t^{-1}}^2 \le 2d \log\left(1 + \frac{T\mathcal{U}L_\varphi^2}{\lambda d}\right)$$

$$(3) \quad \sum_{t=1}^T \max_{s' \in \mathcal{S}_t} \|\widehat{\varphi}_{t,s'}\|_{\Sigma_t^{-1}}^2 \le \frac{8d}{\kappa} \log\left(1 + \frac{T\mathcal{U}L_\varphi^2}{\lambda d}\right).$$

Proof. **Statement (1):** In (11), we argued that for any t and θ,

$$\nabla_\theta^2(\ell_t(\theta)) \succeq \kappa \sum_{s' \in \mathcal{S}_t \setminus \{\varsigma_t\}} \varphi_{t,s'}\varphi_{t,s'}^\top. \tag{32}$$

Note that if two matrices $A, B \in \mathbb{R}^{d\times d}$ satisfy $A \succeq B$, then for any $x \in \mathbb{R}^d$, $\|x\|_A \ge \|x\|_B$ holds because $A - B$ is positive semidefinite and thus $x^\top(A-B)x \ge 0$. Then Lemma 19 implies that $\det(A) \ge \det(B)$. Note that

$$\begin{aligned}\det(\Sigma_{t+1}) &\ge \det\left(\Sigma_t + \kappa \sum_{s' \in \mathcal{S}_t \setminus \{\varsigma_t\}} \varphi_{t,s'}\varphi_{t,s'}^\top\right) \\ &= \det(\Sigma_t) \cdot \det\left(1 + \sum_{s' \in \mathcal{S}_t \setminus \{\varsigma_t\}} \sqrt{\kappa}\Sigma_t^{-1/2}\varphi_{t,s'}\left(\sqrt{\kappa}\Sigma_t^{-1/2}\varphi_{t,s'}\right)^\top\right) \\ &\ge \det(\Sigma_t)\left(1 + \kappa \sum_{s' \in \mathcal{S}_t \setminus \{\varsigma_t\}} \|\varphi_{t,s'}\|_{\Sigma_t^{-1}}^2\right)\end{aligned} \tag{33}$$

where the first inequality holds due to $\Sigma_{t+1} = \Sigma_t + \nabla_\theta^2(\ell_t(\widehat{\theta}_{t+1}))$ and (32), the equality holds because Σ_t is positive definite, and the second inequality is from Lemma 20.

Moreover, since $\Sigma_t \succeq \Sigma_1 = \lambda I_d$, we have $(1/\lambda)I_d = \Sigma_1^{-1} \succeq \Sigma_t^{-1}$. Then

$$\|\varphi_{t,s'}\|_{\Sigma_t^{-1}}^2 \le \|\varphi_{t,s'}\|_{\Sigma_1^{-1}}^2 = \frac{1}{\lambda}\|\varphi_{t,s'}\|_2^2 \le 1$$

where the first inequality holds as $\Sigma_1^{-1} \succeq \Sigma_t^{-1}$ while the second inequality holds because $\lambda \geq L_\varphi^2$. In particular, as $\kappa \leq 1$, we deduce that

$$\kappa \cdot \max_{s' \in \mathcal{S}_t} \|\varphi_{t,s'}\|_{\Sigma_t^{-1}}^2 \leq 1.$$

Note that for any $z \in [0,1]$, we have $z \leq 2\log(1+z)$, which implies that

$$\kappa \sum_{t=1}^{T} \max_{s' \in \mathcal{S}_t} \|\varphi_{t,s'}\|_{\Sigma_t^{-1}}^2 \leq 2 \sum_{t=1}^{T} \log\left(1 + \kappa \cdot \max_{s' \in \mathcal{S}_t} \|\varphi_{t,s'}\|_{\Sigma_t^{-1}}^2\right). \tag{34}$$

Furthermore, as $\det(\Sigma_1) = \lambda^d$, Lemma 21 implies that

$$\sum_{t=1}^{T} \log\left(1 + \kappa \sum_{s' \in \mathcal{S}_t \setminus \{\varsigma_t\}} \|\varphi_{t,s'}\|_{\Sigma_t^{-1}}^2\right) \leq \log\left(\frac{\det(\Sigma_{T+1})}{\det(\Sigma_1)}\right) \leq d \log\left(1 + \frac{T\mathcal{U}L_\varphi^2}{\lambda d}\right). \tag{35}$$

Combining (34) and (35), it follows that

$$\sum_{t=1}^{T} \max_{s' \in \mathcal{S}_t} \|\varphi_{t,s'}\|_{\Sigma_t^{-1}}^2 \leq \frac{2d}{\kappa} \log\left(1 + \frac{T\mathcal{U}L_\varphi^2}{\lambda d}\right),$$

as required.

Statement (2): Note that

$$\begin{aligned}
\nabla_\theta^2(\ell_t(\widehat{\theta}_{t+1})) &= \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \varphi_{t,s'} \varphi_{t,s'}^\top - \sum_{s' \in \mathcal{S}_t} \sum_{s'' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) p_{t,s''}(\widehat{\theta}_{t+1}) \varphi_{t,s'} \varphi_{t,s''}^\top \\
&= \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \varphi_{t,s'} \widehat{\varphi}_{t,s'}^\top \\
&= \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \widehat{\varphi}_{t,s'} \widehat{\varphi}_{t,s'}^\top + \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \sum_{s'' \in \mathcal{S}_t} p_{t,s''}(\widehat{\theta}_{t+1}) \varphi_{t,s''} \widehat{\varphi}_{t,s'}^\top \\
&= \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \widehat{\varphi}_{t,s'} \widehat{\varphi}_{t,s'}^\top
\end{aligned}$$

where the last equality holds because

$$\begin{aligned}
\sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \widehat{\varphi}_{t,s'} &= \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \varphi_{t,s'} - \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \sum_{s'' \in \mathcal{S}_t} p_{t,s''}(\widehat{\theta}_{t+1}) \varphi_{t,s''} \\
&= \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \varphi_{t,s'} - \sum_{s'' \in \mathcal{S}_t} p_{t,s''}(\widehat{\theta}_{t+1}) \varphi_{t,s''} \\
&= 0.
\end{aligned}$$

Therefore, it follows that

$$\begin{aligned}
&\det(\Sigma_{t+1}) \\
&= \det\left(\Sigma_t + \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \widehat{\varphi}_{t,s'} \widehat{\varphi}_{t,s'}^\top\right) \\
&= \det(\Sigma_t) \cdot \det\left(I_d + \sum_{s' \in \mathcal{S}_t} \sqrt{p_{t,s'}(\widehat{\theta}_{t+1})} \Sigma_t^{-1/2} \widehat{\varphi}_{t,s'} \left(\sqrt{p_{t,s'}(\widehat{\theta}_{t+1})} \Sigma_t^{-1/2} \widehat{\varphi}_{t,s'}\right)^\top\right) \\
&\geq \det(\Sigma_t) \left(1 + \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \|\widehat{\varphi}_{t,s'}\|_{\Sigma_t^{-1}}^2\right)
\end{aligned} \tag{36}$$

where the first equality holds due to $\Sigma_{t+1} = \Sigma_t + \nabla_\theta^2(\ell_t(\widehat{\theta}_{t+1}))$ and (32), the second equality holds because Σ_t is positive definite, and the inequality is from Lemma 20. Moreover,

$$\begin{aligned}
&\sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \, \|\widehat{\varphi}_{t,s'}\|^2_{\Sigma_t^{-1}} \\
&\leq \frac{1}{\lambda} \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \, \|\widehat{\varphi}_{t,s'}\|_2^2 \\
&= \frac{1}{\lambda} \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \left(\varphi_{t,s'} - \sum_{s'' \in \mathcal{S}_t} p_{t,s''}(\widehat{\theta}_{t+1}) \varphi_{t,s''} \right)^\top \left(\varphi_{t,s'} - \sum_{s'' \in \mathcal{S}_t} p_{t,s''}(\widehat{\theta}_{t+1}) \varphi_{t,s''} \right) \\
&= \frac{1}{\lambda} \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \, \|\varphi_{t,s'}\|_2^2 - \frac{1}{\lambda} \left\| \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \varphi_{t,s'} \right\|_2^2 \\
&\leq \frac{1}{\lambda} \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \, \|\varphi_{t,s'}\|_2^2 \\
&\leq \frac{1}{\lambda} L_\varphi^2
\end{aligned}$$

where the first inequality holds because $(1/\lambda) I_d = \Sigma_1^{-1} \succeq \Sigma_t^{-1}$. Since $\lambda \geq L_\varphi^2$, we have

$$\sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \, \|\widehat{\varphi}_{t,s'}\|^2_{\Sigma_t^{-1}} \leq 1.$$

Then we deduce that

$$\begin{aligned}
\sum_{t=1}^T \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \, \|\widehat{\varphi}_{t,s'}\|^2_{\Sigma_t^{-1}} &\leq 2 \sum_{t=1}^T \log \left(1 + \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \, \|\widehat{\varphi}_{t,s'}\|^2_{\Sigma_t^{-1}} \right) \\
&\leq 2 \log \left(\frac{\det(\Sigma_{T+1})}{\det(\Sigma_1)} \right) \\
&\leq 2d \log \left(1 + \frac{TUL_\varphi^2}{\lambda d} \right)
\end{aligned}$$

where the first inequality holds because $z \leq 2\log(1+z)$ for any $z \in [0,1]$, the second inequality follows from (36), and the third inequality is due to Lemma 21.

Statement (3): Note that

$$\|\widehat{\varphi}_{t,s'}\|_{\Sigma_t^{-1}} \leq \|\varphi_{t,s'}\|_{\Sigma_t^{-1}} + \sum_{s'' \in \mathcal{S}_t} p_{t,s''}(\widehat{\theta}_{t+1}) \|\varphi_{t,s''}\|_{\Sigma_t^{-1}} \leq 2 \cdot \max_{s'' \in \mathcal{S}_t} \|\varphi_{t,s''}\|_{\Sigma_t^{-1}},$$

which implies that

$$\|\widehat{\varphi}_{t,s'}\|^2_{\Sigma_t^{-1}} \leq 4 \cdot \max_{s'' \in \mathcal{S}_t} \|\varphi_{t,s''}\|^2_{\Sigma_t^{-1}}.$$

Then statement (3) follows from statement (1), as required. □

we prove Lemmas 14 and 15 which provide upper bounds on terms (⋆) and (⋆⋆), respectively.

F.1 Proof of Lemma 14

Consider the kth episode for $k \in \{1, \ldots, K_T\}$. For $t_k \leq t \leq t_{k+1} - 1$, let us use notations $\bar{\varphi}_{t,s'}$ and $\widehat{\varphi}_{t,s'}$ given by

$$\bar{\varphi}_{t,s'} = \varphi_{t,s'} - \sum_{s'' \in \mathcal{S}_t} p_{t,s''}(\widehat{\theta}_{t_k}) \varphi_{t,s''}, \quad \widehat{\varphi}_{t,s'} = \varphi_{t,s'} - \sum_{s'' \in \mathcal{S}_t} p_{t,s''}(\widehat{\theta}_{t+1}) \varphi_{t,s''}.$$

Then we have

$$\begin{aligned}
&\sum_{t=t_k}^{t_{k+1}-1} B^{1,t_k}_{s_t,a_t} \\
&= \sum_{t=t_k}^{t_{k+1}-1} \beta_t \sum_{s'\in\mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t_k})\|\bar{\varphi}_{t,s'}\|_{\Sigma^{-1}_{t_k}} \\
&\le \beta_T \sum_{t=t_k}^{t_{k+1}-1} \sum_{s'\in\mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t_k})\|\bar{\varphi}_{t,s'}\|_{\Sigma^{-1}_{t_k}} \\
&\le \beta_T \sum_{t=t_k}^{t_{k+1}-1} \sum_{s'\in\mathcal{S}_t} \left(p_{t,s'}(\widehat{\theta}_{t_k})\|\bar{\varphi}_{t,s'} - \widehat{\varphi}_{t,s'}\|_{\Sigma^{-1}_{t_k}} + \left(p_{t,s'}(\widehat{\theta}_{t_k}) - p_{t,s'}(\widehat{\theta}_{t+1}) \right) \|\widehat{\varphi}_{t,s'}\|_{\Sigma^{-1}_{t_k}} \right) \\
&\quad + \beta_T \sum_{t=t_k}^{t_{k+1}-1} \sum_{s'\in\mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1})\|\widehat{\varphi}_{t,s'}\|_{\Sigma^{-1}_{t_k}} \\
&\le 2\beta_T \underbrace{\sum_{t=t_k}^{t_{k+1}-1} \sum_{s'\in\mathcal{S}_t} \left| p_{t,s'}(\widehat{\theta}_{t_k}) - p_{t,s'}(\widehat{\theta}_{t+1}) \right| \|\widehat{\varphi}_{t,s'}\|_{\Sigma^{-1}_{t_k}}}_{(a)} + \beta_T \underbrace{\sum_{t=t_k}^{t_{k+1}-1} \sum_{s'\in\mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1})\|\widehat{\varphi}_{t,s'}\|_{\Sigma^{-1}_{t_k}}}_{(b)}
\end{aligned} \tag{37}$$

where the first inequality holds because β_t increases as t gets large and the last inequality follows from

$$\begin{aligned}
\sum_{s'\in\mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t_k})\|\bar{\varphi}_{t,s'} - \widehat{\varphi}_{t,s'}\|_{\Sigma^{-1}_{t_k}} &= \sum_{s'\in\mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t_k}) \left\| \sum_{s''\in\mathcal{S}_t} \left(p_{t,s''}(\widehat{\theta}_{t+1}) - p_{t,s''}(\widehat{\theta}_{t_k}) \right) \varphi_{t,s''} \right\|_{\Sigma^{-1}_{t_k}} \\
&= \left\| \sum_{s''\in\mathcal{S}_t} \left(p_{t,s''}(\widehat{\theta}_{t+1}) - p_{t,s''}(\widehat{\theta}_{t_k}) \right) \varphi_{t,s''} \right\|_{\Sigma^{-1}_{t_k}} \\
&= \left\| \sum_{s''\in\mathcal{S}_t} \left(p_{t,s''}(\widehat{\theta}_{t+1}) - p_{t,s''}(\widehat{\theta}_{t_k}) \right) \widehat{\varphi}_{t,s''} \right\|_{\Sigma^{-1}_{t_k}}
\end{aligned}$$

as we have

$$\sum_{s''\in\mathcal{S}_t} \left(p_{t,s''}(\widehat{\theta}_{t+1}) - p_{t,s''}(\widehat{\theta}_{t_k}) \right) \sum_{s''''\in\mathcal{S}_t} p_{t,s''''}(\widehat{\theta}_{t+1})\varphi_{t,s''''} = 0.$$

Let us first consider term (a). Note that

$$\left| p_{t,s'}(\widehat{\theta}_{t_k}) - p_{t,s'}(\widehat{\theta}_{t+1}) \right| \le \left| p_{t,s'}(\widehat{\theta}_{t_k}) - p_{t,s'}(\theta^*) \right| + \left| p_{t,s'}(\theta^*) - p_{t,s'}(\widehat{\theta}_t) \right| + \left| p_{t,s'}(\widehat{\theta}_t) - p_{t,s'}(\widehat{\theta}_{t+1}) \right|.$$

Moreover, by the triangle inequality,

$$\|\widehat{\varphi}_{t,s'}\|_{\Sigma^{-1}_{t_k}} \le \|\varphi_{t,s'}\|_{\Sigma^{-1}_{t_k}} + \sum_{s''\in\mathcal{S}_t} p_{t,s''}(\widehat{\theta}_{t+1})\|\varphi_{t,s''}\|_{\Sigma^{-1}_{t_k}} \le 2\max_{s'\in\mathcal{S}_t} \|\varphi_{t,s'}\|_{\Sigma^{-1}_{t_k}}.$$

Then it follows that

$$\begin{aligned}
(a) &\le 2 \sum_{t=t_k}^{t_{k+1}-1} \max_{s'\in\mathcal{S}_t} \|\varphi_{t,s'}\|_{\Sigma^{-1}_{t_k}} \sum_{s'\in\mathcal{S}_t} \left(\left| p_{t,s'}(\widehat{\theta}_{t_k}) - p_{t,s'}(\theta^*) \right| + \left| p_{t,s'}(\theta^*) - p_{t,s'}(\widehat{\theta}_t) \right| \right) \\
&\quad + \sum_{t=t_k}^{t_{k+1}-1} \sum_{s'\in\mathcal{S}_t} \left| p_{t,s'}(\widehat{\theta}_t) - p_{t,s'}(\widehat{\theta}_{t+1}) \right| \|\widehat{\varphi}_{t,s'}\|_{\Sigma^{-1}_{t_k}} \\
&\le \underbrace{4\beta_T \sum_{t=t_k}^{t_{k+1}-1} \max_{s'\in\mathcal{S}_t} \|\varphi_{t,s'}\|^2_{\Sigma^{-1}_{t_k}}}_{(a1)} + \underbrace{\sum_{t=t_k}^{t_{k+1}-1} \sum_{s'\in\mathcal{S}_t} \left| p_{t,s'}(\widehat{\theta}_t) - p_{t,s'}(\widehat{\theta}_{t+1}) \right| \|\widehat{\varphi}_{t,s'}\|_{\Sigma^{-1}_{t_k}}}_{(a2)}
\end{aligned} \tag{38}$$

where the second inequality follows because Lemma 11 holds and $\beta_t \leq \beta_T$ for any $t \leq T$. Here, let us consider term $(a2)$. By Taylor's theorem, for any $s'' \in \mathcal{S}_t$, there exists some $\alpha_{s''} \in [0,1]$ such that $\vartheta_{t,s''} = \alpha_{s''}\widehat{\theta}_{t+1} + (1-\alpha_{s''})\widehat{\theta}_t$ satisfying

$$\begin{aligned}
&p_{t,s''}(\widehat{\theta}_{t+1}) - p_{t,s''}(\widehat{\theta}_t)\\
&= \nabla_\theta(p_{t,s''}(\vartheta_{t,s''}))^\top(\widehat{\theta}_{t+1} - \widehat{\theta}_t)\\
&= \left(p_{t,s''}(\vartheta_{t,s''})\varphi_{t,s''} - p_{t,s''}(\vartheta_{t,s''})\sum_{s'''\in\mathcal{S}_t} p_{t,s'''}(\vartheta_{t,s''})\varphi_{t,s'''}\right)^\top(\widehat{\theta}_{t+1} - \widehat{\theta}_t)\\
&= \left(p_{t,s''}(\vartheta_{t,s''})\widehat{\varphi}_{t,s''} - p_{t,s''}(\vartheta_{t,s''})\sum_{s'''\in\mathcal{S}_t} p_{t,s'''}(\vartheta_{t,s''})\widehat{\varphi}_{t,s'''}\right)^\top(\widehat{\theta}_{t+1} - \widehat{\theta}_t)
\end{aligned}$$

where the last equality holds because

$$p_{t,s''}(\vartheta_{t,s''})\left(1 - \sum_{s'''\in\mathcal{S}_t} p_{t,s'''}(\vartheta_{t,s''})\right)\sum_{s''''\in\mathcal{S}_t} p_{t,s''''}(\widehat{\theta}_{t+1})\varphi_{t,s''''} = 0.$$

This implies that

$$\begin{aligned}
&\sum_{s'\in\mathcal{S}_t}\left|p_{t,s'}(\widehat{\theta}_t) - p_{t,s'}(\widehat{\theta}_{t+1})\right|\|\widehat{\varphi}_{t,s'}\|_{\Sigma_{t_k}^{-1}}\\
&\leq \sqrt{2}\sum_{s'\in\mathcal{S}_t}\left|p_{t,s'}(\widehat{\theta}_t) - p_{t,s'}(\widehat{\theta}_{t+1})\right|\|\widehat{\varphi}_{t,s'}\|_{\Sigma_t^{-1}}\\
&\leq \sqrt{2}\sum_{s''\in\mathcal{S}_t} p_{t,s''}(\vartheta_{t,s''})\left(\|\widehat{\varphi}_{t,s''}\|_{\Sigma_t^{-1}} + \sum_{s'''\in\mathcal{S}_t} p_{t,s'''}(\vartheta_{t,s''})\|\widehat{\varphi}_{t,s'''}\|_{\Sigma_t^{-1}}\right)\left\|\widehat{\theta}_{t+1} - \widehat{\theta}_t\right\|_{\Sigma_t}\|\widehat{\varphi}_{t,s''}\|_{\Sigma_t^{-1}}\\
&\leq 2\sqrt{2}\sum_{s''\in\mathcal{S}_t} p_{t,s''}(\vartheta_{t,s''})\left\|\widehat{\theta}_{t+1} - \widehat{\theta}_t\right\|_{\Sigma_t}\max_{s'''\in\mathcal{S}_t}\|\widehat{\varphi}_{t,s'''}\|_{\Sigma_t^{-1}}^2
\end{aligned} \tag{39}$$

where the first inequality is implied by Lemma 19 because $\Sigma_{t_k}^{-1} \succeq \Sigma_t^{-1}$ and

$$\|\widehat{\varphi}_{t,s''}\|_{\Sigma_{t_k}^{-1}}^2 \leq \|\widehat{\varphi}_{t,s''}\|_{\Sigma_t^{-1}}^2\frac{\det(\Sigma_{t_k}^{-1})}{\det(\Sigma_t^{-1})} = \|\widehat{\varphi}_{t,s''}\|_{\Sigma_t^{-1}}^2\frac{\det(\Sigma_t)}{\det(\Sigma_{t_k})} \leq 2\|\widehat{\varphi}_{t,s''}\|_{\Sigma_t^{-1}}^2$$

and the second inequality is due to the Cauchy-Schwarz inequality. Here, due to our choice of $\widehat{\theta}_{t+1}$ in (5), we have

$$\nabla_\theta(\ell_t(\widehat{\theta}_t))^\top\widehat{\theta}_{t+1} + \frac{1}{2\eta}\|\widehat{\theta}_{t+1} - \widehat{\theta}_t\|_{\widehat{\Sigma}_t}^2 \leq \nabla_\theta(\ell_t(\widehat{\theta}_t))^\top\widehat{\theta}_t,$$

implying in turn that

$$\|\widehat{\theta}_{t+1} - \widehat{\theta}_t\|_{\widehat{\Sigma}_t}^2 \leq 2\eta\nabla_\theta(\ell_t(\widehat{\theta}_t))^\top\left(\widehat{\theta}_t - \widehat{\theta}_{t+1}\right) \leq 2\eta\|\nabla_\theta(\ell_t(\widehat{\theta}_t))\|_{\widehat{\Sigma}_t^{-1}}\|\widehat{\theta}_{t+1} - \widehat{\theta}_t\|_{\widehat{\Sigma}_t}.$$

Therefore, it follows that

$$\|\widehat{\theta}_{t+1} - \widehat{\theta}_t\|_{\widehat{\Sigma}_t} \leq 2\eta\|\nabla_\theta(\ell_t(\widehat{\theta}_t))\|_{\widehat{\Sigma}_t^{-1}}.$$

Moreover, recall that for $t \geq 1$

$$\widehat{\Sigma}_t = \Sigma_t + \eta\nabla_\theta^2(\ell_t(\widehat{\theta}_t)) \succeq \Sigma_t + \eta\kappa\sum_{s'\in\mathcal{S}_t}\varphi_{t,s'}\varphi_{t,s'} \succeq \Sigma_t \succeq \Sigma_1 = \lambda I_d$$

where the first inequality is given as in (32). Hence, we have $\widehat{\Sigma}_t \succeq \Sigma_t$ and $(1/\lambda)I_d = \Sigma_1^{-1} \succeq \widehat{\Sigma}_t^{-1}$. Then it follows that

$$\|\widehat{\theta}_{t+1} - \widehat{\theta}_t\|_{\Sigma_t} \leq \|\widehat{\theta}_{t+1} - \widehat{\theta}_t\|_{\widehat{\Sigma}_t} \leq 2\eta\|\nabla_\theta(\ell_t(\widehat{\theta}_t))\|_{\widehat{\Sigma}_t^{-1}} \leq \frac{2\eta}{\sqrt{\lambda}}\|\nabla_\theta(\ell_t(\widehat{\theta}_t))\|_2. \tag{40}$$

Here, we have

$$\begin{aligned}
\|\nabla_\theta(\ell_t(\widehat{\theta}_t))\|_2 &= \left\| - \sum_{s' \in \mathcal{S}_t} \left(y_{t,s'} - p_{t,s'}(\widehat{\theta}_t) \right) \varphi_{t,s'} \right\|_2 \\
&\leq \left\| \sum_{s' \in \mathcal{S}_t} y_{t,s'} \varphi_{t,s'} \right\|_2 + \left\| \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_t) \varphi_{t,s'} \right\|_2 \\
&\leq 2 \cdot \max_{s' \in \mathcal{S}_t} \|\varphi_{t,s'}\|_2 \\
&\leq 2L_\varphi.
\end{aligned} \tag{41}$$

Combining (39), (40), and (41), we may provide an upper bound on term (a) as follows.

$$\begin{aligned}
&\sum_{t=t_k}^{t_{k+1}-1} \sum_{s' \in \mathcal{S}_t} \left| p_{t,s'}(\widehat{\theta}_t) - p_{t,s'}(\widehat{\theta}_{t+1}) \right| \|\widehat{\varphi}_{t,s'}\|_{\Sigma_t^{-1}} \\
&\leq \frac{8\sqrt{2}L_\varphi \eta}{\sqrt{\lambda}} \sum_{t=t_k}^{t_{k+1}-1} \sum_{s'' \in \mathcal{S}_t} p_{t,s''}(\vartheta_{t,s''}) \max_{s''' \in \mathcal{S}_t} \|\widehat{\varphi}_{t,s'''}\|_{\Sigma_t^{-1}}^2 \\
&= \frac{8\sqrt{2}L_\varphi \eta}{\sqrt{\lambda}} \sum_{t=t_k}^{t_{k+1}-1} \max_{s'' \in \mathcal{S}_t} \|\widehat{\varphi}_{t,s''}\|_{\Sigma_t^{-1}}^2
\end{aligned} \tag{42}$$

Moreover, Lemma 19 implies that for term $(a1)$,

$$4\beta_T \sum_{t=t_k}^{t_{k+1}-1} \max_{s' \in \mathcal{S}_t} \|\varphi_{t,s'}\|_{\Sigma_{t_k}^{-1}}^2 \leq 8\beta_T \sum_{t=t_k}^{t_{k+1}-1} \max_{s' \in \mathcal{S}_t} \|\varphi_{t,s'}\|_{\Sigma_t^{-1}}^2 \tag{43}$$

and for term (b),

$$\sum_{t=t_k}^{t_{k+1}-1} \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \|\widehat{\varphi}_{t,s'}\|_{\Sigma_{t_k}^{-1}} \leq \sqrt{2} \sum_{t=t_k}^{t_{k+1}-1} \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \|\widehat{\varphi}_{t,s'}\|_{\Sigma_t^{-1}}. \tag{44}$$

By (37), we may deduce the following upper bound on term $(\star)$.

$$\begin{aligned}
&\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} B_{s_t,a_t}^{1,t_k} \\
&\leq \beta_T \left(16\beta_T \sum_{t=1}^{T} \max_{s'' \in \mathcal{S}_t} \|\varphi_{t,s''}\|_{\Sigma_t^{-1}}^2 + \frac{16\sqrt{2}L_\varphi \eta}{\sqrt{\lambda}} \sum_{t=1}^{T} \max_{s'' \in \mathcal{S}_t} \|\widehat{\varphi}_{t,s''}\|_{\Sigma_t^{-1}}^2 + \sqrt{2} \sum_{t=1}^{T} \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \|\widehat{\varphi}_{t,s'}\|_{\Sigma_t^{-1}} \right) \\
&\leq \beta_T \left(\left(\frac{32\beta_T}{\kappa} + \frac{128\sqrt{2}L_\varphi \eta}{\kappa\sqrt{\lambda}} \right) d \log \left(1 + \frac{T\mathcal{U}L_\varphi^2}{\lambda d} \right) + \sqrt{2 \sum_{t=1}^{T} \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \cdot \sum_{t=1}^{T} \sum_{s' \in \mathcal{S}_t} p_{t,s'}(\widehat{\theta}_{t+1}) \|\widehat{\varphi}_{t,s'}\|_{\Sigma_t^{-1}}^2} \right) \\
&\leq \beta_T \left(\left(\frac{32\beta_T}{\kappa} + \frac{128\sqrt{2}L_\varphi \eta}{\kappa\sqrt{\lambda}} \right) d \log \left(1 + \frac{T\mathcal{U}L_\varphi^2}{\lambda d} \right) + 2\sqrt{dT \log \left(1 + \frac{T\mathcal{U}L_\varphi^2}{\lambda d} \right)} \right)
\end{aligned}$$

where the first inequality is due to (37), (38), (42), (43), and (44), the second inequality is by the Cauchy-Schwarz inequality, and the third inequality follows from Lemma 22.

F.2 Proof of Lemma 15

Note that

$$\begin{aligned}\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} B^{2,t_k}_{s_t,a_t} &= 3\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} \beta_t^2 \max_{s'\in\mathcal{S}_t}\|\varphi_{t,s'}\|^2_{\Sigma^{-1}_{t_k}}\\ &\le 3\beta_T^2\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} \max_{s'\in\mathcal{S}_t}\|\varphi_{t,s'}\|^2_{\Sigma^{-1}_{t_k}}\\ &\le 6\beta_T^2\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} \max_{s'\in\mathcal{S}_t}\|\varphi_{t,s'}\|^2_{\Sigma^{-1}_{t}}\\ &\le \frac{12d}{\kappa}\beta_T^2\log\left(1+\frac{T\mathcal{U}L_\varphi^2}{\lambda d}\right)\end{aligned}$$

where the first inequality holds because β_t increases as t gets large, the second inequality is by Lemma 19, and the third inequality follows from Lemma 22.

G LOWER BOUND PROOF FOR THE FINITE-HORIZON EPISODIC SETTING

Recall that the transition core $\bar\theta_h$ for each step $h\in[H]$ is given by

$$\bar\theta_h=\left(\frac{\theta_h}{\alpha},\frac{1}{\beta}\right)\quad\text{where}\quad \theta_h\in\left\{-\bar\Delta,\bar\Delta\right\}^{d-1},\quad \bar\Delta=\frac{1}{d-1}\log\left(\frac{(1-\delta)(\delta+(d-1)\Delta)}{\delta(1-\delta-(d-1)\Delta)}\right),$$

and $\delta=1/H$ and $\Delta=1/(4\sqrt{2HK})$.

G.1 Linear Approximation of the Multinomial Logit Model

We consider a multinomial logit function given by $f:\mathbb{R}\to\mathbb{R}$ as

$$f(x)=\frac{1}{1+\frac{1-\delta}{\delta}\exp(-x)}.$$

In contrast to the infinite-horizon average-reward case, we take $\delta=1/H$ where H is the horizon of each episode. Recall that the derivative of f is given by

$$f'(x)=\frac{\frac{1-\delta}{\delta}\exp(-x)}{\left(1+\frac{1-\delta}{\delta}\exp(-x)\right)^2}=f(x)-f(x)^2.$$

For simplicity, for $h\in[H]$, we use notation p_{θ_h} given by

$$p_{\theta_h}(x_i\mid x_h,a):=p(x_i\mid x_h,a,\bar\theta_h)=\begin{cases}f(a^\top\theta_h), & \text{if } i=H+2\\ 1-f(a^\top\theta_h), & \text{if } i=h+1.\end{cases}.$$

Note that $-(d-1)\bar\Delta\le a^\top\theta_h\le(d-1)\bar\Delta$ for any $a\in\mathcal{A}$, which means that $f(-(d-1)\bar\Delta)\le p_{\theta_h}(x_{H+2}\mid x_h,a)\le f((d-1)\bar\Delta)$. The following lemma is analogous to Lemma 29.

Lemma 23. *For any $x,y\in[-(d-1)\bar\Delta,(d-1)\bar\Delta]$ with $x\ge y$, we have*

$$0\le f(x)-f(y)\le(\delta+(d-1)\Delta)(x-y).$$

Proof. By the mean value theorem, there exists $y\le z\le x$ such that $f(x)-f(y)=f'(z)(x-y)$. Note that $f'(z)=f(z)-f(z)^2\le f(z)\le f((d-1)\bar\Delta)=\delta+(d-1)\Delta$ where the last equality holds by our choice of $\bar\Delta$. □

G.2 Basic Properties of the Hard Finite-Horizon Episodic MDP Instance

Recall that δ and Δ are given by

$$\delta = \frac{1}{D} \quad \text{and} \quad \Delta = \frac{1}{45\sqrt{(2/5)\log 2}} \cdot \frac{(d-1)}{\sqrt{DT}},$$

respectively. The following lemma characterizes the sizes of parameters δ and Δ under the setting of our hard-t0-learn MDP.

Lemma 24. *Suppose that $T \geq H^3(d-1)^2/32$. Then $(d-1)\Delta \leq \delta/H$*

Proof. Note that $(d-1)\Delta \leq \delta/H$ if and only if $K \geq H^3(d-1)^2/32$. □

The following lemma provides upper bounds on L_φ and L_θ. Moreover,

Lemma 25. *Suppose that $H \geq 3$. For any $\bar{\theta} = (\theta/\alpha, 1/\beta)$, we have $\|\bar{\theta}\|_2 \leq 3/2$. Moreover, for any $a \in \mathcal{A}$ and $(i,j) \in \{(h,h+1) : h \in [H]\} \cup \{(h, H+2) : h \in [H]\}$, $\|\varphi(x_i, a, x_j)\|_2 \leq 1 + \log(H-1)$.*

Proof. Recall that $\alpha = \sqrt{\bar{\Delta}/(1+(d-1)\bar{\Delta})}$ and $\beta = \sqrt{1/(1+(d-1)\bar{\Delta})}$. Moreover,

$$\|\bar{\theta}\|_2^2 = \frac{\|\theta\|_2^2}{\alpha^2} + \frac{1}{\beta^2} = (1+(d-1)\bar{\Delta})^2.$$

Note that

$$(d-1)\bar{\Delta} = \log\left(\frac{(d-1)\Delta + \delta(1-\delta-(d-1)\Delta)}{\delta(1-\delta-(d-1)\Delta)}\right) \leq \frac{(d-1)\Delta}{\delta(1-\delta-(d-1)\Delta)}.$$

Since $(d-1)\Delta \leq \delta/H$ by Lemma 24, it follows that

$$(d-1)\bar{\Delta} \leq \frac{1}{H} \cdot \frac{1}{1-\frac{H+1}{H}\delta} = \frac{1}{H-(H+1)/H} \leq \frac{1}{H-1},$$

which implies that $\|\bar{\theta}\|_2 \leq 1 + \bar{\Delta} \leq 3/2$. Moreover, for any $(i,j) \in \{(h,h+1) : h \in [H]\} \cup \{(h,H+2) : h \in [H]\}$,

$$\begin{aligned}\|\varphi(x_i, a, x_j)\|^2 &\leq \alpha^2\|a\|_2^2 + \beta^2(\log(H-1))^2 \\ &= \frac{(d-1)\bar{\Delta}}{1+(d-1)\bar{\Delta}} + \frac{(\log(H-1))^2}{1+(d-1)\bar{\Delta}} \\ &\leq (1+\log(H-1))^2,\end{aligned}$$

as required. □

G.3 Proof of Theorem 3

Let $\pi = \{\pi_h\}_{h=1}^H$ be a policy for the H-horizon MDP. Recall that the value function V_1^π under policy π is given by

$$V_1^\pi(x_1) = \mathbb{E}_{\theta,\pi}\left[\sum_{h=1}^H r(s_h, a_h) \mid s_1 = x_1\right]$$

where the expectation is taken with respect to the distribution that has dependency on the transition core θ and the policy π. Let N_h denote the event that the process visits state x_h in step h and then enters x_{H+2}, i.e., $N_h = \{s_h = x_h, x_{h+1} = x_{H+2}\}$. Then we have that

$$V_1^\pi(x_1) = \sum_{h=1}^{H-1}(H-h)\mathbb{P}_{\theta,\pi}(N_h \mid s_1 = x_1).$$

Moreover, note that

$$\begin{aligned}
&\mathbb{P}_{\theta,\pi}(s_{h+1} = x_{H+2} \mid s_h = x_h, s_1 = x_1) \\
&= \sum_{a\in\mathcal{A}} \mathbb{P}_{\theta,\pi}(s_{h+1} = x_{H+2} \mid s_h = x_h, a_h = a)\mathbb{P}_{\theta,\pi}(a_h = a \mid s_h = x_h, s_1 = x_1) \\
&= \sum_{a\in\mathcal{A}} f(a^\top\theta_h)\mathbb{P}_{\theta,\pi}(a_h = a \mid s_h = x_h, s_1 = x_1) \\
&= \delta + \underbrace{\sum_{a\in\mathcal{A}} (f(a^\top\theta_h) - \delta)\mathbb{P}_{\theta,\pi}(a_h = a \mid s_h = x_h, s_1 = x_1)}_{a_h}.
\end{aligned}$$

Then it follows that

$$\mathbb{P}_{\theta,\pi}(s_{h+1} = x_{h+1} \mid s_h = x_h, s_1 = x_1) = 1 - \delta - a_h,$$

which implies that

$$\mathbb{P}_{\theta,\pi}(N_h) = (\delta + a_h)\prod_{j=1}^{h-1}(1 - \delta - a_j).$$

Therefore, we deduce that

$$V_1^\pi(x_1) = \sum_{h=1}^{H}(H-h)(\delta + a_h)\prod_{j=1}^{h-1}(1 - \delta - a_j).$$

Note that the optimal policy $\pi^* = \{\pi_h^*\}_{h=1}^H$ deterministically chooses the action maximizing $a^\top\theta_h$ at each step h. Recall that the maximum value of $a^\top\theta_h$ is $(d-1)\bar{\Delta}$ for any h, and moreover, $f((d-1)\bar{\Delta}) = \delta + (d-1)\Delta$. Therefore, under the optimal policy,

$$\mathbb{P}_{\theta,\pi^*}(s_{h+1} = x_{H+2} \mid s_h = x_h, s_1 = x_1) = \delta + (d-1)\Delta$$

This further implies that the value function under the optimal policy is given by

$$V_1^*(x_1) = \sum_{h=1}^{H}(H-h)(\delta + (d-1)\Delta)(1 - \delta - (d-1)\Delta)^{h-1}.$$

Next, let us define S_i and T_i for $i \in [H]$ as follows.

$$S_i = \sum_{h=i}^{H}(H-h)(\delta + a_h)\prod_{j=i}^{h-1}(1 - \delta - a_j) \text{ and } T_i = \sum_{h=i}^{H}(H-h)(\delta + (d-1)\Delta)(1 - \delta - (d-1)\Delta)^{h-i}.$$

Following the induction argument of (Zhou et al., 2021a, Equation (C.25)) we may deduce that

$$T_1 - S_1 = \sum_{h=1}^{H-1}((d-1)\Delta - a_h)(H - h - T_{h+1})\prod_{j=1}^{h-1}(1 - \delta - a_j).$$

Moreover, since $3(d-1)\Delta \le \delta = 1/H$ and $H \ge 3$ by Lemma 24, it follows from (Zhou et al., 2021a, Equations (C.26)) that $H - h - T_{h+1} \ge H/3$ for $h \le H/2$. Moreover, as $a_j \le (d-1)\Delta \le \delta/3$, we have $\delta + a_j \le 4\delta/3$. Since $H \ge 3$, it holds that

$$\prod_{j=1}^{h-1}(1 - \delta - a_j) \ge \left(1 - \frac{4\delta}{3}\right)^H \ge \frac{1}{3}.$$

Consequently, we deduce that

$$V_1^*(x_1) - V_1^\pi(x_1) = T_1 - S_1 \ge \frac{H}{10}\sum_{h=1}^{H/2}((d-1)\Delta - a_h). \tag{45}$$

From the right-hand side of (45), we have that

$$(d-1)\Delta = \max_{a\in\mathcal{A}} \mu_h^\top a \quad \text{where} \quad \mu_h = \frac{\Delta}{\bar{\Delta}}\theta_h \in \{-\Delta, \Delta\}^{d-1}.$$

Moreover, note that

$$f(\theta_h^\top a) - \delta \leq (\delta + (d-1)\Delta)\theta_h^\top a = \frac{\bar{\Delta}(\delta + (d-1)\Delta)}{\Delta}\mu_h^\top a \leq \frac{\delta + (d-1)\Delta}{\delta(1-\delta-(d-1)\Delta)}\mu_h^\top a$$

where the first inequality is due to Lemma 23 and the second inequality holds because

$$\bar{\Delta} = \frac{1}{d-1}\log\left(1 + \frac{(d-1)\Delta}{\delta(1-\delta-(d-1)\Delta)}\right) \leq \frac{1}{d-1}\cdot\frac{(d-1)\Delta}{\delta(1-\delta-(d-1)\Delta)} = \frac{\Delta}{\delta(1-\delta-(d-1)\Delta)}.$$

Furthermore, as $(d-1)\Delta \leq \delta/H$ by Lemma 24, we have

$$\frac{\delta + (d-1)\Delta}{\delta(1-\delta-(d-1)\Delta)} \leq \frac{(1+1/H)\delta}{\delta(1-(1+1/H)\delta)} = \frac{H^2+H}{H^2-H-1} = 1 + \frac{2H+1}{H^2-H-1} \leq 1 + \frac{3}{H}$$

where the first inequality holds because $(d-1)\Delta \leq \delta/H$, the first equality holds due to $\delta = 1/H$, and the last inequality is by $H \geq 3$. Then it follows that

$$f(\theta_h^\top a) - \delta \leq \frac{\delta + (d-1)\Delta}{\delta(1-\delta-(d-1)\Delta)}\mu_h^\top a \leq \mu_h^\top a + \frac{3}{H}\mu_h^\top a \leq \mu_h^\top a + \frac{3(d-1)\Delta}{H}$$

where the last inequality holds because $\mu_h \in \{-\Delta, \Delta\}^{d-1}$ and thus $\mu_h^\top a \leq (d-1)\Delta$. This in turn implies that

$$a_h \leq \frac{3(d-1)\Delta}{H} + \mu_h^\top \underbrace{\sum_{a\in\mathcal{A}} \mathbb{P}_{\theta,\pi}(a_h = a \mid s_h = x_h, s_1 = x_1)\cdot a}_{\bar{a}_h^\pi}$$

Based on (45), we get

$$V_1^*(x_1) - V_1^\pi(x_1) = T_1 - S_1 \geq \frac{H}{10}\sum_{h=1}^{H/2}\left(\max_{a\in\mathcal{A}}\mu_h^\top a - \mu_h^\top \bar{a}_h^\pi\right) - \frac{H(d-1)\Delta}{20}. \tag{46}$$

Let $\mathfrak{A}$ be an algorithm that takes policy $\pi^k = \{\pi_h^k\}_{h=1}^H$ for episodes $k \in [K]$. Then we deduce from (46) that

$$\begin{aligned}\mathbb{E}\left[\text{Regret}(M_\theta, \mathfrak{A}, K)\right] &= \mathbb{E}\left[\sum_{k=1}^K \left(V_1^*(x_1) - V_1^{\pi^k}(x_1)\right)\right] \\ &\geq \frac{H}{10}\sum_{h=1}^{H/2}\underbrace{\mathbb{E}\left[\sum_{k=1}^K\left(\max_{a\in\mathcal{A}}\mu_h^\top a - \mu_h^\top \bar{a}_h^{\pi^k}\right)\right]}_{I_h(\theta,\pi)} - \frac{H(d-1)}{20}K\Delta.\end{aligned} \tag{47}$$

Here, we now argue that the term $I_h(\theta, \pi)$ corresponds to the regret under a bandit algorithm for a linear bandit problem. Let $\mathcal{L}_{\mu_h}$ denote the linear bandit problem parameterized by $\mu_h \in \{-\Delta, \Delta\}^{d-1}$ where the action set is $\mathcal{A} = \{-1, 1\}^{d-1}$ and the reward distribution for taking action $a \in \mathcal{A}$ is a Bernoulli distribution $B(\delta + \mu_h^\top a)$. Recall that $\bar{a}_h^{\pi^k}$ is given by

$$\bar{a}_h^{\pi^k} = \sum_{a\in\mathcal{A}} \mathbb{P}_{\theta,\pi^k}(a_h = a \mid s_h = x_h, s_1 = x_1)\cdot a.$$

Basically, $\mathfrak{A}$ corresponds to a bandit algorithm that takes action $a \in \mathcal{A}$ with probability $\mathbb{P}_{\theta,\pi^k}(a_h = a \mid s_h = x_h, s_1 = x_1)$ in episode k. Let $a_h^{\pi^k}$ denote the random action taken by $\mathfrak{A}$. Then by linearity of expectation,

$$I_h(\theta, \pi) = \mathbb{E}\left[\sum_{k=1}^K\left(\max_{a\in\mathcal{A}}\mu_h^\top a - \mu_h^\top a_h^{\pi^k}\right)\right]$$

where the expectation is taken with respect to the randomness generated by $\mathfrak{A}$ and which is the expected pseudo-regret under $\mathfrak{A}$. The following lemma provides a lower bound on the expected pseudo-regret for the particular linear bandit instance.

Lemma 26. (Zhou et al., 2021a, Lemma C.8). *Suppose that $0 < \delta \leq 1/3$ and $K \geq (d-1)^2/(2\delta)$. Let $\Delta = 4\sqrt{2\delta/K}$ and consider the linear bandit problems $\mathcal{L}_{\mu_h}$ described above. Then for any bandit algorithm $\mathfrak{A}$, there exists a parameter $\mu_h^* \in \{-\Delta, \Delta\}^{d-1}$ such that the expected pseudo-regret of $\mathfrak{A}$ over the first K steps on $\mathcal{L}_{\mu_h^*}$ is at least $(d-1)\sqrt{K\delta}/(8\sqrt{2})$.*

Applying Lemma 26 to (47), we deduce that

$$\begin{aligned}\mathbb{E}\left[\mathrm{Regret}(M_\theta, \mathfrak{A}, K)\right] &\geq \frac{H^{3/2}(d-1)\sqrt{K}}{160\sqrt{2}} - \frac{H^{1/2}(d-1)\sqrt{K}}{80\sqrt{2}}\\ &\geq \frac{H^{3/2}(d-1)\sqrt{K}}{160\sqrt{2}} - \frac{H^{3/2}(d-1)\sqrt{K}}{240\sqrt{2}}\\ &= \frac{H^{3/2}(d-1)\sqrt{K}}{480\sqrt{2}}\end{aligned}$$

where the second inequality holds because $H \geq 3$.

H LOWER BOUND PROOFS FOR THE INFINITE-HORIZON SETTING

Recall that the transition core $\bar{\theta}$ is given by

$$\bar{\theta} = \left(\frac{\theta}{\alpha}, \frac{1}{\beta}\right) \quad \text{where} \quad \theta \in \left\{-\frac{\bar{\Delta}}{d-1}, \frac{\bar{\Delta}}{d-1}\right\}^{d-1}, \quad \bar{\Delta} = \log\left(\frac{(1-\delta)(\delta+\Delta)}{\delta(1-\delta-\Delta)}\right),$$

and $\Delta = (d-1)/(45\sqrt{(2/5)(T/\delta)\log 2})$. Moreover, we set δ as

$$\delta = \begin{cases} 1/D & \text{for the average-reward case,}\\ 1-\gamma & \text{for the discounted-reward case.}\end{cases}$$

The following lemma characterizes the sizes of parameters δ and Δ under the setting of our hard-to-learn MDP.

Lemma 27. *Suppose that $d \geq 2$, $\delta \leq 1/101$, $T \geq 45(d-1)^2/\delta$. Then the following statements hold.*

$$100\Delta \leq \delta, \quad 2\delta + \Delta \leq 1, \quad \Delta \leq \delta(1-\delta), \quad \frac{1}{\delta} \leq \left(\frac{3}{2}\cdot\frac{4}{5}\cdot\left(\frac{99}{101}\right)^4 - 1\right)T.$$

Proof. If $T \geq 45(d-1)^2/\delta$, then $T \geq (100/15)^2(d-1)^2/\delta$. Note that $\sqrt{(2/5)\log 2} > 1/3$. Then $100\Delta < (100/15)(d-1)/\sqrt{T/\delta}$, and as $T \geq (100/15)^2(d-1)^2/\delta$, we get that $100\Delta < \delta$. Moreover, since $\delta \leq 1/3$, we also have that $2\delta + \Delta \leq 1$ and $\Delta \leq \delta(1-\delta)$. Moreover, we know that

$$\frac{3}{2}\cdot\frac{4}{5}\cdot\left(\frac{99}{101}\right)^4 > \frac{11}{10}.$$

Since $T \geq 45(d-1)^2/\delta \geq 10/\delta$, the last inequality holds. □

The following lemma provides upper bounds on L_φ and L_θ.

Lemma 28. *For any $\bar{\theta} = (\theta/\alpha, 1/\beta)$, we have $\|\bar{\theta}\|_2 \leq 100/99$. Moreover, for any $a \in \mathcal{A}$ and $i, j \in \{0, 1\}$, $\|\varphi(x_i, a, x_j)\|_2 \leq 1 + \log((1/\delta) - 1)$.*

Proof. Recall that $\alpha = \sqrt{\bar{\Delta}/((d-1)(1+\bar{\Delta}))}$ and $\beta = \sqrt{1/(1+\bar{\Delta})}$. Moreover,

$$\|\bar{\theta}\|_2^2 = \frac{\|\theta\|_2^2}{\alpha^2} + \frac{1}{\beta^2} = (1+\bar{\Delta})^2.$$

Note that

$$\bar{\Delta} = \log\left(\frac{(1-\delta)(\delta+\Delta)}{\delta(1-\delta-\Delta)}\right) = \log\left(\frac{\Delta + \delta(1-\delta-\Delta)}{\delta(1-\delta-\Delta)}\right) \leq \frac{\Delta}{\delta(1-\delta-\Delta)}.$$

Then it follows from Lemma 27 that

$$\bar{\Delta} \leq \frac{1}{100} \cdot \frac{1}{1 - \frac{101}{100}\delta} = \frac{1}{100 - 101\delta} \leq \frac{1}{99},$$

which implies that $\|\bar{\theta}\|_2 \leq 1 + \bar{\Delta} \leq 100/99$. Moreover, for any $i, j \in \{0, 1\}$,

$$\|\varphi(x_i, a, x_j)\|^2 \leq \alpha^2 \|a\|_2^2 + \beta^2 (\log((1/\delta) - 1))^2 = \frac{\bar{\Delta}}{1 + \bar{\Delta}} + \frac{(\log((1/\delta) - 1))^2}{1 + \bar{\Delta}},$$

in which case, we have $(1 + \log((1/\delta) - 1))^2 \leq (1 + \log((1/\delta) - 1))^2$, as required. □

H.1 Linear Approximation of the Multinomial Logit Model

Let us define a function $f : \mathbb{R} \to \mathbb{R}$ as

$$f(x) = \frac{1}{1 + \frac{1-\delta}{\delta} \exp(-x)}.$$

The derivative of f is given by

$$f'(x) = \frac{\frac{1-\delta}{\delta} \exp(-x)}{\left(1 + \frac{1-\delta}{\delta} \exp(-x)\right)^2} = f(x) - f(x)^2.$$

The following lemma bridges the multinomial logit function x and a linear function based on the mean value theorem.

Lemma 29. *For any $x, y \in [-\bar{\Delta}, \bar{\Delta}]$ with $x \geq y$, we have*

$$0 \leq f(x) - f(y) \leq (\delta + \Delta)(x - y).$$

Proof. By the mean value theorem, there exists $y \leq z \leq x$ such that $f(x) - f(y) = f'(z)(x - y)$. Note that $f'(z) = f(z) - f(z)^2 \leq f(z) \leq f(\bar{\Delta}) = \delta + \Delta$ where the last equality holds by our choice of $\bar{\Delta}$. □

By our choice of feature vector φ and transition core $\bar{\theta} = (\theta/\alpha, 1/\beta)$, we have

$$p(x_1 \mid x_0, a) = \frac{1}{1 + ((1/\delta) - 1)\exp(-a^\top \theta)} = f(a^\top \theta) \quad \text{and} \quad p(x_0 \mid x_1, a) = \delta = f(0).$$

H.2 Upper Bound on the Number of Visits to State 1

For simplicity, we introduce notation p_θ given by

$$p_\theta(x_j \mid x_i, a) := p(x_j \mid x_i, a, \bar{\theta})$$

for any $i, j \in \{0, 1\}$. Note that inducing a higher probability of transitioning to x_1 from x_0 results in a larger reward. This means that the optimal policy chooses action a that maximizes $a^\top \theta$ so that $p(x_1 \mid x_0, a)$ is maximized. Then under the optimal policy, we take action a_θ such that $a_\theta^\top \theta = \bar{\Delta}$. Hence, the transition probability under the optimal policy is given by

$$p_\theta(x_1 \mid x_0, a_\theta) = f(\bar{\Delta}) = \delta + \Delta$$

where the second equality follows from our choice of $\bar{\Delta}$.

To provide a lower bound, it is sufficient to consider deterministic stationary policies (Auer et al., 2002; Puterman, 2014). Let π be a deterministic (non-stationary) policy. Let $\mathcal{P}_\theta$ denote the distribution over $\mathcal{S}^T$ where $s_1 = x_0$, a_t is determined by π, and s_{t+1} is sampled from $p_\theta(\cdot \mid s_t, a_t)$. Let $\mathbb{E}_\theta$ denote the expectation taken over $\mathcal{P}_\theta$. Moreover, we define N_i for $i \in \{0, 1\}$ and N_0^a as the number of times x_i is visited for $i \in \{0, 1\}$ and the number of time steps in which state x_0 is visited and action a is chosen. We also define $N_0^{\mathcal{V}}$ for $\mathcal{V} \subseteq \mathcal{A}$ as the number of time steps in which state x_0 is visited and an action from the set $\mathcal{V}$ is chosen.

In this section, we analyze term $\mathbb{E}_\theta N_1$ and provide an upper bound on it, which is crucial for coming up the desired lower bounds for the both average-reward and discounted-reward settings. We prove the following lemma that is analogous to (Wu et al., 2022, Lemma C.2).

Lemma 30. *Suppose that* $2\delta+\Delta\le 1$, $\Delta\le\delta(1-\delta)$, *and*

$$\frac{1}{\delta}\le\left(\frac{3}{2}\cdot\frac{4}{5}\cdot\left(\frac{99}{101}\right)^4-1\right)T.$$

Then

$$\mathbb{E}_\theta N_1\le\frac{T}{2}+\frac{\delta+\Delta}{2\delta}\sum_{a\in\mathcal{A}}a^\top\theta\cdot\mathbb{E}_\theta N_0^a\quad\textit{and}\quad\mathbb{E}_\theta N_0\le\left(\frac{99}{101}\right)^4\cdot\frac{4}{5}T.$$

Proof. See Lemma H.5. □

Note that since $a\in\{-1,1\}^{d-1}$,

$$(\delta+\Delta)a^\top\theta\le(\delta+\Delta)\frac{\bar{\Delta}}{d-1}\sum_{j=1}^{d-1}\mathbf{1}\left\{\text{sign}(a_j)=\text{sign}(\theta_j)\right\}.$$

Moreover,

$$\bar{\Delta}=\log\left(\frac{(1-\delta)(\delta+\Delta)}{\delta(1-\delta-\Delta)}\right)=\log\left(\frac{\Delta+\delta(1-\delta-\Delta)}{\delta(1-\delta-\Delta)}\right)\le\frac{\Delta}{\delta(1-\delta-\Delta)}$$

where the inequality holds because $1+x\le\exp(x)$ for any $x\in\mathbb{R}$. Moreover, since $100\Delta\le\delta$ and $\delta\le 1/101$ by Lemma 27, we have

$$(\delta+\Delta)\bar{\Delta}\le\frac{(\delta+\Delta)}{\delta(1-\delta-\Delta)}\le\frac{101}{100}\cdot\frac{1}{1-\frac{101}{100}\delta}\cdot\Delta\le\frac{101}{99}\Delta. \tag{48}$$

Then it follows from Lemma 30 that

$$\begin{aligned}\frac{1}{|\Theta|}\sum_{\theta\in\Theta}\mathbb{E}_\theta N_1&\le\frac{T}{2}+\frac{1}{|\Theta|}\sum_{\theta\in\Theta}\frac{\Delta}{\delta(d-1)}\sum_{a\in\mathcal{A}}\sum_{j=1}^{d-1}\mathbf{1}\left\{\text{sign}(a_j)=\text{sign}(\theta_j)\right\}\frac{101\mathbb{E}_\theta N_0^a}{198}\\&\le\frac{T}{2}+\frac{101\Delta}{198\delta|\Theta|(d-1)}\sum_{j=1}^{d-1}\sum_{\theta\in\Theta}\sum_{a\in\mathcal{A}}\mathbb{E}_\theta\left[\mathbf{1}\left\{\text{sign}(a_j)=\text{sign}(\theta_j)\right\}N_0^a\right].\end{aligned} \tag{49}$$

For a given θ and a coordinate $j\in[d-1]$, we consider θ' that differs from θ only in the jth coordinate. Then we have

$$\begin{aligned}&\mathbb{E}_\theta\left[\mathbf{1}\left\{\text{sign}(a_j)=\text{sign}(\theta_j)\right\}N_0^a\right]+\mathbb{E}_{\theta'}\left[\mathbf{1}\left\{\text{sign}(a_j)=\text{sign}(\theta_j')\right\}N_0^a\right]\\&=\mathbb{E}_{\theta'}N_0^a+\mathbb{E}_\theta\left[\mathbf{1}\left\{\text{sign}(a_j)=\text{sign}(\theta_j)\right\}N_0^a\right]-\mathbb{E}_{\theta'}\left[\mathbf{1}\left\{\text{sign}(a_j)=\text{sign}(\theta_j)\right\}N_0^a\right]\end{aligned}$$

because $\mathbf{1}\left\{\text{sign}(a_j)=\text{sign}(\theta_j)\right\}+\mathbf{1}\left\{\text{sign}(a_j)=\text{sign}(\theta_j')\right\}=1$. Summing up this equality for $\theta\in\Theta$ and $a\in\mathcal{A}$, we obtain

$$\begin{aligned}&2\sum_{\theta\in\Theta}\sum_{a\in\mathcal{A}}\mathbb{E}_\theta\left[\mathbf{1}\left\{\text{sign}(a_j)=\text{sign}(\theta_j)\right\}N_0^a\right]\\&=\sum_{\theta\in\Theta}\mathbb{E}_{\theta'}N_0+\sum_{\theta\in\Theta}\mathbb{E}_\theta\left[\sum_{a\in\mathcal{A}}\mathbf{1}\left\{\text{sign}(a_j)=\text{sign}(\theta_j)\right\}N_0^a\right]\\&\quad-\sum_{\theta\in\Theta}\mathbb{E}_{\theta'}\left[\sum_{a\in\mathcal{A}}\mathbf{1}\left\{\text{sign}(a_j)=\text{sign}(\theta_j)\right\}N_0^a\right]\\&=\sum_{\theta\in\Theta}\mathbb{E}_{\theta'}N_0+\sum_{\theta\in\Theta}\left(\mathbb{E}_\theta\left[N_0^{\mathcal{A}_j}\right]-\mathbb{E}_{\theta'}\left[N_0^{\mathcal{A}_j}\right]\right)\end{aligned}$$

where $\mathcal{A}_j$ is the set of all actions a which satisfy $\mathbf{1}\{\text{sign}(a_j)=\text{sign}(\theta_j)\}$. Here, to provide an upper bound on the term $\mathbb{E}_\theta[N_0^{\mathcal{A}_j}]-\mathbb{E}_{\theta'}[N_0^{\mathcal{A}_j}]$, we apply the following version of Pinsker's inequality due to Jaksch et al. (2010).

Lemma 31. (Jaksch et al., 2010, Equation (49)). *Let $s = \{s_1, \ldots, s_T\} \in \mathcal{S}^T$ denote the sequence of the observed states from time step 1 to T. Then for any two distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ over $\mathcal{S}^T$ and any bounded function $f : \mathcal{S}^T \to [0, B]$, we have*

$$\mathbb{E}_{\mathcal{P}_1} f(s) - \mathbb{E}_{\mathcal{P}_2} f(s) \leq \sqrt{\log 2/2} B \sqrt{\mathrm{KL}(\mathcal{P}_2 || \mathcal{P}_1)}$$

where $\mathrm{KL}(\mathcal{P}_2||\mathcal{P}_1)$ is the Kullback–Leibler divergence of $\mathcal{P}_2$ from $\mathcal{P}_1$.

By Lemma 31, it holds that

$$2 \sum_{\theta\in\Theta} \sum_{a\in\mathcal{A}} \mathbb{E}_\theta \left[\mathbf{1}\left\{\mathrm{sign}(a_j) = \mathrm{sign}(\theta_j)\right\} N_0^a\right] \leq \sum_{\theta\in\Theta} \mathbb{E}_{\theta'} N_0 + \sum_{\theta\in\Theta} \sqrt{\log 2/2} T \sqrt{\mathrm{KL}(\mathcal{P}_{\theta'} \parallel \mathcal{P}_\theta)}.$$

Here, we need to provide an upper bound on the KL divergence term $\mathrm{KL}(\mathcal{P}_{\theta'} \parallel \mathcal{P}_\theta)$. For this, we prove the following lemma which is analogous to (Wu et al., 2022, Lemma C.4).

Lemma 32. *Suppose that θ and θ' only differ in the jth coordinate and $100\Delta \leq \delta \leq 1/101$. Then we have the following bound for the KL divergence of $\mathcal{P}_{\theta'}$ from $\mathcal{P}_\theta$.*

$$\mathrm{KL}(\mathcal{P}_{\theta'} \parallel \mathcal{P}_\theta) \leq \left(\frac{101}{99}\right)^2 \frac{16\Delta^2}{(d-1)^2\delta} \mathbb{E}_{\theta'} N_0$$

Proof. See Lemma H.6. □

By Lemma 32, we deduce that

$$\begin{aligned}
&2 \sum_{\theta\in\Theta} \sum_{a\in\mathcal{A}} \mathbb{E}_\theta \left[\mathbf{1}\left\{\mathrm{sign}(a_j) = \mathrm{sign}(\theta_j)\right\} N_0^a\right] \\
&\leq \sum_{\theta\in\Theta} \mathbb{E}_{\theta'} N_0 + \sum_{\theta\in\Theta} \frac{202}{99} \sqrt{2\log 2} \frac{T\Delta}{(d-1)\sqrt{\delta}} \sqrt{\mathbb{E}_{\theta'} N_0} \\
&\leq \sum_{\theta\in\Theta} \mathbb{E}_{\theta} N_0 + \sum_{\theta\in\Theta} \frac{202}{99} \sqrt{2\log 2} \frac{T\Delta}{(d-1)\sqrt{\delta}} \sqrt{\mathbb{E}_{\theta} N_0}.
\end{aligned} \tag{50}$$

Combining (49) and (50), we deduce that

$$\begin{aligned}
\frac{1}{|\Theta|} \sum_{\theta\in\Theta} \mathbb{E}_\theta N_1 &\leq \frac{T}{2} + \frac{101\Delta}{396\delta|\Theta|} \sum_{\theta\in\Theta} \left(\mathbb{E}_\theta N_0 + \frac{202}{99} \sqrt{2\log 2} \frac{T\Delta}{(d-1)\sqrt{\delta}} \sqrt{\mathbb{E}_\theta N_0}\right) \\
&\leq \frac{T}{2} + \frac{\Delta}{4\delta|\Theta|} \sum_{\theta\in\Theta} \left(\frac{4}{5} T + 2\sqrt{2\log 2} \frac{T\Delta}{(d-1)\sqrt{\delta}} \frac{2\sqrt{T}}{\sqrt{5}}\right) \\
&\leq \frac{T}{2} + \frac{\Delta T}{5\delta} + \sqrt{\frac{2}{5}\log 2} \frac{\Delta^2 T^{3/2}}{(d-1)\delta^{3/2}}
\end{aligned} \tag{51}$$

where the second inequality follows from Lemma 30.

H.3 Proof of Theorem 4

Let us first argue that the diameter of M_θ is D.

Lemma 33. *The diameter of M_θ is $1/\delta$.*

Proof. Note that the expected travel time from state x_1 to state x_0 is $1/(\delta + \Delta)$ which is less than $1/\delta$, while the expected travel time from state x_0 to state x_1 is $1/\delta$. Hence, the diameter of our hard-to-learn MDP M_θ is $1/\delta$. □

As we set $\delta = 1/D$ for the average-reward setting, the diameter of M_θ equals D by Lemma 33.

Recall that under the optimal policy, we have $p^*(x_1 \mid x_0, a) = \delta + \Delta$. This means that under the optimal policy, the stationary distribution over states x_0 and x_1 is given by

$$\mu = \left(\frac{\delta}{2\delta + \Delta}, \frac{\delta + \Delta}{2\delta + \Delta}\right).$$

As $r(x_0, a) = 0$ and $r(x_1, a) = 1$ for any $a \in \mathcal{A}$, it follows that the optimal average reward equals

$$J^*(M_\theta) = \frac{\delta + \Delta}{2\delta + \Delta}. \tag{52}$$

For simplicity, we refer to the regret of policy π as $\mathrm{Regret}_\theta(T)$. Then we have

$$\mathbb{E}_\theta\left[\mathrm{Regret}_\theta(T)\right] = TJ^*(M_\theta) - \mathbb{E}_\theta\left[\sum_{t=1}^T r(s_t, a_t)\right] = TJ^*(M_\theta) - \mathbb{E}_\theta N_1.$$

Taking $\Theta = \{-\bar{\Delta}/(d-1), \bar{\Delta}/(d-1)\}^{d-1}$, we deduce that

$$\frac{1}{|\Theta|}\sum_{\theta\in\Theta}\mathbb{E}_\theta\left[\mathrm{Regret}_\theta(T)\right] = TJ^*(M_\theta) - \frac{1}{|\Theta|}\sum_{\theta\in\Theta}\mathbb{E}_\theta N_1. \tag{53}$$

Then it follows from (51) that

$$\begin{aligned}
\frac{1}{|\Theta|}\sum_{\theta\in\Theta}\mathbb{E}_\theta\left[\mathrm{Regret}_\theta(T)\right] &\geq \frac{(\delta+\Delta)T}{2\delta+\Delta} - \frac{T}{2} - \frac{\Delta T}{5\delta} - \sqrt{\frac{2}{5}\log 2}\frac{\Delta^2 T^{3/2}}{(d-1)\delta^{3/2}} \\
&= \frac{\Delta(\delta - 2\Delta)T}{10\delta(2\delta+\Delta)} - \sqrt{\frac{2}{5}\log 2}\frac{\Delta^2 T^{3/2}}{(d-1)\delta^{3/2}} \\
&\geq \frac{2\Delta}{45\delta}T - \sqrt{\frac{2}{5}\log 2}\frac{\Delta^2 T^{3/2}}{(d-1)\delta^{3/2}}
\end{aligned}$$

where the second inequality holds because $0 < 4\Delta \leq \delta$. Setting Δ as

$$\Delta = \frac{1}{45\sqrt{(2/5)\log 2}} \cdot \frac{(d-1)}{\sqrt{DT}},$$

the rightmost side equals

$$\frac{1}{2025\sqrt{(2/5)\log 2}}(d-1)\sqrt{DT}.$$

When $d \geq 2$, we have $2(d-1) \geq d$. Moreover, $\sqrt{(2/5)\log 2} \leq 1/2$. So, we get that

$$\frac{1}{|\Theta|}\sum_{\theta\in\Theta}\mathbb{E}_\theta\left[\mathrm{Regret}_\theta(T)\right] \geq \frac{1}{2025}d\sqrt{DT},$$

as required.

H.4 Proof of Theorem 5

As in the discounted-reward setting, we refer to the regret of policy π as $\mathrm{Regret}_\theta(T)$. Let us show the following lemma that is useful to provide a lower bound on the regret.

Lemma 34. *We have*

$$\mathbb{E}_\theta[\mathrm{Regret}_\theta(T)] \geq \mathbb{E}_\theta\left[\sum_{t=1}^T V^*(s_t) - \frac{1}{1-\gamma}\sum_{t=1}^T r(s_t, a_t) - \frac{\gamma}{(1-\gamma)^2}\right].$$

Proof. By the definition of V_t^π and the regret in Section 2.2, we deduce that

$$\begin{aligned}\mathbb{E}_\theta[\text{Regret}_\theta(T)] &= \mathbb{E}_\theta\left[\sum_{t=1}^T V^*(s_t) - \sum_{t=1}^T\sum_{t'=0}^\infty \gamma^{t'} r(s_{t+t'}, a_{t+t'})\right] \\ &= \mathbb{E}_\theta\left[\sum_{t=1}^T V^*(s_t) - \underbrace{\sum_{t=1}^T r(s_t,a_t)\sum_{t'=0}^{t-1}\gamma^{t'}}_{I_1} - \underbrace{\sum_{t=T+1}^\infty r(s_t,a_t)\sum_{t'=t-T}^{t-1}\gamma^{t'}}_{I_2}\right].\end{aligned}$$

Note that

$$I_1 = \sum_{t=1}^T r(s_t,a_t)\sum_{t'=0}^{t-1}\gamma^{t'} \le \sum_{t=1}^T r(s_t,a_t)\sum_{t'=0}^{\infty}\gamma^{t'} = \sum_{t=1}^T \frac{r(s_t,a_t)}{1-\gamma}.$$

Moreover,

$$I_2 = \sum_{t=T+1}^\infty r(s_t,a_t)\sum_{t'=t-T}^{t-1}\gamma^{t'} \le \sum_{t=T+1}^\infty 1 \sum_{t'=t-T}^{\infty}\gamma^{t'} = \sum_{t=T+1}^\infty 1\cdot\frac{\gamma^{t-T}}{1-\gamma} = \frac{\gamma}{(1-\gamma)^2},$$

where the first equality holds by $r(s_t,a_t)\le 1$. These bounds on I_1 and I_2 lead to the desired lower bound on the expected regret. □

Recall that in state x_0, the optimal policy always takes action a_θ such that $a_\theta^\top\theta = \bar{\Delta}$. Hence, the transition probability under the optimal policy is given by

$$\begin{aligned}p_\theta(x_0|x_0,a_\theta) &= 1 - f(\bar{\Delta}) = 1-\delta-\Delta,\\ p_\theta(x_1|x_0,a_\theta) &= f(\bar{\Delta}) = \delta+\Delta,\\ p_\theta(x_0|x_1,a_\theta) &= f(0) = \delta,\\ p_\theta(x_1|x_1,a_\theta) &= 1-f(0) = 1-\delta.\end{aligned}$$

Then it follows from the Bellman optimality equation (1) that

$$\begin{aligned}V^*(x_0) &= 0 + \gamma(1-\delta-\Delta)V^*(x_0) + \gamma(\delta+\Delta)V^*(x_1),\\ V^*(x_1) &= 1 + \gamma\delta V^*(x_0) + \gamma(1-\delta)V^*(x_1).\end{aligned}$$

Therefore, the optimal value functions are given by

$$V^*(x_0) = \frac{\gamma(\Delta+\delta)}{(1-\gamma)(\gamma(2\delta+\Delta-1)+1)},\quad V^*(x_1) = \frac{\gamma(\Delta+\delta)+1-\gamma}{(1-\gamma)(\gamma(2\delta+\Delta-1)+1)}. \tag{54}$$

Note that

$$\begin{aligned}&\frac{1}{|\Theta|}\sum_{\theta\in\Theta}\left[\mathbb{E}_\theta[\text{Regret}_\theta(T)] + \frac{\gamma}{(1-\gamma)^2}\right]\\ &\ge \frac{1}{|\Theta|}\sum_{\theta\in\Theta}\mathbb{E}_\theta\left[N_0V^*(x_0) + N_1V^*(x_1) - \frac{1}{1-\gamma}N_1\right]\\ &= \frac{1}{(1-\gamma)|\Theta|}\sum_{\theta\in\Theta}\mathbb{E}_\theta\left[N_0\frac{\gamma(\Delta+\delta)}{\gamma(2\delta+\Delta-1)+1} + N_1\frac{-\gamma\delta}{\gamma(2\delta+\Delta-1)+1}\right] \qquad (55)\\ &= \frac{1}{(1-\gamma)|\Theta|}\sum_{\theta\in\Theta}\mathbb{E}_\theta\left[T\frac{\gamma(\Delta+\delta)}{\gamma(2\delta+\Delta-1)+1} - N_1\frac{\gamma(\Delta+2\delta)}{\gamma(2\delta+\Delta-1)+1}\right]\\ &= \frac{\gamma}{(1-\gamma)(\gamma(2\delta+\Delta-1)+1)}\left((\Delta+\delta)T - \frac{\Delta+2\delta}{|\Theta|}\sum_{\theta\in\Theta}\mathbb{E}_\theta[N_1]\right)\end{aligned}$$

where the the inequality is implied by $r(x_0, a) = 0$ and $r(x_1, a) = 1$ for any a and Lemma 34, the first equality is due to (54), and the second equality holds because $T = N_0 + N_1$. Moreover,

$$\begin{aligned}
&(\Delta+\delta)T - \frac{\Delta+2\delta}{|\Theta|}\sum_{\theta\in\Theta}\mathbb{E}_\theta[N_1] \\
&\geq (\Delta+\delta)T - \frac{(\Delta+2\delta)T}{2} - (\Delta+2\delta)\left(\frac{\Delta T}{5\delta} + \sqrt{\frac{2}{5}\log 2}\frac{\Delta^2 T^{3/2}}{(d-1)\delta^{3/2}}\right) \\
&= \left(\frac{\Delta}{2} - \frac{(\Delta+2\delta)\Delta}{5\delta}\right)T - (\Delta+2\delta)\sqrt{\frac{2}{5}\log 2}\frac{\Delta^2 T^{3/2}}{(d-1)\delta^{3/2}}
\end{aligned} \tag{56}$$

Here, by Lemma 27, we know that $100\Delta \leq \delta$, which implies that

$$\Delta + 2\delta \leq \frac{201}{100}\delta, \quad \frac{\Delta}{2} - \frac{(\Delta+2\delta)\Delta}{5\delta} \geq \frac{49}{500}\Delta.$$

Then it follows that

$$\begin{aligned}
&\left(\frac{\Delta}{2} - \frac{(\Delta+2\delta)\Delta}{5\delta}\right)T - (\Delta+2\delta)\sqrt{\frac{2}{5}\log 2}\frac{\Delta^2 T^{3/2}}{(d-1)\delta^{3/2}} \\
&\geq \left(\frac{49}{500} - \frac{201}{100}\sqrt{\frac{2T}{5\delta}\log 2}\frac{\Delta}{(d-1)}\right)\Delta T \\
&\geq \left(\frac{49}{500} - \frac{201}{4500}\right)\Delta T \\
&\geq \frac{240}{4500}\Delta T
\end{aligned} \tag{57}$$

where the second inequality is due to our choice of Δ. Moreover, since $\Delta \leq 100\Delta \leq \delta$, we have

$$\gamma(2\delta + \Delta - 1) + 1 \leq 1 - \gamma + 3\delta\gamma = 1 - \gamma + 3(1-\gamma)\gamma \leq 4(1-\gamma). \tag{58}$$

Combining (55), (56), (57), and (58), we obtain

$$\begin{aligned}
\frac{1}{|\Theta|}\sum_{\theta\in\Theta}\mathbb{E}_\theta[\mathrm{Regret}_\theta(T)] &\geq \frac{24}{1800(1-\gamma)^2}\gamma\Delta T - \frac{\gamma}{(1-\gamma)^2} \\
&= \frac{\gamma(d-1)\sqrt{T}}{3375(1-\gamma)^{3/2}\sqrt{(2/5)\log 2}} - \frac{\gamma}{(1-\gamma)^2} \\
&\geq \frac{\gamma}{3375(1-\gamma)^{3/2}}d\sqrt{T} - \frac{\gamma}{(1-\gamma)^2}
\end{aligned}$$

where the last inequality holds because $2(d-1) \geq 1$ and $\sqrt{(2/5)\log 2} \leq 1/2$.

H.5 Proof of Lemma 30

We have that

$$\begin{aligned}
\mathbb{E}_\theta N_1 &= \sum_{t=2}^{T}\mathcal{P}_\theta(s_t = x_1) \\
&= \underbrace{\sum_{t=2}^{T}\mathcal{P}_\theta(s_t = x_1 \mid s_{t-1} = x_1)\mathcal{P}_\theta(s_{t-1} = x_1)}_{I_1} + \underbrace{\sum_{t=2}^{T}\mathcal{P}_\theta(s_t = x_1, s_{t-1} = x_0)}_{I_2}.
\end{aligned} \tag{59}$$

For I_1, note that $\mathcal{P}_\theta(s_t = x_1 \mid s_{t-1} = 1 - \delta$ regardless of action a_{t-1}, so we have

$$I_1 = (1-\delta)\sum_{t=2}^{T}\mathcal{P}_\theta(s_{t-1} = x_1) = (1-\delta)\mathbb{E}_\theta N_1 - (1-\delta)\mathcal{P}_\theta(s_T = x_1). \tag{60}$$

For I_2, note that

$$\begin{aligned}
I_2 &= \sum_{t=2}^{T}\sum_{a\in\mathcal{A}} \mathcal{P}_\theta(s_t = x_1 \mid s_{t-1}=x_0, a_{t-1}=a)\mathcal{P}_\theta(s_{t-1}=x_0, a_{t-1}=a)\\
&= \sum_{t=2}^{T}\sum_{a\in\mathcal{A}} f(a^\top\theta)\mathcal{P}_\theta(s_{t-1}=x_0, a_{t-1}=a) \qquad (61)\\
&= \sum_{a\in\mathcal{A}} f(a^\top\theta)\left(\mathbb{E}N_0^a - \mathcal{P}_\theta(s_T=x_0, a_T=a)\right).
\end{aligned}$$

Plugging (60) and (61) to (59), we deduce that

$$\begin{aligned}
\mathbb{E}_\theta N_1 &= \sum_{a\in\mathcal{A}} \frac{f(a^\top\theta)}{\delta}\mathbb{E}_\theta N_0^a - \underbrace{\left(\frac{1-\delta}{\delta}\mathcal{P}_\theta(x_T=x_1) + \sum_{a\in\mathcal{A}}\frac{f(a^\top\theta)}{\delta}\mathcal{P}_\theta(s_T=x_0,a_T=a)\right)}_{\psi_\theta} \qquad (62)\\
&= \mathbb{E}_\theta N_0 + \frac{1}{\delta}\sum_{a\in\mathcal{A}}(f(a^\top\theta)-\delta)\mathbb{E}_\theta N_0^a - \psi_\theta.
\end{aligned}$$

Since $T = \mathbb{E}_\theta N_0 + \mathbb{E}_\theta N_1$, it follows that

$$\mathbb{E}_\theta N_1 \le \frac{T}{2} + \frac{1}{2\delta}\sum_{a\in\mathcal{A}}(f(a^\top\theta)-\delta)\mathbb{E}_\theta N_0^a. \qquad (63)$$

Note that

$$f(a^\top\theta) - \delta = f(a^\top\theta) - f(0) \le (\delta+\Delta)a^\top\theta$$

where the first inequality is from Lemma 29.

Next, for $\mathbb{E}_\theta N_0$, since $f(-\bar{\Delta}) \le f(a^\top\theta) \le f(\bar{\Delta}) = \delta+\Delta$, we have from (62) that

$$\begin{aligned}
\mathbb{E}_\theta N_1 &\ge \left(1 + \frac{f(-\bar{\Delta})-f(0)}{\delta}\right)\mathbb{E}_\theta N_0 - \frac{1-\delta}{\delta}\mathcal{P}_\theta(x_T=x_1) - \frac{\delta+\Delta}{\delta}\mathcal{P}_\theta(s_T=x_0)\\
&\ge \left(1 + \frac{f(-\bar{\Delta})-f(0)}{\delta}\right)\mathbb{E}_\theta N_0 - \frac{1-\delta}{\delta} + \frac{1-2\delta-\Delta}{\delta}\mathcal{P}_\theta(s_T=x_0)\\
&\ge \left(1 + \frac{f(-\bar{\Delta})-f(0)}{\delta}\right)\mathbb{E}_\theta N_0 - \frac{1-\delta}{\delta}
\end{aligned}$$

where the second inequality holds because $2\delta+\Delta\le 1$. This implies that

$$\mathbb{E}_\theta N_0 \le \frac{T + \frac{1-\delta}{\delta}}{2 - \frac{1}{\delta}\left(\delta - f(-\bar{\Delta})\right)}.$$

The following lemma provides a lower bound on $f(-\bar{\Delta})$.

Lemma 35. *$f(-\bar{\Delta}) \ge \delta/2$ if and only if $\Delta \le \delta(1-\delta)$.*

Proof. $f(-\bar{\Delta}) \ge \delta/2$ if and only if $1 + \frac{1-\delta}{\delta}\exp(\bar{\Delta}) \le 2/\delta$, which is equivalent to $\exp(-\bar{\Delta}) \ge (1-\delta)/(2-\delta)$. By plugging in the definition of $\bar{\Delta}$ to the inequality, we get that $f(-\bar{\Delta}) \ge \delta/2$ if and only if $\delta(1-\delta-\Delta)/((1-\delta)(\delta+\Delta)) \ge (1-\delta)/(2-\delta)$, which is equivalent to $\Delta \le \delta(1-\delta)$. □

By simple algebra, we may derive from $f(-\bar{\Delta}) \ge \delta/2$ that $2(\delta - f(-\bar{\Delta})) \le \delta$ holds. Since we assumed that $\Delta \le \delta(1-\delta)$, it follows that

$$\mathbb{E}_\theta N_0 \le \frac{T + \frac{1-\delta}{\delta}}{3/2} = \left(\frac{99}{101}\right)^4 \cdot \frac{4}{5}T$$

where the inequality holds because

$$\frac{1-\delta}{\delta} \le \frac{1}{\delta} \le \left(\frac{3}{2}\cdot\frac{4}{5}\cdot\left(\frac{99}{101}\right)^4 - 1\right)T,$$

as required.

H.6 Proof of Lemma 32

First of all, we consider the following lemma.

Lemma 36. (Jaksch et al., 2010, Lemma 20). *Suppose* $0 \leq \delta' \leq 1/2$ *and* $\epsilon' \leq 1 - 2\delta'$, *then*

$$\delta' \log \frac{\delta'}{\delta' + \epsilon'} + (1 - \delta') \log \frac{(1 - \delta')}{1 - \delta' - \epsilon'} \leq \frac{2(\epsilon')^2}{\delta'}.$$

Let $\boldsymbol{s}_t$ denote the sequence of states $\{s_1, \ldots, s_t\}$ from time step 1 to T. By the Markovian property of MDP, we may decompose the KL divergence term of $\mathcal{P}_{\theta'}$ from $\mathcal{P}_\theta$ as follows.

$$\mathrm{KL}\left(\mathcal{P}_{\theta'} \parallel \mathcal{P}_\theta\right) = \sum_{t=1}^{T-1} \mathrm{KL}\left(\mathcal{P}_{\theta'}\left(s_{t+1} \mid \boldsymbol{s}_t\right) \parallel \mathcal{P}_\theta\left(s_{t+1} \mid \boldsymbol{s}_t\right)\right)$$

where the KL divergence of $\mathcal{P}_{\theta'}\left(s_{t+1} \mid \boldsymbol{s}_t\right)$ from $\mathcal{P}_\theta\left(s_{t+1} \mid \boldsymbol{s}_t\right)$ is given by

$$\mathrm{KL}\left(\mathcal{P}_{\theta'}\left(s_{t+1} \mid \boldsymbol{s}_t\right) \parallel \mathcal{P}_\theta\left(s_{t+1} \mid \boldsymbol{s}_t\right)\right) = \sum_{\boldsymbol{s}_{t+1} \in \mathcal{S}^{t+1}} \mathcal{P}_{\theta'}\left(\boldsymbol{s}_{t+1}\right) \log \frac{\mathcal{P}_{\theta'}\left(s_{t+1} \mid \boldsymbol{s}_t\right)}{\mathcal{P}_\theta\left(s_{t+1} \mid \boldsymbol{s}_t\right)}.$$

The right-hand side can be further decomposed as follows.

$$\begin{aligned}
&\sum_{\boldsymbol{s}_{t+1} \in \mathcal{S}^{t+1}} \mathcal{P}_{\theta'}\left(\boldsymbol{s}_{t+1}\right) \log \frac{\mathcal{P}_{\theta'}\left(s_{t+1} \mid \boldsymbol{s}_t\right)}{\mathcal{P}_\theta\left(s_{t+1} \mid \boldsymbol{s}_t\right)} \\
&= \sum_{\boldsymbol{s}_t \in \mathcal{S}^t} \mathcal{P}_{\theta'}\left(\boldsymbol{s}_t\right) \sum_{x \in \mathcal{S}} \mathcal{P}_{\theta'}\left(s_{t+1} = x \mid \boldsymbol{s}_t\right) \log \frac{\mathcal{P}_{\theta'}\left(s_{t+1} = x \mid \boldsymbol{s}_t\right)}{\mathcal{P}_\theta\left(s_{t+1} = x \mid \boldsymbol{s}_t\right)} \\
&= \sum_{\boldsymbol{s}_{t-1} \in \mathcal{S}^{t-1}} \mathcal{P}_{\theta'}\left(\boldsymbol{s}_{t-1}\right) \sum_{x' \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mathcal{P}_{\theta'}\left(s_t = x', a_t = a \mid \boldsymbol{s}_{t-1}\right) \\
&\quad \times \sum_{x \in \mathcal{S}} \mathcal{P}_{\theta'}\left(s_{t+1} = x \mid \boldsymbol{s}_{t-1}, s_t = x', a_t = a\right) \underbrace{\log \frac{\mathcal{P}_{\theta'}\left(s_{t+1} = x \mid \boldsymbol{s}_{t-1}, s_t = x', a_t = a\right)}{\mathcal{P}_\theta\left(s_{t+1} = x \mid \boldsymbol{s}_{t-1}, s_t = x', a_t = a\right)}}_{I_1}.
\end{aligned}$$

Note that at state x_1, the transition probability does not depend on the action taken and the underlying transition core. This implies that $\mathcal{P}_{\theta'}\left(s_{t+1} = x \mid \boldsymbol{s}_{t-1}, s_t = x', a_t = a\right) = \mathcal{P}_\theta\left(s_{t+1} = x \mid \boldsymbol{s}_{t-1}, s_t = x', a_t = a\right)$ for all θ, θ'. This means that if $x' = x_1$, we have $I_1 = 0$. Then it holds that

$$\begin{aligned}
&\sum_{\boldsymbol{s}_{t+1} \in \mathcal{S}^{t+1}} \mathcal{P}_{\theta'}\left(\boldsymbol{s}_{t+1}\right) \log \frac{\mathcal{P}_{\theta'}\left(s_{t+1} \mid \boldsymbol{s}_t\right)}{\mathcal{P}_\theta\left(s_{t+1} \mid \boldsymbol{s}_t\right)} \\
&= \sum_{\boldsymbol{s}_{t-1} \in \mathcal{S}^{t-1}} \mathcal{P}_{\theta'}\left(\boldsymbol{s}_{t+1}\right) \sum_{a} \mathcal{P}_{\theta'}\left(s_t = x_0, a_t = a \mid \boldsymbol{s}_{t-1}\right) \\
&\quad \times \sum_{x \in \mathcal{S}} \mathcal{P}_{\theta'}\left(s_{t+1} = x \mid \boldsymbol{s}_{t-1}, s_t = x_0, a_t = a\right) \log \frac{\mathcal{P}_{\theta'}\left(s_{t+1} = s \mid \boldsymbol{s}_{t-1}, s_t = x_0, a_t = a\right)}{\mathcal{P}_\theta\left(s_{t+1} = s \mid \boldsymbol{s}_{t-1}, s_t = x_0, a_t = a\right)} \\
&= \sum_{a} \mathcal{P}_{\theta'}\left(s_t = x_{0,1}, a_t = a\right) \\
&\quad \times \underbrace{\sum_{x \in \mathcal{S}} \mathcal{P}_{\theta'}\left(s_{t+1} = s \mid s_t = x_0, a_t = a\right) \log \frac{\mathcal{P}_{\theta'}\left(s_{t+1} = x \mid s_t = x_0, a_t = a\right)}{\mathcal{P}_\theta\left(s_{t+1} = x \mid s_t = x_0, a_t = a\right)}}_{I_2}.
\end{aligned}$$

To bound I_2, we know that s_{t+1} follows the Bernoulli distribution over x_0 and x_1 with probability $1 - f(a^\top \theta')$ and $f(a^\top \theta')$. Then, we have

$$I_2 = \left(1 - f(a^\top \theta')\right) \log \frac{1 - f(a^\top \theta')}{1 - f(a^\top \theta)} + f(a^\top \theta') \log \frac{f(a^\top \theta')}{f(a^\top \theta)}.$$

Note that

$$\frac{1}{100} \geq \frac{101}{100}\delta \geq \delta + \Delta = f(\bar{\Delta}) \geq f(a^\top \theta') \geq f(-\bar{\Delta}) \geq \frac{\delta}{2}$$

where the first inequality is due to $\delta \leq 1/101$, the second holds because $100\Delta \leq \delta$, and the last inequality is by Lemma 35. Moreover, since $f(\bar{\Delta}) \leq 1/100$,

$$f(a^\top \theta) - f(a^\top \theta') \leq f(\bar{\Delta}) \leq \frac{1}{100} \leq 1 - f(\bar{\Delta}) \leq 1 - f(a^\top \theta').$$

Then we deduce that

$$I_2 \leq \frac{2\left(f(a^\top \theta') - f(a^\top \theta)\right)^2}{f(a^\top \theta')} \leq \frac{16(\delta + \Delta)^2 \bar{\Delta}^2}{\delta(d-1)^2} \leq \left(\frac{101}{99}\right)^2 \frac{16\Delta^2}{\delta(d-1)^2}$$

where the first inequality is implied by Lemma 36 with $\delta' = f(a^\top \theta')$ and $\epsilon' = f(a^\top \theta) - f(a^\top \theta')$, the second inequality holds because of $f(a^\top \theta') \geq \delta/2$ and Lemma 29. Then

$$\begin{aligned}
\mathrm{KL}\left(\mathcal{P}_{\theta'} \parallel \mathcal{P}_\theta\right) &= \sum_{t=1}^{T-1} \sum_{\boldsymbol{s}_{t+1} \in \mathcal{S}^{t+1}} \mathcal{P}_{\theta'}(\boldsymbol{s}_{t+1}) \log \frac{\mathcal{P}_{\theta'}(s_{t+1} \mid \boldsymbol{s}_t)}{\mathcal{P}_\theta(s_{t+1} \mid \boldsymbol{s}_t)} \\
&\leq \left(\frac{101}{99}\right)^2 \frac{16\Delta^2}{(d-1)^2\delta} \sum_{t=1}^{T-1} \sum_a \mathcal{P}_{\theta'}(s_t = x_0 \mid a_t = a) \\
&= \left(\frac{101}{99}\right)^2 \frac{16\Delta^2}{(d-1)^2\delta} \sum_{t=1}^{T-1} \mathcal{P}_{\theta'}(s_t = x_0) \\
&= \left(\frac{101}{99}\right)^2 \frac{16\Delta^2}{(d-1)^2\delta} \mathbb{E}_{\theta'} N_0,
\end{aligned}$$

as required.