Near-optimal Policy Optimization Algorithms for Learning Adversarial Linear Mixture MDPs

Jiafan He [1] **Dongruo Zhou** [1] **Quanquan Gu** [1]

Abstract

Learning Markov decision processes (MDPs) in the presence of the adversary is a challenging problem in reinforcement learning (RL). In this paper, we study RL in episodic MDPs with adversarial reward and full information feedback, where the unknown transition probability function is a linear function of a given feature mapping, and the reward function can change arbitrarily episode by episode. We propose an optimistic policy optimization algorithm POWER and show that it can achieve $\widetilde{O}(dH\sqrt{T})$ regret, where H is the length of the episode, T is the number of interaction with the MDP, and d is the dimension of the feature mapping. Furthermore, we also prove a matching lower bound of $\widetilde{\Omega}(dH\sqrt{T})$ up to logarithmic factors. Our key technical contributions are two-fold: (1) a new value function estimator based on importance weighting; and (2) a tighter confidence set for the transition kernel. They together lead to the nearly minimax optimal regret.

1. Introduction

The goal of reinforcement learning (RL) is to design a policy to maximize the reward through observation from interaction with the unknown environment. In reinforcement learning, the Markov decision process (MDP) (Puterman, 1994) is a typical model to describe the unknown environment and widely used to analyze the sequential dynamic environment. In this work, we consider episodic MDPs with a finite horizon. Traditional MDPs often assume the unknown transition probability function is fixed and the reward function is stochastic, which means the reward of each state-action pair follows an unknown stationary distribution. Yet, in many real world models, the reward function is not fixed and may change over time. In order to capture the changed or even adversarial reward, Even-Dar et al. (2009) first introduced the concept of adversarial MDP model and proposed MDP-Expert (MDP-E) algorithm, which attains $\widetilde{O}(\tau^2\sqrt{T})$ regret with τ being the mixing time of the MDP, for known transition probability function and full information of the reward function. In a concurrent work, Yu et al. (2009) proposed an algorithm in the same setting and obtained $\widetilde{O}(T^{2/3})$ regret. There is a line of follow up work studying RL for adversarial MDPs (Neu et al., 2010; 2012; Zimin & Neu, 2013; Dekel & Hazan, 2013; Rosenberg & Mansour, 2019a; Efroni et al., 2020), which studies various settings depending on whether the transition probability function is known, and whether the feedback is full-information or bandit. Please see the related work section for a more detailed discussion.

However, most existing works on adversarial MDP are in the tabular MDP setting, where both the number of actions and states are finite, and the action-value function is represented by a table. In many real-world RL problems, the state and action spaces are large or even infinite. A widely used method to overcome the curse of large state and action spaces is function approximation, which reparameterizes the tabular action-value function as a function over some feature mapping that maps the state and action to a low-dimensional space. Learning adversarial MDPs with linear function approximation is still understudied, with Cai et al. (2020) being a notable existing work. In particular, Cai et al. (2020) proposed an optimistic variant of proximal policy optimization algorithm for the linear kernel MDP (Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021b) with unknown transition probability and full reward information in the adversarial setting, which achieves $\widetilde{O}(\sqrt{d^2H^3T})$ regret. Here H is the length of the episode, T is the number of interactions with the MDP and d is the dimension of the feature mapping.

In this paper, we seek a computationally efficient and statistically optimal algorithm for learning adversarial MDPs. The focus of this work is the unknown transition and full information setting. We first propose an algorithm called optimistic **P**olicy **O**ptimization **W**ith B**ER**nstein bonus (POWER) for adversarial linear mixture MDP (See Definition for more details) with full information feedback. At a high level, our algorithm POWER is similar to Optimistic-PPO algorithm

[1]Department of Computer Science, University of California, Los Angeles, USA. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

(Cai et al., 2020), which can also be seen as an extension of MDP-Expert (MDP-E) with linear function. More specifically, POWER consists of two main steps in each round: (1) one-step least-square temporal difference (LSTD) learning along with exploration bonus for policy evaluation; and (2) mirror descent on the policy space for policy improvement. Our key algorithmic contributions include a weighted LSTD algorithm which takes into the variance of the Bellman residue into account, and a Bernstein-type bonus for exploration based on the principle of "optimism-in-the-face-of-uncertainty" (Abbasi-Yadkori et al., 2011). We prove that POWER achieves $\widetilde{O}(dH\sqrt{T})$ regret with high probability, where H is the length of the episode, T is the number of interactions with the MDP and d is the dimension of the feature mapping. We also prove an $\widetilde{\Omega}(dH\sqrt{T})$ lower bound for adversarially learning linear kernel MDPs. Our upper bound matches the lower bound up to logarithmic factors, which suggests that our algorithm is nearly minimax optimal. To the best of our knowledge, our algorithm is the first computationally efficient and statistical (nearly) optimal algorithm for adversarial MDPs in the unknown transition and full reward information setting.

Notation. We use lower case letters to denote scalars, and use lower and upper case boldface letters to denote vectors and matrices respectively. For a vector $\mathbf{x} \in \mathbb{R}^d$ and matrix $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$, we denote by $\|\mathbf{x}\|_2$ the Euclidean norm, $\|\mathbf{x}\|_1 = \sum_{i=1}^d |x_i|$, and $\|\mathbf{x}\|_{\mathbf{\Sigma}} = \sqrt{\mathbf{x}^\top \mathbf{\Sigma}\mathbf{x}}$. For two sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ if there exists an absolute constant C such that $a_n \le Cb_n$, and we write $a_n = \Omega(b_n)$ if there exists an absolute constant C such that $a_n \ge Cb_n$. We use $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to further hide the logarithmic factors. For any $a \le b \in \mathbb{R}$, $x \in \mathbb{R}$, let $[x]_{[a,b]}$ denote $a \cdot \mathbb{1}(x \le a) + x \cdot \mathbb{1}(a \le x \le b) + b \cdot \mathbb{1}(b \le x)$, where $\mathbb{1}(\cdot)$ is the indicator function. For a positive integer n, we use $[n] = \{1, 2, .., n\}$ to denote the set of integers from 1 to n.

2. Preliminaries

Time-inhomogeneous, episodic adversarial MDPs. In this paper, we consider a time-inhomogeneous, episodic Markov decision process (MDP), which is denoted by a tuple $M = M(\mathcal{S}, \mathcal{A}, H, \{r_h^k\}_{h\in[H],k\in[K]}, \{\mathbb{P}_h\}_{h=1}^H)$. Here $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, H is the length of the episode, $r_h^k : \mathcal{S}\times\mathcal{A} \to [0,1]$ is the deterministic reward function at stage h of the k-th episode. $\mathbb{P}_h(s'|s,a)$ is the transition probability function which denotes the probability for state s to transfer to state s' given action a at stage h. For simplicity, we assume the reward function r_h^k is adversarially chosen by the environment at the beginning of the k-th episode and *known after the episode* k. A policy $\pi = \{\pi_h\}_{h=1}^H$ is a collection of functions π_h, where each $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ is a function which maps a state s to distributions over action set $\mathcal{A}$ at stage h. For each state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, we denote the action-value function $Q_{k,h}^\pi$ and the value function $V_{k,h}^\pi$ as follows:

$$Q_{k,h}^\pi(s,a) = r_h^k(s,a) + \mathbb{E}\bigg[\sum_{h'=h+1}^H r_{h'}^k(s_{h'},a_{h'})\bigg| s_h = s, a_h = a\bigg],$$
$$V_{k,h}^\pi(s) = \mathbb{E}_{a\sim\pi_h(\cdot|s)}\big[Q_{k,h}^\pi(s,a)\big], V_{k,H+1}^\pi(s) = 0.$$

In the definition of $Q_{k,h}^\pi$, we denote by $\mathbb{E}[\cdot]$ the expectation over the state-action sequences $(s_h, a_h, s_{h+1}, a_{h+1}, .., s_H, a_H)$, where $s_h = s, a_h = a$ and $s_{h'+1} \sim \mathbb{P}_h(\cdot|s_{h'}, a_{h'})$, $a_{h'+1} \sim \pi_{h'+1}(\cdot|s_{h'+1})(h' = h, h+1, .., H-1)$. For simplicity, for any function $V : \mathcal{S} \to \mathbb{R}$, we denote

$$[\mathbb{P}_h V](s,a) = \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)} V(s'),$$
$$[\mathbb{V}_h V](s,a) = [\mathbb{P}_h V^2](s,a) - \big([\mathbb{P}_h V](s,a)\big)^2, \quad (2.1)$$

where V^2 is a shorthand for the function whose value at state s is $\big(V(s)\big)^2$. Using this notation, for policy π, we have the following Bellman equality $Q_{k,h}^\pi(s,a) = r_h^k(s,a) + [\mathbb{P}_h V_{k,h+1}^\pi](s,a)$.

In the *online learning setting*, for each episode $k \ge 1$, at the beginning of the episode k, the agent determines a policy π^k to be followed in this episode and we assume that the initial state s_1^k is fixed across all episodes. At each stage $h \in [H]$, the agent observe the state s_h^k, choose an action following the policy $a_h^k \sim \pi_h^k(\cdot|s_h^k)$ and observe the next state with $s_{h+1}^k \sim \mathbb{P}_h(\cdot|s_h^k, a_h^k)$. For the adversarial environment, we focus on the expected regret, which is the expected loss of the algorithm relative to the best-fixed policy in hindsight (Cesa-Bianchi & Lugosi, 2006):

$$\text{Regret}(M,K) = \sup_\pi \sum_{k=1}^K \big(V_{k,1}^\pi(s_1^k) - V_{k,1}^{\pi^k}(s_1^k)\big).$$

For simplicity, we denote the optimal policy π^* as $\pi^* = \sup_\pi \sum_{k=1}^K V_{k,1}^\pi(s_1^k)$. Therefore, we have the following Bellman optimally equation $Q_{k,h}^*(s,a) = r_h^k(s,a) + [\mathbb{P}_h V_{k,h}^*](s,a)$, where $Q_{k,h}^*(s,a), V_{k,h}^*(s,a)$ are the corresponding optimal action-value function and value function. For any two policies π and π', we define the Kullback–Leibler divergence between them as follows $D_{KL}(\pi\|\pi') = \sum_{a\in\mathcal{A}} \pi(a)\log\big(\pi(a)/\pi'(a)\big)$.

Linear Mixture MDPs. In this work, we focus on a special class of MDPs called *linear mixture MDPs* (Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021b), where the transition probability function is a linear function of a given feature mapping $\boldsymbol{\phi} : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}^d$. The formal definition of a linear kernel MDP is as follows:

Definition 2.1. $M(\mathcal{S}, \mathcal{A}, H, \{r_h^k\}_{h\in[H],k\in[K]}, \{\mathbb{P}_h\}_{h=1}^H)$ is called a inhomogenous, episode B-bounded linear mixture MDP if there exist a *known* feature mapping $\boldsymbol{\phi}(s'|s,a) : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}^d$ and an *unknown* vector $\boldsymbol{\theta}_h \in \mathbb{R}^d$ with $\|\boldsymbol{\theta}\|_2 \le B$, such that

- For any state-action-next-state triplet $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, we have $\mathbb{P}_h(s'|s, a) = \langle \boldsymbol{\phi}(s'|s, a), \boldsymbol{\theta}_h \rangle$;
- For any bounded function $V : \mathcal{S} \to [0, 1]$ and any tuple $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have $\|\boldsymbol{\phi}_V(s, a)\|_2 \leq 1$, where $\boldsymbol{\phi}_V(s, a) = \sum_{s' \in \mathcal{S}} \boldsymbol{\phi}(s'|s, a)V(s') \in \mathbb{R}^d$.

Based on Definition 2.1, we can see that for any linear mixture MDP M and function $V : \mathcal{S} \to \mathbb{R}$, we have the following properties:

$$\begin{aligned}[\mathbb{P}_h V](s, a) &= \textstyle\sum_{s' \in \mathcal{S}} \mathbb{P}_h(s'|s, a)V(s') \\ &= \textstyle\sum_{s' \in \mathcal{S}} \langle \boldsymbol{\phi}(s'|s, a), \boldsymbol{\theta}_h \rangle V(s') \\ &= \langle \boldsymbol{\phi}_V(s, a), \boldsymbol{\theta}_h \rangle, \end{aligned} \tag{2.2}$$

and

$$\begin{aligned}[\mathbb{V}_h V](s, a) &= \textstyle\sum_{s' \in \mathcal{S}} \mathbb{P}_h(s'|s, a)V^2(s') \\ &\quad - \big[\textstyle\sum_{s' \in \mathcal{S}} \mathbb{P}_h(\cdot|s, a)V(s')\big]^2 \\ &= \langle \boldsymbol{\phi}_{V^2}(s, a), \boldsymbol{\theta}_h \rangle - [\langle \boldsymbol{\phi}_V(s, a), \boldsymbol{\theta}_h \rangle]^2. \end{aligned} \tag{2.3}$$

(2.2) and (2.3) suggest that both the conditional expectation and the variance of a function V can be calculated based on certain linear functions of different feature mappings, i.e., $\boldsymbol{\phi}_V$ and $\boldsymbol{\phi}_{V^2}$. Therefore, we can estimate them by estimating the corresponding linear functions.

3. The Proposed Algorithm

In this section, we propose an algorithm POWER to learn the episodic linear mixture MDP (see Definition 2.1) with adversarial rewards, which is illustrated in Algorithm 1. At a high level, POWER is an improved version of the Optimistic-PPO (OPPO) algorithm (Cai et al., 2020) with a refined estimate of the action-value function $Q_{k,h}(s, a)$. The POWER can be divided into two phases: (1) policy improvement phase and (2) policy evaluation phase.

Policy improvement phase (Line 4 to Line 9): In the policy improvement phase, POWER calculates its policy π^k for the current episode, based on its previous policy π^{k-1} using the proximal policy optimization (PPO) method (Schulman et al., 2017). In detail, let s_1^k be the starting state at the k-th episode, then following PPO, we update π^k as a solution to the following optimization problem:

$$\pi^k \leftarrow \operatorname*{argmin}_{\pi} [L_{k-1}(\pi) - \alpha^{-1} \widetilde{D}_{KL}(\pi, \pi^{k-1})], \tag{3.1}$$

where

$$L_{k-1}(\pi) = \mathbb{E}_{\pi^{k-1}} \bigg[\sum_{h=1}^{H} \langle Q_{k-1,h}(s_h, \cdot), \pi_h^k(\cdot|s_h) \rangle \bigg| s_1 = s_1^k \bigg]$$

is proportional to the first-order Taylor approximation of $V_{k-1,h}^{\pi_h^{k-1}}$ at π^{k-1}, and replaces the action-value function $Q_{k-1,h}^{\pi_h^{k-1}}(\cdot, \cdot)$ by the estimated one $Q_{k-1,h}(\cdot, \cdot)$, and

$$\begin{aligned}&\widetilde{D}_{KL}(\pi, \pi^{k-1}) \\ &= \mathbb{E}_{\pi^{k-1}} \bigg[\textstyle\sum_{h=1}^{H} D_{KL}(\pi_h(\cdot|s_h), \pi_h^k(\cdot|s_h)) \bigg| s_1 = s_1^k \bigg]\end{aligned}$$

is the sum of KL-divergences between π_h and π_h^{k-1}, which encourages π^k to stay close to π^{k-1} to ensure the above first-order Taylor approximation is accurate enough. The closed-form solution to (3.1) is in Line 6. Here $\alpha > 0$ is the step size of the exponential update. Note that the update rule in Line 6 is also the same as the MDP-E algorithm (Even-Dar et al., 2009). After obtaining π^k, POWER chooses action a_h^k based on the new policy π_h^k and the current state s_h^k. It then observes the next state s_{h+1}^k and the adversarial reward function $r_h^k(\cdot, \cdot)$.

Policy evaluation phase (Line 10 to Line 15): In the policy evaluation phase, POWER evaluates the policy π^k by constructing the action-value function $Q_{k,h}$ and the value function $V_{k,h}$ for policy π^k based on the observed data, which are optimistic estimates of the action-value function $Q_{k,h}^{\pi^k}$ and the value function $V_{k,h}^{\pi^k}$ respectively.

Specifically, for each episode $k \in [K]$ and each stage $h \in [H]$, POWER maintains an estimator $\widehat{\boldsymbol{\theta}}_{k,h}$ and an uncentered covariance matrix $\widehat{\boldsymbol{\Sigma}}_{k,h}$ based on the observed data before the k-th episode. Then POWER recursively computes the optimistic $Q_{k,h}, V_{k,h}$ as follows:

$$\begin{aligned}Q_{k,h}(s, a) &= \Big[r_h^k(s, a) + \big\langle \widehat{\boldsymbol{\theta}}_{k,h}, \boldsymbol{\phi}_{V_{k,h+1}}(s, a) \big\rangle \\ &\qquad + \widehat{\beta}_k \big\| \widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2} \boldsymbol{\phi}_{V_{k,h+1}}(s, a) \big\|_2 \Big]_{[0, H-h+1]}, \\ V_{k,h}(s) &= \mathbb{E}_{a \sim \pi_h^k(\cdot|s)} [Q_{k,h}(s, a)],\end{aligned}$$

where $\widehat{\beta_k}$ is the radius of the confidence ball defined as:

$$\widehat{\beta_k} = 8\sqrt{d \log(1 + k/\lambda) \log(4k^2 H/\delta)} + 4\sqrt{d \log(4k^2 H/\delta)} + \sqrt{\lambda} B.$$

Now we illustrate how to construct the estimator $\widehat{\boldsymbol{\theta}}_{k,h}$ and the covariance matrix $\widehat{\boldsymbol{\Sigma}}_{k,h}$, which is the key difference compared with OPPO proposed in Cai et al. (2020). Recall (2.2), we know that the expectation of the random variables $V_{k,h}(s_{k,h+1})$ can be written as a linear function with weight vector $\boldsymbol{\theta}_h$. Therefore, a natural way to estimate $\boldsymbol{\theta}_h$ is to consider it as the unknown weight vector of a stochastic linear bandits problem with context $\boldsymbol{\phi}_{V_{k,h}}(s_h^k, a_h^k)$ and target $V_{k,h}(s_{k,h+1})$, and apply algorithms for linear bandits such as OFUL (Abbasi-Yadkori et al., 2011), to obtain the estimator $\widehat{\boldsymbol{\theta}}_{k,h}$. Such an approach is adopted by Cai et al. (2020).

Algorithm 1 POWER

Require: Regularization parameter λ, learning rate α.

Set initial policy $\{\pi_h^0(\cdot|\cdot)\}_{h=1}^H$ as uniform distribution on the action set $\mathcal{A}$
For $h \in [H+1]$, set the initial value functions $Q_{0,h}(\cdot,\cdot) \leftarrow 0, V_{0,h}(\cdot) \leftarrow 0$
for $k = 1, \ldots, K$ **do**
  Receive state s_1^k
  for $h = 1, \ldots, H$ **do**
    Update the policy by $\pi_h^k(\cdot|\cdot) \propto \pi_h^{k-1}(\cdot|\cdot)\exp\{\alpha Q_{k-1,h}(\cdot,\cdot)\}$
    Take action $a_h^k \sim \pi_h^k(\cdot|s_h^k)$ and receive next state $s_{h+1}^k \sim \mathbb{P}_h(\cdot|s_h^k, a_h^k)$
    Observe the adversarial reward function $r_h^k(\cdot,\cdot)$
  end for
  Set $V_{k,H+1}(\cdot) \leftarrow 0$
  for $h = H, \ldots, 1$ **do**
    Set $Q_{k,h}(\cdot,\cdot) \leftarrow \Big[r_h^k(\cdot,\cdot) + \big\langle \widehat{\boldsymbol{\theta}}_{k,h}, \boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot)\big\rangle + \widehat{\beta}_k \big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot)\big\|_2\Big]_{[0,H-h+1]}$
    Set $V_{k,h}(\cdot) \leftarrow \mathbb{E}_{a\sim\pi_h^k(\cdot|\cdot)}[Q_{k,h}(\cdot,a)]$
    Set the matrix $\widehat{\boldsymbol{\Sigma}}_{k+1,h}$ and vector $\widehat{\boldsymbol{\theta}}_{k+1,h}$ as in (B.1).
  end for
end for

However, OFUL uses the vanilla linear regression to construct $\widehat{\boldsymbol{\theta}}_{k,h}$, which is limited to the *homoscedastic* noises case. For the linear mixture MDP, the noises are actually *heteroscedastic* as each target enjoys different noises. Thus the vanilla linear regression is known as statistically inefficient (Kirschner & Krause, 2018). Inspired by Kirschner & Krause (2018); Zhou et al. (2021a), we adopt the *weighted linear regression* to construct $\widehat{\boldsymbol{\theta}}_{k,h}$, which is the solution to the following weighted regression problem:

$$\widehat{\boldsymbol{\theta}}_{k,h} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \lambda\|\boldsymbol{\theta}\|_2^2 + \textstyle\sum_{i=1}^{k-1}\big[\boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i), \boldsymbol{\theta}_h\rangle - V_{i,h+1}(s_{h+1}^i)\big]/\bar{\sigma}_{i,h}^2,$$

where $\bar{\sigma}_{i,h}^2$ is the upper confidence bound of the variance $[\mathbb{V}_h V_{i,h+1}(s_h^i, a_h^i)]$. $\widehat{\boldsymbol{\Sigma}}_{k,h}$ is the weighted "covariance" matrix of $\boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i)$ weighted by $1/\bar{\sigma}_{i,h}^2$. The formal definition of $\bar{\sigma}_{k,h}^2$, $\widehat{\boldsymbol{\Sigma}}_{k,h}$ and $\widehat{\boldsymbol{\theta}}_{k,h}$ can be found in the Section B.

4. Main Results

In this section, we provide the regret bound for our algorithm POWER. Here, $T = KH$ is the number of interactions with the MDP.

Theorem 4.1. For any linear mixture MDP M, if we set the parameter $\lambda = 1/B^2$ in POWER, then with probability at least $1 - 6\delta$, the regret of POWER is upper bounded as follows:

$$\text{Regret}(M, K) \le \widetilde{O}\big(\alpha T H^2 + \alpha^{-1} H \log|\mathcal{A}| + \sqrt{d^2H^2T} + \sqrt{dH^3T} + d^2H^3 + d^{2.5}H^{2.5}\big).$$

Remark 4.2. If we set the learning rate $\alpha = O\big(\sqrt{\log|\mathcal{A}|/(H^2K)}\big)$ in POWER, then Theorem 4.1 suggests that with high probability, the regret of POWER is upper bounded by $\widetilde{O}(\sqrt{d^2H^2T} + \sqrt{dH^3T} + d^2H^3 + d^{2.5}H^{2.5} + \sqrt{H^3\log|\mathcal{A}|T})$. When $T \ge d^3H^3, d \ge H, d \ge \log|\mathcal{A}|$, the regret can be simplified as $\widetilde{O}(dH\sqrt{T})$. Compared with the result of Cai et al. (2020), the POWER improve the upper bound of regret by a factor of $\sqrt{H}$.

The following theorem gives a lower bound of the regret for any algorithms for learning the adversarial linear mixture MDPs.

Theorem 4.3. Suppose $d \ge 4, H \ge 3, K \ge (d-1)^2H/2$, then for any algorithm, there exist a time-inhomogenous, episodic 2-bounded adversarial linear mixture MDP M, such that the expected regret for this MDP is lower bounded by $\Omega(dH\sqrt{T})$.

Remark 4.4. Theorem 4.3 suggests that when the number of episodes K is large enough, for any algorithm, the regret of learning time-inhomogenous episodic adversarial linear mixture MDPs is at least $\Omega(dH\sqrt{T})$. Furthermore, the lower bound of regret in Theorem 4.3 matches the upper bound in Theorem 4.1 up to logarithmic factors, which suggests that POWER is nearly minimax optimal for learning adversarial linear mixture MDPs.

5. Conclusion and Future Work

In this work, we considered learning adversarial Markov decision processes under the linear mixture MDP assumption. We proposed a novel algorithm POWER and proved that with high probability, the regret of POWER is upper bounded by $\widetilde{O}(dH\sqrt{T})$, which matches the lower bound

up to logarithmic factors. Currently, our work requires the full information feedback of the reward and it remains an open problem that if there exists a provably efficient algorithm for learning adversarial linear mixture MDPs with bandit-feedback on the reward. We leave it as future work.

References

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

Abbasi-Yadkori, Y., Bartlett, P., Bhatia, K., Lazic, N., Szepesvari, C., and Weisz, G. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pp. 3692–3702. PMLR, 2019a.

Abbasi-Yadkori, Y., Lazic, N., Szepesvari, C., and Weisz, G. Exploration-enhanced politex. *arXiv preprint arXiv:1908.10479*, 2019b.

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

Bagnell, J. A. and Schneider, J. Covariant policy search. 2003.

Baxter, J. and Bartlett, P. L. Direct gradient-based reinforcement learning. In *2000 IEEE International Symposium on Circuits and Systems (ISCAS)*, volume 3, pp. 271–274. IEEE, 2000.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Dekel, O. and Hazan, E. Better rates for any adversarial deterministic mdp. In *International Conference on Machine Learning*, pp. 675–683. PMLR, 2013.

Du, S., Krishnamurthy, A., Jiang, N., Agarwal, A., Dudik, M., and Langford, J. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pp. 1665–1674. PMLR, 2019.

Efroni, Y., Shani, L., Rosenberg, A., and Mannor, S. Optimistic policy optimization with bandit feedback. *arXiv preprint arXiv:2002.08243*, 2020.

Even-Dar, E., Kakade, S. M., and Mansour, Y. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Gergely Neu, A. G., Szepesvári, C., and Antos, A. Online markov decision processes under bandit feedback. In *Proceedings of the Twenty-Fourth Annual Conference on Neural Information Processing Systems*, 2010.

Hao, B., Lazic, N., Abbasi-Yadkori, Y., Joulani, P., and Szepesvari, C. Provably efficient adaptive approximate policy iteration. *arXiv preprint arXiv:2002.03069*, 2020.

He, J., Zhou, D., and Gu, Q. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*. PMLR, 2021.

Jia, Z., Yang, L., Szepesvari, C., and Wang, M. Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*, pp. 666–686. PMLR, 2020.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1704–1713. JMLR. org, 2017.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pp. 4868–4878, 2018.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, pp. 4860–4869. PMLR, 2020a.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143, 2020b.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

Kakade, S. M. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Kakade, S. M. et al. *On the sample complexity of reinforcement learning*. PhD thesis, 2003.

Kirschner, J. and Krause, A. Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pp. 358–384. PMLR, 2018.

Modi, A., Jiang, N., Tewari, A., and Singh, S. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Neu, G., György, A., and Szepesvári, C. The online loop-free stochastic shortest-path problem. In *COLT*, volume 2010, pp. 231–243. Citeseer, 2010.

Neu, G., Gyorgy, A., and Szepesvári, C. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Artificial Intelligence and Statistics*, pp. 805–813. PMLR, 2012.

Neu, G., Jonsson, A., and Gómez, V. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Puterman, M. L. Markov decision processes: Discrete stochastic dynamic programming, 1994.

Rosenberg, A. and Mansour, Y. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, pp. 5478–5486. PMLR, 2019a.

Rosenberg, A. and Mansour, Y. Online stochastic shortest path with bandit feedback and unknown transition function. *Advances in Neural Information Processing Systems*, 32:2212–2221, 2019b.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory*, pp. 2898–2933. PMLR, 2019.

Sutton, R. S., McAllester, D. A., Singh, S. P., Mansour, Y., et al. Policy gradient methods for reinforcement learning with function approximation. In *NIPs*, volume 99, pp. 1057–1063. Citeseer, 1999.

Wang, Y., Wang, R., Du, S. S., and Krishnamurthy, A. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Wu, Y., Zhou, D., and Gu, Q. Nearly minimax optimal regret for learning infinite-horizon average-reward mdps with linear function approximation. *arXiv preprint arXiv:2102.07301*, 2021.

Yang, L. and Wang, M. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pp. 6995–7004, 2019a.

Yang, L. F. and Wang, M. Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, 2019b.

Yu, J. Y., Mannor, S., and Shimkin, N. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, 34(3):737–757, 2009.

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

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, 2020b.

Zhou, D., Gu, Q., and Szepesvari, C. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *COLT*, 2021a.

Zhou, D., He, J., and Gu, Q. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*. PMLR, 2021b.

Zimin, A. and Neu, G. Online learning in episodic markovian decision processes by relative entropy policy search. In *Neural Information Processing Systems 26*, 2013.

A. Related Work

RL with adversarial reward. There is a long line of research on learning adversarial MDPs, where the reward function is adversarially chosen at the beginning of each episode and can change arbitrarily across different episodes (Even-Dar et al., 2009; Yu et al., 2009; Gergely Neu et al., 2010; Neu et al., 2010; Zimin & Neu, 2013; Neu et al., 2012; Rosenberg & Mansour, 2019a;b; Wang et al., 2019; Cai et al., 2020; Efroni et al., 2020). The seminal works by Even-Dar et al. (2009); Yu et al. (2009) are in the known transition probability and full reward information setting. In the known transition and bandit feedback on the reward setting, Gergely Neu et al. (2010) proposed MDP-EXP3 algorithm and obtained $\widetilde{O}(T^{2/3})$ regret. Neu et al. (2010) proposed Bandit O-SSP algorithm which achieves $\widetilde{O}(\sqrt{T}/\alpha)$ regret with an addition assumption that all states are reachable with probability $\alpha > 0$ for any policy. Zimin & Neu (2013) further proposed O-REPS algorithm, which improves the regret from $\widetilde{O}(T^{2/3})$ to $\widetilde{O}(\sqrt{T})$ without any additional assumption. In the unknown transition but full reward information setting, Neu et al. (2012) proposed FPOP algorithm that achieves $\widetilde{O}(SA\sqrt{T})$ regret. Rosenberg & Mansour (2019a) proposed UC-O-REP algorithm and improved the regret to $\widetilde{O}(S\sqrt{AT})$. In the most challenging unknown transition and bandit reward feedback setting, Rosenberg & Mansour (2019b) proposed Shifted Bandit UC-O-REPS algorithm which achieves $\widetilde{O}(T^{3/4})$ regret. Rosenberg & Mansour (2019b) also proposed Bounded Bandit UC-O-REPS algorithm and obtained $\widetilde{O}(\sqrt{T}/\alpha)$ regret under the assumption that all states are reachable with probability $\alpha > 0$ for any policy. Jin et al. (2020a) proposed UOB-REPS algorithm that achieves $\widetilde{O}(\sqrt{T})$ regret without the additional assumption made by Rosenberg & Mansour (2019b). The focus of this paper is the unknown transition but full reward information setting.

RL with linear function approximation. Recently, there emerges a large body of literature on solving MDP with linear function approximation. These works can be generally divided into three lines based on the specific assumption on the underlying MDP. The first line of work (Sun et al., 2019; Du et al., 2019) is based on the low Bellman rank assumption (Jiang et al., 2017), which assumes a low-rank factorization of the Bellman error matrix. The second line of work (Wang et al., 2019; He et al., 2021; Zanette et al., 2020a) focuses on the linear MDP (Yang & Wang, 2019a; Jin et al., 2020b), where the transition probability function and reward function are parameterized as a linear function of a feature mapping $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. Later, Zanette et al. (2020b) made a weaker assumption called low inherent Bellman error and proposed Eleanor algorithm. The last line of work (Cai et al., 2020; Yang & Wang, 2019b; He et al., 2021; Modi et al., 2019; Zhou et al., 2021a) is based on the linear mixture/kernel MDP (Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021b; Wu et al., 2021), where the transition probability function can be parameterized as a linear function of a feature mapping $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$. Note that none of the above work with linear function approximation can handle adversarially chosen reward with Cai et al. (2020) being a notable exception. Our paper also considers the linear kernel MDP but with an adversarial reward function.

RL with policy gradient. Our work is also related to policy optimization and policy gradient methods (Williams, 1992; Baxter & Bartlett, 2000; Sutton et al., 1999; Kakade, 2001; Kakade & Langford, 2002; Kakade et al., 2003; Bagnell & Schneider, 2003; Schulman et al., 2015; 2017; Abbasi-Yadkori et al., 2019a;b; Cai et al., 2020; Hao et al., 2020; Efroni et al., 2020), among which the most related works to ours are trust-region policy optimization (Schulman et al., 2015), proximal policy optimization (Schulman et al., 2017), Politex (Abbasi-Yadkori et al., 2019a), EE-Politex (Abbasi-Yadkori et al., 2019b), AAPI (Hao et al., 2020) and OPPO (Cai et al., 2020). More specifically, Cai et al. (2020) proposed the optimistic variant of the Proximal Policy Optimization algorithm for adversarial linear kernel MDP, which can be seen as an extension of MDP-E. Abbasi-Yadkori et al. (2019a) proposed Politex algorithm with least-squares policy evaluation for infinite-horizon average-reward MDPs, which can be seen a generalization of the MDP-E (Even-Dar et al., 2009). In fact, MDP-E is equivalent to TRPO/PPO (Schulman et al., 2015; 2017) as shown by Neu et al. (2017). Our algorithm can be seen as a nontrivial extension of OPPO and MDP-E.

B. Complete Algorithm

In this section, we propose the full algorithm and show the formal definition of $\bar{\sigma}_{k,h}^2$, $\widehat{\boldsymbol{\Sigma}}_{k,h}$ and $\widehat{\mathbf{b}}_{k,h}$. For $\widehat{\boldsymbol{\Sigma}}_{k,h}$ and $\widehat{\mathbf{b}}_{k,h}$, we denote them as

$$\widehat{\boldsymbol{\Sigma}}_{k,h} = \lambda \mathbf{I} + \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i) \boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i)^\top$$

$$\widehat{\mathbf{b}}_{k,h} = \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i) V_{i,h+1}(s_{h+1}^i)$$

Algorithm 2 POWER

Require: Regularization parameter λ, learning rate α.

1: Set initial policy $\{\pi_h^0(\cdot|\cdot)\}_{h=1}^H$ as uniform distribution on the action set $\mathcal{A}$
2: For $h \in [H+1]$, set the initial value functions $Q_{0,h}(\cdot,\cdot) \leftarrow 0, V_{0,h}(\cdot) \leftarrow 0$
3: For $h \in [H]$, set $\widehat{\boldsymbol{\Sigma}}_{1,h}, \widetilde{\boldsymbol{\Sigma}}_{1,h} \leftarrow \lambda\mathbf{I}, \widehat{\mathbf{b}}_{1,h}, \widetilde{\mathbf{b}}_{1,h} \leftarrow \mathbf{0}, \widehat{\boldsymbol{\theta}}_{1,h}, \widetilde{\boldsymbol{\theta}}_{1,h} \leftarrow \mathbf{0}$
4: **for** $k = 1, \ldots, K$ **do**
5: Receive state s_1^k
6: **for** $h = 1, \ldots, H$ **do**
7: Update the policy by $\pi_h^k(\cdot|\cdot) \propto \pi_h^{k-1}(\cdot|\cdot)\exp\{\alpha Q_{k-1,h}(\cdot,\cdot)\}$
8: Take action $a_h^k \sim \pi_h^k(\cdot|s_h^k)$ and receive next state $s_{h+1}^k \sim \mathbb{P}_h(\cdot|s_h^k, a_h^k)$
9: Observe the adversarial reward function $r_h^k(\cdot,\cdot)$
10: **end for**
11: Set $V_{k,H+1}(\cdot) \leftarrow 0$
12: **for** $h = H, \ldots, 1$ **do**
13: Set $Q_{k,h}(\cdot,\cdot) \leftarrow \Big[r_h^k(\cdot,\cdot) + \big\langle \widehat{\boldsymbol{\theta}}_{k,h}, \boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot)\big\rangle + \widehat{\beta}_k \big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot)\big\|_2\Big]_{[0,H-h+1]}$
14: Set $V_{k,h}(\cdot) \leftarrow \mathbb{E}_{a\sim\pi_h^k(\cdot|\cdot)}[Q_{k,h}(\cdot,a)]$
15: Set the estimated variance $[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k)$ as in (B.2)
16: Set the bonus term $E_{k,h}$ as in (B.3)
17: $\bar{\sigma}_{k,h} \leftarrow \sqrt{\max\big\{H^2/d, [\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k) + E_{k,h}\big\}}$
18: $\widehat{\boldsymbol{\Sigma}}_{k+1,h} \leftarrow \widehat{\boldsymbol{\Sigma}}_{k,h} + \bar{\sigma}_{k,h}^{-2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)^\top$
19: $\widehat{\mathbf{b}}_{k+1,h} \leftarrow \widehat{\mathbf{b}}_{k,h} + \bar{\sigma}_{k,h}^{-2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)V_{k,h+1}(s_{h+1}^k)$
20: $\widetilde{\boldsymbol{\Sigma}}_{k+1,h} \leftarrow \widetilde{\boldsymbol{\Sigma}}_{k,h} + \boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)^\top$
21: $\widetilde{\mathbf{b}}_{k+1,h} \leftarrow \widetilde{\mathbf{b}}_{k,h} + \boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)V_{k,h+1}^2(s_{h+1}^k)$
22: $\widehat{\boldsymbol{\theta}}_{k+1,h} \leftarrow \widehat{\boldsymbol{\Sigma}}_{k+1,h}^{-1}\widehat{\mathbf{b}}_{k+1,h}, \widetilde{\boldsymbol{\theta}}_{k+1,h} \leftarrow \widetilde{\boldsymbol{\Sigma}}_{k+1,h}^{-1}\widetilde{\mathbf{b}}_{k+1,h}$
23: **end for**
24: **end for**

$$\widehat{\boldsymbol{\theta}}_{k,h} = \widehat{\boldsymbol{\Sigma}}_{k,h}^{-1}\widehat{\mathbf{b}}_{k,h}, \tag{B.1}$$

where the estimated variance $\bar{\sigma}_{k,h}^2$ will be defined later. For estimated variance $\bar{\sigma}_{k,h}^2$, due to (2.3), it suffices to estimate $\langle\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \boldsymbol{\theta}_h\rangle$ and $\langle\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k), \boldsymbol{\theta}_h\rangle$. For the first one, we use $\langle\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widehat{\boldsymbol{\theta}}_{k,h}\rangle$ to estimate it. For the second one, we use $\langle\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widetilde{\boldsymbol{\theta}}_{k,h}\rangle$, where $\widetilde{\boldsymbol{\theta}}_{k,h}$ is the linear regression estimator with contexts $\boldsymbol{\phi}_{V_{i,h+1}^2}(s_h^i, a_h^i)$ and targets $V_{i,h+1}^2(s_h^{i+1})$. Its update rule is shown in Line 22. Then we can specify the choice of $\bar{\sigma}_{k,h}^2$ in the following lemma:

Lemma B.1. Let the estimated variance $[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k)$ be defined as

$$[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k) = \Big[\big\langle\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k), \widetilde{\boldsymbol{\theta}}_{k,h}\big\rangle\Big]_{[0,H^2]} - \Big[\big\langle\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widehat{\boldsymbol{\theta}}_{k,h}\big\rangle_{[0,H]}\Big]^2, \tag{B.2}$$

then with probability at least $1-3\delta$, for all $k \in [K], h \in [H]$, we have

$$\big|[\bar{\mathbb{V}}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k)\big| \le E_{k,h},$$

where $E_{k,h}$ is defined as

$$\begin{aligned}
E_{k,h} &= \min\Big\{\widetilde{\beta}_k\big\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)\big\|_2, H^2\Big\} + \min\Big\{2H\bar{\beta}_k\big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\big\|_2, H^2\Big\}, \\
\widetilde{\beta}_k &= 8H^2\sqrt{d\log\big(1 + kH^4/(d\lambda)\big)\log(4k^2H/\delta)} + 4H^2\log(4k^2H/\delta) + \sqrt{\lambda}B, \\
\bar{\beta}_k &= 8d\sqrt{\log(1+k/\lambda)\log(4k^2H/\delta)} + 4\sqrt{d}\log(4k^2H/\delta) + \sqrt{\lambda}B.
\end{aligned} \tag{B.3}$$

For the estimator $\widehat{\boldsymbol{\theta}}_{k,h}$, we have

$$\boldsymbol{\theta}_h \in \mathcal{C}_{k,h} = \{\boldsymbol{\theta} : \|\widehat{\boldsymbol{\Sigma}}_{k,h}^{1/2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{k,h})\| \leq \widehat{\beta}_k\}. \tag{B.4}$$

By Lemma B.1, we know that in order to guarantee $\bar{\sigma}_{k,h}^2$ is an upper bound of the variance, it suffices to set it as $[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k) + E_{k,h}$. Finally, due to the technical reason, we set $\bar{\sigma}_{k,h}^2$ as

$$\bar{\sigma}_{k,h} = \sqrt{\max\big\{H^2/d, [\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k) + E_{k,h}\big\}}.$$

Furthermore, according to (B.4), we have

$$\begin{aligned}
Q_{k,h}(s,a) &= \Big[r_h^k(s,a) + \max_{\boldsymbol{\theta}\in\mathcal{C}_{k,h}} \big\langle \boldsymbol{\phi}_{V_{k,h+1}}, \boldsymbol{\theta}\big\rangle\Big]_{[0,H-h+1]} \\
&\geq \Big[r_h^k(s,a) + \big\langle \boldsymbol{\phi}_{V_{k,h+1}}, \boldsymbol{\theta}_h\big\rangle\Big]_{[0,H-h+1]} \\
&= \big[r_h^k(s,a) + [\mathbb{P}_h V_{k,h+1}](s,a)\big]_{[0,H-h+1]},
\end{aligned}$$

and it is easy to show that the optimistic action-value function $Q_{k,h}(s,a)$ and the optimistic value function $V_{k,h}(s)$ are indeed upper bounds of the true action-value function $Q_{k,h}^{\pi^k}$ and the true value function $V_{k,h}^{\pi^k}$, respectively.

C. Proof of the Main Results

C.1. Proof of Theorem 4.1

In this section, we provide the proof of Theorems 4.1 and we first propose the following lemmas.

Lemma C.1. On the event $\mathcal{E}$, for all $k \in [K], h \in [H], s \in \mathcal{S}, a \in \mathcal{A}$, we have

$$Q_{k,h}(s,a) \geq r_h^k(s,a) + \big[\mathbb{P}_h V_{k,h+1}\big](s,a).$$

Furthermore, on the event $\mathcal{E}$, for all $k \in [K], h \in [H], s \in \mathcal{S}, a \in \mathcal{A}$, we have

$$Q_{k,h}(s,a) - Q_{k,h}^{\pi^k}(s,a) \geq 0, V_{k,h}(s) - V_{k,h}^{\pi^k}(s) \geq 0.$$

Lemma C.2. On the event $\mathcal{E}$, for all $k \in [K]$, we have

$$V_{k,1}^*(s_1^k) - V_{k,1}(s_1^k) \leq \mathbb{E}\bigg[\sum_{h=1}^H \Big\{\mathbb{E}_{a\sim\pi_h^*(\cdot|s_h)}\big[Q_{k,h}(s_h,a)\big] - \mathbb{E}_{a\sim\pi_h^k(\cdot|s_h)}\big[Q_{k,h}(s_h,a)\big]\big|s_1 = s_1^k\Big\}\bigg].$$

Here $\mathbb{E}[\cdot|s = s_1^k]$ is the expectation with respect to the randomness of the state-action sequence $\{(s_h, a_h)\}_{h=1}^H$, where $a_h \sim \pi_h^*(\cdot|s_h)$ and $s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h)$.

Lemma C.2 suggests that the regret can be decomposed as the sum of the advantages at different stages h. Furthermore, since the initial state s_1^k and the optimal policy π^* is fixed across different episode k, the state-action sequence $(s_1, a_1, .., s_H, a_H)$ induced by the policy π^* follows the same distribution across different episode k.

Lemma C.3. On the event $\mathcal{E}$, for all $k \in [K]$, $h \in [H]$, $s \in \mathcal{S}$, we have

$$\mathbb{E}_{a\sim\pi_h^*(\cdot|s)}\big[Q_{k,h}(s,a)\big] - \mathbb{E}_{a\sim\pi_h^k(\cdot|s)}\big[Q_{k,h}(s,a)\big] \leq \frac{\alpha H^2}{2} + \alpha^{-1}\Big(D_{KL}\big(\pi_h^*(\cdot|s)\|\pi_h^k(\cdot|s)\big)D_{KL}\big(\pi_h^*(\cdot|s)\|\pi_h^{k+1}(\cdot|s)\big)\Big).$$

Lemma C.4. On the event $\mathcal{E}$, for all $k \in [K], h \in [H]$, we have

$$Q_{k,h}(s_h^k, a_h^k) - Q_{k,h}^{\pi^k}(s_h^k, a_h^k) \leq \big[\mathbb{P}_h(V_{k,h+1} - V_{k,h+1}^{\pi^k})\big](s_h^k, a_h^k) + 2\widehat{\beta}_k\bar{\sigma}_{k,h}\min\Big\{\big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)/\bar{\sigma}_{k,h}\big\|_2, 1\Big\}.$$

Lemma C.4 suggests that the difference between the true action-value function $Q_{k,h}^{\pi^k}(s_h^k, a_h^k)$ and the estimated value function $Q_{k,h}(s_h^k, a_h^k)$ can be bounded by the expected difference at the next-stage difference. Furthermore, for the expected difference at the next-stage and the exact difference at the next-stage, we have the following equation

$$\big[\mathbb{P}_h(V_{k,h+1} - V_{k,h+1}^{\pi^k})\big](s_h^k, a_h^k) = Q_{k,h+1}(s_{h+1}^k, a_{h+1}^k) - Q_{k,h+1}^{\pi^k}(s_{h+1}^k, a_{h+1}^k) + A_{h,k} + B_{h+1,k},$$

where $A_{h,k} = \big[\mathbb{P}_h(V_{k,h+1} - V_{k,h+1}^{\pi^k})\big](s_h^k, a_h^k) - \big(V_{k,h+1}(s_{h+1}^k) - V_{k,h+1}^{\pi^k}(s_{h+1}^k)\big)$ is the noise from the state transition and $B_{h,k} = \mathbb{E}_{a\sim\pi_h^k(\cdot|s_h^k)}\big[Q_{k,h}(s_h^k, a) - Q_{k,h}^{\pi^k}(s_h^k, a)\big] - \big(Q_{k,h}(s_h^k, a_h^k) - Q_{k,h}^{\pi^k}(s_h^k, a_h^k)\big)$ is the noise from the stochastic policy. These noises form a martingale difference sequence and we define two high probability events for them:

$$\mathcal{E}_1 = \Big\{\forall h \in [H], \sum_{k=1}^{K}\sum_{h'=h}^{H} A_{h',k} + B_{h',k} \le 4H\sqrt{T\log(H/\delta)}\Big\},$$
$$\mathcal{E}_2 = \Big\{\sum_{k=1}^{K}\sum_{h=1}^{H} A_{h,k} \le 2H\sqrt{2T\log(1/\delta)}\Big\}.$$

Then according to the Azuma–Hoeffding inequality, we have $\Pr(\mathcal{E}_1) \ge 1-\delta$ and $\Pr(\mathcal{E}_2) \ge 1-\delta$. Furthermore, on the events $\mathcal{E}_1$ and $\mathcal{E}_2$, we can telescope the inequality in Lemma C.4 over the K episodes, and obtain the following lemma.

Lemma C.5. On the event $\mathcal{E}\cap\mathcal{E}_1\cap\mathcal{E}_2$, for all $h \in [H]$, we have

$$\sum_{k=1}^{K}\big(V_{k,h}(s_h^k) - V_{k,h}^{\pi^k}(s_h^k)\big) \le 2\widehat{\beta}_K\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2}\sqrt{2Hd\log(1+K/\lambda)} + 4H\sqrt{T\log(H/\delta)},$$

and

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\big[\mathbb{P}_h(V_{k,h+1} - V_{k,h+1}^{\pi^k})\big](s_h^k, a_h^k) \le 2H\widehat{\beta}_K\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2}\sqrt{2Hd\log(1+K/\lambda)} + 4H^2\sqrt{T\log(H/\delta)}.$$

Lemma C.5 shows that the regret can be upper bounded by the total estimated variance $\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2$. For the total variance $\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}_h V_{k,h+1}^{\pi^k}](s_h^k, a_h^k)$, we introduce the high probability event $\mathcal{E}_3 = \Big\{\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}_h V_{k,h+1}^{\pi^k}](s_h^k, a_h^k) \le 3HT + 3H^3\log(1/\delta)\Big\}$. Then Lemma C.5 in Jin et al. (2018) suggests that $\Pr(\mathcal{E}_3) \ge 1-\delta$ and on the event $\mathcal{E}\cap\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3$, the following lemma gives a upper bound of the total estimated variance.

Lemma C.6. On the event $\mathcal{E}\cap\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3$, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2 \le 2HT/d + 179HT + 165d^3H^4\log^2(4K^2H/\delta)\log^2(1+KH^4/\lambda)$$
$$+ 2062H^5d^2\log^2(4K^2H/\delta)\log^2(1+K/\lambda).$$

Lemma C.7 (Azuma–Hoeffding inequality, Cesa-Bianchi & Lugosi 2006)**.** Let $\{x_i\}_{i=1}^n$ be a martingale difference sequence with respect to a filtration $\{\mathcal{G}_i\}$ satisfying $|x_i| \le M$ for some constant M, x_i is $\mathcal{G}_{i+1}$-measurable, $\mathbb{E}[x_i|\mathcal{G}_i] = 0$. Then for any $0 < \delta < 1$, with probability at least $1-\delta$, we have

$$\sum_{i=1}^{n} x_i \le M\sqrt{2n\log(1/\delta)}.$$

Lemma C.8. [Lemma 11, Abbasi-Yadkori et al. 2011] Let $\{\mathbf{x}_t\}_{t=1}^{\infty}$ be a sequence in $\mathbb{R}^d$, $\mathbf{V}_0 = \lambda\mathbf{I}$ and define $\mathbf{V}_t = \mathbf{V}_0 + \sum_{i=1}^{t}\mathbf{x}_i\mathbf{x}_i^\top$. If $\|\mathbf{x}_i\|_2 \le L$ holds for each i, then for each t, we have

$$\sum_{i=1}^{t}\min\big\{1, \|\mathbf{x}_i\|_{\mathbf{V}_{i-1}^{-1}}\big\} \le 2d\log\bigg(\frac{d\lambda + tL^2}{d\lambda}\bigg).$$

Proof of Theorem 4.1. For the regret, we have

$$\begin{aligned}\text{Regret}(K) &= \sup_{\pi}\sum_{k=1}^{K}\big(V_{k,1}^{\pi}(s_1^k) - V_{k,1}^{\pi^k}(s_1^k)\big)\\ &= \sum_{k=1}^{K}\big(V_{k,1}^{*}(s_1^k) - V_{k,1}^{\pi^k}(s_1^k)\big)\\ &= \underbrace{\sum_{k=1}^{K}\big(V_{k,1}^{*}(s_1^k) - V_{k,1}(s_1^k)\big)}_{I_1} + \underbrace{\sum_{k=1}^{K}\big(V_{k,1}(s_1^k) - V_{k,1}^{\pi^k}(s_1^k)\big)}_{I_2}.\end{aligned}$$

For the term I_1, applying Lemma C.2, we have

$$\begin{aligned}I_1 &= \sum_{k=1}^{K}\big(V_{k,1}^{*}(s_1^k) - V_{k,1}(s_1^k)\big)\\ &\le \sum_{k=1}^{K}\mathbb{E}\bigg[\sum_{h=1}^{H}\Big\{\mathbb{E}_{a\sim\pi_h^*(\cdot|s_h)}\big[Q_{k,h}(s_h,a)\big] - \mathbb{E}_{a\sim\pi_h^k(\cdot|s_h)}\big[Q_{k,h}(s_h,a)\big]\Big\}\big|s_1 = s_1^k\bigg]\\ &\le \sum_{k=1}^{K}\mathbb{E}\bigg[\sum_{h=1}^{H}\Big\{\frac{\alpha H^2}{2} + \alpha^{-1}\Big(D_{KL}\big(\pi_h^*(\cdot|s_h)\|\pi_h^k(\cdot|s_h)\big) - D_{KL}\big(\pi_h^*(\cdot|s_h)\|\pi_h^{k+1}(\cdot|s_h)\big)\Big)\Big\}\big|s_1 = s_1^k\bigg]\\ &= \frac{\alpha K H^3}{2} + \sum_{k=1}^{K}\alpha^{-1}\mathbb{E}\bigg[\sum_{h=1}^{H}\Big\{D_{KL}\big(\pi_h^*(\cdot|s_h)\|\pi_h^k(\cdot|s_h)\big) - D_{KL}\big(\pi_h^*(\cdot|s_h)\|\pi_h^{k+1}(\cdot|s_h)\big)\Big\}\big|s_1 = s_1^k\bigg],\end{aligned}\tag{C.1}$$

where $a_h \sim \pi_h^*(\cdot|s_h), s_{h+1}\sim\mathbb{P}_h(\cdot|s_h,a_h)$, the first inequality holds due to Lemma C.2 and the second inequality holds due to Lemma C.3. For Kullback–Leibler divergence $D_{KL}\big(\pi_h^*(\cdot|s_h)\|\pi_h^1(\cdot|s_h)\big)$, we have $0 \le D_{KL}\big(\pi_h^*(\cdot|s_h)\|\pi_h^1(\cdot|s_h)\big)$ and

$$\begin{aligned}D_{KL}\big(\pi_h^*(\cdot|s_h)\|\pi_h^1(\cdot|s_h)\big) &= \sum_{a\in\mathcal{A}}\pi_h^*(a|s_h)\log\bigg(\frac{\pi_h^*(a|s_h)}{\pi_h^1(a|s_h)}\bigg)\\ &= \sum_{a\in\mathcal{A}}\pi_h^*(a|s_h)\log\big(\pi_h^*(a|s_h)\times|\mathcal{A}|\big)\\ &= \log|\mathcal{A}| + \sum_{a\in\mathcal{A}}\pi_h^*(a|s_h)\log\big(\pi_h^*(a|s_h)\big)\\ &\le \log|\mathcal{A}|,\end{aligned}\tag{C.2}$$

where the first equation holds due to $\pi_h^1(a|s_h) = 1/|\mathcal{A}|$ and the inequality holds on due to $0 \le \pi_h^*(a|s_h) \le 1$. Substituting (C.2) into (C.1), we have

$$\begin{aligned}I_1 &\le \frac{\alpha K H^3}{2} + \sum_{k=1}^{K}\alpha^{-1}\mathbb{E}\bigg[\sum_{h=1}^{H}\Big\{D_{KL}\big(\pi_h^*(\cdot|s_h)\|\pi_h^k(\cdot|s_h)\big) - D_{KL}\big(\pi_h^*(\cdot|s_h)\|\pi_h^{k+1}(\cdot|s_h)\big)\Big\}\bigg]\\ &= \frac{\alpha K H^3}{2} + \alpha^{-1}\mathbb{E}\bigg[\sum_{h=1}^{H}\Big\{D_{KL}\big(\pi_h^*(\cdot|s_h)\|\pi_h^1(\cdot|s_h)\big) - D_{KL}\big(\pi_h^*(\cdot|s_h)\|\pi_h^{K+1}(\cdot|s_h)\big)\Big\}\bigg]\\ &\le \frac{\alpha K H^3}{2} + \alpha^{-1}\mathbb{E}\bigg[\sum_{h=1}^{H}\Big\{D_{KL}\big(\pi_h^*(\cdot|s_h)\|\pi_h^1(\cdot|s_h)\big)\Big\}\bigg]\\ &\le \frac{\alpha K H^3}{2} + \alpha^{-1}H\log|\mathcal{A}|,\end{aligned}\tag{C.3}$$

where s_1 is the fixed initial state, $a_h \sim \pi_h^*(\cdot|s_h), s_{h+1}\sim\mathbb{P}_h(\cdot|s_h,a_h)$, the first inequality holds due to (C.1) and the second inequality holds due to Kullback–Leibler divergence is non-negative and the third inequality holds due to (C.2). For the

term I_2, we have

$$
\begin{aligned}
I_2 &= \sum_{k=1}^{K} \big(V_{k,1}(s_1^k) - V_{k,1}^{\pi^k}(s_1^k)\big) \\
&\le 2\widehat{\beta}_K \sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H} \bar{\sigma}_{k,h}^2} \sqrt{2Hd\log(1+K/\lambda)} + 4H\sqrt{T\log(H/\delta)} \\
&\le 56\sqrt{dH^3T}\log(4K^2H/\delta)\log(1+K/\lambda) + 492\sqrt{d^2H^2T}\log(4K^2H/\delta)\log(1+K/\lambda) \\
&\quad + 1670d^2H^3\log^2(4K^2H/\delta)\log^2(1+K/\lambda) + 473d^{2.5}H^{2.5}\log^2(4K^2H/\delta)\log^2(1+KH^4/\lambda), \qquad \text{(C.4)}
\end{aligned}
$$

where the first inequality holds due to Lemma C.5 and the second inequality holds due to Lemma C.6 with the fact that $\sqrt{a+b+c+d} \le \sqrt{a} + \sqrt{b} + \sqrt{c} + \sqrt{d}$. Substituting (C.3) and (C.4) into (C.2), we finish the proof of Theorem 4.1. □

C.2. Proof of Theorem 4.3

In this section, we provide the proof of the lower bounds of the regret and the lower bound is based on previous work (Zhou et al., 2021b;a).

Proof of Theorem 4.3. To prove the lower bound, we construct a series of hard-to-learn adversarial MDPs introduced by Zhou et al. (2021b;a). To be more specific, the state space $\mathcal{S}$ consist of state $s_1, .., s_{H+2}$, where s_{H+1} and s_{H+2} are absorbing states. The action space $\mathcal{A} = \{-1, 1\}^{d-1}$ consists of 2^{d-1} different actions. The adversarial reward function r_h^k satisfies that $r_h^k(s_h, \boldsymbol{a}) = 0 (1 \le h \le H+1)$ and $r_h^k(s_{H+2}, \boldsymbol{a}) = 1$. For the transition probability function $\mathbb{P}_h$, s_{H+1} and s_{H+2} are absorbing states, which will always stay at the same state, and for other state $s_h (1 \le h \le H)$, we have

$$
\begin{aligned}
\mathbb{P}_h(s_{h+1}|s_h, \mathbf{a}) &= 1 - \delta - \langle \boldsymbol{\mu}_h, \mathbf{a} \rangle, \\
\mathbb{P}_h(s_{H+2}|s_h, \mathbf{a}) &= \delta + \langle \boldsymbol{\mu}_h, \mathbf{a} \rangle,
\end{aligned}
$$

where each $\boldsymbol{\mu}_h \in \{-\Delta, \Delta\}^d$ and $\delta = 1/H$. Furthermore, these hard-to-learn adversarial MDPs can be represented as linear mixture MDPs with the following feature mapping $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and vector $\boldsymbol{\theta}_h$:

$$
\begin{aligned}
&\boldsymbol{\phi}(s_{h+1}|s_h, \mathbf{a}) = \big(\alpha(1-\delta), -\beta\mathbf{a}\big), h \in [H], \\
&\boldsymbol{\phi}(s_{H+2}|s_h, \mathbf{a}) = \big(\alpha\delta, \beta\mathbf{a}\big), h \in [H], \\
&\boldsymbol{\phi}(s_{h+1}|s_h, \mathbf{a}) = \big(\alpha, \mathbf{0}\big), h \in [H], \\
&\boldsymbol{\phi}(s_{h+1}|s_h, \mathbf{a}) = (0, \mathbf{0}), h \in [H], \\
&\boldsymbol{\theta}_h = (1/\alpha, \boldsymbol{\mu}_h/\beta), h \in [H],
\end{aligned}
$$

where $\mathbf{0} = 0^{d-1}$ is a $(d-1)$-dimensional vector of all zeros, $\alpha = \sqrt{1/\big(1+(d-1)\Delta\big)}$ and $\beta = \sqrt{\Delta/\big(1+(d-1)\Delta\big)}$. Then, we have $\|\boldsymbol{\theta}_h\|_2 \le 2$ and these hard-to-learn MDP are $2-$bounded linear mixutre MDPs.

Since the adversarial reward function is fixed across different episode k, the value function $V_{k,h}^{\pi}$ is also fixed for each policy π and the optimal policy π^* is pick the action $\boldsymbol{a}^* = \boldsymbol{\mu}_{h'}/\Delta$ at state $s_{h'} (1 \le h' \le H)$. Therefore, the adversarial MDP will degenerate to non-adversarial MDP and the adversarial regret is the same as the non-adversarial regret. For the lower bound of the non-adversarial regret, Theorem 5.6 (Zhou et al., 2021a) shows that for any algorithm, if $H \ge 3, d \ge 4$ and $K \ge (d-1)^2 H/2$, then there exist a parameter $\boldsymbol{\mu}^* = \{\boldsymbol{\mu}_1^*, .., \boldsymbol{\mu}_H^*\}$ such that the expected regret is lower bounded by

$$
\text{Regret(K)} = \Omega(dH\sqrt{T}).
$$

Therefore, we finish the proof of Theorem 4.3. □

D. Proof of Lemmas in Appendix

D.1. Proof of Lemma B.1

We need the following Lemmas:

Lemma D.1 (Bernstein inequality for vector-valued martingales, Zhou et al. 2021a)**.** Let $\{\mathcal{F}_t\}_{t=1}^{\infty}$ be a filtration and $(\mathbf{x}_t, \eta_t)_{t\geq 1}$ be a stochastic process so that $\mathbf{x}_t \in \mathbb{R}^d$ is $\mathcal{F}_t$-measurable and $\eta_t \in \mathbb{R}$ is $\mathcal{F}_{t+1}$-measurable. For constant $R, L, \sigma, \lambda > 0$, $\boldsymbol{\mu}^* \in \mathbb{R}^d$, let $y_t = \langle \mathbf{x}_t, \boldsymbol{\mu}^* \rangle + \eta_t$ and suppose that

$$|\eta_t| \leq R, \mathbb{E}[\eta_t|\mathcal{F}_t] = 0, \mathbb{E}[\eta_t|\mathcal{F}_t] \leq \sigma^2, \|\mathbf{x}_t\|_2 \leq L.$$

Then, for any $0 \leq \delta \leq 1$, with probability at least $1-\delta$, we have

$$\forall t > 0, \left\| \sum_{i=1}^{t} \mathbf{x}_i \eta_i \right\|_{\boldsymbol{\Sigma}_t^{-1}} \leq \beta_t, \|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\boldsymbol{\Sigma}_t} \leq \beta_t + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2,$$

where $\boldsymbol{\Sigma}_t = \lambda \mathbf{I} + \sum_{i=1}^{t} \mathbf{x}\mathbf{x}^\top$, $\mathbf{b}_t = \sum_{i=1}^{t} \mathbf{x}_i y_i$, $\boldsymbol{\mu}_t = \boldsymbol{\Sigma}_t^{-1}\mathbf{b}_t$ and

$$\beta_t = 8\sigma\sqrt{d\log\big(1 + tL^2/(d\lambda)\big)\log(4t^2/\delta)} + 4R\log(4t^2/\delta).$$

Proof of Lemma B.1. For each $h \in [H]$, by the definition of $[\bar{\mathbb{V}}_h V_{k,h+1}](s_h^k, a_h^k)$ in (B.2) and $[\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k)$ in (2.1), we have

$$\begin{aligned}
&[\bar{\mathbb{V}}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k) \\
&= \Big[\big\langle \boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k), \widetilde{\boldsymbol{\theta}}_{k,h} \big\rangle, H^2\Big]_{[0,H^2]} - \Big[\big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widehat{\boldsymbol{\theta}}_{k,h} \big\rangle_{[0,H]}\Big]^2 \\
&\quad - \Big\{[\mathbb{P}_h V_{k,h+1}^2](s_h^k, a_h^k) - \big([\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k)\big)^2\Big\} \\
&= \underbrace{\Big[\big\langle \boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k), \widetilde{\boldsymbol{\theta}}_{k,h} \big\rangle\Big]_{[0,H^2]} - [\mathbb{P}_h V_{k,h+1}^2](s_h^k, a_h^k)}_{I_1} \\
&\quad + \underbrace{\big([\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k)\big)^2 - \Big[\big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widehat{\boldsymbol{\theta}}_{k,h} \big\rangle_{[0,H]}\Big]^2}_{I_2}.
\end{aligned} \tag{D.1}$$

For the term I_1, we have

$$\begin{aligned}
|I_1| &= \Big|\Big[\big\langle \boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k), \widetilde{\boldsymbol{\theta}}_{k,h} \big\rangle\Big]_{[0,H^2]} - [\mathbb{P}_h V_{k,h+1}^2](s_h^k, a_h^k)\Big| \\
&\leq \Big|\big\langle \boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k), \widetilde{\boldsymbol{\theta}}_{k,h} \big\rangle - [\mathbb{P}_h V_{k,h+1}^2](s_h^k, a_h^k)\Big| \\
&= \Big|\big\langle \boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k), \widetilde{\boldsymbol{\theta}}_{k,h} \big\rangle - \big\langle \boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k), \boldsymbol{\theta}_h \big\rangle\Big| \\
&= \Big|\big\langle \boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k), \widetilde{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h \big\rangle\Big| \\
&\leq \big\|\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)\big\|_{\widetilde{\boldsymbol{\Sigma}}_{k,h}^{-1}} \big\|\widetilde{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h\big\|_{\widetilde{\boldsymbol{\Sigma}}_{k,h}},
\end{aligned} \tag{D.2}$$

where the first inequality holds due to $0 \leq [\mathbb{P}_h V_{k,h+1}^2](s_h^k, a_h^k) \leq H^2$ and the second inequality holds due to Cauchy-Schwarz inequality. For the term $\|\widetilde{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h\|_{\widetilde{\boldsymbol{\Sigma}}_{k,h}}$, we apply Lemma B.1 with $\mathbf{x}_t = \boldsymbol{\phi}_{V_{t,h+1}^2}(s_h^t, a_h^t)$, $\eta_t = V_{t,h+1}^2(s_{h+1}^t) - [\mathbb{P}_h V_{t,h+1}^2](s_h^t, a_h^t)$. For $\mathbf{x}_t, \eta_t$, we have the following property

$$\begin{aligned}
\|\mathbf{x}_t\|_2 = \big\|\boldsymbol{\phi}_{V_{t,h+1}^2}(s_h^t, a_h^t)\big\|_2 \leq \max_{s'} V_{t,h+1}^2(s') \leq H^2, \\
\mathbb{E}[\eta_t|\mathcal{F}_t] = 0, |\eta_t| = \big|V_{t,h+1}^2(s_{h+1}^t) - [\mathbb{P}_h V_{t,h+1}^2](s_h^t, a_h^t)\big| \leq H^2, \\
\mathbb{E}[\eta_t^2|\mathcal{F}_t] \leq H^4.
\end{aligned}$$

Therefore, with probability at least $1 - \delta/H$, for all $k \in [K]$, we have

$$\big\|\widetilde{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h\big\|_{\widetilde{\boldsymbol{\Sigma}}_{k,h}} \leq 8H^2\sqrt{d\log\big(1 + kH^4/(d\lambda)\big)\log(4k^2H/\delta)} + 4H^2\log(4k^2H/\delta) + \sqrt{\lambda}B. \tag{D.3}$$

Substituting (D.3) into (D.2), we have

$$
\begin{aligned}
|I_1| &\leq \big\|\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)\big\|_{\widetilde{\boldsymbol{\Sigma}}_{k,h}^{-1}} \Big(8H^2\sqrt{d\log\big(1+kH^4/(d\lambda)\big)\log(4k^2H/\delta)} + 4H^2\log(4k^2H/\delta) + \sqrt{\lambda}B\Big) \\
&= \widetilde{\beta}_k \big\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)\big\|_2.
\end{aligned}
$$

Since both two terms of I_1 belong to the interval $[0, H^2]$, we have

$$
|I_1| \leq \min\Big\{\widetilde{\beta}_k \big\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)\big\|_2, H^2\Big\}. \tag{D.4}
$$

For the term I_2, we have

$$
\begin{aligned}
|I_2| &= \Big|\big([\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k)\big)^2 - \Big[\big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widehat{\boldsymbol{\theta}}_{k,h}\big\rangle_{[0,H]}\Big]^2\Big| \\
&= \Big|\Big[\big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widehat{\boldsymbol{\theta}}_{k,h}\big\rangle\Big]_{[0,H]} - [\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k)\Big| \\
&\quad \times \Big|\Big[\big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widehat{\boldsymbol{\theta}}_{k,h}\big\rangle\Big]_{[0,H]} + [\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k)\Big| \\
&\leq 2H\Big|\Big[\big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widehat{\boldsymbol{\theta}}_{k,h}\big\rangle\Big]_{[0,H]} - [\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k)\Big| \\
&\leq 2H\Big|\big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widehat{\boldsymbol{\theta}}_{k,h}\big\rangle - \big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \boldsymbol{\theta}_h\big\rangle\Big| \\
&= 2H\Big|\big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h\big\rangle\Big| \\
&\leq 2H\big\|\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\big\|_{\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1}} \big\|\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h\big\|_{\widehat{\boldsymbol{\Sigma}}_{k,h}},
\end{aligned} \tag{D.5}
$$

where the first inequality and second inequality holds due to $0 \leq [\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k) \leq H$ and the third inequality holds due to Cauchy-Schwarz inequality. For the term $\|\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h\|_{\widehat{\boldsymbol{\Sigma}}_{k,h}}$, we apply Lemma B.1 with $\mathbf{x}_t = \bar{\sigma}_{k,h}^{-1}\boldsymbol{\phi}_{V_{t,h+1}}(s_h^t, a_h^t)$, $\eta_t = \bar{\sigma}_{k,h}^{-1}V_{t,h+1}(s_{h+1}^t) - \bar{\sigma}_{k,h}^{-1}[\mathbb{P}_h V_{t,h+1}](s_h^t, a_h^t)$. For $\mathbf{x}_t, \eta_t$, we have following property

$$
\begin{aligned}
&\|\mathbf{x}_t\|_2 = \big\|\bar{\sigma}_{k,h}^{-1}\boldsymbol{\phi}_{V_{t,h+1}}(s_h^t, a_h^t)\big\|_2 \leq \bar{\sigma}_{k,h}^{-1}\max_{s'}|V_{t,h+1}(s')| \leq \sqrt{d}, \\
&\mathbb{E}[\eta_t|\mathcal{F}_t] = 0, |\eta_t| = \big|\bar{\sigma}_{k,h}^{-1}V_{t,h+1}(s_{h+1}^t) - \bar{\sigma}_{k,h}^{-1}[\mathbb{P}_h V_{t,h+1}](s_h^t, a_h^t)\big| \leq \sqrt{d}, \\
&\mathbb{E}[\eta_t^2|\mathcal{F}_t] \leq \sup\eta_t^2 \leq d.
\end{aligned}
$$

Therefore, with probability at least $1 - \delta/H$, for all $k \in [K]$, we have

$$
\big\|\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h\big\|_{\widehat{\boldsymbol{\Sigma}}_{k,h}} \leq 8d\sqrt{\log\big(1+kH^4/(d\lambda)\big)\log(4k^2H/\delta)} + 4\sqrt{d}\log(4k^2H/\delta) + \sqrt{\lambda}B. \tag{D.6}
$$

Substituting (D.6) into (D.5), we have

$$
\begin{aligned}
|I_2| &\leq 2H\big\|\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)\big\|_{\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1}} \Big(8d\sqrt{\log\big(1+kH^4/(d\lambda)\big)\log(4k^2H/\delta)} + 4\sqrt{d}\log(4k^2H/\delta) + \sqrt{\lambda}B\Big) \\
&= \widehat{\beta}_k \big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)\big\|_2.
\end{aligned}
$$

Since both two terms of I_2 belong to the interval $[0, H^2]$, we have

$$
|I_2| \leq \min\Big\{2H\bar{\beta}_k \big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)\big\|_2, H^2\Big\}. \tag{D.7}
$$

Substituting (D.4) and (D.7) into (D.1), with probability at least $1 - 2\delta/H$, we have

$$
\big|[\bar{\mathbb{V}}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k)\big| = |I_1 + I_2| \leq |I_1| + |I_2| \leq E_{k,h}, \tag{D.8}
$$

where

$$E_{k,h} = \min\Big\{\widetilde{\beta}_k\big\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)\big\|_2, H^2\Big\} + \min\Big\{2H\bar{\beta}_k\big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\big\|_2, H^2\Big\}.$$

We apply Lemma B.1 again with $\mathbf{x}_t = \bar{\sigma}_{k,h}^{-1}\boldsymbol{\phi}_{V_{t,h+1}}(s_h^t, a_h^t), \eta_t = \bar{\sigma}_{k,h}^{-1}V_{t,h+1}(s_{h+1}^t) - \bar{\sigma}_{k,h}^{-1}[\mathbb{P}_h V_{t,h+1}](s_h^t, a_h^t)$. For $\mathbf{x}_t, \eta_t$, we have following property

$$\|\mathbf{x}_t\|_2 = \big\|\bar{\sigma}_{k,h}^{-1}\boldsymbol{\phi}_{V_{t,h+1}}(s_h^t, a_h^t)\big\|_2 \le \bar{\sigma}_{k,h}^{-1}\max_{s'} V_{t,h+1}(s') \le \sqrt{d}$$
$$\mathbb{E}[\eta_t|\mathcal{F}_t] = 0, |\eta_t| = \big|\bar{\sigma}_{k,h}^{-1}V_{t,h+1}(s_{h+1}^t) - \bar{\sigma}_{k,h}^{-1}[\mathbb{P}_h V_{t,h+1}](s_h^t, a_h^t)\big| \le \sqrt{d}.$$

With probability at least $1 - 2\delta/H$, for all $t \in [K]$, we have

$$\mathbb{E}[\eta_t^2|\mathcal{F}_t] = \bar{\sigma}_{t,h}^{-1}[\mathbb{V}_h V_{t,h+1}](s_h^t, a_h^t), \le \bar{\sigma}_{t,h}^{-1}\big([\bar{\mathbb{V}}_h V_{t,h+1}](s_h^t, a_h^t) + E_{t,h}\big) \le 1, \tag{D.9}$$

where the first inequality holds due to (D.8) and the second inequality holds due to the definition of $\bar{\sigma}_{t,h}^{-1}$. Therefore, with probability at least $1 - 3\delta/H$, for all $k \in [K]$, we have

$$\big\|\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h\big\|_{\widehat{\boldsymbol{\Sigma}}_{k,h}} \le 8\sqrt{d\log\big(1 + kH^4/(d\lambda)\big)\log(4k^2H/\delta)} + 4\sqrt{d}\log(4k^2H/\delta) + \sqrt{\lambda}B = \widehat{\beta}_k.$$

Taking union bound for all $h \in [H]$, we finish the proof. □

D.2. Proof of Lemma C.1

Proof of Lemma C.1. For the lower bound of $Q_{k,h}(s,a) - Q_{k,h}^{\pi^k}(s,a)$, we have

$$\begin{aligned}
&r_h^k(s,a) + \big\langle\widehat{\boldsymbol{\theta}}_{k,h}, \boldsymbol{\phi}_{V_{k,h+1}}(s,a)\big\rangle + \widehat{\beta}_k\big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s,a)\big\|_2 \\
&= r_h^k(s,a) + \big[\mathbb{P}_h V_{k,h+1}\big](s,a) + \big\langle\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h, \boldsymbol{\phi}_{V_{k,h+1}}(s,a)\big\rangle + \widehat{\beta}_k\big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s,a)\big\|_2 \\
&\ge r_h^k(s,a) + \big[\mathbb{P}_h V_{k,h+1}\big](s,a) + \widehat{\beta}_k\big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s,a)\big\|_2 - \big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{1/2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{k,h})\big\|_2\big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s,a)\big\|_2 \\
&\ge r_h^k(s,a) + \big[\mathbb{P}_h V_{k,h+1}\big](s,a),
\end{aligned} \tag{D.10}$$

where the first inequality holds due to Cauchy-Schwarz inequality and the second inequality holds due to the definition of event $\mathcal{E}$. Therefore, we have

$$\begin{aligned}
Q_{k,h}(s,a) &= \Big[r_h^k(s,a) + \big\langle\widehat{\boldsymbol{\theta}}_{k,h}, \boldsymbol{\phi}_{V_{k,h+1}}(s,a)\big\rangle + \widehat{\beta}_k\big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s,a)\big\|_2\Big]_{[0,H-h+1]} \\
&\ge \min\Big\{r_h^k(s,a) + \big[\mathbb{P}_h V_{k,h+1}\big](s,a), H - h + 1\Big\} \\
&\ge r_h^k(s,a) + \big[\mathbb{P}_h V_{k,h+1}\big](s,a),
\end{aligned} \tag{D.11}$$

where the first inequality holds due to (D.10) and the second inequality holds due to $r_h^k(s,a) + \big[\mathbb{P}_h V_{k,h+1}\big](s,a) \le 1 + (H - h) = H - h + 1$. Now, we prove the second part of Lemma C.1 by induction. The statement holds for stage $h = H + 1$, since

$$V_{k,h}(s) = V_{k,h}^{\pi^k}(s) = 0.$$

When the second part of Lemma C.1 holds for stage $h + 1$, we have

$$\begin{aligned}
Q_{k,h}(s,a) &\ge r_h^k(s,a) + \big[\mathbb{P}_h V_{k,h+1}\big](s,a) \\
&\ge r_h^k(s,a) + \big[\mathbb{P}_h V_{k,h+1}^{\pi^k}\big](s,a) \\
&= Q^{\pi^k}(s,a),
\end{aligned} \tag{D.12}$$

where the first inequality holds due to (D.11) and the second inequality holds due to the induction assumption. Furthermore, we have

$$V_{k,h}(s) = \mathbb{E}_{a\sim\pi_h^k(\cdot|s)}[Q_{k,h}(s,a)] \ge \mathbb{E}_{a\sim\pi_h^k(\cdot|s)}[Q_{k,h}^{\pi^k}(s,a)] = V_{k,h}^{\pi^k}(s).$$

Therefore, we finish the inductive step and complete the proof of Lemma C.1. □

D.3. Proof of Lemma C.2

Proof of Lemma C.2. For each $h \in [H]$ and $s \in \mathcal{S}$, we have

$$\begin{aligned} V_{k,h}^*(s) - V_{k,h}(s) &= \mathbb{E}_{a\sim\pi_h^*(\cdot|s)}\big[Q_{k,h}^*(s,a)\big] - \mathbb{E}_{a\sim\pi_h^k(\cdot|s)}\big[Q_{k,h}(s,a)\big] \\ &= \underbrace{\mathbb{E}_{a\sim\pi_h^*(\cdot|s)}\big[Q_{k,h}^*(s,a)\big] - \mathbb{E}_{a\sim\pi_h^*(\cdot|s)}\big[Q_{k,h}(s,a)\big]}_{I} \\ &\quad + \mathbb{E}_{a\sim\pi_h^*(\cdot|s)}\big[Q_{k,h}(s,a)\big] - \mathbb{E}_{a\sim\pi_h^k(\cdot|s)}\big[Q_{k,h}(s,a)\big]. \end{aligned} \tag{D.13}$$

For the term I, we have

$$\begin{aligned} I &= \mathbb{E}_{a\sim\pi_h^*(\cdot|s)}\big[Q_{k,h}^*(s,a) - Q_{k,h}(s,a)\big] \\ &\le \mathbb{E}_{a\sim\pi_h^*(\cdot|s)}\Big[\big[\mathbb{P}_h(V_{k,h+1}^* - V_{k,h+1})\big](s,a)\Big] \\ &= \mathbb{E}_{a\sim\pi_h^*(\cdot|s),s'\sim\mathbb{P}_h(\cdot|s,a)}\big[V_{k,h+1}^*(s') - V_{k,h+1}(s')\big], \end{aligned} \tag{D.14}$$

where the inequality holds due to Lemma C.1. Recursively using (D.14) with all $h \in [H]$, for all $k \in [K]$, we have

$$V_{k,1}^*(s_1^k) - V_{k,1}(s_1^k) \le \mathbb{E}\bigg[\sum_{h=1}^H \Big\{\mathbb{E}_{a\sim\pi_h^*(\cdot|s_h)}\big[Q_{k,h}(s_h,a)\big] - \mathbb{E}_{a\sim\pi_h^k(\cdot|s_h)}\big[Q_{k,h}(s_h,a)\big]\Big\}\big|s_1 = s_1^k\bigg],$$

where $a_h \sim \pi_h^*(\cdot|s_h), s_{h+1} \sim \mathbb{P}_h(\cdot|s_h,a_h)$. Therefore, we finish the proof. □

D.4. Proof of Lemma C.3

Proof of Lemma C.3. By the update rule of the policy π_h^k, for all $k \in [K]$,$h \in [H]$, $s \in \mathcal{S}$, we have

$$\exp\big(\alpha Q_{k,h}(s,a)\big) = \frac{\pi_h^k(a|s)\exp\big\{\alpha Q_{h,k}(a|s)\big\}}{\pi_h^k(a|s)} = \frac{\rho\pi_h^{k+1}(a|s)}{\pi_h^k(a|s)}, \tag{D.15}$$

where $\rho = \sum_{a\in\mathcal{A}} \pi_h^k(a|s)\exp\big\{\alpha Q_{h,k}(a|s)\big\}$ is fixed for all action a. Thus, we have

$$\begin{aligned} &\sum_{a\in\mathcal{A}} \alpha Q_{k,h}(s,a)\big(\pi_h^*(a|s) - \pi_h^{k+1}(a|s)\big) \\ &= \sum_{a\in\mathcal{A}} \big(\log\rho + \log\pi_h^{k+1}(a|s) - \log\pi_h^k(a|s)\big)\big(\pi_h^*(a|s) - \pi_h^{k+1}(a|s)\big) \\ &= \sum_{a\in\mathcal{A}} \pi_h^*(a|s)\big(\log\pi_h^*(a|s) - \log\pi_h^k(a|s)\big) \\ &= \sum_{a\in\mathcal{A}} \pi_h^*(a|s)\big(\log\pi_h^*(a|s) - \log\pi_h^{k+1}(a|s)\big) - \sum_{a\in\mathcal{A}} \pi_h^*(a|s)\big(\log\pi_h^{k+1}(a|s) - \log\pi_h^k(a|s)\big) \\ &\quad - \sum_{a\in\mathcal{A}} \pi_h^{k+1}(a|s)\big(\log\pi_h^{k+1}(a|s) - \log\pi_h^k(a|s)\big) \\ &= D_{KL}\big(\pi_h^*(\cdot|s)\|\pi_h^{k+1}(\cdot|s)\big) - D_{KL}\big(\pi_h^*(\cdot|s)\|\pi_h^k(\cdot|s)\big) - D_{KL}\big(\pi_h^{k+1}(\cdot|s)\|\pi_h^k(\cdot|s)\big), \end{aligned} \tag{D.16}$$

where the first equation holds due to (D.15) and the second equation holds due to $\sum_{a\in\mathcal{A}}\big(\pi_h^*(a|s) - \pi_h^{k+1}(a|s)\big) = 0$. Therefore, we have

$$\begin{aligned} &\mathbb{E}_{a\sim\pi_h^*(\cdot|s_h)}\big[Q_{k,h}(s,a)\big] - \mathbb{E}_{a\sim\pi_h^k(\cdot|s)}\big[Q_{k,h}(s,a)\big] \\ &= \sum_{a\in\mathcal{A}} Q_{k,h}(s,a)\big(\pi_h^*(a|s) - \pi_h^k(a|s)\big) \\ &= \sum_{a\in\mathcal{A}} Q_{k,h}(s,a)\big(\pi_h^*(a|s) - \pi_h^{k+1}(a|s)\big) + \sum_{a\in\mathcal{A}} Q_{k,h}(s,a)\big(\pi_h^{k+1}(a|s) - \pi_h^k(a|s)\big) \end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{a\in\mathcal{A}} Q_{k,h}(s,a)\big(\pi_h^*(a|s) - \pi_h^{k+1}(a|s)\big) + H\big\|\pi_h^{k+1}(\cdot|s) - \pi_h^k(\cdot|s)\big\|_1 \\
&= \alpha^{-1}\Big(D_{KL}\big(\pi_h^*(\cdot|s)\|\pi_h^{k+1}(\cdot|s)\big) - D_{KL}\big(\pi_h^*(\cdot|s)\|\pi_h^k(\cdot|s)\big) - D_{KL}\big(\pi_h^{k+1}(\cdot|s)\|\pi_h^k(\cdot|s)\big)\Big) \\
&\quad + H\big\|\pi_h^{k+1}(\cdot|s) - \pi_h^k(\cdot|s)\big\|_1 \\
&\leq \alpha^{-1}\Big(D_{KL}\big(\pi_h^*(\cdot|s)\|\pi_h^{k+1}(\cdot|s)\big) - D_{KL}\big(\pi_h^*(\cdot|s)\|\pi_h^k(\cdot|s)\big)\Big) \\
&\quad + H\big\|\pi_h^{k+1}(\cdot|s) - \pi_h^k(\cdot|s)\big\|_1 - \frac{\big\|\pi_h^{k+1}(\cdot|s) - \pi_h^k(\cdot|s)\big\|_1^2}{2\alpha} \\
&\leq \frac{\alpha H^2}{2} + \alpha^{-1}\Big(D_{KL}\big(\pi_h^*(\cdot|s_h)\|\pi_h^k(\cdot|s_h)\big) - D_{KL}\big(\pi_h^*(\cdot|s_h)\|\pi_h^{k+1}(\cdot|s_h)\big)\Big),
\end{aligned} \tag{D.17}
$$

where the first inequality holds due to the fact that $0 \leq Q_{k,h}^{\pi^k}(s,a) \leq Q_{k,h}(s,a) \leq H$, the second inequality holds due to Pinsker's inequality and the last inequality holds due to the fact that $ax - bx^2 \leq a^2/4b$. Therefore, we finish the proof.

□

D.5. Proof of Lemma C.4

Proof of Lemma C.4.

$$
\begin{aligned}
&Q_{k,h}(s_h^k,a_h^k) - Q_{k,h}^{\pi^k}(s_h^k,a_h^k) \\
&= \Big[r_h^k(s_h^k,a_h^k) + \big\langle\widehat{\boldsymbol{\theta}}_{k,h}, \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\big\rangle + \widehat{\beta}_k\big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\big\|_2\Big]_{[0,H-h+1]} \\
&\quad - r_h^k(s_h^k,a_h^k) - [\mathbb{P}_h V_{k,h}^{\pi^k}](s_h^k,a_h^k) \\
&\leq \Big|\big\langle\widehat{\boldsymbol{\theta}}_{k,h}, \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\big\rangle + \widehat{\beta}_k\big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\big\|_2\Big| - [\mathbb{P}_h V_{k,h}^{\pi^k}](s_h^k,a_h^k) \\
&\leq \big[\mathbb{P}_h V_{k,h+1}\big](s_h^k,a_h^k) + \Big|\big\langle\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h, \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\big\rangle\Big| \\
&\quad + \widehat{\beta}_k\big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\big\|_2 - \big[\mathbb{P}_h V_{k,h+1}^{\pi^k}\big](s_h^k,a_h^k) \\
&\leq \big[\mathbb{P}_h(V_{k,h+1} - V_{k,h+1}^{\pi^k})\big](s_h^k,a_h^k) + 2\widehat{\beta}_k\big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\big\|_2 \\
&= \big[\mathbb{P}_h(V_{k,h+1} - V_{k,h+1}^{\pi^k})\big](s_h^k,a_h^k) + 2\widehat{\beta}_k\bar{\sigma}_{k,h}\big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)/\bar{\sigma}_{k,h}\big\|_2,
\end{aligned} \tag{D.18}
$$

where the first inequality holds due to the fact that $x_{[0,z]} - y \leq |x-y|$ when $y \geq 0$, the second inequality holds due to the fact that $|x+y+z| \leq |x| + |y| + |z|$ and the third inequality holds due to event $\mathcal{E}$. Furthermore, we have

$$
Q_{k,h}(s_h^k,a_h^k) - Q_{k,h}^{\pi^k}(s_h^k,a_h^k) \leq H - Q_{k,h}^{\pi^k} \leq H \leq 2\widehat{\beta}_k\bar{\sigma}_{k,h}, \tag{D.19}
$$

where the first inequality holds due to $Q_{k,h}(s_h^k,a_h^k) \leq H$, the second inequality holds due to $Q_{k,h}^{\pi^k}(s_h^k,a_h^k) \geq 0$ and the last inequality holds due to $2\widehat{\beta}_k\bar{\sigma}_{k,h} \geq \sqrt{d}H/\sqrt{d} = H$. Combined (D.18) and (D.19), we have

$$
Q_{k,h}(s_h^k,a_h^k) - Q_{k,h}^{\pi^k}(s_h^k,a_h^k) \leq \big[\mathbb{P}_h(V_{k,h+1} - V_{k,h+1}^{\pi^k})\big](s_h^k,a_h^k) + 2\widehat{\beta}_k\bar{\sigma}_{k,h}\min\Big\{\big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)/\bar{\sigma}_{k,h}\big\|_2, 1\Big\}.
$$

Therefore, we finish the proof. □

D.6. Proof of Lemma C.5

Proof of Lemma C.5.

$$
\begin{aligned}
&V_{k,h}(s_h^k) - V_{k,h}^{\pi^k}(s_h^k) \\
&= \mathbb{E}_{a\sim\pi_h^k(\cdot|s_h^k)}\big[Q_{k,h}(s_h^k,a) - Q_{k,h}^{\pi^k}(s_h^k,a)\big] \\
&= \mathbb{E}_{a\sim\pi_h^k(\cdot|s_h^k)}\big[Q_{k,h}(s_h^k,a) - Q_{k,h}^{\pi^k}(s_h^k,a)\big] - \big(Q_{k,h}(s_h^k,a_h^k) - Q_{k,h}^{\pi^k}(s_h^k,a_h^k) + Q_{k,h}(s_h^k,a_h^k) - Q_{k,h}^{\pi^k}(s_h^k,a_h^k)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \mathbb{E}_{a\sim\pi_h^k(\cdot|s_h^k)}\big[Q_{k,h}(s_h^k,a) - Q_{k,h}^{\pi^k}(s_h^k,a)\big] - \big(Q_{k,h}(s_h^k,a_h^k) - Q_{k,h}^{\pi^k}(s_h^k,a_h^k)\big) + \big[\mathbb{P}_h(V_{k,h+1} - V_{k,h+1}^{\pi^k})\big](s_h^k,a_h^k) \\
&\quad + 2\widehat{\beta}_k\bar{\sigma}_{k,h}\min\Big\{\big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)/\bar{\sigma}_{k,h}\big\|_2, 1\Big\} \\
&= \mathbb{E}_{a\sim\pi_h^k(\cdot|s_h^k)}\big[Q_{k,h}(s_h^k,a) - Q_{k,h}^{\pi^k}(s_h^k,a)\big] - \big(Q_{k,h}(s_h^k,a_h^k) - Q_{k,h}^{\pi^k}(s_h^k,a_h^k)\big) \\
&\quad + V_{k,h+1}(s_{h+1}^k) - V_{k,h+1}^{\pi^k}(s_{h+1}^k) + \big[\mathbb{P}_h(V_{k,h+1} - V_{k,h+1}^{\pi^k})\big](s_h^k,a_h^k) - \big(V_{k,h+1}(s_{h+1}^k) - V_{k,h+1}^{\pi^k}(s_{h+1}^k)\big) \\
&\quad + 2\widehat{\beta}_k\bar{\sigma}_{k,h}\min\Big\{\big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)/\bar{\sigma}_{k,h}\big\|_2, 1\Big\}, \qquad \text{(D.20)}
\end{aligned}
$$

where the inequality holds due to Lemma C.4. Furthermore, on the event $\mathcal{E}$ and $\mathcal{E}_1$, for all $h \in [H]$, we have

$$
\begin{aligned}
&\sum_{k=1}^K \big(V_{k,h}(s_h^k) - V_{k,h}^{\pi^k}(s_h^k)\big) \\
&\leq \sum_{k=1}^K\sum_{h'=h}^H 2\widehat{\beta}_k\bar{\sigma}_{k,h}\min\Big\{\big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)/\bar{\sigma}_{k,h}\big\|_2, 1\Big\} \\
&\quad + \sum_{k=1}^K\sum_{h'=h}^H\Big(\mathbb{E}_{a\sim\pi_h^k(\cdot|s_h^k)}\big[Q_{k,h}(s_h^k,a) - Q_{k,h}^{\pi^k}(s_h^k,a)\big] - \big(Q_{k,h}(s_h^k,a_h^k) - Q_{k,h}^{\pi^k}(s_h^k,a_h^k)\big)\Big) \\
&\quad + \sum_{k=1}^K\sum_{h'=h}^H\Big(\big[\mathbb{P}_h(V_{k,h+1} - V_{k,h+1}^{\pi^k})\big](s_h^k,a_h^k) - \big(V_{k,h+1}(s_{h+1}^k) - V_{k,h+1}^{\pi^k}(s_{h+1}^k)\big)\Big) \\
&\leq \sum_{k=1}^K\sum_{h'=h}^H 2\widehat{\beta}_k\bar{\sigma}_{k,h}\min\Big\{\big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)/\bar{\sigma}_{k,h}\big\|_2, 1\Big\} + 4H\sqrt{T\log(H/\delta)} \\
&\leq 2\widehat{\beta}_K\sum_{k=1}^K\sum_{h'=h}^H\bar{\sigma}_{k,h}\min\Big\{\big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)/\bar{\sigma}_{k,h}\big\|_2, 1\Big\} + 4H\sqrt{T\log(H/\delta)} \\
&\leq 2\widehat{\beta}_K\sqrt{\sum_{k=1}^K\sum_{h=1}^H\bar{\sigma}_{k,h}^2}\sqrt{\sum_{k=1}^K\sum_{h=1}^H\min\Big\{\big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)/\bar{\sigma}_{k,h}\big\|_2, 1\Big\}} + 4H\sqrt{T\log(H/\delta)} \\
&\leq 2\widehat{\beta}_K\sqrt{\sum_{k=1}^K\sum_{h=1}^H\bar{\sigma}_{k,h}^2}\sqrt{2Hd\log(1+K/\lambda)} + 4H\sqrt{T\log(H/\delta)}, \qquad \text{(D.21)}
\end{aligned}
$$

where the first inequality holds by taking the summation of (D.20) for $k \in [K]$ and $h \leq h' \leq H$, the second inequality holds due to the definition of event $\mathcal{E}_1$, the third inequality holds due to $\widehat{\beta}_k \leq \widehat{\beta}_K$, the fourth inequality holds due to Cauchy-Schwarz inequality and the last inequality holds due to Lemma C.8. Furthermore, taking the summation of (D.21), we have

$$
\begin{aligned}
&\sum_{k=1}^K\sum_{h=1}^H\big[\mathbb{P}_h(V_{k,h+1} - V_{k,h+1}^{\pi^k})\big](s_h^k,a_h^k) \\
&= \sum_{k=1}^K\sum_{h=1}^H\big(V_{k,h+1}(s_{h+1}^k) - V_{k,h+1}^{\pi^k}(s_{h+1}^k)\big) \\
&\quad + \sum_{k=1}^K\sum_{h=1}^H\Big(\big[\mathbb{P}_h(V_{k,h+1} - V_{k,h+1}^{\pi^k})\big](s_h^k,a_h^k) - \big(V_{k,h+1}(s_{h+1}^k) - V_{k,h+1}^{\pi^k}(s_{h+1}^k)\big)\Big) \\
&\leq \sum_{k=1}^K\sum_{h=1}^H\big(V_{k,h+1}(s_{h+1}^k) - V_{k,h+1}^{\pi^k}(s_{h+1}^k)\big) + 2H\sqrt{2T\log(1/\delta)}
\end{aligned}
$$

$$\leq 2H\widehat{\beta}_K\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2}\sqrt{2Hd\log(1+K/\lambda)}+4H^2\sqrt{T\log(H/\delta)},$$

where the first inequality holds due to the definition of event $\mathcal{E}_2$ and the last inequality holds due (D.21). Therefore, we finish the proof. □

D.7. Proof of Lemma C.6

Proof of Lemma C.6. On the event $\mathcal{E}$, by Lemma B.1, for all $k\in[K], h\in[H]$, we have

$$[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k,a_h^k)+E_{k,h}\geq[\mathbb{V}_hV_{k,h+1}](s_h^k,a_h^k)\geq 0.$$

Therefore, we have

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2 &= \sum_{k=1}^{K}\sum_{h=1}^{H}\max\left\{H^2/d,[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k,a_h^k)+E_{k,h}\right\}\\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\frac{H^2}{d}+\sum_{k=1}^{K}\sum_{h=1}^{H}[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k,a_h^k)+\sum_{k=1}^{K}\sum_{h=1}^{H}E_{k,h}\\
&= \frac{H^2T}{d}+\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\left([\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k,a_h^k)-[\mathbb{V}_hV_{k,h+1}](s_h^k,a_h^k)\right)}_{I_1}+\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}E_{k,h}}_{I_2}\\
&\quad+\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\left([\mathbb{V}_hV_{k,h+1}](s_h^k,a_h^k)-[\mathbb{V}_hV_{k,h+1}^{\pi^k}](s_h^k,a_h^k)\right)}_{I_3}+\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}_hV_{k,h+1}^{\pi^k}](s_h^k,a_h^k)}_{I_4},
\end{aligned}\tag{D.22}$$

where the inequality holds due to the fact that $\max\{a,b\}\leq a+b$, when $a,b\geq 0$. For the term I_1, we have

$$I_1=\sum_{k=1}^{K}\sum_{h=1}^{H}\left([\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k,a_h^k)-[\mathbb{V}_hV_{k,h+1}](s_h^k,a_h^k)\right)\leq\sum_{k=1}^{K}\sum_{h=1}^{H}E_{k,h}=I_2,\tag{D.23}$$

where the inequality holds due to the definition of event $\mathcal{E}$. For the term I_2, we have

$$\begin{aligned}
I_2 &= \sum_{k=1}^{K}\sum_{h=1}^{H}E_{k,h}\\
&= \sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{\widetilde{\beta}_k\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k,a_h^k)\|_2,H^2\right\}+\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{2H\bar{\beta}_k\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\|_2,H^2\right\}\\
&\leq \widetilde{\beta}_K\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k,a_h^k)\|_2,1\right\}\\
&\quad+2H\bar{\beta}_K\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}\min\left\{\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\|_2,1\right\}\\
&\leq \widetilde{\beta}_K\sqrt{T}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k,a_h^k)\|_2^2,1\right\}}\\
&\quad+2\sqrt{3}H^2\bar{\beta}_K\sqrt{T}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\|_2,1\right\}}\\
&\leq \widetilde{\beta}_K\sqrt{T}\sqrt{2dH\log\left(1+kH^4/(d\lambda)\right)}+2H^2\bar{\beta}_K\sqrt{3T}\sqrt{2dH\log(1+K/\lambda)},
\end{aligned}\tag{D.24}$$

where the first inequality holds on due to $\widetilde{\beta}_K \geq \widetilde{\beta}_k \geq H^2$ and $\bar{\beta}_K \bar{\sigma}_{k,h} \geq \bar{\beta}_k \bar{\sigma}_{k,h} \geq H$, the second inequality holds due to Cauchy-Schwarz inequality with the fact that $\bar{\sigma}_{k,h} \leq \max\{H^2/d, H^2 + 2H^2\} \leq 3H^2$ and the last inequality holds due to Lemma C.8.

On the event $\mathcal{E} \cap \mathcal{E}_1 \cap \mathcal{E}_2$, for the term I_3, we have

$$
\begin{aligned}
I_3 &= \sum_{k=1}^{K}\sum_{h=1}^{H}\big([\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{k,h+1}^{\pi^k}](s_h^k, a_h^k)\big) \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H}\Big([\mathbb{P}_h V_{k,h+1}^2](s_h^k, a_h^k) - \big([\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k)\big)^2 - [\mathbb{P}_h (V_{k,h+1}^{\pi^k})^2](s_h^k, a_h^k) + \big([\mathbb{P}_h V_{k,h+1}^{\pi^k}](s_h^k, a_h^k)\big)^2\Big) \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\big([\mathbb{P}_h V_{k,h+1}^2](s_h^k, a_h^k) - [\mathbb{P}_h (V_{k,h+1}^{\pi^k})^2](s_h^k, a_h^k)\big) \\
&\leq 2H\sum_{k=1}^{K}\sum_{h=1}^{H}\big([\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{P}_h V_{k,h+1}^{\pi^k}](s_h^k, a_h^k)\big) \\
&\leq 4H^2\widehat{\beta}_K\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2}\sqrt{2Hd\log(1+K/\lambda)} + 8H^3\sqrt{T\log(H/\delta)},
\end{aligned} \tag{D.25}
$$

where the first inequality holds due to the fact that $V_{k,h+1}^{\pi^k}(s') \leq V_{k,h+1}(s')$, the second inequality holds due to $0 \leq V_{k,h+1}(s'), V_{k,h+1}^{\pi^k}(s') \leq H$ and the last inequality holds due to Lemma C.5.

On the event $\mathcal{E}_3$, for the term I_4, we have

$$
I_4 = \sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}_h V_{k,h+1}^{\pi^k}](s_h^k, a_h^k) \leq 3\big(HT + H^3\log(1/\delta)\big). \tag{D.26}
$$

Substituting (D.23), (D.24), (D.25) and (D.26) into (D.22), we have

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^{-2} \leq{}& H^2T/d + 3\big(HT + H^3\log(1/\delta)\big) \\
&+ 2\widetilde{\beta}_K\sqrt{T}\sqrt{2dH\log\big(1 + KH^4/(d\lambda)\big)} + 4H^2\bar{\beta}_K\sqrt{3T}\sqrt{2dH\log(1+K/\lambda)} \\
&+ 4H^2\widehat{\beta}_K\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2}\sqrt{2Hd\log(1+K/\lambda)} + 8H^3\sqrt{T\log(H/\delta)},
\end{aligned}
$$

where

$$
\begin{aligned}
\bar{\beta}_K &= 8d\sqrt{\log(1+K/\lambda)\log(4K^2H/\delta)} + 4\sqrt{d}\log(4K^2H/\delta) + \sqrt{\lambda}B, \\
\widehat{\beta_K} &= 8\sqrt{d\log(1+K/\lambda)\log(4K^2H/\delta)} + 4\sqrt{d}\log(4K^2H/\delta) + \sqrt{\lambda}B, \\
\widetilde{\beta}_K &= 8H^2\sqrt{d\log\big(1 + KH^4/(d\lambda)\big)\log(4K^2H/\delta)} + 4H^2\log(4K^2H/\delta) + \sqrt{\lambda}B.
\end{aligned}
$$

Therefore, by the fact that $x \leq a\sqrt{x} + b$ implies $x \leq a^2 + 2b$, we have

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^{-2} \leq{}& 2H^2T/d + 6\big(HT + H^3\log(1/\delta)\big) \\
&+ 4\widetilde{\beta}_K\sqrt{T}\sqrt{2dH\log\big(1 + KH^4/(d\lambda)\big)} + 8H^2\bar{\beta}_K\sqrt{3T}\sqrt{2dH\log(1+K/\lambda)} \\
&+ 32H^5d(\widehat{\beta}_K)^2\log(1+K/\lambda) + 16H^3\sqrt{T\log(H/\delta)}
\end{aligned}
$$

$$
\begin{aligned}
&\leq 2H^2T/d + 6\big(HT + H^3\log(1/\delta)\big) + 330\sqrt{d^3H^5T}\log(4K^2H/\delta)\log(1+KH^4/\lambda) \\
&\quad + 2048H^5d^2\log^2(4K^2H/\delta)\log^2(1+K/\lambda) + 16H^3\sqrt{T\log(H/\delta)} \\
&\leq 2HT/d + 179HT + 165d^3H^4\log^2(4K^2H/\delta)\log^2(1+KH^4/\lambda) \\
&\quad + 2062H^5d^2\log^2(4K^2H/\delta)\log^2(1+K/\lambda),
\end{aligned}
$$

where the second inequality holds due to the definition of parameter $\bar{\beta}_K, \widehat{\beta_K}, \widetilde{\beta}_K$ with the fact that $\lambda = 1/B^2 \leq 1$ and the third inequality holds due to Young's inequality. Therefore, we finish the proof. □

E. Computational complexity

The computational complexity of POWER is related to the property of the given feature mapping $\boldsymbol{\phi}(s'|s,a)$ and we consider a special class of linear mixture MDPs studied by Yang & Wang (2019b); Zhou et al. (2021b;a). For this special class of linear mixture MDPs, we have

$$
[\boldsymbol{\phi}(s'|s,a)]_i = [\boldsymbol{\psi}(s')]_i \cdot [\boldsymbol{\mu}(s,a)]_i, \forall i \in [d],
$$

where $\boldsymbol{\psi} : \mathcal{S} \to \mathbb{R}^d$ and $\boldsymbol{\mu} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. Under this setting, for each function V, the vector $\boldsymbol{\phi}_V(s,a)$ can be written as the product of $\boldsymbol{\mu}(s,a)$ and $\sum_{s'\in\mathcal{S}} \boldsymbol{\psi}(s')V(s')$. Furthermore, the term $\sum_{s'\in\mathcal{S}} \boldsymbol{\psi}(s')V(s')$ can be estimated by Monte Carlo method and in this work, we assume an access to the oracle $\mathcal{O}$ which can compute the term $\sum_{s'\in\mathcal{S}} \boldsymbol{\psi}(s')V(s')$. We also assume the size of action space is finite ($|\mathcal{A}| < \infty$) and analyze the computational complexity of POWER in the sequel.

Recall that POWER can be divided into two phases: (1) policy improvement; and (2) policy evaluation. For the policy evaluation phase, in order to compute the vector $\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)$ and the vector $\boldsymbol{\phi}_{V^2_{k,h+1}}(s_h^k, a_h^k)$, POWER needs to compute the term $\sum_{s'\in\mathcal{S}} \boldsymbol{\psi}(s')\boldsymbol{\phi}_{V_{k,h+1}}(s')$ and $\sum_{s'\in\mathcal{S}} \boldsymbol{\psi}(s')V^2_{k,h+1}(s')$, which need two accesses to the oracle $\mathcal{O}$. Given the vector $\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)$ and $\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)$, the covariance matrix can be computed in $O(d^2)$ time, and the estimators $\widehat{\boldsymbol{\theta}}_{k+1,h}$ and $\widetilde{\boldsymbol{\theta}}_{k+1,h}$ can be computed in $O(d^3)$ time. Therefore, the policy evaluation phase can be computed in $O(d^3HK)$ time with $O(HK)$ accesses to the oracle $\mathcal{O}$.

For the policy improvement phase, by the update rule of the policy π_h^k, we have

$$
\pi_h^{k+1}(a|s) \propto \pi_h^k(a|s)\exp\big\{\alpha Q_{k,h}(s,a)\big\}.
$$

When the size of the state space $\mathcal{S}$ is too large or even infinite, it is not computationally efficient. However, in Line 7, we do not need to calculate the π_h^k for all state s and we only need the value of policy π_h^k for state s_h^k , which can be calculated as follows

$$
\begin{aligned}
\pi_h^k(a|s_h^k) &\propto \exp\bigg\{\alpha\sum_{i=1}^{k-1} Q_{i,h}(s_h^k, a)\bigg\} \\
&\propto \exp\bigg\{\alpha\sum_{i=1}^{k-1}\Big[r_h^i(s_h^k,a) + \big\langle\widehat{\boldsymbol{\theta}}_{i,h}, \boldsymbol{\phi}_{V_{i,h+1}}(s_h^k,a)\big\rangle + \widehat{\beta}_i\big\|\widehat{\boldsymbol{\Sigma}}_{i,h}^{-1/2}\boldsymbol{\phi}_{V_{i,h+1}}(s_h^k,a)\big\|_2\Big]_{[0,H-h+1]}\bigg\}.
\end{aligned}
$$

Thus, given the covariance matrix $\widehat{\boldsymbol{\Sigma}}_{i,h}$, the estimator $\widehat{\boldsymbol{\theta}}_{i,h}$ and the term $\sum_{s'\in\mathcal{S}} \boldsymbol{\psi}(s')V_{i,h+1}(s')$, the policy $\pi_h^k(\cdot|s_h^k)$ can be computed in $O\big(d^3K|\mathcal{A}|\big)$ time and it will take in total $O\big(d^3HK^2|\mathcal{A}|\big)$ time to compute all policies $\pi_h^k(\cdot|s_h^k)$ for $k \in [K]$ and $h \in [H]$. Combining the time complexity of the two phases, the total time complexity of POWER is $O\big(d^3HK^2|\mathcal{A}|\big)$ with $O(HK)$ accesses to the oracle.