OPTIMISTIC POLICY OPTIMIZATION IS PROVABLY EFFICIENT IN NON-STATIONARY MDPS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study episodic reinforcement learning (RL) in non-stationary linear kernel Markov decision processes (MDPs). In this setting, both the reward function and the transition kernel are linear with respect to the given feature maps and are allowed to vary over time, as long as their respective parameter variations do not exceed certain variation budgets. We propose the periodically restarted optimistic policy optimization algorithm (PROPO), which is an optimistic policy optimization algorithm with linear function approximation. PROPO features two mechanisms: sliding-window-based policy evaluation and periodic-restart-based policy improvement, which are tailored for policy optimization in a non-stationary environment. In addition, only utilizing the technique of sliding window, we propose a value-iteration algorithm. We establish dynamic upper bounds for the proposed methods and a matching minimax lower bound which shows the (near-) optimality of the proposed methods. To our best knowledge, PROPO is the first provably efficient policy optimization algorithm that handles non-stationarity.

1 INTRODUCTION

Reinforcement Learning (RL) (Sutton & Barto, 2018), coupled with powerful function approximators such as deep neural network, has demonstrated great potential in solving complicated sequential decision-making tasks such as games (Silver et al., 2016; 2017; Vinyals et al., 2019) and robotic control (Kober et al., 2013; Gu et al., 2017; Akkaya et al., 2019; Andrychowicz et al., 2020). Most of these empirical successes are driven by deep policy optimization methods such as trust region policy optimization (TRPO) (Schulman et al., 2015) and proximal policy optimization (PPO) (Schulman et al., 2017), whose performance has been extensively studied recently (Agarwal et al., 2019; Liu et al., 2019; Shani et al., 2020; Mei et al., 2020; Cen et al., 2020).

While classical RL assumes that an agent interacts with a time-invariant (stationary) environment, when deploying RL to real-world applications, both the reward function and Markov transition kernel can be time-varying. For example, in autonomous driving (Sallab et al., 2017), the vehicle needs to handle varying conditions of weather and traffic. When the environment changes with time, the agent must quickly adapt its policy to maximize the expected total rewards in the new environment. Meanwhile, another example of such a non-stationary scenario is when the environment is subject to adversarial manipulations, which is the case of adversarial attacks (Pinto et al., 2017; Huang et al., 2017; Pattanaik et al., 2017). In this situation, it is desired that the RL agent is robust against the malicious adversary.

Although there is a huge body of literature on developing provably efficient RL methods, most the existing works focus on the classical stationary setting, with a few exceptions include Jaksch et al. (2010); Gajane et al. (2018); Cheung et al. (2019a;c; 2020); Fei et al. (2020); Mao et al. (2020); Ortner et al. (2020); Domingues et al. (2020); Zhou et al. (2020b); Touati & Vincent (2020). However, these works all focus on value-based methods which only output greedy policies, and mostly focus on the tabular case where the state space is finite. Thus, the following problem remains open:

How can we design a provably efficient policy optimization algorithm for non-stationary environment in the context of function approximation?

There are four intertwined challenges associated with this problem: (i) bandit feedbacks from non-stationary reward and transition kernel, (ii) exploration-exploitation tradeoff that is inherent to online RL, (iii) incorporating function approximation in the algorithm, and (iv) characterizing the convergence and optimality of policy optimization. Existing works merely address a subset of these four challenges and it remains open how to tackle all of them simultaneously. For example, a line of research develops optimism-based value iteration algorithms that successfully handle (ii) and (iii), e.g., (Jiang et al., 2017; Jin et al., 2019b; Wang et al., 2019b; Zanette et al., 2020; Wang et al., 2020; Ayoub et al., 2020; Zhou et al., 2020a). Besides, Cai et al. (2019); Agarwal et al. (2020); Efroni et al. (2020) address challenges (ii)–(iv) but fail to consider (i), and Zhou et al. (2020b); Touati & Vincent (2020) tackle (i)–(iii) but leave (iv) open. More importantly, these four challenges are coupled together, which requires sophisticated algorithm design. In particular, due to challenges (i) and (iii), we need to track the non-stationary reward function and transition kernel by function estimation based on the bandit feedbacks. The estimated model is also time-varying and thus the corresponding policy optimization problem (challenge (iv)) has a non-stationary objective function. Moreover, to obtain sample efficiency, we need to strike a balance between exploration and exploitation in the policy update steps (challenge (i)).

In this work, we propose a periodically restarted optimistic policy optimization algorithm (PROPO) which successfully tackle the four challenges above. Specifically, we focus on the model of episodic linear kernel MDP (Ayoub et al., 2020; Zhou et al., 2020a) where both the reward and transition functions are parameterized by linear functions. Besides, we focus on the non-stationary setting and adopt the dynamic regret as the performance metric. Moreover, PROPO performs a policy evaluation step and a policy improvement step in each iteration. To handle challenges (i)–(iii), we propose a novel optimistic policy evaluation method that incorporates the technique of sliding window to handle non-stationarity. Specifically, based on the non-stationary bandit feedbacks, we propose to estimate the time-varying model via a sliding-window-based least-squares regression problem, where we only keep a subset of recent samples in regression. Based on the model estimator, we construct an optimistic value function by implementing model-based policy evaluation and adding an exploration bonus. Then, using such an optimistic value function as the update direction, in the policy improvement step, we propose to obtain a new policy by solving a Kullback-Leibler (KL) divergence regularized problem, which can be viewed as a mirror descent step. Moreover, as the underlying optimal policy is time-varying (challenge (iv)), we additionally restart the policy periodically by setting it to uniform policy every τ episodes. The two novel mechanisms, sliding window and periodic restart, respectively enable us to track the non-stationary MDP based on bandit feedbacks and handle the time-varying policy optimization problem.

Finally, to further exhibit effect of these two mechanisms, we propose an optimism-based value iteration algorithm, dubbed as SW-LSVI-UCB, which only utilize the sliding window and does not restart the policy as challenge (iv) disappears.

Our Contributions Our contribution is four-fold. First, we propose PROPO, a policy optimization algorithm designed for non-stationary linear kernel MDPs. This algorithm features two novel mechanisms, namely sliding window and periodic restart, and also incorporates linear function approximation and a bonus function to incentivize exploration. Second, we prove that PROPO achieves a sublinear dynamic regret, where d is the feature dimension, Δ is the total variation budget, H is the episode horizon, and T is the total number of steps. Third, to separately demonstrate the effect of sliding window, we propose a value-iteration algorithm, SW-LSVI-UCB, which adopts sliding-window-based regression to handle non-stationarity. Such an algorithm is shown to achieve a $\widetilde{O}(d^{5/6}\Delta^{1/3}HT^{2/3})$ dynamic regret. Finally, we establish a $\Omega(d^{5/6}\Delta^{1/3}H^{2/3}T^{2/3})$ lower bound on the dynamic regret, which shows the (near-)optimality of the proposed algorithms. To our best knowledge, PROPO is the first provably efficient policy optimization algorithm under the non-stationary environment.

Related Work Our work adds to the vast body of existing literature on non-stationary MDPs. A line of work studies non-stationary RL in the tabular setting. See Jaksch et al. (2010); Gajane et al. (2018); Cheung et al. (2019a;c; 2020); Fei et al. (2020); Mao et al. (2020); Ortner et al. (2020) and the references therein for details. Recently, Domingues et al. (2020) consider the non-stationary RL in continuous environments and proposes a kernel-based algorithm. More related works are Zhou et al. (2020b); Touati & Vincent (2020), which study non-stationary linear MDPs, but their setting is not directly comparable with ours since linear MDPs cannot imply linear kernel MDPs. Moreover, Zhou et al. (2020b); Touati & Vincent (2020) do not incorporate policy optimization methods, which are more difficult because we need to handle the variation of the optimal policies of adjacent episodes and value-based methods only need to handle the non-stationarity drift of reward functions and transition kernels. Fei et al. (2020) also makes an attempt to investigate policy optimization algorithm for non-stationary environments. However, this work requires full-information feedback and only focuses on the tabular MDPs with time-varying reward functions and time-invariant transition kernels.

As a special case of MDP problems with unit horizon, bandit problems have been the subject of intense recent interest. See Besbes et al. (2014; 2019); Russac et al. (2019); Cheung et al. (2019a); Chen et al. (2019) and the references therein for details.

Another line closely related to our work is policy optimization. As proved in Yang et al. (2019); Agarwal et al. (2019); Liu et al. (2019); Wang et al. (2019a), policy optimization enjoys computational efficiency. Recently Cai et al. (2019); Efroni et al. (2020); Agarwal et al. (2020) proposed optimistic policy optimization methods which simultaneously attain computational efficiency and sample efficiency. Our work is also related to the value-based methods, especially LSVI (Bradtke & Barto, 1996; Jiang et al., 2017; Jin et al., 2019b; Wang et al., 2019b; Zanette et al., 2020; Wang et al., 2020; Ayoub et al., 2020; Zhou et al., 2020a).

Broadly speaking, our work is also related to a line of research on adversarial MDPs (Even-Dar et al., 2009; Neu et al., 2010; 2012; Zimin & Neu, 2013; Rosenberg & Mansour, 2019; Jin et al., 2019a).

Notaion See §A for details.

2 Preliminaries

2.1 Non-stationary MDPs

An episodic non-stationary MDP is defined by a tuple $(\mathcal{S}, \mathcal{A}, H, P, r)$, where $\mathcal{S}$ is a state space, $\mathcal{A}$ is an action space, H is the length of each episode, $P = \{P_h^k\}_{(k,h)\in[K]\times[H]}$, $r = \{r_h^k\}_{(k,h)\in[K]\times[H]}$, where $P_h^k : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to [0,1]$ is the probability transition kernel at the h-th step of the k-th episode, and $r_h^k : \mathcal{S}\times\mathcal{A} \to [0,1]$ is the reward function at the h-th step of the k-th episode. We consider an agent which iteratively interacts with a non-stationary MDP in a sequence of K episodes. At the beginning of the k-th episode, the initial state s_1^k is adversarially given to the agent, and the agent determines a policy $\pi^k = \{\pi_h^k\}_{h=1}^H$. Then, at each step $h \in [H]$, the agent observes the state s_h^k, takes an action following the policy $a_h^k \sim \pi_h^k(\cdot \,|\, s_h^k)$ and receives a reward $r_h^k(s_h^k, a_h^k)$. Meanwhile, the MDP evolves into next state $s_{h+1}^k \sim P_h^k(\cdot \,|\, x_h^k, a_h^k)$. The k-th episode ends at state s_{H+1}^k, when this happens, no control action is taken and reward is equal to zero. We define the state and state-action value functions of policy $\pi = \{\pi_h\}_{h=1}^H$ recursively via the following Bellman equation:

$$Q_h^{\pi,k}(s,a) = r_h^k(s,a) + (\mathbb{P}_h^k V_{h+1}^{\pi,k})(s,a), \quad V_h^{\pi,k}(s) = \langle Q_h^{\pi,k}(s,\cdot), \pi_h(\cdot\,|\,s)\rangle_{\mathcal{A}}, \quad V_{H+1}^{\pi,k} = 0, \tag{2.1}$$

where $\mathbb{P}_h^k$ is the operator form of the transition kernel $P_h^k(\cdot\,|\,\cdot,\cdot)$, which is defined as

$$(\mathbb{P}_h^k f)(s,a) = \mathbb{E}[f(s') \,|\, s' \sim P_h^k(s' \,|\, s,a)] \tag{2.2}$$

for any function $f : \mathcal{S} \to \mathbb{R}$. Here $\langle\cdot,\cdot\rangle_{\mathcal{A}}$ denotes the inner product over $\mathcal{A}$.

In the literature of optimization and reinforcement learning, the performance of the agent is measured by its dynamic regret, which measures the difference between the agent's policy and the benchmark policy $\pi^* = \{\pi^{*,k}\}_{k=1}^K$. Specifically, the dynamic regret is defined as

$$\text{D-Regret}(T, \pi^*) = \sum_{k=1}^K \big(V_1^{\pi^{*,k},k}(s_1^k) - V_1^{\pi^k,k}(s_1^k)\big), \tag{2.3}$$

where $T = HK$ is the number of steps taken by agent and $\pi^{*,k}$ is the benchmark policy of episode k. It is worth mentioning that when the benchmark policy is the optimal policy of each individual episode, that is, $\pi^{*,k} = \operatorname{argmax}_\pi V_1^{\pi,k}(s_1^k)$, the dynamic regret reaches the maximum, and this special case is widely considered in previous works (Cheung et al., 2020; Mao et al., 2020; Domingues et al., 2020). Throughout the paper, when π^* is clear from the context, we may omit π^* from D-Regret(T, π^*).

2.2 Model Assumptions

We focus on the linear setting of Markov decision process, where the reward functions and transition kernels are assumed to be linear. We formally make the following assumption.

Assumption 2.1 (Non-stationary Linear Kernel MDP)**.** MDP $(\mathcal{S}, \mathcal{A}, H, P, r)$ is a linear kernel MDP with known feature maps $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and $\psi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$, if for any $(k, h) \in [K] \times [H]$, there exist unknown vectors $\theta_h^k \in \mathbb{R}^d$ and $\xi_h \in \mathbb{R}^d$, such that

$$r_h^k(s,a) = \phi(s,a)^\top \theta_h^k, \qquad P_h^k(s' \mid s,a) = \psi(s,a,s')^\top \xi_h^k$$

for any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Without loss of generality, we assume that

$$\|\phi(s,a)\|_2 \le 1, \qquad \|\theta_h^k\|_2 \le \sqrt{d}, \qquad \|\xi_h^k\|_2 \le \sqrt{d}$$

for any $(k, h) \in [K] \times [H]$. Moreover, we assume that

$$\int_{\mathcal{S}} \|\psi(s,a,s')\|_2 \, \mathrm{d}s' \le \sqrt{d}$$

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

Our assumption consists of two parts. One is about reward functions, which follows the setting of linear bandits (Abbasi-Yadkori et al., 2011; Agrawal & Goyal, 2013; Besbes et al., 2014; 2015; Cheung et al., 2019a;b). The other part is about transition kernels. As shown in Cai et al. (2019); Ayoub et al. (2020); Zhou et al. (2020a), linear kernel MDPs as defined above cover several other MDPs studied in previous works, as special cases. For example, tabular MDPs with canonical basis (Cai et al., 2019; Ayoub et al., 2020; Zhou et al., 2020a), feature embedding of transition models (Yang & Wang, 2019a) and linear combination of base models (Modi et al., 2020) are special cases. However, it is worth mentioning that Jin et al. (2019b); Yang & Wang (2019b) studied another "linear MDPs", which assumes the transition kernels can be represented as $\mathbb{P}_h(s' \mid s,a) = \psi'(s,a)^\top \mu_h(s')$ for any $h \in [H]$ and $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Here $\psi'(\cdot,\cdot)$ is a known feature map and $\mu(\cdot)$ is an unknown measure. It is worth noting that linear MDPs studied in our paper and linear MDPs (Jin et al., 2019b; Yang & Wang, 2019b) are two different classes of MDPs since their feature maps $\psi(\cdot,\cdot,\cdot)$ and $\psi'(\cdot,\cdot)$ are different and neither class of MDPs includes the other.

To facilitate the following analysis, we denote by $P_h^{k,\pi}$ the Markov kernel of policy at the h-step of the k-th episode, that is, for $s \in \mathcal{S}$, $P_h^{k,\pi}(\cdot \mid s) = \sum_{a \in \mathcal{A}} P_h^k(\cdot \mid s,a) \cdot \pi_h(a \mid s)$. Also, we define

$$\|\pi_h - \pi'_h\|_{\infty,1} = \max_{s \in \mathcal{S}} \|\pi_h(\cdot \mid s) - \pi'_h(\cdot \mid s)\|_1,$$
$$\|P_h^{k,\pi} - P_h^{k,\pi'}\|_{\infty,1} = \max_{s \in \mathcal{S}} \|P_h^{k,\pi}(\cdot \mid s) - P_h^{k,\pi'}(\cdot \mid s)\|_1.$$

Next, we introduce several measures of change in MDPs. First, we denote by P_T the total variation in the benchmark policies of adjacent episodes:

$$P_T = \sum_{k=1}^{K} \sum_{h=1}^{H} \|\pi_h^{*,k} - \pi_h^{*,k-1}\|_{\infty,1}, \tag{2.4}$$

where we choose $\pi_h^{*,0} = \pi_h^{*,1}$ for any $h \in [H]$.

Next, we assume the drifting environment (Besbes et al., 2014; 2015; Cheung et al., 2019a; Russac et al., 2019), that is, θ_h^k and ξ_h^k can change over different indexes (k, h), with the constraint that the sum of the Euclidean distances between consecutive θ_h^k and ξ_h^k are bounded by variation budgets B_T and B_P, that is,

$$\sum_{h=1}^{H} \sum_{k=1}^{K} \|\theta_h^{k-1} - \theta_h^k\|_2 \le B_T, \quad \sum_{h=1}^{H} \sum_{k=1}^{K} \|\xi_h^{k-1} - \xi_h^k\|_2 \le B_P, \quad \Delta = B_T + B_P, \tag{2.5}$$

where H is the length of each episode, K is the total number of episodes, and $T = HK$ is the total number of steps taken by the agent. Here Δ is the total variation budget, which quantifies the non-stationarity of a linear kernel MDP.

3 MINIMAX LOWER BOUND

In this section, we provide the information-theoretical lower bound result. The following theorem shows a minimax lower bound of dynamic regret for any algorithm to learn non-stationary linear kernel MDPs.

Theorem 3.1 (Minimax lower bound)**.** Fix $\Delta > 0$, $H > 0$, $d \geq 2$, and $T = \Omega(d^{5/2}\Delta H^{1/2})$. Then, there exists a non-stationary linear kernel MDP with a d-dimensional feature map and maximum total variation budget Δ, such that,

$$\min_{\mathbb{A}} \max_{\pi^*} \text{D-Regret}(T, \pi^*) \geq \Omega(d^{5/6}\Delta^{1/3}H^{2/3}T^{2/3}),$$

where $\mathbb{A}$ denotes the learning algorithm.

Proof sketch. As mentioned above, we only need to establish the lower bound of the dynamic regret when the benchmark policy is the optimal policy of each individual episode. The proof of lower bound relies on the construction of a hard-to-learn non-stationary linear kernel MDP instance. To handle the non-stationarity, we need to divide the total T steps into L segments, where each segment contains $T_0 = \lfloor \frac{T}{L} \rfloor$ steps and has $K_0 = \lfloor \frac{K}{L} \rfloor$ episodes. Within each segment, the construction of MDP is similar to the hard-to-learn instance constructed in stationary RL problems (Jaksch et al., 2010; Lattimore & Hutter, 2012; Osband & Van Roy, 2016). Then, we can derive a lower bound of $\Omega(dH\sqrt{T_0})$ for the stationary RL problem. Meanwhile, the transition kernel of this hard-to-learn MDP changes abruptly between two consecutive segments, which forces the agent to learn a new stationary MDP in each segment. Finally, by optimization L subject to the total budget constraint, we obtain the lower bound of $\Omega(d^{5/6}\Delta^{1/3}H^{2/3}T^{2/3})$. See Appendix C for details. $\square$

4 ALGORITHM AND THEORY

4.1 PROPO

Now we present Periodically Restarted Optimistic Policy Optimization (PROPO) in Algorithm 1, which includes a policy improvement step and a policy evaluation step.

Policy Improvement Step. At k-th episode, Model-Based OPPO updates $\pi^k = \{\pi_h^k\}_{h=1}^H$ according to $\pi^{k-1} = \{\pi_h^{k-1}\}_{h=1}^H$. Motivated by the policy improvement step in NPG (Kakade, 2002), TRPO (Schulman et al., 2015), and PPO (Schulman et al., 2017), we consider the following policy improvement step

$$\pi^k = \operatorname*{argmax}_{\pi} L_{k-1}(\pi), \tag{4.1}$$

where $L_{k-1}(\pi)$ is defined as

$$\begin{aligned} L_{k-1}(\pi) = \mathbb{E}_{\pi^{k-1}}\Bigg[\sum_{h=1}^{H} \langle Q_h^{k-1}(s_h, \cdot), \pi_h(\cdot \,|\, s_h) - \pi_h^{k-1}(\cdot \,|\, s_h)\rangle\Bigg] \qquad (4.2) \\ - \alpha^{-1} \cdot \mathbb{E}_{\pi^{k-1}}\Bigg[\sum_{h=1}^{H} \mathrm{KL}\big(\pi_h(\cdot \,|\, s_h) \,\big\|\, \pi_h^{k-1}(\cdot \,|\, s_h)\big)\Bigg], \end{aligned}$$

where $\alpha > 0$ is a stepsize and Q_h^{k-1} which is obtained in Line 10 of Algorithm 2 is the estimator of $Q_h^{\pi^{k-1},k-1}$. Here the expectation $\mathbb{E}_{\pi^{k-1}}$ is taken over the random state-action pairs $\{(s_h, a_h)\}_{h=1}^H$, where the initial state $s_1 = s_1^k$, the distribution of action a_h follows $\pi(\cdot \,|\, s_h)$, and the distribution of the next state s_{h+1} follows the transition dynamics $P_h^k(\cdot \,|\, s_h, a_h)$. Such a policy improvement step can also be regarded as one iteration of infinite-dimensional mirror descent (Nemirovsky & Yudin, 1983; Liu et al., 2019; Wang et al., 2019a).

By the optimality condition, policy update in (4.1) admits a closed-form solution

$$\pi_h^k(\cdot \,|\, s) \propto \pi_h^{k-1}(\cdot \,|\, s) \cdot \exp\{\alpha \cdot Q_h^{k-1}(s, \cdot)\} \tag{4.3}$$

for any $s \in \mathcal{S}$ and $(k, h) \in [K] \times [H]$.

Policy Evaluation Step. At the end of the k-th episode, Model-Based OPPO evaluates the policy π^k based on the $(k-1)$ historical trajectories. Then, we show the details of estimating the reward functions and transition kernels, respectively.

(i) Estimating Reward. To estimate the reward functions, we use the sliding window regularized least squares estimator (SW-RLSE) (Garivier & Moulines, 2011; Cheung et al., 2019a;b), which is a key tool in estimating the unknown parameters online. At h-th step of k-th episode, we aim to estimate the unknown parameter θ_h^k based on the historical observation $\{(s_h^\tau, a_h^\tau), r_h^\tau(s_h^\tau, a_h^\tau)\}_{\tau=1}^{k-1}$. The design of SW-RLSE is based on the "forgetting principle" (Garivier & Moulines, 2011), that is, under non-stationarity, the historical observations far in the past are obsolete, and they do not contain relevant information for the evaluation of the current policy. Therefore, we could estimate θ_h^k using only $\{(s_h^\tau, a_h^\tau), r_h^\tau(s_h^\tau, a_h^\tau)\}_{\tau=1\vee(k-w)}^{k-1}$, the observations during the sliding window $1 \vee (k-w)$ to $k-1$,

$$\widehat{\theta}_h^k = \underset{\theta}{\mathrm{argmin}}\bigg(\sum_{\tau=1\vee(k-w)}^{k-1}\big(r_h^\tau(s_h^\tau, a_h^\tau) - \phi(s_h^k, a_h^k)^\top\theta\big)^2 + \lambda\cdot\|\theta\|_2^2\bigg), \tag{4.4}$$

where λ is the regularization parameter and w is the length of a sliding window. By solving (4.4), we obtain the estimator of θ_h^k:

$$\widehat{\theta}_h^k = (\Lambda_h^k)^{-1}\sum_{\tau=1\vee(k-w)}^{k-1}\phi(s_h^\tau, a_h^\tau)r_h^\tau(s_h^\tau, a_h^\tau), \tag{4.5}$$
$$\text{where } \Lambda_h^k = \sum_{\tau=1\vee(k-w)}^{k-1}\phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top + \lambda I_d.$$

(ii) Estimating Transition. Similar to the estimation of reward functions, for any $(k, h) \in [K] \times [H]$, we define the sliding window empirical mean-squared Bellman error (SW-MSBE) as

$$M_h^k(\xi) = \sum_{\tau=1\vee(k-w)}^{k-1}\big(V_{h+1}^\tau(s_{h+1}^\tau) - \eta_h^\tau(s_h^\tau, a_h^\tau)^\top\xi\big)^2,$$

where we denote $\eta_h^\tau(\cdot,\cdot)$ as

$$\eta_h^\tau(\cdot,\cdot) = \int_{\mathcal{S}}\psi(\cdot,\cdot,s')\cdot V_{h+1}^\tau(s')\mathrm{d}s'. \tag{4.6}$$

By Assumption 2.1, we have

$$\|\eta_h^k(\cdot,\cdot)\|_2 \le H\sqrt{d}$$

for any $(k, h) \in [K] \times [H]$. Then we estimate ξ_h^k by solving the following problem:

$$\widehat{\xi}_h^k = \underset{w\in\mathbb{R}^d}{\mathrm{argmin}}(M_h^k(w) + \lambda'\cdot\|w\|_2^2), \tag{4.7}$$

where λ' is the regularization parameter. By solving (4.7), we obtain

$$\widehat{\xi}_h^k = (A_h^k)^{-1}\bigg(\sum_{\tau=1\vee(k-w)}^{k-1}\eta_h^\tau(s_h^\tau, a_h^\tau)\cdot V_{h+1}^\tau(s_{h+1}^\tau)\bigg) \tag{4.8}$$
$$\text{where } A_h^k = \sum_{\tau=1\vee(k-w)}^{k-1}\eta_h^\tau(s_h^\tau, a_h^\tau)\eta_h^\tau(s_h^\tau, a_h^\tau)^\top + \lambda' I_d.$$

The policy evaluation step is iteratively updating the estimated Q-function $Q^k = \{Q_h^k\}_{h=1}^H$ by

$$Q_h^k(\cdot,\cdot) = \min\{\phi(\cdot,\cdot)^\top\widehat{\theta}_h^k + \eta_h^k(\cdot,\cdot)^\top\widehat{\xi}_h^k + B_h^k(\cdot,\cdot) + \Gamma_h^k(\cdot,\cdot), H-h+1\}^+, \tag{4.9}$$
$$V_h^k(s) = \langle Q_h^k(s,\cdot), \pi_h^k(\cdot|s)\rangle_{\mathcal{A}}$$

in the order of $h = H, H-1, \cdots, 1$. Here bonus functions $B_h^k(\cdot,\cdot) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^+$ and $\Gamma_h^k(\cdot,\cdot) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^+$ are used to quantify the uncertainty in estimating reward r_h^k and quantity $\mathbb{P}_h^k V_{h+1}^k$ respectively, defined as

$$B_h^k(\cdot,\cdot) = \beta\big(\phi(\cdot,\cdot)^\top (\Lambda_h^k)^{-1}\phi(\cdot,\cdot)\big)^{1/2}, \quad \Gamma_h^k(\cdot,\cdot) = \beta'\big(\eta_h^k(\cdot,\cdot)^\top (A_h^k)^{-1}\eta_h^k(\cdot,\cdot)\big)^{1/2}, \tag{4.10}$$

where $\beta > 0$ and $\beta' > 0$ are parameters depend on d, H and K, which are specified in Theorem 4.2.

To handle the non-stationary drift incurred by the different optimal policies in different episodes, Algorithm 1 also includes a periodic restart mechanism, which resets the policy estimates every τ episodes. We call the τ episodes between every two resets a segment. In each segment, each episode is approximately the same as the first episode, which means that we can regard it as a stationary MDP. Then we can use the method of solving the stationary MDP to analyze each segment with a small error, and finally combine each segment and choose the value of τ to get the desired result. Such a restart mechanism is widely used in RL (Auer et al., 2009; Ortner et al., 2020), bandits (Besbes et al., 2014; Zhao et al., 2020), and non-stationary optimization (Besbes et al., 2015; Jadbabaie et al., 2015).

The pseudocode of the PROPO algorithm is given in Algorithm 1.

Algorithm 1 Periodically Restarted Optimistic Policy Optimization (PROPO)

Require: Reset cycle length τ, sliding window length w, stepsize α, regularization factors λ and λ', and bonus multipliers β and β'.

Initialize $\{\pi_h^0(\cdot\,|\,\cdot)\}_{h=1}^H$ as uniform distribution policies, $\{Q_h^0(\cdot,\cdot)\}_{h=1}^H$ as zero functions.
for $k = 1, 2, \ldots, K$ **do**
  Receive the initial state s_1^k.
  if $k \mod \tau = 1$ **then**
    Set $\{Q_h^{k-1}\}_{h\in[H]}$ as zero functions and $\{\pi_h^{k-1}\}_{h\in[H]}$ as uniform distribution on $\mathcal{A}$.
  end if
  for $h = 1, 2, \ldots, H$ **do**
    $\pi_h^k(\cdot\,|\,\cdot) \propto \pi_h^{k-1}(\cdot\,|\,\cdot) \cdot \exp\{\alpha \cdot Q_h^{k-1}(\cdot,\cdot)\}$.
    Take action $a_h^k \sim \pi_h^k(\cdot\,|\,s_h^k)$.
    Observe the reward $r_h^k(s_h^k, a_h^k)$ and receive the next state s_{h+1}^k.
  end for
  Compute Q_h^k by SWOPE($k, \{\pi_h^k\}, \lambda, \lambda', \beta, \beta'$) (Algorithm 2).
end for

Algorithm 2 Sliding Window Optimistic Policy Evaluation (SWOPE)

Require: Episode index k, policies $\{\pi_h\}$, regularization factors λ and λ', and bonus multipliers β and β'.

Initialize V_{H+1}^k as a zero function.
for $h = H, H-1, \ldots, 0$ **do**
  $\eta_h^k(\cdot,\cdot) = \int_{\mathcal{S}} \psi(\cdot,\cdot,s') \cdot V_{h+1}^k(s')\mathrm{d}s'$.
  $\Lambda_h^k = \sum_{\tau=1\vee(k-w)}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top + \lambda I_d$.
  $\widehat{\theta}_h^k = (\Lambda_h^k)^{-1} \sum_{\tau=1\vee(k-w)}^{k-1} \phi(s_h^\tau, a_h^\tau) r_h^\tau(s_h^\tau, a_h^\tau)$.
  $A_h^k = \sum_{\tau=1\vee(k-w)}^{k-1} \eta_h^\tau(s_h^\tau, a_h^\tau)\eta_h^\tau(s_h^\tau, a_h^\tau)^\top + \lambda' I_d$..
  $\widehat{\xi}_h^k = (A_h^k)^{-1}\big(\sum_{\tau=1\vee(k-w)}^{k-1} \eta_h^\tau(s_h^\tau, a_h^\tau) \cdot V_{h+1}^\tau(s_{h+1}^\tau)\big)$.
  $B_h^k(\cdot,\cdot) = \beta\big(\phi(\cdot,\cdot)^\top(\Lambda_h^k)^{-1}\phi(\cdot,\cdot)\big)^{1/2}$.
  $\Gamma_h^k(\cdot,\cdot) = \beta'\big(\eta_h^k(\cdot,\cdot)^\top(A_h^k)^{-1}\eta_h^k(\cdot,\cdot)\big)^{1/2}$.
  $Q_h^k(\cdot,\cdot) = \min\{\phi(\cdot,\cdot)^\top\widehat{\theta}_h^k + \eta_h^k(\cdot,\cdot)^\top\widehat{\xi}_h^k + B_h^k(\cdot,\cdot) + \Gamma_h^k(\cdot,\cdot), H-h+1\}^+$.
  $V_h^k(s) = \langle Q_h^k(s,\cdot), \pi_h^k(\cdot|s)\rangle_{\mathcal{A}}$.
end for

4.2 SW-LSVI-UCB

In this subsection, we present the details of Sliding Window Least-Square Value Iteration with UCB (SW-LSVI-UCB) in Algorithm 3 (cf. Appendix B).

Similar to Least-Square Value Iteration with UCB (LSVI-UCB) in Jin et al. (2019b), SW-LSVI-UCB is also an optimistic modification of Least-Square Value Iteration (LSVI) (Bradtke & Barto, 1996), where the optimism is realized by Upper-Confidence Bounds (UCB). Specifically, the optimism is achieved due to the bonus functions B_h^k and Γ_h^k, which quantify the uncertainty of reward functions and transition kernels, respectively. It is worth noting that in order to handle the non-stationarity, SW-LSVI-UCB also uses the sliding window method (Garivier & Moulines, 2011; Cheung et al., 2019a;b).

In detail, at k-th episode, SW-LSVI-UCB consists of two steps. In the first step, by solving the sliding window least-square problems (4.4) and (4.7), SW-LSVI-UCB updates the parameters Λ_h^k in (4.5), $\widehat{\theta}_h^k$ in (4.5), A_h^k in (4.8), and $\widehat{\xi}_h^k$ in (4.8), which are used to form the Q-function Q_h^k . In the second step, SW-LSVI-UVB obtains the greedy policy with respect to the Q-function Q_h^k gained in the first step. See Algorithm 3 in Appendix B for more details.

4.3 Regret Analysis

In this subsection, we analyze the dynamic regret incurred by Algorithms 1 and 3 and compare the theoretical regret upper bounds derived for these two algorithms.

To derive sharp dynamic regret bounds, we impose the following technical assumption.

Assumption 4.1. There exists an orthonormal basis $\Psi = (\Psi_1, \cdots, \Psi_d)$ such that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, there exists a vector $z \in \mathbb{R}^d$ satisfying that $\phi(s,a) = \Psi z$. We also assume the existence of another orthonormal basis $\Psi' = (\Psi_1', \cdots, \Psi_d')$ such that for any $(s,a,k,h) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H]$ such that $\eta_h^k(s,a) = \Psi' z'$ for some $z' \in \mathbb{R}^d$.

It is not difficult to show that this assumption holds in the tabular setting. Similar assumption is also adopted by previous work in non-stationary optimization (Cheung et al., 2019a). We will provide more comments on this technical assumption after showing main results.

First, we establish an upper bound on the dynamic regret of PROPO. Recall that the dynamic regret is defined in (2.3) and d is the dimension of the feature maps ϕ and ψ. Also, $|\mathcal{A}|$ is the cardinality of $\mathcal{A}$. We also define $\rho = \lceil K/\tau \rceil$ to be the number of restarts that take place in Algorithm 1.

Theorem 4.2 (Upper bound for Algorithm 1)**.** Suppose Assumptions 2.1 and 4.1 hold. Let $\tau = \Pi_{[1,K]}(\lfloor (\frac{T\sqrt{\log|\mathcal{A}|}}{H(P_T+\sqrt{d}\Delta)})^{2/3} \rfloor)$, $\alpha = \sqrt{\rho \log|\mathcal{A}|/(H^2K)}$ in (4.2), $w = \Theta(d^{1/3}\Delta^{-2/3}T^{2/3})$ in (4.4), $\lambda = \lambda' = 1$ in (4.4) and (4.9), $\beta = \sqrt{d}$ in (4.10), and $\beta' = C'\sqrt{dH^2 \cdot \log(dT/\zeta)}$ in (4.10), where $C' > 1$ is an absolute constant and $\zeta \in (0,1]$. We have

$$\text{D-Regret}(T) \lesssim d^{5/6}\Delta^{1/3}HT^{2/3} \cdot \log(dT/\zeta) + \begin{cases} \sqrt{H^3T\log|\mathcal{A}|}, & \text{if } 0 \le P_T + \sqrt{d}\Delta \le \sqrt{\frac{\log|\mathcal{A}|}{K}}, \\ (H^2T\sqrt{\log|\mathcal{A}|})^{2/3}(P_T+\sqrt{d}\Delta)^{1/3}, & \text{if } \sqrt{\frac{\log|\mathcal{A}|}{K}} \le P_T + \sqrt{d}\Delta \lesssim K\sqrt{\log|\mathcal{A}|}, \\ H^2(P_T+\sqrt{d}\Delta), & \text{if } P_T + \sqrt{d}\Delta \gtrsim K\sqrt{\log|\mathcal{A}|}, \end{cases}$$

with probability at least $1-\zeta$.

Proof. See Appendix D for a proof sketch and Appendix G for a detailed proof. □

Then we discuss the regret bound throughout three regimes of $P_T + \sqrt{d}\Delta$:

- Small $P_T + \sqrt{d}\Delta$: when $0 \le P_T + \sqrt{d}\Delta \le \sqrt{\frac{\log|\mathcal{A}|}{K}}$, restart period $\tau = K$, which means that we do not need to periodically restart in this case. Assuming that $\log|\mathcal{A}| = \mathcal{O}(d^{5/3}\Delta^{2/3}H^{-1}T^{1/3})$, Algorithm 1 attains a $\widetilde{\mathcal{O}}(d^{5/6}\Delta^{1/3}HT^{2/3})$ dynamic regret. Combined with the lower bound established in Theorem 3.1, our result matches the lower bound in d, Δ and T up to logarithmic factors. Hence, we can conclude that Algorithm 1 is a near-optimal algorithm;

- Moderate $P_T + \sqrt{d}\Delta$: when $\sqrt{\frac{\log|\mathcal{A}|}{K}} \leq P_T + \sqrt{d}\Delta \lesssim K\sqrt{\log|\mathcal{A}|}$, restart period $\tau = (\frac{T\sqrt{\log|\mathcal{A}|}}{H(P_T+\sqrt{d}\Delta)})^{2/3} \in [2, K]$. Algorithm 2 incurs a $\widetilde{\mathcal{O}}(T^{2/3})$ dynamic regret if $\Delta = \mathcal{O}(1)$ and $P_T = \mathcal{O}(1)$;
- Large $P_T + \sqrt{d}\Delta$: when $P_T + \sqrt{d}\Delta \gtrsim K\sqrt{\log|\mathcal{A}|}$, restart period $\tau = K$. Since the model is highly non-stationary, we only obtain a linear regret in T.

In the following theorem, we establish the upper bound of dynamic regret incurred by SW-LSVI-UCB (Algorithm 3).

Theorem 4.3 (Upper bound for Algorithm 3)**.** Suppose Assumption 2.1 and 4.1 hold. Let $w = \Theta(d^{1/3}\Delta^{-2/3}T^{2/3})$ in (4.4), $\lambda = \lambda' = 1$ in (4.4) and (4.9), $\beta = \sqrt{d}$ in (4.10), and $\beta' = C'\sqrt{dH^2 \cdot \log(dT/\zeta)}$ in (4.10), where $C' > 1$ is an absolute constant and $\zeta \in (0, 1]$. We have

$$\text{D-Regret}(T) \lesssim d^{5/6}\Delta^{1/3}HT^{2/3} \cdot \log(dT/\zeta)$$

with probability at least $1 - \zeta$.

Proof. See Appendix H for a detailed proof. □

Regarding Assumption 4.1. Due to some technical issue (Touati & Vincent, 2020; Zhao & Zhang, 2021), without this assumption and the knowledge of locally variation budget (Touati & Vincent, 2020), previous work can only obtain the bound $\widetilde{\mathcal{O}}(T^{3/4})$ (Cheung et al., 2020; Zhao & Zhang, 2021; Zhao et al., 2020; Russac et al., 2019; Zhou et al., 2020b; Touati & Vincent, 2020). Thanks to Assumption 4.1, we derive sharper regret bounds at the order $\widetilde{\mathcal{O}}(T^{2/3})$. We also remark that we can establish slightly worse regret bounds for Algorithms 1 and 3 without Assumption 4.1. See Appendix I for details.

Optimality of the Bounds. Notably, the term $\widetilde{\mathcal{O}}(d^{5/6}\Delta^{1/3}HT^{2/3})$ appears in both the results in Theorems 4.2 and 4.3. Ignoring logarithmic factors, there is only a gap of $H^{1/3}$ between this upper bound and the lower bound $\Omega(d^{5/6}\Delta^{1/3}H^{2/3}T^{2/3})$ established in Theorem 3.1. We conjecture that this gap can be bridged by using the "Bernstein" type bonus functions Azar et al. (2017); Jin et al. (2018). Since our focus is on designing a provably efficient policy optimization algorithm for non-stationary linear kernel MDPs, we don't use this technique for the clarity of our analysis.

Comparison. Compared with PROPO, SW-LSVI-UCB achieves a slightly better regret without the help of the periodic restart mechanism. Especially in the highly non-stationary case, that is $P_T + \sqrt{d}\Delta \gtrsim K\sqrt{\log|\mathcal{A}|}$, SW-LSVI-UCB achieves a $\widetilde{\mathcal{O}}(T^{2/3})$ regret, where PROPO only attains a linear regret in T. However, PROPO achieves the same $\widetilde{\mathcal{O}}(T^{2/3})$ regret as SW-LSVI-UCB when $P_T + \sqrt{d}\Delta \lesssim K\sqrt{\log|\mathcal{A}|}$, which suggests that PROPO is provably efficient for solving slightly or even moderately non-stationary MDPs. Therefore, it is important to investigate whether it is possible to bridge this gap between policy and value based methods, or alternatively to show that this gap is actually a true drawback of policy optimization methods in the non-stationary case.

5 Conclusion

In this work, we have proposed a probably efficient policy optimization algorithm, dubbed as PROPO, for non-stationary linear kernel MDPs. Such an algorithm incorporates a bonus function to incentivize exploration, and more importantly, adopts sliding-window-based regression in policy evaluation and periodic restart in policy update to handle the challenge of non-stationarity. Moreover, as a byproduct, we establish an optimistic value iteration algorithm, SW-LSVI-UCB, by combining UCB and sliding-window. We prove that PROPO and SW-LSVI-UCB both achieve sample efficiency by having sublinear dynamic regret. We also establish a dynamic regret lower bound which shows that PROPO and SW-LSVI-UCB are near-optimal. To our best knowledge, we propose the first provably efficient policy optimization method that successfully handles non-stationarity.

REFERENCES

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

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *arXiv preprint arXiv:1908.00261*, 2019.

Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. PC-PG: Policy cover directed exploration for provable policy gradient learning. *arXiv preprint arXiv:2007.08459*, 2020.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pp. 127–135, 2013.

Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in neural information processing systems*, pp. 89–96, 2009.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin F Yang. Model-based reinforcement learning with value-targeted regression. *arXiv preprint arXiv:2006.01107*, 2020.

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

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in neural information processing systems*, pp. 199–207, 2014.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Optimal exploration–exploitation in a multi-armed bandit problem with non-stationary rewards. *Stochastic Systems*, 9(4):319–337, 2019.

Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1-3):33–57, 1996.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *arXiv preprint arXiv:2007.06558*, 2020.

Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Conference on Learning Theory*, pp. 696–726. PMLR, 2019.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Hedging the drift: Learning to optimize under non-stationarity. *arXiv preprint arXiv:1903.01461*, 2019a.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1079–1087, 2019b.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Non-stationary reinforcement learning: The blessing of (more) optimism. *Available at SSRN 3397818*, 2019c.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism. *arXiv preprint arXiv:2006.14389*, 2020.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. A kernel-based approach to non-stationary reinforcement learning in metric spaces. *arXiv preprint arXiv:2007.05078*, 2020.

Yonathan Efroni, Lior Shani, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. *arXiv preprint arXiv:2002.08243*, 2020.

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

Yingjie Fei, Zhuoran Yang, Zhaoran Wang, and Qiaomin Xie. Dynamic regret of policy optimization in non-stationary environments. *arXiv preprint arXiv:2007.00148*, 2020.

Pratik Gajane, Ronald Ortner, and Peter Auer. A sliding-window algorithm for markov decision processes with arbitrarily changing rewards and transitions. *arXiv preprint arXiv:1805.10066*, 2018.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pp. 174–188. Springer, 2011.

Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 3389–3396. IEEE, 2017.

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017.

Ali Jadbabaie, Alexander Rakhlin, Shahin Shahrampour, and Karthik Sridharan. Online optimization: Competing with dynamic comparators. In *Artificial Intelligence and Statistics*, pp. 398–406, 2015.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010.

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

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pp. 4863–4873, 2018.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial mdps with bandit feedback and unknown transition. *arXiv preprint arXiv:1912.01192*, 2019a.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*, 2019b.

Sham M Kakade. A natural policy gradient. In *Advances in neural information processing systems*, pp. 1531–1538, 2002.

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

Tor Lattimore and Marcus Hutter. Pac bounds for discounted mdps. In *International Conference on Algorithmic Learning Theory*, pp. 320–334. Springer, 2012.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural trust region/proximal policy optimization attains globally optimal policy. In *Advances in Neural Information Processing Systems*, pp. 10565–10576, 2019.

Weichao Mao, Kaiqing Zhang, Ruihao Zhu, David Simchi-Levi, and Tamer Başar. Near-optimal regret bounds for model-free rl in non-stationary episodic mdps, 2020.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pp. 6820–6829. PMLR, 2020.

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

Arkadiĭ Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

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

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

Ronald Ortner, Pratik Gajane, and Peter Auer. Variational regret bounds for reinforcement learning. In *Uncertainty in Artificial Intelligence*, pp. 81–90. PMLR, 2020.

Ian Osband and Benjamin Van Roy. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

Anay Pattanaik, Zhenyi Tang, Shuijing Liu, Gautham Bommannan, and Girish Chowdhary. Robust deep reinforcement learning with adversarial attacks. *arXiv preprint arXiv:1712.03632*, 2017.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pp. 2817–2826. PMLR, 2017.

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

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems*, pp. 12017–12026, 2019.

Ahmad EL Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017(19):70–76, 2017.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5668–5675, 2020.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

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

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Ahmed Touati and Pascal Vincent. Efficient learning in non-stationary linear markov decision processes. *arXiv preprint arXiv:2010.12870*, 2020.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019a.

Ruosong Wang, Ruslan Salakhutdinov, and Lin F Yang. Provably efficient reinforcement learning with general value function approximation. *arXiv preprint arXiv:2005.10804*, 2020.

Yining Wang, Ruosong Wang, Simon S Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019b.

Lin F Yang and Mengdi Wang. Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389*, 2019a.

Lin F Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. *arXiv preprint arXiv:1902.04779*, 2019b.

Zhuoran Yang, Yongxin Chen, Mingyi Hong, and Zhaoran Wang. On the global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. *arXiv preprint arXiv:1907.06246*, 2019.

Andrew Chi-Chin Yao. Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)*, pp. 222–227. IEEE, 1977.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020.

Peng Zhao and Lijun Zhang. Non-stationary linear bandits revisited. *arXiv preprint arXiv:2103.05324*, 2021.

Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. A simple approach for non-stationary linear bandits. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 2020, 2020.

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted MDPs with feature mapping. *arXiv preprint arXiv:2006.13165*, 2020a.

Huozhi Zhou, Jinglin Chen, Lav R Varshney, and Ashish Jagmohan. Nonstationary reinforcement learning with linear function approximation. *arXiv preprint arXiv:2010.04244*, 2020b.

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

A NOTATION

We denote by $\|\cdot\|_2$ the ℓ_2-norm of a vector or the spectral norm of a matrix. Furthermore, for a positive definite matrix A, we denote by $\|x\|_A$ the matrix norm $\sqrt{x^\top Ax}$ of a vector x. For any number a, we denote $\lceil a \rceil$ the smallest integer that is no smaller than a, and $\lfloor a \rfloor$ the largest integer no larger than a. Also, for any two numbers a and b, let $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For some positive integer K, $[K]$ denotes the index set $\{1, 2, \cdots, K\}$. When logarithmic factors are omitted, we use $\widetilde{O}$ to denote function growth.

B PSEUDOCODE OF SW-LSVI-UCB

Algorithm 3 Sliding Window Least-Square Value Iteration with UCB (SW-LSVI-UCB)

Require: Sliding window length w, stepsize α, regularization factors λ and λ', and bonus multipliers β and β'.

Initialize $\{\pi_h^0(\cdot \mid \cdot)\}_{h=1}^H$ as uniform distribution policies, $\{Q_h^0(\cdot,\cdot)\}_{h=1}^H$ as zero functions.
for $k = 1, 2, \dots, K$ **do**
  Receive the initial state s_1^k.
  Initialize V_{H+1}^k as a zero function.
  for $h = H, H-1, \dots, 0$ **do**
    $\eta_h^k(\cdot,\cdot) = \int_{\mathcal{S}} \psi(\cdot,\cdot,s') \cdot V_{h+1}^k(s')\mathrm{d}s'$.
    $\Lambda_h^k = \sum_{\tau=1\vee(k-w)}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top + \lambda I_d$.
    $\widehat{\theta}_h^k = (\Lambda_h^k)^{-1} \sum_{\tau=1\vee(k-w)}^{k-1} \phi(s_h^\tau, a_h^\tau) r_h^\tau(s_h^\tau, a_h^\tau)$.
    $A_h^k = \sum_{\tau=1\vee(k-w)}^{k-1} \eta_h^\tau(s_h^\tau, a_h^\tau)\eta_h^\tau(s_h^\tau, a_h^\tau)^\top + \lambda' I_d$..
    $\widehat{\xi}_h^k = (A_h^k)^{-1}\big(\sum_{\tau=1\vee(k-w)}^{k-1} \eta_h^\tau(s_h^\tau, a_h^\tau) \cdot V_{h+1}^\tau(s_{h+1}^\tau)\big)$.
    $B_h^k(\cdot,\cdot) = \beta\big(\phi(\cdot,\cdot)^\top(\Lambda_h^k)^{-1}\phi(\cdot,\cdot)\big)^{1/2}$.
    $\Gamma_h^k(\cdot,\cdot) = \beta'\big(\eta_h^k(\cdot,\cdot)^\top(A_h^k)^{-1}\eta_h^k(\cdot,\cdot)\big)^{1/2}$.
    $Q_h^k(\cdot,\cdot) = \min\{\phi(\cdot,\cdot)^\top\widehat{\theta}_h^k + \eta_h^k(\cdot,\cdot)^\top\widehat{\xi}_h^k + B_h^k(\cdot,\cdot) + \Gamma_h^k(\cdot,\cdot), H-h+1\}^+$.
    $V_h^k(s) = \max_a Q_h^k(s,a)$.
    $\pi_h^k(s) = \mathrm{argmax}_a\, Q_h^k(s,a)$.
  end for
end for

C PROOF OF THEOREM 3.1

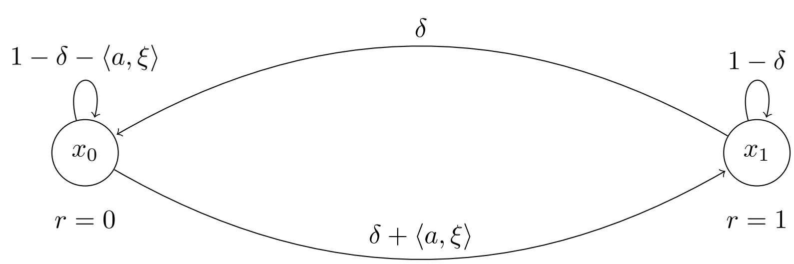

Figure 1: The hard-to-learn linear kernel MDP constructed in the proof of Theorem 3.1. Note that the probability of state x_0 to state x_1 depends on the choice of action a.

Proof. To handle the non-stationarity, we divide the total T steps into L segments, where each segment has $K_0 = \lfloor \frac{K}{L} \rfloor$ episodes and contains $T_0 = HK_0 = H\lfloor \frac{K}{L} \rfloor$ steps. Now we show the construction of a hard-to-learn MDP within each segment, the construction is similar to that used in previous works (Jaksch et al., 2010; Lattimore & Hutter, 2012; Osband & Van Roy, 2016; Zhou et al., 2020a). Consider an MDP as depicted in Figure 1. The state space $\mathcal{S}$ consists of two states x_0 and x_1. The action space $\mathcal{A}$ consists of 2^{d-1} vectors $a \in \{-1, 1\}^{d-1}$, where $d \geq 2$ is the dimension of feature map ψ defined in Assumption 2.1. The reward function does not depend on actions: state x_0 always gives reward 0, and state x_1 always gives reward 1, that is, for any $a \in \mathcal{A}$,

$$r(x_0, a) = 0, \qquad r(x_1, a) = 1.$$

Choosing,

$$\theta = (1/d, 1/d, \cdots, 1/d)^\top \in \mathbb{R}^d, \ \phi(x_0, a) = (0, 0, \cdots, 0)^\top \in \mathbb{R}^d, \phi(x_1, a) = (1, 1, \cdots, 1)^\top \in \mathbb{R}^d,$$

for any $a \in \mathcal{A}$, it follows that $r(s, a) = \phi(s, a)^\top \theta$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, and thus this reward is indeed linear. The probability transition P_ξ is parameterized by a $(d-1)$-dimensional vector $\xi \in \Xi = \{-\epsilon/(d-1), \epsilon/(d-1)\}^{d-1}$, which is defined as

$$\begin{aligned} P_\xi(x_0 \,|\, x_0, a) &= 1 - \delta - \langle a, \xi \rangle, & P_\xi(x_1 \,|\, x_0, a) &= \delta + \langle a, \xi \rangle, \\ P_\xi(x_0 \,|\, x_1, a) &= \delta, & P_\xi(x_1 \,|\, x_1, a) &= 1 - \delta, \end{aligned}$$

where $\delta > 0$ and $\epsilon \in [0, d-1]$ are parameters which satisfy that $2\epsilon \leq \delta \leq 1/3$. This MDP is indeed a linear kernel MDP with the d-dimensional vector $\widetilde{\xi} = (\xi^\top, 1)^\top$. Specifically, we can define the feature map $\psi(s, a, s')$ as

$$\begin{aligned} \psi(x_0, a, x_0) &= (-a^\top, 1 - \delta)^\top, & \psi(x_0, a, x_1) &= (a^\top, \delta)^\top, \\ \psi(x_1, a, x_0) &= (\mathbf{0}^\top, \delta)^\top, & \psi(x_1, a, x_1) &= (\mathbf{0}^\top, 1 - \delta)^\top, \end{aligned}$$

and it is not difficult to verify that $P_\xi(s' \,|\, s, a) = \psi(s, a, s')^\top \widetilde{\xi}$.

Now we are ready to establish the lower bound in Theorem 3.1. By Yao's minimax principle (Yao, 1977), it is sufficient to consider deterministic policies. Hence, we assume that the policy π obtained by the algorithm maps from a sequence of observations to an action deterministically. To facilitate the following proof, we introduce some notations. Let N_0, N_1, N_0^a and $N_0^{\mathcal{A}'}$ denote the total number of visits to state x_0, the total number of visits to x_1, the total number of visits to state x_0 followed by taking action a, and the total number of visits to state x_0 followed by taking an action in $\mathcal{A}' \subseteq \mathcal{A}$, respectively. Let $\mathcal{P}_\xi(\cdot)$ denote the distribution over $\mathcal{S}^{T_0}$, where $s_1^k = x_0$, $s_{h+1}^k \sim P_\xi(\cdot \,|\, s_h^k, a_h^k)$, a_h^k is decided by π_h^k. We use $\mathbb{E}_\xi$ to denote the expectation with respect to $\mathcal{P}_\xi$.

Now we consider a segment that consists of K_0 episodes and each episode starts from state x_0. Let s_h^k denote the state in the h-th state of the k-th episode. Fix $\xi \in \Xi$. We have,

$$\begin{aligned} \mathbb{E}_\xi N_1 &= \sum_{k=1}^{K_0} \sum_{h=2}^{H} \mathcal{P}_\xi(s_h^k = x_1) = \sum_{k=1}^{K_0} \sum_{h=2}^{H} \mathcal{P}_\xi(s_h^k = x_1, s_{h-1}^k = x_1) + \sum_{k=1}^{K_0} \sum_{h=2}^{H} \mathcal{P}_\xi(s_h^k = x_1, s_{h-1}^k = x_0) \\ &= \underbrace{\sum_{k=1}^{K_0} \sum_{h=2}^{H} \mathcal{P}_\xi(s_h^k = x_1 \,|\, s_{h-1}^k = x_1) \mathcal{P}_\xi(s_{h-1}^k = x_1)}_{\text{(i)}} + \underbrace{\sum_{k=1}^{K_0} \sum_{h=2}^{H} \mathcal{P}_\xi(s_h^k = x_1, s_{h-1}^k = x_0)}_{\text{(ii)}}. \end{aligned} \tag{C.1}$$

By the construction of this hard-to-learn MDP, we have $\mathcal{P}_\xi(s_h^k = x_1 \,|\, s_{h-1}^k = x_1) = 1 - \delta$, which implies that

$$\begin{aligned} \text{(i)} &= (1 - \delta) \cdot \sum_{k=1}^{K_0} \sum_{h=2}^{H} \mathcal{P}_\xi(s_{h-1}^k = x_1) \\ &= (1 - \delta) \cdot \mathbb{E}_\xi N_1 - (1 - \delta) \cdot \sum_{k=1}^{K_0} \mathcal{P}_\xi(s_H^k = x_1). \end{aligned} \tag{C.2}$$

Meanwhile, we have

$$\text{(ii)} = \sum_{k=1}^{K_0} \sum_{h=2}^{H} \sum_{a} \mathcal{P}_\xi(s_h^k = x_1 \,|\, s_{h-1}^k = x_0, a_{h-1}^k = a) \cdot \mathcal{P}_\xi(s_{h-1}^k = x_0, a_{h-1}^k = a).$$

By the fact that $\mathcal{P}_\xi(s_h^k = x_1 \mid s_{h-1}^k = x_0, a_{h-1}^k = a) = \delta + \langle a, \xi \rangle$, we further obtain

$$\begin{aligned}
\text{(ii)} &= \sum_{k=1}^{K_0} \sum_{h=2}^{H} \sum_a (\delta + \langle a, \xi \rangle) \cdot \mathcal{P}_\xi(s_{h-1}^k = x_0, a_{h-1}^k = a) \\
&= \sum_a (\delta + \langle a, \xi \rangle) \cdot \big(\mathbb{E}_\xi N_0^a - \sum_{k=1}^{K_0} \mathcal{P}_\xi(s_H^k = x_0, a_H^k = a) \big). \qquad \text{(C.3)}
\end{aligned}$$

Plugging (C.2) and (C.3) into (C.1) and rearranging gives

$$\begin{aligned}
\mathbb{E}_\xi N_1 &= \sum_a (1 + \langle a, \xi \rangle / \delta) \cdot \mathbb{E}_\xi N_0^a - \underbrace{\sum_{k=1}^{K_0} \big(\frac{1-\delta}{\delta} \cdot \mathcal{P}_\xi(s_H^k = x_1) + \sum_a (1 + \frac{\langle a, \xi \rangle}{\delta}) \cdot \mathcal{P}_\xi(s_H^k = x_0, a_H^k = a) \big)}_{\Phi_\xi} \\
&= \mathbb{E}_\xi N_0 + \delta^{-1} \cdot \sum_a \langle a, \xi \rangle \mathbb{E}_\xi N_0^a - \Phi_\xi. \qquad \text{(C.4)}
\end{aligned}$$

By (C.4) and the fact that $\langle a, \xi \rangle \leq \epsilon$, we further have

$$\begin{aligned}
\mathbb{E}_\xi N_1 &= \mathbb{E}_\xi N_0 + \delta^{-1} \cdot \sum_a \langle a, \xi \rangle \cdot \mathbb{E}_\xi N_0^a - \Phi_\xi \\
&\geq \mathbb{E}_\xi N_0 - \frac{\epsilon}{\delta} \cdot \mathbb{E}_\xi N_0 - \sum_{k=1}^{K_0} \big(\frac{1-\delta}{\delta} \mathcal{P}_\xi(s_H^k = x_1) + (1 + \frac{\epsilon}{\delta}) \cdot \mathcal{P}_\xi(s_H^k = x_0) \big) \\
&= (1 - \frac{\epsilon}{\delta}) \cdot \mathbb{E}_\xi N_0 - \sum_{k=1}^{K_0} \big(\frac{1-\delta}{\delta} + \frac{\epsilon + 2\delta - 1}{\delta} \cdot \mathcal{P}_\xi(s_H^k = x_0) \big) \\
&\geq (1 - \frac{\epsilon}{\delta}) \cdot \mathbb{E}_\xi N_0 - \frac{1-\delta}{\delta} \cdot K_0, \qquad \text{(C.5)}
\end{aligned}$$

where the second equality uses the fact that $\mathcal{P}_\xi(s_H^k = x_0) + \mathcal{P}_\xi(s_H^k = x_1) = 1$, and the last inequality holds since $\frac{\epsilon + 2\delta - 1}{\delta} \cdot \mathcal{P}_\xi(s_H^k = x_0)$ is negative. Together with $N_0 + N_1 = T_0$, (C.5) implies that

$$\mathbb{E}_\xi N_0 \leq \frac{T_0 + (1-\delta)/\delta \cdot K_0}{2 - \epsilon/\delta} \leq \frac{2T_0}{3} + \frac{2}{3\delta} K_0,$$

where the last inequality follows from $2\epsilon \leq \delta$ and $\delta > 0$. Meanwhile, note that Φ_ξ is non-negative because $\langle a, \xi \rangle \geq -\epsilon \geq -\delta$. Combined with the fact that $N_0 + N_1 = T_0$, (C.4) and $\Phi_\xi \geq 0$ imply that

$$\mathbb{E}_\xi N_1 \leq T_0/2 + \delta^{-1} \cdot \sum_a \langle a, \xi \rangle \cdot \mathbb{E}_\xi N_0^a / 2. \qquad \text{(C.6)}$$

Hence, we have

$$\begin{aligned}
\frac{1}{|\Xi|} \sum_\xi \mathbb{E}_\xi N_1 &\leq \frac{T_0}{2} + \frac{1}{2\delta |\Xi|} \sum_\xi \sum_a \langle a, \xi \rangle \mathbb{E}_\xi N_0^a \\
&\leq \frac{T_0}{2} + \frac{\epsilon}{2\delta (d-1) |\Xi|} \sum_{j=1}^{d-1} \sum_a \sum_\xi \mathbb{E}_\xi \big(\mathbf{1}\{\mathrm{sgn}(\xi_j) = \mathrm{sgn}(a_j)\} \big) N_0^a, \qquad \text{(C.7)}
\end{aligned}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. Here the last inequality uses the fact that $\langle a, \xi \rangle \leq \frac{\epsilon}{d-1} \sum_{j=1}^{d-1} \mathbf{1}\{\mathrm{sgn}(\xi_j) = \mathrm{sgn}(a_j)\}$ for any $a \in \mathcal{A}$ and $\xi \in \Xi$. Fix $j \in [d-1]$. We define a new vector $g(\xi)$ as

$$g(\xi)_i = \begin{cases} \xi_i, & \text{if } i \neq j, \\ -\xi_i, & \text{if } i = j. \end{cases}$$

Then, for any $a \in \mathcal{A}$ and $\xi \in \Xi$, we have

$$\begin{aligned}
&\mathbb{E}_\xi \mathbf{1}\{\mathrm{sgn}(\xi_j) = \mathrm{sgn}(a_j)\} N_0^a + \mathbb{E}_{g(\xi)} \mathbf{1}\{\mathrm{sgn}(g(\xi)_j) = \mathrm{sgn}(a_j)\} N_0^a \\
&\quad = \mathbb{E}_{g(\xi)} N_0^a + \mathbb{E}_\xi \mathbf{1}\{\mathrm{sgn}(\xi_j) = \mathrm{sgn}(a_j)\} N_0^a - \mathbb{E}_{g(\xi)} \mathbf{1}\{\mathrm{sgn}(\xi_j) = \mathrm{sgn}(a_j)\} N_0^a. \qquad \text{(C.8)}
\end{aligned}$$

Taking summation of (C.8) over a and ξ, and because $g(\xi)$ is uniformly distributed over Ξ when ξ is uniformly distributed over Ξ, we have

$$
\begin{aligned}
&2\sum_{a}\sum_{\xi}\mathbb{E}_{\xi}\big(\mathbf{1}\{\mathrm{sgn}(\xi_j)=\mathrm{sgn}(a_j)\}\big)N_0^a \\
&\quad=\sum_{\xi}\sum_{a}\big(\mathbb{E}_{g(\xi)}N_0^a+\mathbb{E}_{\xi}\mathbf{1}\{\mathrm{sgn}(\xi_j)=\mathrm{sgn}(a_j)\}N_0^a-\mathbb{E}_{g(\xi)}\mathbf{1}\{\mathrm{sgn}(\xi_j)=\mathrm{sgn}(a_j)\}N_0^a\big) \\
&\quad=\sum_{\xi}\big(\mathbb{E}_{g(\xi)}N_0+\mathbb{E}_{\xi}N_0^{\mathcal{A}_j^{\xi}}-\mathbb{E}_{g(\xi)}N_0^{\mathcal{A}_j^{\xi}}\big),
\end{aligned}
\tag{C.9}
$$

where $\mathcal{A}_j^{\xi}=\{a:\mathrm{sgn}(\xi_j)=\mathrm{sgn}(a_j)\}$. By Lemma J.4 and the fact that $N_0^{\mathcal{A}_j^{\xi}}\le T_0$, we have

$$
\mathbb{E}_{\xi}N_0^{\mathcal{A}_j^{\xi}}-\mathbb{E}_{g(\xi)}N_0^{\mathcal{A}_j^{\xi}}\le\frac{cT_0}{16}\sqrt{\mathrm{KL}(\mathcal{P}_{g(\xi)}\|\mathcal{P}_{\xi})},
\tag{C.10}
$$

where $c=8\sqrt{\log 2}$. Moreover, by Lemma J.5, we have

$$
\mathrm{KL}(\mathcal{P}_{\xi'}\|\mathcal{P}_{\xi})\le\frac{16\epsilon^2}{(d-1)^2\delta}\mathbb{E}_{\xi}N_0.
\tag{C.11}
$$

Plugging (C.8), (C.9), (C.10) and (C.11) into (C.7), we obtain

$$
\begin{aligned}
\frac{1}{|\Xi|}\sum_{\xi}\mathbb{E}_{\xi}N_1&\le\frac{T_0}{2}+\frac{\epsilon}{4\delta(d-1)|\Xi|}\sum_{j=1}^{d-1}\sum_{\xi}\Big(\mathbb{E}_{\xi'}N_0+\frac{cT_0\epsilon}{2d\sqrt{\delta}}\sqrt{\mathbb{E}_{\xi}N_0}\Big) \\
&\le\frac{T_0}{2}+\frac{\epsilon}{4\delta(d-1)|\Xi|}\sum_{j=1}^{d-1}\sum_{\xi}\left(\frac{2T_0}{3}+\frac{2}{3\delta}K_0+\frac{cT_0\epsilon}{2d\sqrt{\delta}}\sqrt{\frac{2T_0}{3}+\frac{2}{3\delta}K_0}\right) \\
&=\frac{T_0}{2}+\frac{\epsilon T_0}{6\delta}+\frac{\epsilon K_0}{6\delta^2}+\frac{cT_0\epsilon^2}{8d\delta\sqrt{\delta}}\sqrt{\frac{2T_0}{3}+\frac{2}{3\delta}K_0}.
\end{aligned}
\tag{C.12}
$$

Note that for a given ξ, whether in state x_0 or x_1, the optimal policy is to choose $a_{\xi}=[\mathrm{sgn}(\xi_i)]_{i=1}^{d-1}$. Hence, we can calculate the stationary distribution and find that the optimal average reward is $\frac{\delta+\epsilon}{2\delta+\epsilon}$. Recall the definition of dynamic regret in (2.3), we have

$$
\begin{aligned}
\frac{1}{|\Xi|}\sum_{\xi}\mathbb{E}_{\xi}\text{D-Regret}(T_0)&\ge\frac{\delta+\epsilon}{2\delta+\epsilon}\cdot T_0-\frac{1}{|\Xi|}\sum_{\xi}\mathbb{E}_{\xi}N_1 \\
&\ge\frac{\delta+\epsilon}{2\delta+\epsilon}\cdot T_0-\frac{T_0}{2}-\frac{\epsilon T_0}{6\delta}-\frac{\epsilon K_0}{6\delta^2}-\frac{cT_0\epsilon^2}{8d\delta\sqrt{\delta}}\sqrt{\frac{2T_0}{3}+\frac{2}{3\delta}K_0}.
\end{aligned}
\tag{C.13}
$$

Setting $\delta=\Theta(\frac{1}{H})$ and $\epsilon=\Theta(\frac{d}{\sqrt{HT_0}})$, we have

$$
\frac{1}{|\Xi|}\sum_{\xi}\mathbb{E}_{\xi}\text{D-Regret}(T_0)\ge\Omega(d\sqrt{HT_0}).
$$

Recall that in our episodic setting, the transition kernels $\mathbb{P}_1,\mathbb{P}_2,\cdots,\mathbb{P}_H$ may be different. By the same argument in Jin et al. (2018) (consider H distinct hard-to-learn MDPs and set $\delta=\Theta(\frac{1}{H})$ and $\epsilon=\Theta(\frac{d}{\sqrt{T_0}})$), we obtain a dynamic regret lower bound of $\Omega(dH\sqrt{T_0})$ in the stationary linear kernel MDPs. For non-stationary linear kernel MDPs, the number of segments L is under budget constraint $2\epsilon HL/\sqrt{d}\le\Delta$. By choosing $L=\Theta(d^{-1/3}\Delta^{2/3}H^{-2/3}T^{1/3})$, we have

$$
\frac{1}{|\Xi|}\sum_{\xi}\mathbb{E}_{\xi}\text{D-Regret}(T)\ge\Omega(L\cdot dH\sqrt{T/L})=\Omega(d^{5/6}\Delta^{1/3}H^{2/3}T^{2/3}),
$$

which concludes the proof of Theorem 3.1.

□

D Proof Sketch of Theorem 4.2

In this section, we sketch the proof of Theorem 4.2.

To facilitate the following analysis, we define the model prediction error as

$$l_h^k = r_h^k + \mathbb{P}_h^k V_{h+1}^k - Q_h^k, \tag{D.1}$$

which characterizes the error using V_h^k to replace $V_h^{\pi_k,k}$ in the Bellman equation (2.1).

D.1 Proof Sketch of Theorem 4.2

Proof Sketch of Theorem 4.2. First, we decompose the regret of Algorithm 1 into two terms

$$\text{D-Regret}(T) = \mathcal{R}_1 + \mathcal{R}_2,$$

where $\mathcal{R}_1 = \sum_{k=1}^K V_1^{\pi^*,k}(s_1^k) - V_1^k(s_1^k)$ and $\mathcal{R}_2 = \sum_{k=1}^K V_1^k(s_1^k) - V_1^{\pi^k,k}(s_1^k)$. Then we analyze $\mathcal{R}_1$ and $\mathcal{R}_2$ respectively. By Lemma E.1, we have

$$\begin{aligned}
\mathcal{R}_1 = &\sum_{i=1}^{\rho} \sum_{k=(i-1)\tau+1}^{i\tau} \sum_{h=1}^{H} \mathbb{E}_{\pi^*,k}\big[\langle Q_h^k(s_h,\cdot), \pi_h^{*,k}(\cdot \,|\, s_h) - \pi_h^k(\cdot \,|\, s_h)\rangle\big] \\
&+ \sum_{i=1}^{\rho} \sum_{k=(i-1)\tau+1}^{i\tau} \sum_{h=1}^{H} \mathbb{E}_{\pi^*,k}[l_h^k(s_h,a_h)].
\end{aligned}$$

Applying Lemma D.2 to the first term, we obtain

$$\begin{aligned}
&\sum_{i=1}^{\rho} \sum_{k=(i-1)\tau+1}^{i\tau} \sum_{h=1}^{H} \mathbb{E}_{\pi^*,k}\big[\langle Q_h^k(s_h,\cdot), \pi_h^{*,k}(\cdot \,|\, s_h) - \pi_h^k(\cdot \,|\, s_h)\rangle\big] \\
&\quad \le \sqrt{2H^3 T\rho \log|\mathcal{A}|} + \tau H^2(P_T + \sqrt{d}\Delta).
\end{aligned}$$

Meanwhile, as shown in Lemma E.1, we have

$$\mathcal{R}_2 = \mathcal{M}_{K,H,2} - \sum_{i=1}^{\rho} \sum_{k=(i-1)\tau+1}^{i\tau} \sum_{h=1}^{H} l_h^k(s_h^k, a_h^k).$$

Here $\mathcal{M}_{K,H,2}$ is a martingale defined in Appendix E. Then by the Azuma-Hoeffding inequality, we obtain $|\mathcal{M}_{K,H,2}| \le \sqrt{16H^2T\cdot\log(4/\zeta)}$ with probability at least $1-\zeta/2$. Here $\zeta \in (0,1]$ is a constant.

Now we only need to derive the bound of the quantity $\sum_{i=1}^{\rho}\sum_{k=(i-1)\tau+1}^{i\tau}\sum_{h=1}^{H}(\mathbb{E}_{\pi^*,k}[l_h^k(s_h,a_h)] - l_h^k(s_h^k,a_h^k))$. Applying the bound of l_h^k in Lemma D.3 to this quantity, it holds with probability at least $1-\zeta/2$ that

$$\begin{aligned}
&\sum_{i=1}^{\rho} \sum_{k=(i-1)\tau+1}^{i\tau} \sum_{h=1}^{H} \big((\mathbb{E}_{\pi^*,k}[l_h^k(s_h,a_h)] - l_h^k(s_h^k,a_h^k)\big) \\
&\quad \le 2\sum_{k=1}^{K}\sum_{h=1}^{H}\Big(\sum_{i=1\vee(k-w)}^{k-1} \|\theta_h^i - \theta_h^{i+1}\|_2 + \sum_{i=1\vee(k-w)}^{k-1} \|\xi_h^i - \xi_h^{i+1}\|_2 + B_h^k(s,a) + \Gamma_h^k(s,a)\Big).
\end{aligned}$$

Then we apply Lemmas J.1 and J.2 to bound this quantity by $2w\Delta H\sqrt{d} + 8dT\sqrt{\log(w)/w} + 8C'dTH\cdot\sqrt{\log(wH^2d)/w}\cdot\log(dT/\zeta)$, where C' is a constant specified in the detailed proof.

With the help of these bounds, we derive the regret bound in Theorem 4.2. $\square$

D.2 Online Mirror Descent Term

In this subsection, we establish the upper bound of the online mirror descent term.

The following lemma characterizes the policy improvement step defined in (4.1), where the updated policy π^k takes the closed form in (4.3).

Lemma D.1 (One-Step Descent)**.** For any distribution π on $\mathcal{A}$ and $\{\pi^k\}_{k=1}^K$ obtained in Algorithm 1, it holds that

$$\alpha \cdot \langle Q_h^k, \pi_h(\cdot \,|\, s) - \pi^k(\cdot \,|\, s)\rangle \\ \leq \mathrm{KL}\big(\pi_h(\cdot \,|\, s) \,\|\, \pi_h^k(\cdot \,|\, s)\big) - \mathrm{KL}\big(\pi_h(\cdot \,|\, s) \,\|\, \pi_h^{k+1}(\cdot \,|\, s)\big) + \alpha^2 H^2/2.$$

Proof. See Appendix F.1 for a detailed proof. □

Based on Lemma D.1, we establish an upper bound of online mirror descent term in the following lemma.

Lemma D.2 (Online Mirror Descent Term)**.** For the Q-functions $\{Q_h^k\}_{(k,h)\in[K]\times[H]}$ obtained in (4.9) and the policies $\{\pi_h^k\}_{(k,h)\in[K]\times[H]}$ obtained in (4.3), we have

$$\sum_{i=1}^{\rho} \sum_{k=(i-1)\tau+1}^{i\tau} \sum_{h=1}^{H} \mathbb{E}_{\pi^{*,k}}\big[\langle Q_h^k(s_h, \cdot), \pi_h^{*,k}(\cdot \,|\, s_h) - \pi_h^k(\cdot \,|\, s_h)\rangle\big] \\ \leq \sqrt{2H^3 T\rho \log|\mathcal{A}|} + \tau H^2 (P_T + \sqrt{d}\Delta).$$

Proof. See Appendix F.2 for a detailed proof. □

D.3 Model Prediction Error Term

In this subsection, we characterize the model prediction errors arising from estimating reward functions and transition kernels.

Lemma D.3 (Upper Confidence Bound)**.** Under Assumptions 2.1 and 4.1, it holds with probability at least $1 - \zeta/2$ that

$$-2B_h^k(s,a) - 2\Gamma_h^k(s,a) - \sum_{i=1\vee(k-w)}^{k-1} \|\theta_h^i - \theta_h^{i+1}\|_2 - H\sqrt{d} \cdot \sum_{i=1\vee(k-w)}^{k-1} \|\xi_h^i - \xi_h^{i+1}\|_2 \\ \leq l_h^k(s,a) \leq \sum_{i=1\vee(k-w)}^{k-1} \|\theta_h^i - \theta_h^{i+1}\|_2 + H\sqrt{d} \cdot \sum_{i=1\vee(k-w)}^{k-1} \|\xi_h^i - \xi_h^{i+1}\|_2$$

for any $(k,h) \in [K] \times [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, where w is the length of a sliding window defined in (4.5), $B_h^k(\cdot,\cdot)$ is the bonus function of reward defined in (4.10) and $\Gamma_h^k(\cdot,\cdot)$ is the bonus function of transition kernel defined in (4.10).

Proof. See Appendix F.3 for a detailed proof. □

Since our model is non-stationary, we cannot ensure that the estimated Q-function is "optimistic in the face of uncertainty" as $l_h^k \leq 0$ like the previous work (Jin et al., 2019b; Cai et al., 2019) in the stationary case. Thanks to the sliding window method, the model prediction error here can be upper bounded by the slight changes of parameters in the sliding window. Specifically, within the sliding window, the reward functions and transition kernels can be considered unchanged, which encourages us to estimate the Q-function by regression and UCB bonus, and thus achieve the optimism like the stationary case. However, reward functions and transition kernels are actually different in the sliding window, which leads to additional errors caused by parameter changes.

By giving the bound of the model prediction error l_h^k defined in (D.1), Lemma D.3 quantifies uncertainty and thus realizes sample-efficient. In detail, uncertainty is because we can only observe finite historical data and many state-action pairs (s,a) are less visited or even unseen. The model prediction error of these state-action pairs may be large. However, as is shown in Lemma D.3, the model prediction error l_h^k can be bounded by the variation of sequences $\{\theta_h^i\}_{i=1\vee(k-w)}^k$ and $\{\xi_h^i\}_{i=1\vee(k-w)}^k$, together with the bonus functions B_h^k and Γ_h^k defined in (4.10), which helps us to derive the bound of the regret. See Appendix G for details.

E Regret Decomposition

Recall the definition of model prediction error in (D.1)

$$l_h^k = r_h^k + \mathbb{P}_h^k V_{h+1}^k - Q_h^k.$$

Meanwhile, for any $(k,h) \in [K]\times[H]$, we define $\mathcal{F}_{k,h,1}$ as the σ-algebra generated by the following state-action sequence and reward functions,

$$\{(s_i^\tau, a_i^\tau)\}_{(\tau,i)\in[k-1]\times[H]} \cup \{r^\tau\}_{\tau\in[k]} \cup \{(s_i^k, a_i^k)\}_{i\in[h]}.$$

Similarly, we define $\mathcal{F}_{k,h,2}$ as the σ-algebra generated by

$$\{(s_i^\tau, a_i^\tau)\}_{(\tau,i)\in[k-1]\times[H]} \cup \{r^\tau\}_{\tau\in[k]} \cup \{(s_i^k, a_i^k)\}_{i\in[h]} \cup \{s_{h+1}^k\},$$

where s_{H+1}^k is a null state for any $k \in [K]$. The σ-algebra sequence $\{\mathcal{F}_{k,h,m}\}_{(k,h,m)\in[K]\times[H]\times[2]}$ is a filtration with respect to the timestep index $t(k,h,m) = (k-1)\cdot 2H + (h-1)\cdot 2 + m$. It holds that $\mathcal{F}_{k,h,m} \subseteq \mathcal{F}_{k',h',m'}$ for any $t(k,h,m) \le t(k',h',m')$.

Lemma E.1 (Dynamic Regret Decomposition)**.** For the policies $\{\pi^k\}_{k=1}^K$ obtained in Algorithm 1 and the optimal policies $\pi^{*,k}$ in k-th episode, we have the following decomposition

$$\begin{aligned}
\text{D-Regret}(T) &= \sum_{k=1}^K \big(V_1^{\pi^{*,k},k}(s_1^k) - V_1^{\pi^k,k}(s_1^k)\big) \qquad \text{(E.1)}\\
&= \underbrace{\sum_{i=1}^{\rho}\sum_{k=(i-1)\tau+1}^{i\tau}\sum_{h=1}^H \mathbb{E}_{\pi^{*,k}}\big[\langle Q_h^k(s_h,\cdot), \pi_h^{*,k}(\cdot\,|\,s_h) - \pi_h^k(\cdot\,|\,s_h)\rangle\big]}_{\text{(i)}} + \underbrace{\mathcal{M}_{K,H,2}}_{\text{(ii)}}\\
&\quad + \underbrace{\sum_{i=1}^{\rho}\sum_{k=(i-1)\tau+1}^{i\tau}\sum_{h=1}^H \mathbb{E}_{\pi^{*,k}}[l_h^k(s_h,a_h)]}_{\text{(iii)}} + \underbrace{\sum_{i=1}^{\rho}\sum_{k=(i-1)\tau+1}^{i\tau}\sum_{h=1}^H -l_h^k(s_h^k,a_h^k)}_{\text{(iv)}},
\end{aligned}$$

Proof. Recall the definition of dynamic regret in (2.3), we have

$$\begin{aligned}
\text{D-Regret}(T) &= \sum_{k=1}^K \big(V_1^{\pi^{*,k},k}(s_1^k) - V_1^{\pi^k,k}(s_1^k)\big)\\
&= \sum_{i=1}^{\rho}\sum_{k=(i-1)\tau+1}^{i\tau} \big(V_1^{\pi^{*,k},k}(s_1^k) - V_1^{\pi^k,k}(s_1^k)\big). \qquad \text{(E.2)}
\end{aligned}$$

Note that

$$V_1^{\pi^{*,k},k}(s_1^k) - V_1^{\pi^k,k}(s_1^k) = \underbrace{V_1^{\pi^{*,k},k}(s_1^k) - V_1^k(s_1^k)}_{\text{(i)}} + \underbrace{V_1^k(s_1^k) - V_1^{\pi^k,k}(s_1^k)}_{\text{(ii)}}. \qquad \text{(E.3)}$$

Term (i): By Bellman equation we have

$$\begin{aligned}
V_h^{\pi^{*,k},k}(s) - V_h^k(s) &= \langle Q_h^{\pi^k,k}(s,\cdot), \pi_h^{*,k}(\cdot\,|\,s)\rangle_{\mathcal{A}} - \langle Q_h^k(s,\cdot), \pi_h^k(\cdot\,|\,s)\rangle_{\mathcal{A}} \qquad \text{(E.4)}\\
&= \langle Q_h^{\pi^k,k}(s,\cdot) - Q_h^k(s,\cdot), \pi_h^{*,k}(\cdot\,|\,s)\rangle_{\mathcal{A}} + \langle Q_h^k(s,\cdot), \pi_h^{*,k}(\cdot\,|\,s) - \pi_h^k(\cdot\,|\,s)\rangle_{\mathcal{A}}
\end{aligned}$$

for any $(s, h, k) \in \mathcal{S} \times [H] \times [K]$. Meanwhile, by the definition of the model prediction error in (D.1), we have

$$Q_h^k = r_h^k + \mathbb{P}_h^k V_{h+1}^k - l_h^k.$$

Combining with Bellman equation in (2.1), we further obtain

$$Q_h^{\pi^k,k} - Q_h^k = \mathbb{P}_h^k (V_{h+1}^{\pi^k,k} - V_{h+1}^k) + l_h^k. \tag{E.5}$$

Plugging (E.4) into (E.5), we obtain

$$\begin{aligned} V_h^{\pi^*,k}(s) - V_h^k(s) &= \langle \mathbb{P}_h^k (V_h^{\pi^k,k} - V_h^k)(s), \pi_h^{*,k}(\cdot \,|\, s)\rangle_{\mathcal{A}} + \langle l_h^k(s,\cdot), \pi_h^{*,k}(\cdot \,|\, s)\rangle_{\mathcal{A}} \\ &\quad + \langle Q_h^k(s,\cdot), \pi_h^{*,k}(\cdot \,|\, s) - \pi_h^k(\cdot \,|\, s)\rangle_{\mathcal{A}}. \end{aligned} \tag{E.6}$$

For notational simplicity, for any $(k, h) \in [K] \times [H]$ and function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. we define the operators $\mathbb{I}_h^k$ and $\mathbb{I}_{k,h}$ respectively by

$$(\mathbb{I}_h^k f)(s) = \langle f(x,\cdot), \pi_h^{*,k}(\cdot \,|\, s)\rangle, \quad (\mathbb{I}_{k,h} f)(s) = \langle f(x,\cdot), \pi_h^k(\cdot \,|\, s)\rangle. \tag{E.7}$$

Also, we define

$$\mu_h^k(s) = (\mathbb{I}_h^k Q_h^k)(s) - (\mathbb{I}_{k,h} Q_h^k)(s) = \langle Q_h^k(s,\cdot), \pi_h^{*,k}(\cdot \,|\, s) - \pi_h^k(\cdot \,|\, s)\rangle \tag{E.8}$$

With this notation, recursively expanding (E.6) over $h \in [H]$, we have

$$\begin{aligned} V_1^{\pi^*,k} - V_1^k &= \Big(\prod_{h=1}^H \mathbb{I}_h^k \mathbb{P}_h^k\Big)(V_{H+1}^{\pi^*,k} - V_{H+1}^k) + \sum_{h=1}^H \Big(\prod_{i=1}^{h-1} \mathbb{I}_i^k \mathbb{P}_i^k\Big)\mathbb{I}_h^k l_h^k + \sum_{h=1}^H \Big(\prod_{i=1}^{h-1} \mathbb{I}_i^k \mathbb{P}_i^k\Big)\mu_h^k \\ &= \sum_{h=1}^H \Big(\prod_{i=1}^{h-1} \mathbb{I}_i^k \mathbb{P}_i^k\Big)\mathbb{I}_h^k l_h^k + \sum_{h=1}^H \Big(\prod_{i=1}^{h-1} \mathbb{I}_i^k \mathbb{P}_i^k\Big)\mu_h^k, \end{aligned}$$

where the last inequality follows from $V_{H+1}^{\pi^*,k} = V_{H+1}^k = 0$. By the definitions of $\mathbb{P}_h^k$ in (2.2), $\mathbb{I}_h^k$ in (E.7), and μ_h^k in (E.8), we further obtain

$$\text{Term(i)} = \sum_{h=1}^H \mathbb{E}_{\pi^*,k}\big[\langle Q_h^k(s_h,\cdot), \pi_h^{*,k}(\cdot \,|\, s_h) - \pi_h^k(\cdot \,|\, s_h)\rangle\big] + \sum_{h=1}^H \mathbb{E}_{\pi^*,k}[l_h^k(s_h, a_h)]. \tag{E.9}$$

Term (ii): Recall the definition of value function $V_h^{\pi^k,k}$ in (2.1), the estimated function V_h^k in (4.9) and the operator $\mathbb{I}_h^k$ in (E.7), we expand the model prediction error l_h^k into

$$\begin{aligned} l_h^k(s_h^k, a_h^k) &= r_h^k(s_h^k, a_h^k) + (\mathbb{P}_h^k V_{h+1}^k)(s_h^k, a_h^k) - Q_h^k(s_h^k, a_h^k) \\ &= \big(r_h^k(s_h^k, a_h^k) + (\mathbb{P}_h^k V_{h+1}^k)(s_h^k, a_h^k) - Q_h^{\pi^k,k}(s_h^k, a_h^k)\big) + Q_h^{\pi^k,k}(s_h^k, a_h^k) - Q_h^k(s_h^k, a_h^k) \\ &= \big(\mathbb{P}_h^k (V_{h+1}^k - V_{h+1}^{\pi^k,k})\big)(s_h^k, a_h^k) + (Q_h^{\pi^k,k} - Q_h^k)(s_h^k, a_h^k), \end{aligned}$$

where the last equality follows from the Bellman equation in (2.1). Then we can expand $V_h^k(s_h^k) - V_h^{\pi^k,k}(s_h^k)$ into

$$\begin{aligned} V_h^k(s_h^k) - V_h^{\pi^k,k}(s_h^k) &= \big(\mathbb{I}_{k,h}(Q_h^k - Q_h^{\pi^k,k})\big)(s_h^k) + l_h^k(s_h^k, a_h^k) - l_h^k(s_h^k, a_h^k) \\ &= \big(\mathbb{I}_{k,h}(Q_h^k - Q_h^{\pi^k,k})\big)(s_h^k) + (Q_h^{\pi^k,k} - Q_h^k)(s_h^k, a_h^k) \\ &\quad + \big(\mathbb{P}_h^k (V_{h+1}^k - V_{h+1}^{\pi^k,k})\big)(s_h^k, a_h^k) - l_h^k(s_h^k, a_h^k). \end{aligned}$$

To facilitate our analysis, we define

$$\begin{aligned} D_{k,h,1} &= \big(\mathbb{I}_{k,h}(Q_h^k - Q_h^{\pi^k,k})\big)(s_h^k) - Q_h^k - Q_h^{\pi^k,k}, \\ D_{k,h,2} &= \big(\mathbb{P}_h^k (V_{h+1}^k - V_{h+1}^{\pi^k,k})\big)(s_h^k, a_h^k) - (V_{h+1}^k - V_{h+1}^{\pi^k,k})(s_{h+1}^k). \end{aligned} \tag{E.10}$$

Hence, we have

$$V_h^k(s_h^k) - V_h^{\pi^k,k}(s_h^k) = D_{k,h,1} + D_{k,h,2} + (V_{h+1}^k - V_{h+1}^{\pi^k,k})(s_{h+1}^k) - l_h^k(s_h^k, a_h^k) \tag{E.11}$$

for any $(k, h) \in [K] \times [H]$. For any $k \in [K]$, recursively expanding (E.11) across $h \in [H]$ yields

$$\begin{aligned} \text{Term(ii)} &= \sum_{h=1}^{H}(D_{k,h,1} + D_{k,h,2}) - \sum_{h=1}^{H} l_h^k(s_h^k, a_h^k) + \big(V_{H+1}^k(s_{H+1}^k) - V_{H+1}^{\pi^k,k}(s_{H+1}^k)\big) \\ &= \sum_{h=1}^{H}(D_{k,h,1} + D_{k,h,2}) - \sum_{h=1}^{H} l_h^k(s_h^k, a_h^k), \end{aligned} \tag{E.12}$$

where the last equality uses the fact that $V_{H+1}^k(s_{H+1}^k) = V_{H+1}^{\pi^k,k}(s_{H+1}^k) = 0$. By the definitions of $\mathcal{F}_{k,h,1}$ and $\mathcal{F}_{k,h,2}$, we have the $D_{k,h,1} \in \mathcal{F}_{k,h,1}$ and $D_{k,h,2} \in \mathcal{F}_{k,h,2}$. Hence, for any $(k, h) \in [K] \times [H]$,

$$\mathbb{E}[D_{k,h,1}|\mathcal{F}_{k,h-1,2}] = 0, \quad \mathbb{E}[D_{k,h,2}|\mathcal{F}_{k,h,1}] = 0.$$

Notice that $\mathcal{F}_{k,0,2} = \mathcal{F}_{k-1,H,2}$ for any $k \geq 2$, which implies the corresponding timestep index $t(k, 0, 2) = t(k-1, H, 2) = 2H(k-1)$. Meanwhile, we define $\mathcal{F}_{1,0,2}$ to be empty. Thus we can define the following martingale

$$\begin{aligned} \mathcal{M}_{k,h,m} &= \sum_{\tau=1}^{k-1}\sum_{i=1}^{H}(D_{\tau,i,1} + D_{\tau,i,2}) + \sum_{i=1}^{h-1}(D_{k,i,1} + D_{k,i,2}) + \sum_{\ell=1}^{m} D_{k,h,\ell} \\ &= \sum_{\substack{(\tau,i,\ell)\in[K]\times[H]\times[2], \\ t(\tau,i,\ell)\leq t(k,h,m)}} D_{\tau,i,\ell}, \end{aligned} \tag{E.13}$$

where $t(k, h, m) = 2(k-1)H + 2(h-1) + m$ is the timestep index. This martingale is obviously adapted to the filtration $\{\mathcal{F}_{k,h,m}\}_{(k,h,m)\in[K]\times[H]\times[2]}$, and particularly we have

$$\mathcal{M}_{K,H,2} = \sum_{k=1}^{K}\sum_{h=1}^{H}(D_{k,h,1} + D_{k,h,2}). \tag{E.14}$$

Plugging (E.9) and (E.12) into (E.2), we conclude the proof of Lemma E.1. $\square$

F Proofs of Lemmas in Section D

F.1 Proof of Lemma D.1

Proof. For any $(k, h) \in [K] \times [H]$, let $z_k(s) = \sum_{a'\in\mathcal{A}} p(a') \cdot \exp(\alpha \cdot \pi_h^k(a' \,|\, s))$. Since $z_k(s)$ is a constant function, it holds that, for any $s \in \mathcal{S}$,

$$\langle \log z_k(s), \pi_h(\cdot \,|\, s) - \pi_h^{k+1}(\cdot \,|\, s)\rangle = 0$$

Hence, for any $s \in \mathcal{S}$, it holds that

$$\begin{aligned} &\mathrm{KL}\big(\pi_h(\cdot \,|\, s) \,\|\, \pi_h^k(\cdot \,|\, s)\big) - \mathrm{KL}\big(\pi_h(\cdot \,|\, s) \,\|\, \pi_h^{k+1}(\cdot \,|\, s)\big) \\ &\quad = \big\langle \log\big(\pi_h^{k+1}(\cdot \,|\, s)/\pi_h^k(\cdot \,|\, s)\big), \pi_h(\cdot \,|\, s)\big\rangle \\ &\quad = \big\langle \log\big(\pi_h^{k+1}(\cdot \,|\, s)/\pi_h^k(\cdot \,|\, s)\big), \pi_h(\cdot \,|\, s) - \pi_h^{k+1}(\cdot \,|\, s)\big\rangle + \mathrm{KL}\big(\pi_h^{k+1}(\cdot \,|\, s) \,\|\, \pi_h^k(\cdot \,|\, s)\big) \\ &\quad = \big\langle \log z_k(s) + \log\big(\pi_h^{k+1}(\cdot \,|\, s)/\pi_h^k(\cdot \,|\, s)\big), \pi_h(\cdot \,|\, s) - \pi_h^{k+1}(\cdot \,|\, s)\big\rangle + \mathrm{KL}\big(\pi_h^{k+1}(\cdot \,|\, s) \,\|\, \pi_h^k(\cdot \,|\, s)\big). \end{aligned}$$

Recall that $\pi_h^{k+1}(\cdot \,|\, \cdot) \propto \pi_h^k(\cdot \,|\, \cdot) \cdot \exp\{\alpha \cdot Q_h^k(\cdot \,|\, \cdot)\}$, we have

$$\begin{aligned} &\mathrm{KL}\big(\pi_h(\cdot \,|\, s) \,\|\, \pi_h^k(\cdot \,|\, s)\big) - \mathrm{KL}\big(\pi_h(\cdot \,|\, s) \,\|\, \pi_h^{k+1}(\cdot \,|\, s)\big) \\ &\quad = \alpha \cdot \langle Q_h^k, \pi_h(\cdot \,|\, s) - \pi_h^{k+1}(\cdot \,|\, s)\rangle + \mathrm{KL}\big(\pi_h^{k+1}(\cdot \,|\, s) \,\|\, \pi_h^k(\cdot \,|\, s)\big). \end{aligned}$$

Thus,

$$\begin{aligned}
&\alpha \cdot \langle Q_h^k, \pi_h(\cdot \,|\, s) - \pi^k(\cdot \,|\, s)\rangle \qquad \text{(F.1)} \\
&\quad = \alpha \cdot \langle Q_h^k, \pi_h(\cdot \,|\, s) - \pi_h^{k+1}(\cdot \,|\, s)\rangle + \alpha \cdot \langle Q_h^k, \pi_h^{k+1}(\cdot \,|\, s) - \pi^k(\cdot \,|\, s)\rangle \\
&\quad \le \mathrm{KL}\big(\pi_h(\cdot \,|\, s) \,\|\, \pi_h^k(\cdot \,|\, s)\big) - \mathrm{KL}\big(\pi_h(\cdot \,|\, s) \,\|\, \pi_h^{k+1}(\cdot \,|\, s)\big) - \mathrm{KL}\big(\pi_h^{k+1}(\cdot \,|\, s) \,\|\, \pi_h^k(\cdot \,|\, s)\big) \\
&\qquad + \alpha \cdot \|Q_h^k(s, \cdot)\|_\infty \cdot \|\pi_h^k(\cdot \,|\, s) - \pi_h^{k+1}(\cdot \,|\, s)\|_1,
\end{aligned}$$

where the last inequality uses Cauchy-Schwartz inequality. Meanwhile, by Pinsker's inequality, it holds that

$$\mathrm{KL}\big(\pi_h^{k+1}(\cdot \,|\, s) \| \pi_h^k(\cdot \,|\, s)\big) \ge \|\pi_h^k(\cdot \,|\, s) - \pi_h^{k+1}(\cdot \,|\, s)\|_1^2/2. \qquad \text{(F.2)}$$

Plugging (F.2) into (F.1), combined with the fact that $\|Q_h^k(s, \cdot)\|_\infty \le H$ for any $s \in \mathcal{S}$, we have

$$\begin{aligned}
&\alpha \cdot \langle Q_h^k, \pi_h(\cdot \,|\, s) - \pi^k(\cdot \,|\, s)\rangle \\
&\quad \le \mathrm{KL}\big(\pi_h(\cdot \,|\, s) \,\|\, \pi_h^k(\cdot \,|\, s)\big) - \mathrm{KL}\big(\pi_h(\cdot \,|\, s) \,\|\, \pi_h^{k+1}(\cdot \,|\, s)\big) \\
&\qquad - \|\pi_h^k(\cdot \,|\, s) - \pi_h^{k+1}(\cdot \,|\, s)\|_1^2/2 + \alpha H \|\pi_h^k(\cdot \,|\, s) - \pi_h^{k+1}(\cdot \,|\, s)\|_1 \\
&\quad \le \mathrm{KL}\big(\pi_h(\cdot \,|\, s) \,\|\, \pi_h^k(\cdot \,|\, s)\big) - \mathrm{KL}\big(\pi_h(\cdot \,|\, s) \,\|\, \pi_h^{k+1}(\cdot \,|\, s)\big) + \alpha^2 H^2/2,
\end{aligned}$$

which completes the proof of Lemma D.1. ☐

F.2 Proof of Lemma D.2

Proof. Recall that $\rho = \lceil K/\tau \rceil$. First, we have the decomposition

$$\begin{aligned}
&\sum_{i=1}^{\rho} \sum_{k=(i-1)\tau+1}^{i\tau} \sum_{h=1}^{H} \mathbb{E}_{\pi^{*,k}} \big[\langle Q_h^k(s_h, \cdot), \pi_h^{*,k}(\cdot \,|\, s_h) - \pi_h^k(\cdot \,|\, s_h)\rangle\big] \qquad \text{(F.3)} \\
&\quad + \underbrace{\sum_{i=1}^{\rho} \sum_{k=(i-1)\tau+1}^{i\tau} \sum_{h=1}^{H} \mathbb{E}_{\pi^{*,(i-1)\tau+1}} \big[\langle Q_h^k(s_h, \cdot), \pi_h^{*,k}(\cdot \,|\, s_h) - \pi_h^k(\cdot \,|\, s_h)\rangle\big]}_{\text{(A)}} \\
&\quad + \underbrace{\sum_{i=1}^{\rho} \sum_{k=(i-1)\tau+1}^{i\tau} \sum_{h=1}^{H} (\mathbb{E}_{\pi^{*,k}} - \mathbb{E}_{\pi^{*,(i-1)\tau+1}}) \big[\langle Q_h^k(s_h, \cdot), \pi_h^{*,k}(\cdot \,|\, s_h) - \pi_h^k(\cdot \,|\, s_h)\rangle\big]}_{\text{(B)}}.
\end{aligned}$$

We can further decompose Term A as

$$\begin{aligned}
\mathrm{Term(A)} &= \underbrace{\sum_{i=1}^{\rho} \sum_{k=(i-1)\tau+1}^{i\tau} \sum_{h=1}^{H} \mathbb{E}_{\pi^{*,(i-1)\tau+1}} \big[\langle Q_h^k(s_h, \cdot), \pi_h^{*,(i-1)\tau+1}(\cdot \,|\, s_h) - \pi_h^k(\cdot \,|\, s_h)\rangle\big]}_{A_1} \\
&\quad + \underbrace{\sum_{i=1}^{\rho} \sum_{k=(i-1)\tau+1}^{i\tau} \sum_{h=1}^{H} \mathbb{E}_{\pi^{*,(i-1)\tau+1}} \big[\langle Q_h^k(s_h, \cdot), \pi_h^{*,k}(\cdot \,|\, s_h) - \pi_h^{*,(i-1)\tau+1}(\cdot \,|\, s_h)\rangle\big]}_{A_2}.
\end{aligned}$$

By Lemma D.1, we have

$$
\begin{aligned}
A_1 &\le \alpha K H^3/2 + \sum_{h=1}^{H}\sum_{i=1}^{\rho}\frac{1}{\alpha} \\
&\quad \times \sum_{k=(i-1)\tau+1}^{i\tau}\Big(\mathbb{E}_{\pi_h^{*,(i-1)\tau+1}}\big[\mathrm{KL}(\pi_h^{*,(i-1)\tau+1}(\cdot\,|\,s_h)\,\|\,\pi_h^{k}(\cdot\,|\,s_h)) - \mathrm{KL}\big(\pi_h^{*,(i-1)\tau+1}(\cdot\,|\,s_h)\,\|\,\pi_h^{k+1}(\cdot\,|\,s_h)\big)\big]\Big) \\
&= \frac{1}{2}\alpha K H^3 + \sum_{h=1}^{H}\sum_{i=1}^{\rho}\frac{1}{\alpha}\cdot \\
&\quad \times \Big(\mathbb{E}_{\pi_h^{*,(i-1)\tau+1}}\big[\mathrm{KL}\big(\pi_h^{*,(i-1)\tau+1}(\cdot\,|\,s_h)\,\|\,\pi_h^{(i-1)\tau+1}(\cdot\,|\,s_h)\big) - \mathrm{KL}\big(\pi_h^{*,(i-1)\tau+1}(\cdot\,|\,s_h)\,\|\,\pi_h^{i\tau+1}(\cdot\,|\,s_h)\big)\big]\Big) \\
&\le \frac{1}{2}\alpha K H^3 + \frac{1}{\alpha}\cdot\sum_{h=1}^{H}\sum_{i=1}^{\rho}\Big(\mathbb{E}_{\pi_h^{*,(i-1)\tau+1}}\big[\mathrm{KL}\big(\pi_h^{*,(i-1)\tau+1}(\cdot\,|\,s_h)\,\|\,\pi_h^{(i-1)\tau+1}(\cdot\,|\,s_h)\big)\big]\Big).
\end{aligned} \tag{F.4}
$$

Here the second inequality is obtained by the fact that the KL-divergence is non-negative. Note that $\pi_h^{(i-1)\tau+1}$ is the uniform policy, that is, $\pi_h^{(i-1)\tau+1}(a\,|\,s_h) = \frac{1}{|\mathcal{A}|}$ for any $a \in \mathcal{A}$. Hence, for any policy π and $i \in [\rho]$, we have

$$
\begin{aligned}
\mathrm{KL}\big(\pi_h(\cdot\,|\,s_h)\,\|\,\pi_h^{(i-1)\tau+1}(\cdot\,|\,s_h)\big) &= \sum_{a\in\mathcal{A}}\pi_h(a\,|\,s_h)\cdot\log\big(|\mathcal{A}|\cdot\pi_h(a\,|\,s_h)\big) \\
&= \log|\mathcal{A}| + \sum_{a\in\mathcal{A}}\pi_h(a\,|\,s_h)\cdot\log\big(\pi_h(a\,|\,s_h)\big) \le \log|\mathcal{A}|,
\end{aligned} \tag{F.5}
$$

where the last inequality follows from the fact that the entropy of $\pi_h(\cdot\,|\,s_h)$ is non-negative. Plugging (F.5) into (F.4), we have

$$
A_1 \le \alpha H^3 K/2 + \rho H \log|\mathcal{A}|/\alpha = \sqrt{2H^3 T\rho\log|\mathcal{A}|}, \tag{F.6}
$$

where the last inequality holds since we set $\alpha = \sqrt{2\rho\log|\mathcal{A}|/(H^2K)}$ in (4.2). Meanwhile,

$$
\begin{aligned}
A_2 &\le \sum_{i=1}^{\rho}\sum_{k=(i-1)\tau+1}^{i\tau}\sum_{h=1}^{H}\mathbb{E}_{\pi^{*,(i-1)\tau+1}}\big[H\cdot\|\pi_h^{*,k}(\cdot\,|\,s_h) - \pi_h^{*,(i-1)\tau+1}(\cdot\,|\,s_h)\|_1\big] \\
&\le H\cdot\sum_{i=1}^{\rho}\sum_{k=(i-1)\tau+1}^{i\tau}\sum_{h=1}^{H}\sum_{t=(i-1)\tau+2}^{k}\mathbb{E}_{\pi^{*,(i-1)\tau+1}}\big[\|\pi_h^{*,k}(\cdot\,|\,s_h) - \pi_h^{*,t-1}(\cdot\,|\,s_h)\|_1\big] \\
&\le H\cdot\sum_{i=1}^{\rho}\sum_{k=(i-1)\tau+1}^{i\tau}\sum_{t=(i-1)\tau+2}^{i\tau}\sum_{h=1}^{H}\max_{s\in\mathcal{S}}\|\pi_h^{*,t}(\cdot\,|\,s) - \pi_h^{*,t-1}(\cdot\,|\,s)\|_1 \\
&= H\tau\cdot\sum_{t=1}^{K}\sum_{h=1}^{H}\max_{s\in\mathcal{S}}\|\pi_h^{*,t}(\cdot\,|\,s) - \pi_h^{*,t-1}(\cdot\,|\,s)\|_1 = H\tau P_T,
\end{aligned} \tag{F.7}
$$

where the first inequality follows by Holder's inequality and the fact that $\|Q_h^k(s,\cdot)\|_\infty \le H$, the second inequality follows from triangle inequality, and the last inequality is obtained by the definition of P_T in (2.4). Combing (F.6) and (F.7), we have

$$
\mathrm{Term(A)} \le \sqrt{2H^3T\rho\log|\mathcal{A}|} + H\tau P_T. \tag{F.8}
$$

By Lemma J.6 and the same proof of Lemma 4 in Fei et al. (2020), we have

$$
\mathrm{Term(B)} \le \tau H^2(P_T + \Delta_P), \tag{F.9}
$$

where $\Delta_P = \sum_{k=1}^K \sum_{h=1}^H \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|P_h^k(\cdot \,|\, s_j, a) - P_h^{k+1}(\cdot \,|\, s, a)\|_1$. By Assumption 2.1, we further obtain

$$
\begin{aligned}
\Delta_P &= \sum_{k=1}^K \sum_{h=1}^H \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|P_h^k(\cdot \,|\, s_j, a) - P_h^{k+1}(\cdot \,|\, s, a)\|_1 \\
&= \sum_{k=1}^K \sum_{h=1}^H \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{s'\in\mathcal{S}} |P_h^k(s' \,|\, s_j, a) - P_h^{k+1}(s' \,|\, s, a)| \\
&= \sum_{k=1}^K \sum_{h=1}^H \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{s'\in\mathcal{S}} |\psi(s,a,s')^\top(\xi_h^k - \xi_h^{k+1})| \\
&\le \sum_{k=1}^K \sum_{h=1}^H \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{s'\in\mathcal{S}} \|\psi(s,a,s')\|_2 \cdot \|\xi_h^k - \xi_h^{k+1}\|_2,
\end{aligned} \tag{F.10}
$$

where the last inequality follows from Cauchy-Schwarz inequality . Recall the assumption that $\sum_{s'\in\mathcal{S}} \|\psi(s,a,s')\|_2 \le \sqrt{d}$ for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, we have

$$
\begin{aligned}
&\sum_{k=1}^K \sum_{h=1}^H \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{s'\in\mathcal{S}} \|\psi(s,a,s')\|_2 \cdot \|\xi_h^k - \xi_h^{k+1}\|_2 \\
&\qquad \le \sqrt{d} \sum_{k=1}^K \sum_{h=1}^H \|\xi_h^k - \xi_h^{k+1}\|_2 = \sqrt{d} B_P \le \sqrt{d}\Delta.
\end{aligned} \tag{F.11}
$$

Combining (F.8), (F.9), (F.10) and (F.11), we have

$$
\sum_{i=1}^{\rho} \sum_{k=(i-1)\tau+1}^{i\tau} \sum_{h=1}^H \mathbb{E}_{\pi^{*,k}}\big[\langle Q_h^k(s_h,\cdot), \pi_h^{*,k}(\cdot \,|\, s_h) - \pi_h^k(\cdot \,|\, s_h)\rangle\big] \le \sqrt{2H^3T\rho\log|\mathcal{A}|} + \tau H^2(P_T + \sqrt{d}\Delta),
$$

which concludes the proof.

□

F.3 Proof of Lemma D.3

Proof. We first derive the upper bound of $-l_h^k(\cdot,\cdot)$. As defined in (D.1), for any $(k,h)\in[K]\times[H]$ and $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$
-l_h^k(s,a) = Q_h^k(s,a) - (r_h^k + \mathbb{P}_h^k V_{h+1}^k)(s,a).
$$

Meanwhile, by the definition of Q_h^k in (4.9), we have

$$
\begin{aligned}
Q_h^k(\cdot,\cdot) &= \min\{\phi(\cdot,\cdot)^\top \widehat{\theta}_h^k + \eta_h^k(\cdot,\cdot)^\top \widehat{\xi}_h^k + B_h^k(\cdot,\cdot) + \Gamma_h^k(\cdot,\cdot), H-h+1\}^+ \\
&\le \phi(s,a)^\top \widehat{\theta}_h^k + \eta_h^k(s,a)^\top \widehat{\xi}_h^k + B_h^k(s,a) + \Gamma_h^k(s,a)
\end{aligned}
$$

for any $(k,h)\in[K]\times[H]$ and $(s,a)\in\mathcal{S}\times\mathcal{A}$. Hence, we obtain

$$
\begin{aligned}
-l_h^k(s,a) &= Q_h^k(s,a) - (r_h^k + \mathbb{P}_h^k V_{h+1}^k)(s,a) \\
&\le \phi(s,a)^\top \widehat{\theta}_h^k + \eta_h^k(s,a)^\top \widehat{\xi}_h^k + B_h^k(s,a) + \Gamma_h^k(s,a) - (r_h^k + \mathbb{P}_h^k V_{h+1}^k)(s,a) \\
&= \underbrace{\phi(s,a)^\top \widehat{\theta}_h^k + B_h^k(s,a) - r_h^k(s,a)}_{\text{(i)}} + \underbrace{\eta_h^k(s,a)^\top \widehat{\xi}_h^k + \Gamma_h^k(s,a) - \mathbb{P}_h^k V_{h+1}^k(s,a)}_{\text{(ii)}}.
\end{aligned}
$$

Term (i): By the definition of $\widehat{\theta}_h^k$ in (4.5), we have

$$\begin{aligned}
\widehat{\theta}_h^k - \theta_h^k &= (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1\vee(k-w)}^{k-1} \phi(s_h^\tau, a_h^\tau) r_h^\tau(s_h^\tau, a_h^\tau)\Big) - \theta_h^k \\
&= (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1\vee(k-w)}^{k-1} \phi(s_h^\tau, a_h^\tau) r_h^\tau(s_h^\tau, a_h^\tau) - \Lambda_h^k \theta_h^k\Big) \\
&= (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1\vee(k-w)}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top (\theta_h^\tau - \theta_h^k) - \lambda \cdot \theta_h^k\Big),
\end{aligned}$$

where the last equality is obtained by the definition of Λ_h^k in (4.5) and the assumption that $r_h^\tau(s_h^\tau, a_h^\tau) = \phi(s_h^\tau, a_h^\tau)^\top \theta_h^\tau$ for any $(\tau, h) \in [K] \times [H]$. Hence, for any $(k, h) \in [K] \times [H]$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\begin{aligned}
&|\phi(s,a)^\top(\widehat{\theta}_h^k - \theta_h^k)| \qquad \text{(F.12)} \\
&\quad \le \underbrace{\Big|\phi(s,a)^\top(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1\vee(k-w)}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top(\theta_h^\tau - \theta_h^k)\Big)\Big|}_{\text{(i.1)}} + \underbrace{|\phi(s,a)^\top(\Lambda_h^k)^{-1}(\lambda \cdot \theta_h^k)|}_{\text{(i.2)}}.
\end{aligned}$$

Then we derive the upper bound of term (i.1) and term (i.2), respectively.

Term (i.1): By Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\Big|\phi(s,a)^\top(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1\vee(k-w)}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top(\theta_h^\tau - \theta_h^k)\Big)\Big| \\
&\qquad \le \|\phi(s,a)\|_2 \cdot \Big\|(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1\vee(k-w)}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top(\theta_h^\tau - \theta_h^k)\Big)\Big\|_2 \\
&\qquad \le \Big\|(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1\vee(k-w)}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top(\theta_h^\tau - \theta_h^k)\Big)\Big\|_2, \qquad \text{(F.13)}
\end{aligned}$$

where the last inequality follows from the fact that $\|\phi(s,a)\|_2 \le 1$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. Moreover, we have

$$\begin{aligned}
&\Big\|(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1\vee(k-w)}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top(\theta_h^\tau - \theta_h^k)\Big)\Big\|_2 \\
&\qquad = \Big\|(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1\vee(k-w)}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top\Big(\sum_{i=\tau}^{k-1}(\theta_h^i - \theta_h^{i+1})\Big)\Big)\Big\|_2, \qquad \text{(F.14)}
\end{aligned}$$

where the last equality follows from the fact that $\theta_h^\tau - \theta_h^k = \sum_{i=\tau}^{k-1}(\theta_h^i - \theta_h^{i+1})$. By exchanging the order of summation, we further have

$$\begin{aligned}
&\Big\|(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1\vee(k-w)}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top\Big(\sum_{i=\tau}^{k-1}(\theta_h^i - \theta_h^{i+1})\Big)\Big)\Big\|_2 \\
&\qquad = \Big\|(\Lambda_h^k)^{-1}\Big(\sum_{i=1\vee(k-w)}^{k-1}\Big(\sum_{\tau=1\vee(k-w)}^{i} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top(\theta_h^i - \theta_h^{i+1})\Big)\Big)\Big\|_2 \\
&\qquad \le \sum_{i=1\vee(k-w)}^{k-1} \Big\|(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1\vee(k-w)}^{i} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top\Big)(\theta_h^i - \theta_h^{i+1})\Big\|_2. \qquad \text{(F.15)}
\end{aligned}$$

By the fact that for any matrix $A \in \mathbb{R}^{d\times d}$ and a vector $x \in \mathbb{R}^d$, $\|Ax\|_2 \le \lambda_{\max}(A^\top A)\|x\|_2$, we have

$$\begin{aligned}
&\sum_{i=1\vee(k-w)}^{k-1} \left\|(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1\vee(k-w)}^{i} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top\Big)(\theta_h^i - \theta_h^{i+1})\right\|_2 \\
&\quad\le \sum_{i=1\vee(k-w)}^{k-1} \lambda_{\max}\bigg(\Big(\sum_{\tau=1\vee(k-w)}^{i} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top\Big)(\Lambda_h^k)^{-2} \\
&\qquad\qquad \Big(\sum_{\tau=1\vee(k-w)}^{i} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top\Big)\bigg)\|(\theta_h^i - \theta_h^{i+1})\|_2. \qquad \text{(F.16)}
\end{aligned}$$

Meanwhile, by Assumption 4.1, we assume $\phi(s_h^\tau, a_h^\tau) = \Psi z_h^\tau$. For simplicity, we define $M_1 = \sum_{\tau=1\vee(k-w)}^{k-1}(z_h^\tau)(z_h^\tau)^\top + \lambda I_d$ and $M_2 = \sum_{\tau=1\vee(k-w)}^{k-1}(z_h^\tau)(z_h^\tau)^\top$. Then, we have

$$\begin{aligned}
&\lambda_{\max}\bigg(\Big(\sum_{\tau=1\vee(k-w)}^{i} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top\Big)(\Lambda_h^k)^{-2}\Big(\sum_{\tau=1\vee(k-w)}^{i} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top\Big)\bigg) \\
&\quad = \lambda_{\max}\big(\Psi M_2 \Psi^\top (\Psi M_1 \Psi^\top)^{-2} \Psi M_2 \Psi^\top\big) = \lambda_{\max}(M_2 M_1^{-2} M_2). \qquad \text{(F.17)}
\end{aligned}$$

Given the eigenvalue decomposition $M_2 = P\mathrm{diag}(\lambda_1, \cdots, \lambda_d)P^\top$ where P is an orthogonal matrix and λ_i is the i-th eigenvalue of M_2, we have $M_1 = P\mathrm{diag}(\lambda_1 + \lambda, \cdots, \lambda_d + \lambda)P^\top$. Thus $M_2 M_1^{-2} M_2 = \mathrm{diag}(\lambda_1^2/(\lambda_1+\lambda)^2, \cdots, \lambda_d^2/(\lambda_d+\lambda)^2)$, which further implies that

$$\lambda_{\max}(M_2 M_1^{-2} M_2) \le 1. \qquad \text{(F.18)}$$

Combined with (F.13), (F.14), (F.15), (F.16), and (F.17), we obtain

$$\left|\phi(s,a)^\top(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1\vee(k-w)}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top(\theta_h^\tau - \theta_h^k)\Big)\right| \le \sum_{i=1\vee(k-w)}^{k-1} \|\theta_h^i - \theta_h^{i+1}\|_2. \qquad \text{(F.19)}$$

Term (i.2): By Cauchy-Schwarz inequality, we obtain

$$|\phi(s,a)^\top(\Lambda_h^k)^{-1}(\lambda\cdot\theta_h^k)| \le \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} \cdot \|\lambda\cdot\theta_h^k\|_{(\Lambda_h^k)^{-1}}.$$

Note the fact that $\Lambda_h^k \succeq \lambda I_d$, which implies $\lambda_{\min}((\Lambda_h^k)^{-1}) \ge \lambda$. We further obtain

$$\|\lambda\cdot\theta_h^k\|_{(\Lambda_h^k)^{-1}}^2 \le \frac{1}{\lambda_{\min}\big((\Lambda_h^k)^{-1}\big)} \cdot \||\lambda\cdot\theta_h^k\|_2^2 \le \frac{1}{\lambda}\cdot\lambda^2 d = \lambda d.$$

Hence, we have

$$|\phi(s,a)^\top(\Lambda_h^k)^{-1}(\lambda\cdot\theta_h^k)| \le \sqrt{\lambda d}\cdot\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}. \qquad \text{(F.20)}$$

Setting $\beta_k = \sqrt{\lambda d}$ for any $k \in [K]$ in the bonus function B_h^k defined in (4.10). Plugging (F.19) and (F.20) into (F.12), we obtain

$$|\phi(s,a)^\top(\widehat{\theta}_h^k - \theta_h^k)| \le B_h^k(s,a) + \sum_{i=1\vee(k-w)}^{k-1} \|\theta_h^i - \theta_h^{i+1}\|_2 \qquad \text{(F.21)}$$

for any $(k,h) \in [K]\times[H]$. Hence, for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have

$$\phi(s,a)\widehat{\theta}_h^k + B_h^k(s,a) - r_h^k(s,a) \le 2B_h^k(s,a) + \sum_{i=1\vee(k-w)}^{k-1} \|\theta_h^i - \theta_h^{i+1}\|_2. \qquad \text{(F.22)}$$

Term (ii): Recall that η_h^k defined in (4.6) takes the form

$$\eta_h^k(\cdot,\cdot) = \int_{\mathcal{S}} \psi(\cdot,\cdot,s')\cdot V_{h+1}^k(s')\mathrm{d}s'$$

for any $(k,h) \in [K] \times [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$. Meanwhile, by Assumption 2.1, we obtain

$$\begin{aligned}(\mathbb{P}_h^k V_{h+1}^k)(s,a) &= \int_{\mathcal{S}} \psi(s,a,s')^\top \xi_h^k \cdot V_{h+1}^k(s')\mathrm{d}s' \\ &= \eta_h^k(s,a)^\top \xi_h^k = \eta_h^k(s,a)^\top (A_h^k)^{-1} A_h^k \xi_h^k\end{aligned} \tag{F.23}$$

for any $(k,h) \in [K] \times [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$. Recall the definition of A_h^k in (4.8), we have

$$\begin{aligned}(\mathbb{P}_h^k V_{h+1}^k)(s,a) &= \eta_h^k(s,a)^\top (A_h^k)^{-1}\Bigg(\sum_{\tau=1\vee(k-w)}^{k-1} \eta_h^\tau(s_h^\tau,a_h^\tau)\eta_h^\tau(s_h^\tau,a_h^\tau)^\top \xi_h^k + \lambda' \cdot \xi_h^k\Bigg) \\ &= \eta_h^k(s,a)^\top (A_h^k)^{-1}\Bigg(\sum_{\tau=1\vee(k-w)}^{k-1} \eta_h^\tau(s_h^\tau,a_h^\tau) \cdot (\mathbb{P}_h^k V_{h+1}^\tau)(s_h^\tau,a_h^\tau) + \lambda' \cdot \xi_h^k\Bigg)\end{aligned}$$

for any $(k,h) \in [K] \times [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$. Here the second equality is obtained by (F.23). Recall the definition of $\widehat{\xi}_h^k$ in (4.8), we have

$$\begin{aligned}&\eta_h^k(\cdot,\cdot)^\top \widehat{\xi}_h^k - (\mathbb{P}_h^k V_{h+1}^k)(s,a) \\ &= \underbrace{\eta_h^k(s,a)^\top (A_h^k)^{-1}\Bigg(\sum_{\tau=1\vee(k-w)}^{k-1} \eta_h^\tau(s_h^\tau,a_h^\tau) \cdot \big(V_{h+1}^\tau(s_{h+1}^\tau) - (\mathbb{P}_h^k V_{h+1}^\tau)(s_h^\tau,a_h^\tau)\big)\Bigg)}_{\text{(ii.1)}} \\ &\quad - \underbrace{\lambda' \cdot \eta_h^k(s,a)^\top (A_h^k)^{-1}\xi_h^k}_{\text{(ii.2)}}\end{aligned} \tag{F.24}$$

for any $(k,h) \in [K] \times [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Term (ii.1): We can decompose $\mathrm{Term(ii.1)}$ as

$$\begin{aligned}\text{Term (ii.1)} =& \underbrace{\eta_h^k(s,a)^\top (A_h^k)^{-1}\Bigg(\sum_{\tau=1\vee(k-w)}^{k-1} \eta_h^\tau(s_h^\tau,a_h^\tau) \cdot \big(V_{h+1}^\tau(s_{h+1}^\tau) - (\mathbb{P}_h^\tau V_{h+1}^\tau)(s_h^\tau,a_h^\tau)\big)\Bigg)}_{\text{(ii.1.1)}} \\ &+ \underbrace{\eta_h^k(s,a)^\top (A_h^k)^{-1}\Bigg(\sum_{\tau=1\vee(k-w)}^{k-1} \eta_h^\tau(s_h^\tau,a_h^\tau) \cdot \big((\mathbb{P}_h^\tau V_{h+1}^\tau)(s_h^\tau,a_h^\tau) - (\mathbb{P}_h^k V_{h+1}^\tau)(s_h^\tau,a_h^\tau)\big)\Bigg)}_{\text{(ii.1.2)}}\end{aligned} \tag{F.25}$$

By the definition of A_h^k in (4.8), $(A_h^k)^{-1}$ is a positive definite matrix. Hence, by Cauchy-Schwarz inequality,

$$\begin{aligned}&|\text{Term (ii.1.1)}| \\ &\le \sqrt{\eta_h^k(s,a)^\top (A_h^k)^{-1}\eta_h^k(s,a)} \cdot \Bigg\|\sum_{\tau=1\vee(k-w)}^{k-1} \eta_h^\tau(s_h^\tau,a_h^\tau) \cdot \big(V_{h+1}^\tau(s_{h+1}^\tau) - (\mathbb{P}_h^\tau V_{h+1}^\tau)(s_h^\tau,a_h^\tau)\big)\Bigg\|_{(A_h^k)^{-1}}\end{aligned} \tag{F.26}$$

for any $(k,h) \in [K] \times [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$. Under the event $\mathcal{E}$ defined in (J.4) of Lemma J.3, which happens with probability at least $1-\zeta/2$, it holds that

$$|\text{Term (ii.1.1)}| \le C''\sqrt{dH^2 \cdot \log(dT/\zeta)} \cdot \sqrt{\eta_h^k(s,a)^\top (A_h^k)^{-1}\eta_h^k(s,a)} \tag{F.27}$$

for any $(k,h) \in [K] \times [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$. Here $C'' > 0$ is an absolute constant defined in Lemma J.3. Meanwhile, by (F.23), we have $(\mathbb{P}_h^k V_{h+1}^\tau)(s,a) = \eta_h^\tau(s,a)^\top \xi_h^k$ and $\mathbb{P}_h^\tau V_{h+1}^\tau(s,a) =$

$\eta_h^\tau(s,a)^\top \xi_h^\tau$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, which implies

$$\begin{aligned}|\text{Term (ii.1.2)}| &= \bigg|\eta_h^k(s,a)^\top (A_h^k)^{-1}\bigg(\sum_{\tau=1\vee(k-w)}^{k-1} \eta_h^\tau(s_h^\tau, a_h^\tau)\eta_h^\tau(s_h^\tau, a_h^\tau)^\top(\xi_h^\tau - \xi_h^k)\bigg)\bigg| \\ &\le \|\eta_h^k(s,a)\|_2 \cdot \bigg\|(A_h^k)^{-1}\bigg(\sum_{\tau=1\vee(k-w)}^{k-1} \eta_h^\tau(s_h^\tau, a_h^\tau)\eta_h^\tau(s_h^\tau, a_h^\tau)^\top(\xi_h^\tau - \xi_h^k)\bigg)\bigg\|_2 \\ &\le H\sqrt{d} \cdot \bigg\|(A_h^k)^{-1}\bigg(\sum_{\tau=1\vee(k-w)}^{k-1} \eta_h^\tau(s_h^\tau, a_h^\tau)\eta_h^\tau(s_h^\tau, a_h^\tau)^\top(\xi_h^\tau - \xi_h^k)\bigg)\bigg\|_2,\end{aligned}$$

where the last inequality is obtained by Assumption 2.1. Then, by the same derivation of (F.19), we have

$$|\text{Term (ii.1.2)}| \le H\sqrt{d} \cdot \sum_{i=1}^{k-1} \|\xi_h^i - \xi_h^{i+1}\|_2. \tag{F.28}$$

Plugging (F.27) and (F.28) into (F.25), we obtain

$$|\text{Term (ii.1)}| \le C''\sqrt{dH^2 \cdot \log(dT/\zeta)} \cdot \sqrt{\eta_h^k(s,a)^\top (A_h^k)^{-1}\eta_h^k(s,a)} + H\sqrt{d} \cdot \sum_{i=1\vee(k-w)}^{k-1} \|\xi_h^i - \xi_h^{i+1}\|_2. \tag{F.29}$$

Term (ii.2): For any $(k,h) \in [K] \times [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\begin{aligned}|\text{Term (ii.2)}| &\le \lambda' \cdot \sqrt{\eta(s,a)^\top (A_h^k)^{-1}\eta(s,a)} \cdot \|\xi_h^k\|_{(A_h^k)^{-1}} \\ &\le \sqrt{\lambda'} \cdot \sqrt{\eta(s,a)^\top (A_h^k)^{-1}\eta(s,a)} \cdot \|\xi_h^k\|_2 \\ &\le \sqrt{\lambda' d} \cdot \sqrt{\eta(s,a)^\top (A_h^k)^{-1}\eta(s,a)},\end{aligned} \tag{F.30}$$

where the first inequality follows from Cauchy-Schwarz inequality, the second inequality follows from the fact that $A_h^k \succeq \lambda' \cdot I_d$ and the last inequality is obtained by Assumption 2.1. Plugging (F.29) and (F.30) into (F.24), we have

$$\begin{aligned}&|\eta_h^k(\cdot,\cdot)^\top \widehat{\xi}_h^k - (\mathbb{P}_h^k V_{h+1}^k)(s,a)| \\ &\quad \le C'\sqrt{dH^2 \cdot \log(dT/\zeta)} \cdot \sqrt{\eta_h^k(s,a)^\top (A_h^k)^{-1}\eta_h^k(s,a)} + H\sqrt{d} \cdot \sum_{i=1\vee(k-w)}^{k-1} \|\xi_h^i - \xi_h^{i+1}\|_2\end{aligned} \tag{F.31}$$

for any $(k,h) \in [K] \times [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$. Here $C' > 1$ is another absolute constant. Setting

$$\beta' = C'\sqrt{dH^2 \cdot \log(dT/\zeta)}$$

in the bonus function Γ_h^k defined in (4.10). Hence, by (F.31), we have

$$|\eta_h^k(s,a)^\top \widehat{\xi}_h^k - (\mathbb{P}_h^k V_{h+1}^k)(s,a)| \le \Gamma_h^k(s,a) + H\sqrt{d} \cdot \sum_{i=1\vee(k-w)}^{k-1} \|\xi_h^i - \xi_h^{i+1}\|_2 \tag{F.32}$$

for any $(k,h) \in [K] \times [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$ under event $\mathcal{E}$. Hence,

$$\eta_h^k(s,a)^\top \widehat{\xi}_h^k + \Gamma_h^k(s,a) - \mathbb{P}_h^k V_{h+1}^k(s,a) \le 2\Gamma_h^k(s,a) + H\sqrt{d} \cdot \sum_{i=1\vee(k-w)}^{k-1} \|\xi_h^i - \xi_h^{i+1}\|_2 \tag{F.33}$$

for any $(k,h) \in [K] \times [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$ under event $\mathcal{E}$. Combining (F.22) and (F.33), we have

$$\begin{aligned} -l_h^k(s,a) &= Q_h^k(s,a) - (r_h^k + \mathbb{P}_h^k V_{h+1}^k)(s,a) \\ &\le 2B_h^k(s,a) + 2\Gamma_h^k(s,a) + \sum_{i=1\vee(k-w)}^{k-1} \|\theta_h^i - \theta_h^{i+1}\|_2 + H\sqrt{d}\cdot \sum_{i=1\vee(k-w)}^{k-1} \|\xi_h^i - \xi_h^{i+1}\|_2. \end{aligned} \tag{F.34}$$

Then, we show that $l_h^k(s,a) \le \sum_{i=1\vee(k-w)}^{k-1} \|\theta_h^i - \theta_h^{i+1}\|_2 + H\sqrt{d}\cdot \sum_{i=1\vee(k-w)}^{k-1} \|\xi_h^i - \xi_h^{i+1}\|_2$ for any $(k,h) \in [K] \times [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$ under event $\mathcal{E}$.

$$\begin{aligned} l_h^k(s,a) &= (r_h^k + \mathbb{P}_h^k V_{h+1}^k)(s,a) - Q_h^k \\ &= (r_h^k + \mathbb{P}_h^k V_{h+1}^k)(s,a) - \min\{\phi(s,a)\widehat{\theta}_h^k + \eta_h^k(s,a)^\top \widehat{\xi}_h^k + B_h^k(s,a) + \Gamma_h^k(s,a), H-h+1\} \\ &= \max\{r_h^k(s,a) - \phi(s,a)\widehat{\theta}_h^k - B_h^k(s,a) + (\mathbb{P}_h^k V_{h+1}^k)(s,a) - \eta_h^k(\cdot,\cdot)^\top \widehat{\xi}_h^k - \Gamma_h^k(s,a), \\ &\qquad (r_h^k + \mathbb{P}_h^k V_{h+1}^k)(s,a) - (H-h+1)\}. \end{aligned} \tag{F.35}$$

By (F.21) and (F.32), we have

$$\begin{aligned} &r_h^k(s,a) - \phi(s,a)\widehat{\theta}_h^k - B_h^k(s,a) + (\mathbb{P}_h^k V_{h+1}^k)(s,a) - \eta_h^k(\cdot,\cdot)^\top \widehat{\xi}_h^k - \Gamma_h^k(s,a) \\ &\quad \le \sum_{i=1\vee(k-w)}^{k-1} \|\theta_h^i - \theta_h^{i+1}\|_2 + H\sqrt{d}\cdot \sum_{i=1\vee(k-w)}^{k-1} \|\xi_h^i - \xi_h^{i+1}\|_2. \end{aligned} \tag{F.36}$$

Also, we note the fact that $V_{h+1}^k \le H-h$, it is not difficult to show that

$$(r_h^k + \mathbb{P}_h^k V_{h+1}^k)(s,a) - (H-h+1) \le 0. \tag{F.37}$$

Plugging (F.36) and (F.37) into (F.35), we obtain

$$l_h^k(s,a) \le \sum_{i=1\vee(k-w)}^{k-1} \|\theta_h^i - \theta_h^{i+1}\|_2 + H\sqrt{d}\cdot \sum_{i=1\vee(k-w)}^{k-1} \|\xi_h^i - \xi_h^{i+1}\|_2 \tag{F.38}$$

for any $(k,h) \in [K] \times [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$ under event $\mathcal{E}$. Combining (F.34) and (F.38), we finish the proof of Lemma D.3.

□

G PROOF OF THEOREM 4.2

Proof. By Lemma E.1, we decompose dynamic regret of Algorithm 1 into four parts:

$$\begin{aligned} \text{D-Regret}(T) &= \sum_{k=1}^{K} \big(V_1^{\pi^{*,k},k}(s_1^k) - V_1^{\pi^k,k}(s_1^k)\big) \\ &= \underbrace{\sum_{i=1}^{\rho} \sum_{k=(i-1)\tau+1}^{i\tau} \sum_{h=1}^{H} \mathbb{E}_{\pi^{*,k}}\big[\langle Q_h^k(s_h,\cdot), \pi_h^{*,k}(\cdot\,|\,s_h) - \pi_h^k(\cdot\,|\,s_h)\rangle\big]}_{\text{(i)}} + \underbrace{\mathcal{M}_{K,H,2}}_{\text{(ii)}} \\ &\quad + \underbrace{\sum_{i=1}^{\rho} \sum_{k=(i-1)\tau+1}^{i\tau} \sum_{h=1}^{H} \mathbb{E}_{\pi^{*,k}}[l_h^k(s_h,a_h)]}_{\text{(iii)}} + \underbrace{\sum_{i=1}^{\rho} \sum_{k=(i-1)\tau+1}^{i\tau} \sum_{h=1}^{H} -l_h^k(s_h^k,a_h^k)}_{\text{(iv)}}, \end{aligned} \tag{G.1}$$

Now we establish the upper bound of these four parts, respectively.

Upper Bounding (i):

By Lemma D.2, we have

$$\text{Term(i)} \leq \sqrt{2H^3T\rho\log|\mathcal{A}|} + \tau H^2(P_T + \sqrt{d}\Delta). \tag{G.2}$$

Then we discuss several cases.

- If $0 \leq P_T + \sqrt{d}\Delta \leq \sqrt{\frac{\log|\mathcal{A}|}{K}}$, then $\tau = \Pi_{[1,K]}(\lfloor(\frac{T\sqrt{\log|\mathcal{A}|}}{H(P_T+\sqrt{d}\Delta)})^{2/3}\rfloor) = K$, which implies that $\rho = 1$. Then (G.2) yields
$$\text{Term(i)} \leq 2H^2\sqrt{K\log|\mathcal{A}|} + \cdot H^2\sqrt{K\log|\mathcal{A}|} = 3\sqrt{H^3T\log|\mathcal{A}|}. \tag{G.3}$$

- If $\sqrt{\frac{\log|\mathcal{A}|}{K}} \leq P_T + \sqrt{d}\Delta \leq 2^{-3/2}\cdot K\sqrt{\log|\mathcal{A}|}$, we have $\tau \in [2, K]$ and (G.2) yields
$$\begin{aligned}\text{Term(i)} &\leq 2\cdot\frac{1}{\sqrt{\tau}}H^2K\sqrt{\log|\mathcal{A}|} + \cdot\tau H^2\sqrt{K\log|\mathcal{A}|}\\ &\leq 5(H^2T\sqrt{\log|\mathcal{A}|})^{2/3}(P_T+\sqrt{d}\Delta)^{1/3}.\end{aligned} \tag{G.4}$$

- If $P_T > 2^{-3/2}\cdot K\sqrt{\log|\mathcal{A}|}$, we have $\tau = 1$ and therefore $\rho = K$. Then (G.2) implies
$$\text{Term(i)} \leq 2H^2K\sqrt{\log|\mathcal{A}|} + \cdot H^2P_T \leq 9H^2(P_T+\sqrt{d}\Delta). \tag{G.5}$$

Combining (G.3), (G.4) and (G.5), we have

$$\text{Term(i)} \leq \begin{cases} \sqrt{H^3T\log|\mathcal{A}|}, & \text{if } 0 \leq P_T + \sqrt{d}\Delta \leq \sqrt{\frac{\log|\mathcal{A}|}{K}},\\ (H^2T\sqrt{\log|\mathcal{A}|})^{2/3}(P_T+\sqrt{d}\Delta)^{1/3}, & \text{if } \sqrt{\frac{\log|\mathcal{A}|}{K}} \leq P_T + \sqrt{d}\Delta \lesssim K\sqrt{\log|\mathcal{A}|},\\ H^2(P_T+\sqrt{d}\Delta), & \text{if } P_T+\sqrt{d}\Delta \gtrsim K\sqrt{\log|\mathcal{A}|}, \end{cases} \tag{G.6}$$

Upper Bounding (ii): Recall that

$$\mathcal{M}_{K,H,2} = \sum_{k=1}^{K}\sum_{h=1}^{H}(D_{k,h,1} + D_{k,h,2}).$$

Here the $D_{k,h,1}$ and $D_{k,h,2}$ defined in (E.10) take the following forms,

$$\begin{aligned} D_{k,h,1} &= \big(\mathbb{I}_h^k(Q_h^k - Q_h^{\pi^k,k})\big)(s_h^k) - Q_h^k - Q_h^{\pi^k,k},\\ D_{k,h,2} &= \big(\mathbb{P}_h^k(V_{h+1}^k - V_{h+1}^{\pi^k,k})\big)(s_h^k,a_h^k) - (V_{h+1}^k - V_{h+1}^{\pi^k,k})(s_{h+1}^k). \end{aligned}$$

By the truncation of $\phi(\cdot,\cdot)\widehat{\theta}_h^k + \eta_h^k(\cdot,\cdot)^\top\widehat{\xi}_h^k + B_h^k(\cdot,\cdot) + \Gamma_h^k(\cdot,\cdot)$ into range $[0, H-h+1]$ in (4.9), we know that $Q_h^k, Q_h^{\pi^k,k}, V_{h+1}^k, V_{h+1}^{\pi^k,k} \in [0,H]$, which implies that $|D_{k,h,1}| \leq 2H$ and $|D_{k,h,2}| \leq 2H$ for any $(k,h) \in [H]\times[K]$. Applying the Azuma-Hoeffding inequality to the martingale $\mathcal{M}_{K,H,2}$, we obtain

$$P(|\mathcal{M}_{K,H,2}| > \varepsilon) \leq 2\exp\bigg(\frac{-\varepsilon^2}{16H^3K}\bigg).$$

For any $\zeta \in (0,1)$, if we set $\varepsilon = \sqrt{16H^3K\cdot\log(4/\zeta)}$, we have

$$|\mathcal{M}_{K,H,2}| \leq \sqrt{16H^2T\cdot\log(4/\zeta)} \tag{G.7}$$

with probability at least $1-\zeta/2$.

Upper Bounding (iii): By Lemma D.3, it holds with probability at least $1-\zeta/2$ that

$$l_h^k(s,a) \leq \sum_{i=1\vee(k-w)}^{k-1}\|\theta_h^i - \theta_h^{i+1}\|_2 + H\sqrt{d}\cdot\sum_{i=1\vee(k-w)}^{k-1}\|\xi_h^i - \xi_h^{i+1}\|_2$$

for any $(k, h) \in [K] \times [H]$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$, which implies that

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}[l_h^k(s_h, a_h) \,|\, s_1 = s_1^k] \\
&\quad \leq \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1\vee(k-w)}^{k-1} \|\theta_h^i - \theta_h^{i+1}\|_2 + H\sqrt{d}\cdot \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1\vee(k-w)}^{k-1} \|\xi_h^i - \xi_h^{i+1}\|_2 \\
&\quad = \sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{i=1\vee(k-w)}^{k-1} \|\theta_h^i - \theta_h^{i+1}\|_2 + H\sqrt{d}\cdot \sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{i=1\vee(k-w)}^{k-1} \|\xi_h^i - \xi_h^{i+1}\|_2 \\
&\quad \leq \sum_{h=1}^{H}\sum_{k=1}^{K} w\cdot \|\theta_h^k - \theta_h^{k+1}\|_2 + H\sqrt{d}\cdot \sum_{h=1}^{H}\sum_{k=1}^{K} w\cdot \|\xi_h^k - \xi_h^{k+1}\|_2 \\
&\quad \leq wB_T + wH\sqrt{d}B_P \leq w\Delta H\sqrt{d}.
\end{aligned} \tag{G.8}
$$

Here the last inequality follows from the definition of total variation budget in (2.5).

Upper Bounding (iv): As is shown in Lemma D.3, it holds with probability at least $1 - \zeta/2$ that

$$
-l_h^k(s, a) \leq 2B_h^k(s, a) + 2\Gamma_h^k(s, a) + \sum_{i=1\vee(k-w)}^{k-1} \|\theta_h^i - \theta_h^{i+1}\|_2 + H\sqrt{d}\cdot \sum_{i=1\vee(k-w)}^{k-1} \|\xi_h^i - \xi_h^{i+1}\|_2
$$

for any $(k, h) \in [K] \times [H]$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$. Meanwhile, by the definitions of Q_h^k and l_h^k in (4.9) and (D.1), we have that $|l_h^k(s, a)| \leq 2H$. Hence,

$$
-l_h^k(s, a) \leq 2B_h^k(s, a) + 2H \wedge 2\Gamma_h^k(s, a) + \sum_{i=1\vee(k-w)}^{k-1} \|\theta_h^i - \theta_h^{i+1}\|_2 + H\sqrt{d}\cdot \sum_{i=1\vee(k-w)}^{k-1} \|\xi_h^i - \xi_h^{i+1}\|_2
$$

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H} -l_h^k(s_h^k, a_h^k) \leq\ & 2\sum_{k=1}^{K}\sum_{h=1}^{H} B_h^k(s, a) + 2\sum_{k=1}^{K}\sum_{h=1}^{H} H \wedge \Gamma_h^k(s, a) \\
&+ \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1\vee(k-w)}^{k-1} \|\theta_h^i - \theta_h^{i+1}\|_2 + H\sqrt{d}\cdot \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1\vee(k-w)}^{k-1} \|\xi_h^i - \xi_h^{i+1}\|_2.
\end{aligned} \tag{G.9}
$$

By Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
\sum_{h=1}^{H}\sum_{k=1}^{K} B_h^k(s_h^k, a_h^k) &\leq \beta\cdot \sum_{h=1}^{H}\sum_{k=1}^{K} \sqrt{\phi(s_h^k, a_h^k)^\top (\Lambda_h^k)^{-1} \phi_h^k(s_h^k, a_h^k)} \\
&\leq \beta\cdot \sum_{h=1}^{H}\Big(K\cdot \sum_{k=1}^{K} \phi(s_h^k, a_h^k)^\top (\Lambda_h^k)^{-1} \phi_h^k(s_h^k, a_h^k)\Big)^{1/2} \\
&= \beta\sqrt{K}\cdot \sum_{h=1}^{H}\sqrt{\sum_{k=1}^{K} \|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}}}.
\end{aligned} \tag{G.10}
$$

As we set $\lambda = 1$, we have that $\Lambda_h^k \succeq I_d$, which implies

$$
\|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}} \leq \|\phi(s_h^k, a_h^k)\|_2 \leq 1
$$

for any $(k, h) \in [K] \times [H]$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$. By Lemma J.2, we further have

$$
\sum_{k=1}^{K} \|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}} \leq 2d\lceil K/w\rceil \log\big((w+\lambda)/\lambda\big) \leq 4dK\log(w)/w. \tag{G.11}
$$

Combining (G.10) and (G.11), we further obtain

$$\sum_{h=1}^{H}\sum_{k=1}^{K} B_h^k(s_h^k,a_h^k) \le 4dT\sqrt{\log(w)/w}. \tag{G.12}$$

Meanwhile, by the definition of Γ_h^k in (4.10), we have

$$\sum_{h=1}^{H}\sum_{k=1}^{K} H\wedge \Gamma_h^k(s_h^k,a_h^k) = \beta'\cdot \sum_{h=1}^{H}\sum_{k=1}^{K} H/\beta' \wedge \sqrt{\eta_h^k(s_h^k,a_h^k)^\top (A_h^k)^{-1}\eta_h^k(s_h^k,a_h^k)}.$$

Recall that

$$\beta' = C'\sqrt{dH^2\cdot \log(dT/\zeta)},$$

which implies that $\beta' > H$. Thus, we have

$$\begin{aligned}\sum_{h=1}^{H}\sum_{k=1}^{K} H\wedge \Gamma_h^k(s_h^k,a_h^k) &\le \beta'\cdot \sum_{h=1}^{H}\sum_{k=1}^{K} 1 \wedge \sqrt{\eta_h^k(s_h^k,a_h^k)^\top (A_h^k)^{-1}\eta_h^k(s_h^k,a_h^k)} \\ &\le \beta'\cdot \sum_{h=1}^{H}\bigg(K\cdot \sum_{k=1}^{K} 1\wedge \|\eta_h^k(s_h^k,a_h^k)\|_{(A_h^k)^{-1}}\bigg)^{1/2}\end{aligned} \tag{G.13}$$

where the second inequality follows from Cauchy-Schwarz inequality. Note the facts that $A_h^1 = \lambda' I_d$ and $\|\eta_h^k(s,a)\|_2 \le \sqrt{d}H$ for any $(k,h)\in[K]\times[H]$ and $(s,a)\in\mathcal{S}\times\mathcal{A}$. By the same proof of Lemma J.2, we have

$$\sum_{k=1}^{K} 1\wedge \|\eta_h^k(s_h^k,a_h^k)\|_{(A_h^k)^{-1}} \le 2d\lceil K/w\rceil \log\big((wH^2d+\lambda')/\lambda'\big) \le 4dK\log(wH^2d)/w. \tag{G.14}$$

Combining (G.13) and (G.14), we have

$$\begin{aligned}\sum_{h=1}^{H}\sum_{k=1}^{K} \Gamma_h^k(s_h^k,a_h^k) &\le 2\beta'\sqrt{dT^2\cdot \log(wH^2d)/w} \\ &= 4C'dTH\cdot \sqrt{\log(wH^2d)/w}\cdot \log(dT/\zeta).\end{aligned} \tag{G.15}$$

where $C' > 1$ is an absolute constant and $T = HK$. By the same proof in (G.8), we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1\vee(k-w)}^{k-1} \|\theta_h^i - \theta_h^{i+1}\|_2 + H\sqrt{d}\cdot \sum_{i=1\vee(k-w)}^{k-1} \|\xi_h^i - \xi_h^{i+1}\|_2 \le w\Delta H\sqrt{d}. \tag{G.16}$$

Plugging (G.12), (G.15) and (G.16) into (G.9), we have

$$\sum_{h=1}^{H}\sum_{k=1}^{K} -l_h^k(s_h^k,a_h^k) \le w\Delta H\sqrt{d} + 4dT\sqrt{\log(w)/w} + 4C'dTH\cdot\sqrt{\log(wH^2d)/w}\cdot \log(dT/\zeta). \tag{G.17}$$

Meanwhile, by (G.7), (G.8) and (G.17), it holds with probability at least $1-\zeta$ that

$$\begin{aligned}\text{Term(ii)} + \text{Term(iii)} + \text{Term(iv)} &\le \sqrt{16H^2T\cdot \log(4/\zeta)} + 2w\Delta H\sqrt{d} \\ &\quad + 8dT\sqrt{\log(w)/w} + 8C'dTH\cdot \sqrt{\log(wH^2d)/w}\cdot \log(dT/\zeta) \\ &\lesssim d^{5/6}\Delta^{1/3}HT^{2/3}\cdot \log(dT/\zeta).\end{aligned} \tag{G.18}$$

Here we uses the facts that $w = \Theta(d^{1/3}\Delta^{-2/3}T^{2/3})$. Plugging (G.6) and (G.18) into (G.1), we finish the proof of Theorem 4.2. $\square$

H PROOF OF THEOREM 4.3

Proof. Let $\tau = K$ in Lemma E.1, we have

$$\begin{aligned}\text{D-Regret}(T) &= \sum_{k=1}^{K}\big(V_1^{\pi^{*,k},k}(s_1^k) - V_1^{\pi^k,k}(s_1^k)\big) \qquad \text{(H.1)}\\ &= \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^{*,k}}\big[\langle Q_h^k(s_h,\cdot), \pi_h^{*,k}(\cdot\,|\,s_h) - \pi_h^k(\cdot\,|\,s_h)\rangle\big]}_{\text{(i)}} + \underbrace{\mathcal{M}_{K,H,2}}_{\text{(ii)}}\\ &\quad + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^{*,k}}[l_h^k(s_h,a_h)]}_{\text{(iii)}} + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} -l_h^k(s_h^k,a_h^k)}_{\text{(iv)}},\end{aligned}$$

Since policies π_h^k are greedy with respect to Q_h^k for any $(k,h)\in[K]\times[H]$, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^{*,k}}\big[\langle Q_h^k(s_h,\cdot), \pi_h^{*,k}(\cdot\,|\,s_h) - \pi_h^k(\cdot\,|\,s_h)\rangle\big] \le 0. \qquad \text{(H.2)}$$

By the same derivation of (G.18) in the proof of Theorem 4.2, we have

$$\begin{aligned}\text{Term(ii)} + \text{Term(iii)} + \text{Term(iv)} &\le \sqrt{16H^2T\cdot\log(4/\zeta)} + 2w\Delta H\sqrt{d}\\ &\quad + 8dT\sqrt{\log(w)/w} + 8C'dTH\cdot\sqrt{\log(wH^2d)/w}\cdot\log(dT/\zeta)\\ &\lesssim d^{5/6}\Delta^{1/3}HT^{2/3}\cdot\log(dT/\zeta). \qquad \text{(H.3)}\end{aligned}$$

Here we uses the facts that $w = \Theta(d^{1/3}\Delta^{-2/3}T^{2/3})$. Plugging (H.2) and (H.3) into (H.1), we finish the proof of Theorem 4.3. □

I RESULTS WITHOUT ASSUMPTION 4.1

Theorem I.1 (Upper bound for Algorithm 1)**.** Suppose Assumptions 2.1 holds. Let $\tau = \Pi_{[1,K]}(\lfloor(\frac{T\sqrt{\log|\mathcal{A}|}}{H(P_T+\sqrt{d}\Delta)})^{2/3}\rfloor)$, $\alpha = \sqrt{\rho\log|\mathcal{A}|/(H^2K)}$ in (4.2), $w = \Theta(\Delta^{-1/4}T^{1/4})$ in (4.4), $\lambda = \lambda' = 1$ in (4.4) and (4.9), $\beta = \sqrt{d}$ in (4.10), and $\beta' = C'\sqrt{dH^2\cdot\log(dT/\zeta)}$ in (4.10), where $C' > 1$ is an absolute constant and $\zeta\in(0,1]$. We have

$$\text{D-Regret}(T) \lesssim d\Delta^{1/4}HT^{3/4}\cdot\log(dT/\zeta) + \begin{cases}\sqrt{H^3T\log|\mathcal{A}|}, & \text{if } 0\le P_T+\sqrt{d}\Delta\le\sqrt{\frac{\log|\mathcal{A}|}{K}},\\ (H^2T\sqrt{\log|\mathcal{A}|})^{2/3}(P_T+\sqrt{d}\Delta)^{1/3}, & \text{if } \sqrt{\frac{\log|\mathcal{A}|}{K}}\le P_T+\sqrt{d}\Delta\lesssim K\sqrt{\log|\mathcal{A}|},\\ H^2(P_T+\sqrt{d}\Delta), & \text{if } P_T+\sqrt{d}\Delta\gtrsim K\sqrt{\log|\mathcal{A}|},\end{cases}$$

with probability at least $1-\zeta$.

Proof. In the previous proof, we only use Assumption 4.1 to derive (F.19) and (F.28) in the proof of Lemma D.3 (§F.3). Then we give a slightly loose bound without Assumption 4.1. For any

$(k,h) \in [K] \times [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$
\begin{aligned}
&\left|\phi(s,a)^\top (\Lambda_h^k)^{-1}\Bigg(\sum_{\tau=1\vee(k-w)}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top(\theta_h^\tau - \theta_h^k)\Bigg)\right| \\
&\quad \le \sum_{\tau=1\vee(k-w)}^{k-1} \left|\phi(s,a)^\top (\Lambda_h^k)^{-1}\phi(s_h^\tau, a_h^\tau)\right| \cdot \left|\phi(s_h^\tau, a_h^\tau)^\top(\theta_h^\tau - \theta_h^k)\right| \\
&\quad \le \sum_{\tau=1\vee(k-w)}^{k-1} \left|\phi(s,a)^\top (\Lambda_h^k)^{-1}\phi(s_h^\tau, a_h^\tau)\right| \cdot \left\|\phi(s_h^\tau, a_h^\tau)\right\|_2 \cdot \left\|\theta_h^\tau - \theta_h^k\right\|_2 \\
&\quad \le \sum_{\tau=1\vee(k-w)}^{k-1} \left|\phi(s,a)^\top (\Lambda_h^k)^{-1}\phi(s_h^\tau, a_h^\tau)\right| \cdot \sum_{i=\tau}^{k-1}\left\|\theta_h^i - \theta_h^{i+1}\right\|_2,
\end{aligned}
$$

where the second inequality is obtained by Cauchy-Schwarz inequality and the last last inequality follows from the facts that $\|\phi(\cdot,\cdot)\|_2 \le 1$ and $\left\|\theta_h^\tau - \theta_h^k\right\|_2 \le \left\|\sum_{i=\tau}^{k-1}(\theta_h^i - \theta_h^{i+1})\right\|_2 \le \sum_{i=\tau}^{k-1}\left\|\theta_h^i - \theta_h^{i+1}\right\|_2$. Note that $\sum_{\tau=1\vee(k-w)}^{k-1}\sum_{i=\tau}^{k-1} = \sum_{i=1\vee(k-w)}^{k-1}\sum_{\tau=1\vee(k-w)}^{i}$, we further obtain that

$$
\begin{aligned}
&\sum_{\tau=1\vee(k-w)}^{k-1} \left|\phi(s,a)^\top (\Lambda_h^k)^{-1}\phi(s_h^\tau, a_h^\tau)\right| \cdot \sum_{i=\tau}^{k-1}\left\|\theta_h^i - \theta_h^{i+1}\right\|_2 \\
&\quad = \sum_{i=1\vee(k-w)}^{k-1}\sum_{\tau=1\vee(k-w)}^{i} \left|\phi(s,a)^\top (\Lambda_h^k)^{-1}\phi(s_h^\tau, a_h^\tau)\right| \cdot \left\|\theta_h^i - \theta_h^{i+1}\right\|_2 \\
&\quad \le \sum_{i=1\vee(k-w)}^{k-1}\sqrt{\sum_{\tau=1\vee(k-w)}^{i}\|\phi(s,a)\|^2_{(\Lambda_h^k)^{-1}} \cdot \sum_{\tau=1\vee(k-w)}^{i}\|\phi(s_h^\tau, a_h^\tau)\|^2_{(\Lambda_h^k)^{-1}}} \cdot \left\|\theta_h^i - \theta_h^{i+1}\right\|_2.
\end{aligned}
\tag{I.1}
$$

Note that $\Lambda_h^k \succeq I_d$, which further implies

$$
\sum_{\tau=1\vee(k-w)}^{i}\|\phi(s,a)\|^2_{(\Lambda_h^k)^{-1}} \le \sum_{\tau=1\vee(k-w)}^{i}\|\phi(s,a)\|_2^2 \le \sum_{\tau=1\vee(k-w)}^{i} 1 \le w. \tag{I.2}
$$

Meanwhile, we have

$$
\begin{aligned}
\sum_{\tau=1\vee(k-w)}^{i}\|\phi(s_h^\tau, a_h^\tau)\|^2_{(\Lambda_h^k)^{-1}} &= \sum_{\tau=1\vee(k-w)}^{i}\mathrm{Tr}\big(\phi(s_h^\tau, a_h^\tau)^\top(\Lambda_h^k)^{-1}\phi(s_h^\tau, a_h^\tau)\big) \\
&= \mathrm{Tr}\Bigg((\Lambda_h^k)^{-1}\sum_{\tau=1\vee(k-w)}^{i}\phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top\Bigg).
\end{aligned}
\tag{I.3}
$$

Similar to the derivation of (F.18), we have

$$
\mathrm{Tr}\Bigg((\Lambda_h^k)^{-1}\sum_{\tau=1\vee(k-w)}^{i}\phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top\Bigg) = \sum_{i=1}^{d}\frac{\lambda_i}{\lambda_i+\lambda} \le d, \tag{I.4}
$$

where λ_i is the i-th eigenvalue of $\sum_{\tau=1\vee(k-w)}^{i}\phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top$. Plugging (I.2), (I.3) and (I.4) into (I.1), we have

$$
\begin{aligned}
&\left|\phi(s,a)^\top (\Lambda_h^k)^{-1}\Bigg(\sum_{\tau=1\vee(k-w)}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top(\theta_h^\tau - \theta_h^k)\Bigg)\right| \\
&\quad \le \sqrt{dw}\cdot\sum_{i=1\vee(k-w)}^{k-1}\left\|\theta_h^i - \theta_h^{i+1}\right\|_2.
\end{aligned}
\tag{I.5}
$$

Similarly, for any $(k,h)\in[K]\times[H]$ and $(s,a)\in\mathcal{S}\times\mathcal{A}$, we have

$$\begin{aligned}&\bigg|\eta_h^k(s,a)^\top(A_h^k)^{-1}\bigg(\sum_{\tau=1\vee(k-w)}^{k-1}\eta_h^\tau(s_h^\tau,a_h^\tau)\eta_h^\tau(s_h^\tau,a_h^\tau)^\top(\xi_h^\tau-\xi_h^k)\bigg)\bigg|\\&\quad\le Hd\sqrt{w}\cdot\sum_{i=1\vee(k-w)}^{k-1}\|\xi_h^i-\xi_h^{i+1}\|_2.\end{aligned}\tag{I.6}$$

Replacing (F.19) and (F.28) by (I.5) and (I.6), we can obtain that

$$\begin{aligned}&-2B_h^k(s,a)-2\Gamma_h^k(s,a)-\sqrt{dw}\cdot\sum_{i=1\vee(k-w)}^{k-1}\|\theta_h^i-\theta_h^{i+1}\|_2-Hd\sqrt{w}\cdot\sum_{i=1\vee(k-w)}^{k-1}\|\xi_h^i-\xi_h^{i+1}\|_2\\&\quad\le l_h^k(s,a)\le\sqrt{dw}\cdot\sum_{i=1\vee(k-w)}^{k-1}\|\theta_h^i-\theta_h^{i+1}\|_2+Hd\sqrt{w}\cdot\sum_{i=1\vee(k-w)}^{k-1}\|\xi_h^i-\xi_h^{i+1}\|_2.\end{aligned}$$

Plugging this inequality in the original proof of Theorem 4.2 (§G) and choosing $w=\Theta(\Delta^{-1/4}T^{1/4})$, we conclude the proof. ☐

Theorem I.2 (Upper bound for Algorithm 3)**.** Suppose Assumption 2.1 holds. Let $w=\Theta(\Delta^{-1/4}T^{1/4})$ in (4.4), $\lambda=\lambda'=1$ in (4.4) and (4.9), $\beta=\sqrt{d}$ in (4.10), and $\beta'=C'\sqrt{dH^2\cdot\log(dT/\zeta)}$ in (4.10), where $C'>1$ is an absolute constant and $\zeta\in(0,1]$. We have

$$\text{D-Regret}(T)\lesssim d\Delta^{1/4}HT^{3/4}\cdot\log(dT/\zeta)$$

with probability at least $1-\zeta$.

Proof. The proof is similar to the proof of Theorem I.1, and we omit it here. ☐

J Useful Lemmas

Lemma J.1. Let $\{\phi_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued sequence with $\|\phi_t\|_2\le1$. Also, let $\Lambda_0\in\mathbb{R}^{d\times d}$ be a positive-definite matrix with $\lambda_{\min}(\Lambda_0)\ge1$ and $\Lambda_t=\Lambda_0+\sum_{j=1}^{t-1}\phi_j\phi_j^\top$. For any $t\in\mathbb{Z}_+$, it holds that

$$\log\bigg(\frac{\det(\Lambda_{t+1})}{\det(\Lambda_1)}\bigg)\le\sum_{j=1}^t\phi_j^\top\Lambda_j^{-1}\phi_j\le2\log\bigg(\frac{\det(\Lambda_{t+1})}{\det(\Lambda_1)}\bigg).$$

Proof. See Dani et al. (2008); Rusmevichientong & Tsitsiklis (2010); Jin et al. (2019b); Cai et al. (2019) for a detailed proof. ☐

Lemma J.2. For the Λ_h^k defined in (4.5), we have

$$\sum_{k=1}^K1\wedge\|\phi(s_h^k,a_h^k)\|_{(\Lambda_h^k)^{-1}}\le2d\lceil K/w\rceil\log\big((w+\lambda)/\lambda\big)$$

for any $h\in[H]$.

Proof. First, we rewrite the sums as follows.

$$\sum_{k=1}^K1\wedge\|\phi(s_h^k,a_h^k)\|_{(\Lambda_h^k)^{-1}}=\sum_{t=0}^{\lceil K/w\rceil-1}\sum_{k=tw+1}^{(t+1)w}1\wedge\|\phi(s_h^k,a_h^k)\|_{(\Lambda_h^k)^{-1}}.\tag{J.1}$$

For the t-th block of length w we define the matrix

$$W_h^{k,t}=\sum_{\tau=tw+1}^{k-1}\phi(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)^\top+\lambda I_d.$$

Recall the Λ_h^k in (4.5)

$$\Lambda_h^k = \sum_{\tau=1\vee(k-w)}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top + \lambda I_d.$$

Note that Λ_h^k contains extra terms which are positive definite matrices for any $(k, h) \in [tw, (t+1)w] \times [H]$, we have $\Lambda_h^k \succeq W_h^{k,t}$ for any $(k, h) \in [tw, (t+1)w] \times [H]$. Hence,

$$(\Lambda_h^k)^{-1} \preceq (W_h^{k,t})^{-1}$$

for any $(k, h) \in [tw, (t+1)w] \times [H]$, which implies that

$$\begin{aligned}\sum_{t=0}^{\lceil K/w\rceil-1} \sum_{k=tw+1}^{(t+1)w} 1 \wedge \|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}} &\le \sum_{t=0}^{\lceil K/w\rceil-1} \sum_{k=tw+1}^{(t+1)w} 1 \wedge \|\phi(s_h^k, a_h^k)\|_{(W_h^{k,t})^{-1}} \\ &\le \sum_{t=0}^{\lceil K/w\rceil-1} 2\log\bigg(\frac{\det(W_h^{(t+1)w+1,t})}{\det(W_h^{tw,t})}\bigg), \end{aligned} \tag{J.2}$$

where the last inequality follows from Lemma J.1. Moreover, we have $\|\phi(s, a)\|_2 \le 1$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, which implies

$$W_h^{(t+1)w+1,t} = \sum_{\tau=tw+1}^{(t+1)w} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top + \lambda I_d \preceq (w+\lambda)\cdot I_d$$

for any $h \in [H]$. It holds for any $h \in [H]$ that

$$2\log\bigg(\frac{\det(W_h^{(t+1)w+1,t})}{\det(W_h^{tw,t})}\bigg) \le 2d\log\big((w+\lambda)/\lambda\big). \tag{J.3}$$

Plugging (J.3) and (J.2) into (J.1), we conclude the proof of Lemma J.2. □

Lemma J.3. Let $\lambda' = 1$ in (4.9). For any $\zeta \in (0, 1]$, the event $\mathcal{E}$ that, for any $(k, h) \in [K] \times [H]$,

$$\bigg\|\sum_{\tau=1}^{k-1} \eta_h^\tau(s_h^\tau, a_h^\tau)\cdot\big(V_{h+1}^k(s_{h+1}^\tau) - (\mathbb{P}_h^\tau V_{h+1}^k)(s_h^\tau, a_h^\tau)\big)\bigg\|_{(A_h^k)^{-1}} \le C''\sqrt{dH^2\cdot\log(dT/\zeta)} \tag{J.4}$$

happens with probability at least $1 - \zeta/2$, where $C'' > 0$ is an absolute constant that is independent of C.

Proof. See Lemma B.3 of Jin et al. (2019b) or Lemma D.1 of Cai et al. (2019) for a detailed proof. □

Lemma J.4 (Pinsker's inequality)**.** Denote $s \in \{s_1, s_2, \cdots, s_T\} \in \mathcal{S}$ be the observed states from step 1 to T. For any two distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ over $\mathcal{S}$ and any bounded function $f : \mathcal{S}^\top \to [0, B]$, we have

$$\mathbb{E}_1 f(s) - \mathbb{E}_2 f(s) \le \frac{\sqrt{\log 2}B}{2}\cdot\sqrt{\mathrm{KL}(\mathcal{P}_2\|\mathcal{P}_1)},$$

where $\mathbb{E}_1$ and $\mathbb{E}_2$ denote expectations with respect to $\mathcal{P}_1$ and $\mathcal{P}_2$.

Proof. See Lemma 13 in Jaksch et al. (2010) or Lemma B.4 in Zhou et al. (2020a) for a detailed proof. □

Lemma J.5. Suppose ξ and ξ' have the same entries except for j-th coordinate. We also assume that $2\epsilon \le \delta \le 1/3$, then we have

$$\mathrm{KL}(\mathcal{P}_{\xi'}\|\mathcal{P}_\xi) \le \frac{16\epsilon^2}{(d-1)^2\delta}\mathbb{E}_\xi N_0.$$

Proof. See Lemma 6.8 in Zhou et al. (2020a) for a detailed proof. □

Lemma J.6. For any $(h, k') \in [H] \times [K]$, $\{k_j\}_{j=1}^{h-1} \in [K]$, $j \in [h-1]$, $(s_1, s_h) \in \mathcal{S} \times \mathcal{S}$, and policies $\{\pi^i\}_{i \in [H]} \cup \{\pi'\}$, we have

$$\begin{aligned}
&|P_1^{k_1,\pi(1)} \cdots P_j^{k_j,\pi(j)} \cdots P_{h-1}^{k_{h-1},\pi(h-1)}(s_h \,|\, s_1) - P_1^{k_1,\pi(1)} \cdots P_j^{k',\pi'} \cdots P_{h-1}^{k_{h-1},\pi(h-1)}(s_h \,|\, s_1)| \\
&\quad \le \|\pi_j^{(j)} - \pi_j'\|_{\infty,1} + \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|P_j^{k_j}(\cdot \,|\, s_j, a) - P_j^{k'}(\cdot \,|\, s, a)\|_1.
\end{aligned}$$

Proof. First, we have

$$\begin{aligned}
&|P_1^{k_1,\pi(1)} \cdots P_j^{k_j,\pi(j)} \cdots P_{h-1}^{k_{h-1},\pi(h-1)}(s_h \,|\, s_1) - P_1^{k_1,\pi(1)} \cdots P_j^{k',\pi'} \cdots P_{h-1}^{k_{h-1},\pi(h-1)}(s_h \,|\, s_1)| \\
&\quad \le |P_1^{k_1,\pi(1)} \cdots P_j^{k_j,\pi(j)} \cdots P_{h-1}^{k_{h-1},\pi(h-1)}(s_h \,|\, s_1) - P_1^{k_1,\pi(1)} \cdots P_j^{k',\pi(j)} \cdots P_{h-1}^{k_{h-1},\pi(h-1)}(s_h \,|\, s_1)| \\
&\qquad + |P_1^{k_1,\pi(1)} \cdots P_j^{k',\pi(j)} \cdots P_{h-1}^{k_{h-1},\pi(h-1)}(s_h \,|\, s_1) - P_1^{k_1,\pi(1)} \cdots P_j^{k',\pi'} \cdots P_{h-1}^{k_{h-1},\pi(h-1)}(s_h \,|\, s_1)|.
\end{aligned} \tag{J.5}$$

By the definition of Markov kernel, we have

$$\begin{aligned}
&|P_1^{k_1,\pi(1)} \cdots P_j^{k_j,\pi(j)} \cdots P_{h-1}^{k_{h-1},\pi(h-1)}(s_h \,|\, s_1) - P_1^{k_1,\pi(1)} \cdots P_j^{k',\pi(j)} \cdots P_{h-1}^{k_{h-1},\pi(h-1)}(s_h \,|\, s_1)| \qquad \text{(J.6)} \\
&\quad \le \sum_{s_2,s_3,\cdots,s_{h-1}} |P_j^{k_j,\pi(j)}(s_{j+1} \,|\, s_j) - P_j^{k',\pi(j)}(s_{j+1} \,|\, s_j)| \cdot \prod_{i\in[h-1]\backslash j} P_i^{k_i,\pi(i)}(s_{i+1} \,|\, s_i) \\
&\quad \le \sum_{s_2,\cdots,s_j,s_{j+2},\cdots,s_{h-1}} \sum_{s_{j+1}} |P_j^{k_j,\pi(j)}(s_{j+1} \,|\, s_j) - P_j^{k',\pi(j)}(s_{j+1} \,|\, s_j)| \cdot \max_{s_{j+1}\in\mathcal{S}} \prod_{i\in[h-1]\backslash j} P_i^{k_i,\pi(i)}(s_{i+1} \,|\, s_i) \\
&\quad \le \sum_{s_2,\cdots,s_{j-1},s_{j+2},\cdots,s_{h-1}} \max_{s_j\in\mathcal{S}} \sum_{s_{j+1}} |P_j^{k_j,\pi(j)}(s_{j+1} \,|\, s_j) - P_j^{k',\pi(j)}(s_{j+1} \,|\, s_j)| \cdot \sum_{s_j} \max_{s_{j+1}\in\mathcal{S}} \prod_{i\in[h-1]\backslash j} P_i^{k_i,\pi(i)}(s_{i+1} \,|\, s_i),
\end{aligned}$$

where the last two inequalities is obtained by Hölder's inequality. By the definition of Markov kernel, we further have

$$\begin{aligned}
|P_j^{k_j,\pi(j)}(s_{j+1} \,|\, s_j) - P_j^{k',\pi(j)}(s_{j+1} \,|\, s_j)| &= \Big|\sum_a \pi^{(j)}(a \,|\, s_j)\big(P_j^{k_j}(s_{j+1} \,|\, s_j, a) - P_j^{k_j}(s_{j+1} \,|\, s_j, a)\big)\Big| \\
&\le \max_a |P_j^{k_j}(s_{j+1} \,|\, s_j, a) - P_j^{k_j}(s_{j+1} \,|\, s_j, a)|.
\end{aligned} \tag{J.7}$$

Hence, we obtain

$$\begin{aligned}
&|P_1^{k_1,\pi(1)} \cdots P_j^{k_j,\pi(j)} \cdots P_{h-1}^{k_{h-1},\pi(h-1)}(s_h \,|\, s_1) - P_1^{k_1,\pi(1)} \cdots P_j^{k',\pi(j)} \cdots P_{h-1}^{k_{h-1},\pi(h-1)}(s_h \,|\, s_1)| \\
&\quad \le \max_{(s_j,a)\in\mathcal{S}\times\mathcal{A}} \|P_j^{k_j}(\cdot \,|\, s_j, a) - P_j^{k'}(\cdot \,|\, s_j, a)\|_1 \cdot \sum_{s_2,\cdots,s_j,s_{j+2},\cdots,s_{h-1}} \max_{s_{j+1}\in\mathcal{S}} \prod_{i\in[h-1]\backslash j} P_i^{k_i,\pi(i)}(s_{i+1} \,|\, s_i) \\
&\quad = \max_{(s_j,a)\in\mathcal{S}\times\mathcal{A}} \|P_j^{k_j}(\cdot \,|\, s_j, a) - P_j^{k'}(\cdot \,|\, s_j, a)\|_1 \\
&\qquad \times \sum_{s_{j+2},\cdots,s_{h-1}} \max_{s_{j+1}\in\mathcal{S}} \prod_{i=j+1}^{h-1} P_i^{k_i,\pi(i)}(s_{i+1} \,|\, s_i) \cdot \sum_{s_2,\cdots,s_j} \prod_{i=1}^{j-1} P_i^{k_i,\pi(i)}(s_{i+1} \,|\, s_i) \\
&\quad \le \max_{(s_j,a)\in\mathcal{S}\times\mathcal{A}} \|P_j^{k_j}(\cdot \,|\, s_j, a) - P_j^{k'}(\cdot \,|\, s_j, a)\|_1,
\end{aligned} \tag{J.8}$$

where the last inequality follows from the facts that $\sum_{s_{j+2},\cdots,s_{h-1}} \max_{s_{j+1}\in\mathcal{S}} \prod_{i=j+1}^{h-1} P_i^{k_i,\pi(i)}(s_{i+1} \,|\, s_i) \le 1$ and $\sum_{s_2,\cdots,s_j} \prod_{i=1}^{j-1} P_i^{k_i,\pi(i)}(s_{i+1} \,|\, s_i) \le 1$. Moreover, by Lemma 5 in Fei et al. (2020), we have

$$\begin{aligned}
&|P_1^{k_1,\pi(1)} \cdots P_j^{k',\pi(j)} \cdots P_{h-1}^{k_{h-1},\pi(h-1)}(s_h \,|\, s_1) - P_1^{k_1,\pi(1)} \cdots P_j^{k',\pi'} \cdots P_{h-1}^{k_{h-1},\pi(h-1)}(s_h \,|\, s_1)| \\
&\quad \le \|\pi_j^{(j)} - \pi_j'\|_{\infty,1}.
\end{aligned} \tag{J.9}$$

Plugging (J.8) and (J.9) into (J.5), we conclude the proof. □