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

C PROOF OF THEOREM 3.1

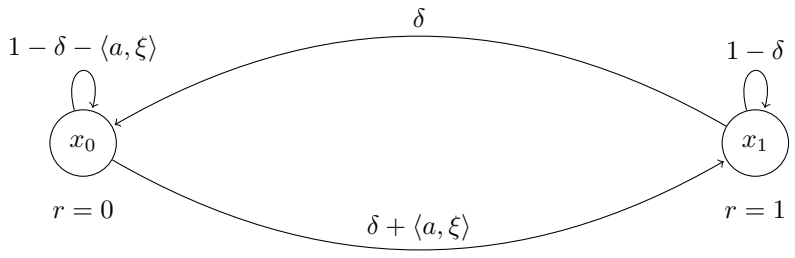

Figure 1: The hard-to-learn linear kernel MDP constructed in the proof of Theorem 3.1. Note that the probability of state x_0 to state x_1 depends on the choice of action a.

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

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

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

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

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

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

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

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

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

$$\mathbb{E}_\xi N_1 \le T_0/2 + \delta^{-1} \cdot \sum_a \langle a, \xi \rangle \cdot \mathbb{E}_\xi N_0^a / 2. \tag{C.6}$$

Hence, we have

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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. □

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

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

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

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

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

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

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 $l_h^k\le 0$ 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

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

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. □

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

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

By Lemma D.1, we have

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned}
&\lambda_{\max}\Bigg(\Big(\sum_{\tau=1\vee(k-w)}^{i}\phi(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)^\top\Big)(\Lambda_h^k)^{-2}\Big(\sum_{\tau=1\vee(k-w)}^{i}\phi(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)^\top\Big)\Bigg)\\
&=\lambda_{\max}\big(\Psi M_2\Psi^\top(\Psi M_1\Psi^\top)^{-2}\Psi M_2\Psi^\top\big)=\lambda_{\max}(M_2M_1^{-2}M_2). \qquad \text{(F.17)}
\end{aligned}$$

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 $M_2M_1^{-2}M_2=\mathrm{diag}(\lambda_1^2/(\lambda_1+\lambda)^2,\cdots,\lambda_d^2/(\lambda_d+\lambda)^2)$, which further implies that

$$\lambda_{\max}(M_2M_1^{-2}M_2)\le 1. \qquad \text{(F.18)}$$

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

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

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

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

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

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

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

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

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

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

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

Term (ii.1): We can decompose Term(ii.1) as

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

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

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

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

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

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

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

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

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

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

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

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:

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

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

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

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

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

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 $|l_h^k(s,a)| \le 2H$. Hence,

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

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

By Cauchy-Schwarz inequality, we have

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

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

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

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

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

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. □

H Proof of Theorem 4.3

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

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

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

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

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

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

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

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

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

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

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 $\big\|\theta_h^\tau - \theta_h^k\big\|_2 \le \big\|\sum_{i=\tau}^{k-1}(\theta_h^i - \theta_h^{i+1})\big\|_2 \le \sum_{i=\tau}^{k-1}\big\|\theta_h^i - \theta_h^{i+1}\big\|_2$. 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

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

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

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

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

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

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

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

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

$$\log\bigg(\frac{\det(\Lambda_{t+1})}{\det(\Lambda_1)}\bigg) \le \sum_{j=1}^{t}\phi_j^\top\Lambda_j^{-1}\phi_j \le 2\log\bigg(\frac{\det(\Lambda_{t+1})}{\det(\Lambda_1)}\bigg).$$

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

$$\sum_{k=1}^{K} 1\wedge\|\phi(s_h^k,a_h^k)\|_{(\Lambda_h^k)^{-1}} \le 2d\lceil K/w\rceil\log\big((w+\lambda)/\lambda\big)$$

for any $h \in [H]$.

Proof. First, we rewrite the sums as follows.

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

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 $(k,h) \in [K] \times [H]$,

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

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

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

Proof. First, we have

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

By the definition of Markov kernel, we have

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

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

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

Hence, we obtain

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

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

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

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