Supplementary Materials for "Near-Optimal Model-Free Reinforcement Learning in Non-Stationary Episodic MDPs"

A. Applications to Sequential Transfer, Multi-Task, and Multi-Agent RL

One area that could benefit from non-stationary RL is sequential transfer in RL (Tirinzoni et al., 2020) or multi-task RL (Brunskill & Li, 2013), which itself is conceptually related to continual RL (Kaplanis et al., 2018) and life-long RL (Abel et al., 2018). In the setting of sequential transfer/multi-task RL, the agent encounters a sequence of tasks over time with different system dynamics, and seeks to bootstrap learning by transferring knowledge from previously-solved tasks. Typical solutions in this area (Brunskill & Li, 2013; Tirinzoni et al., 2020; Sun et al., 2020) need to assume that there are *finitely many* candidate tasks, and every task should be *sufficiently different* from the others[5]. Only under this assumption can the agent quickly identify the current task it is operating on, by essentially comparing the system dynamics it observes with the dynamics it has memorized for each candidate task. After identifying the current task with high confidence, the agent then invokes the policy that it learned through previous interactions with this specific task. This transfer learning paradigm in turn causes another problem—it "cold switches" between policies that are most likely very different, which might lead to unstable and inconsistent behaviors of the agent over time. Fortunately, non-stationary RL can help alleviate both the finite-task assumption and the cold-switching problem. First, non-stationary RL algorithms do not need the candidate tasks to be sufficiently different in order to correctly identify each of them, because the algorithm itself can tolerate some variations in the task environment. There will also be no need to assume the finiteness of the candidate task set anymore, and the candidate tasks can be drawn from a continuous space. Second, since we are running the same non-stationary RL algorithm for a series of tasks, it improves its policy gradually over time, instead of cold-switching to a completely independent policy for each task. This could largely help with the unstable behavior issues.

Multi-agent reinforcement learning (MARL) (Littman, 1994) studies the problem where a set of agents collaborate or compete in a shared environment. In MARL, since the transition and reward functions of the agents are coupled, the environment is non-stationary from each agent's own perspective, especially when the agents learn and update their policies simultaneously. See Zhang et al. (2019) for a more detailed discussion. The non-stationarity in MARL is a setting where non-stationary RL can play a role. As advocated earlier in Bowling & Veloso (2001); Busoniu et al. (2008), a good MARL algorithm should be both *rational* and *convergent*, where the former means that the algorithm converges to its opponent's *best response* if its opponent converges to a stationary policy, and the latter means that if all agents use the same algorithm, the algorithm converges to a stationary policy. As such, a non-stationary RL algorithm can be viewed as a *rational* MARL algorithm, thanks to its dynamic regret guarantees, although its *convergence* property in MARL settings is still worth further investigation. In fact, developing algorithms that are both *rational* and *convergent* in general MARL settings is still relatively open. In addition, non-stationary RL algorithms also apply to the MARL setting to achieve low regret against *slowly-changing* opponents (Lee et al., 2020), which is discussed in detail in our Section 7. Finally, dynamic regret is also pertinent to the notion of *exploitability* of strategies in two-player zero-sum games (Davis et al., 2014).

[5] Needless to say, this assumption itself also to some extent contradicts the primary motivation of transfer learning. After all, we only want to transfer knowledge among tasks that are essentially similar to each other.

B. Proofs of the Technical Lemmas

B.1. Proof of Lemma 1

Proof. For each $d \in [D]$, define $\Delta_r^{(d)}$ to be the *local variation* of the mean reward function within epoch d. By definition, we have $\sum_{d=1}^{D} \Delta_r^{(d)} \leq \Delta_r$. Further, for each $d \in [D]$ and $h \in [H]$, define $\Delta_{r,h}^{(d)}$ to be the variation of the mean reward at step h in epoch d, i.e.,

$$\Delta_{r,h}^{(d)} \overset{\text{def}}{=} \sum_{m=(d-1)K+1}^{\min\{dK,M\}-1} \sup_{s,a} \left| r_h^m(s,a) - r_h^{m+1}(s,a) \right|.$$

It also holds that $\sum_{h=1}^{H} \Delta_{r,h}^{(d)} = \Delta_r^{(d)}$ by definition. Define $\Delta_p^{(d)}$ and $\Delta_{p,h}^{(d)}$ analogously.

In the following, we will prove a stronger statement: $\left| Q_h^{k_1,\star}(s,a) - Q_h^{k_2,\star}(s,a) \right| \leq \sum_{h'=h}^{H} \Delta_{r,h'}^{(1)} + H \sum_{h'=h}^{H} \Delta_{p,h'}^{(1)}$, which implies the statement of the lemma because $\sum_{h'=h}^{H} \Delta_{r,h'}^{(1)} \leq \Delta_r^{(1)}$ and $\sum_{h'=h}^{H} \Delta_{p,h'}^{(1)} \leq \Delta_p^{(1)}$ by definition. Our proof relies on backward induction on h. First, the statement holds for $h = H$ because for any (s,a), by definition

$$\begin{aligned} \left| Q_H^{k_1,\star}(s,a) - Q_H^{k_2,\star}(s,a) \right| = \left| r_H^{k_1}(s,a) - r_H^{k_2}(s,a) \right| &\leq \sum_{k=k_1}^{k_2-1} \left| r_H^{k+1}(s,a) - r_H^{k}(s,a) \right| \\ &\leq \sum_{k=1}^{K-1} \left| r_H^{k+1}(s,a) - r_H^{k}(s,a) \right| \leq \Delta_{r,H}^{(1)}, \end{aligned} \tag{2}$$

where we have used the triangle inequality. Now suppose the statement holds for $h+1$; by the Bellman optimality equation,

$$\begin{aligned} & Q_h^{k_1,\star}(s,a) - Q_h^{k_2,\star}(s,a) \\ =& P_h^{k_1} V_{h+1}^{k_1,\star}(s,a) - P_h^{k_2} V_{h+1}^{k_2,\star}(s,a) + r_h^{k_1}(s,a) - r_h^{k_2}(s,a) \\ \leq& P_h^{k_1} V_{h+1}^{k_1,\star}(s,a) - P_h^{k_2} V_{h+1}^{k_2,\star}(s,a) + \Delta_{r,h}^{(1)} \qquad (3) \\ =& \sum_{s' \in \mathcal{S}} P_h^{k_1}(s' \mid s,a) V_{h+1}^{k_1,\star}(s') - \sum_{s' \in \mathcal{S}} P_h^{k_2}(s' \mid s,a) V_{h+1}^{k_2,\star}(s') + \Delta_{r,h}^{(1)} \\ =& \sum_{s' \in \mathcal{S}} \Big(P_h^{k_1}(s' \mid s,a) Q_{h+1}^{k_1,\star}(s', \pi_{h+1}^{k_1,\star}(s')) - P_h^{k_2}(s' \mid s,a) Q_{h+1}^{k_2,\star}(s', \pi_{h+1}^{k_2,\star}(s')) \Big) + \Delta_{r,h}^{(1)}, \qquad (4) \end{aligned}$$

where inequality (3) holds due to a similar reasoning as in (2), and in (4) $\pi^{k_1,\star}$ and $\pi^{k_2,\star}$ denote the optimal policy in episode k_1 and k_2, respectively. Then by our induction hypothesis on $h+1$, for any $s' \in \mathcal{S}$,

$$\begin{aligned} Q_{h+1}^{k_1,\star}(s', \pi_{h+1}^{k_1,\star}(s')) \leq& Q_{h+1}^{k_2,\star}(s', \pi_{h+1}^{k_1,\star}(s')) + \sum_{h'=h+1}^{H} \Delta_{r,h'}^{(1)} + H \sum_{h'=h+1}^{H} \Delta_{p,h'}^{(1)} \\ \leq& Q_{h+1}^{k_2,\star}(s', \pi_{h+1}^{k_2,\star}(s')) + \sum_{h'=h+1}^{H} \Delta_{r,h'}^{(1)} + H \sum_{h'=h+1}^{H} \Delta_{p,h'}^{(1)}, \end{aligned} \tag{5}$$

where inequality (5) is due to the optimality of the policy $\pi^{k_2,\star}$ in episode k_2 over $\pi^{k_1,\star}$. Then,

$$Q_h^{k_1,\star}(s,a) - Q_h^{k_2,\star}(s,a)$$

$$\leq \sum_{s' \in \mathcal{S}} (P_h^{k_1}(s' \mid s,a) - P_h^{k_2}(s' \mid s,a)) Q_{h+1}^{k_2,\star}(s', \pi_{h+1}^{k_2,\star}(s')) + \sum_{h'=h}^{H} \Delta_{r,h'}^{(1)} + H \sum_{h'=h+1}^{H} \Delta_{p,h'}^{(1)}$$

$$\leq \left\| P_h^{k_1}(\cdot|s,a) - P_h^{k_2}(\cdot|s,a) \right\|_1 \left\| Q_{h+1}^{k_2,\star}(\cdot, \pi_{h+1}^{k_2,\star}(\cdot)) \right\|_\infty + \sum_{h'=h}^{H} \Delta_{r,h'}^{(1)} + H \sum_{h'=h+1}^{H} \Delta_{p,h'}^{(1)} \tag{6}$$

$$\leq \Delta_{p,h}^{(1)}(H-h) + \sum_{h'=h}^{H} \Delta_{r,h'}^{(1)} + H \sum_{h'=h+1}^{H} \Delta_{p,h'}^{(1)} \tag{7}$$

$$\leq \sum_{h'=h}^{H} \Delta_{r,h'}^{(1)} + H \sum_{h'=h}^{H} \Delta_{p,h'}^{(1)},$$

where (6) is by Hölder's inequality, and (7) is by the definition of $\Delta_{p,h}^{(1)}$ and by the definition of optimal Q-values that $Q_{h+1}^{k_2,\star}(s,a) \leq H-h, \forall (s,a) \in \mathcal{S} \times \mathcal{A}$. Repeating a similar process gives us $Q_h^{k_2,\star}(s,a) - Q_h^{k_1,\star}(s,a) \leq \sum_{h'=h}^{H} \Delta_{r,h'}^{(1)} + H \sum_{h'=h}^{H} \Delta_{p,h'}^{(1)}$. This completes our proof. □

B.2. Proof of Lemma 2

Proof. It should be clear from the way we update $Q_h(s,a)$ that $Q_h^k(s,a)$ is monotonically decreasing in k. We now prove $Q_h^{k,\star}(s,a) \leq Q_h^{k+1}(s,a)$ for all s,a,h,k by induction on k. First, it holds for $k=1$ by our initialization of $Q_h(s,a)$. For $k \geq 2$, now suppose $Q_h^{j,\star}(s,a) \leq Q_h^{j+1}(s,a) \leq Q_h^j(s,a)$ for all s,a,h and $1 \leq j \leq k$. For a fixed triple (s,a,h), we consider the following two cases.

Case 1: $Q_h(s,a)$ is updated in episode k. Then with probability at least $1-2\delta$

$$Q_h^{k+1}(s,a) = \frac{\check{r}_h(s,a)}{\check{N}_h^k(s,a)} + \frac{\check{v}_h(s,a)}{\check{N}_h^k(s,a)} + b_h^k + 2b_\Delta$$

$$\geq \frac{\check{r}_h(s,a)}{\check{n}} + \frac{1}{\check{n}} \sum_{i=1}^{\check{n}} V_{h+1}^{\check{l}_i,\star}(s_{h+1}^{\check{l}_i}) + \sqrt{\frac{H^2}{\check{n}}\iota} + \sqrt{\frac{\iota}{\check{n}}} + 2b_\Delta \tag{8}$$

$$\geq \frac{\check{r}_h(s,a)}{\check{n}} + \frac{1}{\check{n}} \sum_{i=1}^{\check{n}} P_h^{\check{l}_i} V_{h+1}^{\check{l}_i,\star}(s,a) + \sqrt{\frac{\iota}{\check{n}}} + 2b_\Delta \tag{9}$$

$$= \frac{\check{r}_h(s,a)}{\check{n}} + \frac{1}{\check{n}} \sum_{i=1}^{\check{n}} \left(Q_h^{\check{l}_i,\star}(s,a) - r_h^{\check{l}_i}(s,a) \right) + \sqrt{\frac{\iota}{\check{n}}} + 2b_\Delta \tag{10}$$

$$\geq Q_h^{k,\star}(s,a) + b_\Delta. \tag{11}$$

Inequality (8) is by the induction hypothesis that $Q_{h+1}^{\check{l}_i}(s_{h+1}^{\check{l}_i}, a) \geq Q_{h+1}^{\check{l}_i,\star}(s_{h+1}^{\check{l}_i}, a), \forall a \in \mathcal{A}$, and hence $V_{h+1}^{\check{l}_i}(s_{h+1}^{\check{l}_i}) \geq V_{h+1}^{\check{l}_i,\star}(s_{h+1}^{\check{l}_i})$. Inequality (9) follows from the Azuma-Hoeffding inequality. (10) uses the Bellman optimality equation. Inequality (11) is by the Hoeffding's inequality that $\frac{1}{\check{n}} \left(\sum_{i=1}^{\check{n}} r_h^{\check{l}_i}(s,a) - \check{r}_h(s,a) \right) \leq \sqrt{\frac{\iota}{\check{n}}}$ with high probability, and by Lemma 1 that $Q_h^{\check{l}_i,\star}(s,a) + b_\Delta \geq Q_h^{k,\star}(s,a)$. According to the monotonicity of $Q_h^k(s,a)$, we know that $Q_h^{k,\star}(s,a) \leq Q_h^{k+1}(s,a) \leq Q_h^k(s,a)$. In fact, we have proved the stronger statement $Q_h^{k+1}(s,a) \geq Q_h^{k,\star}(s,a) + b_\Delta$ that will be useful in Case 2 below.

Case 2: $Q_h(s,a)$ is not updated in episode k. Then there are two possibilities:

1. If $Q_h(s,a)$ has never been updated from episode 1 to episode k: It is easy to see that $Q_h^{k+1}(s,a) = Q_h^k(s,a) = \cdots = Q_h^1(s,a) = H-h+1 \geq Q_h^{k,\star}(s,a)$ holds.

2. If $Q_h(s,a)$ has been updated at least once from episode 1 to episode k: Let j be the index of the latest episode that $Q_h(s,a)$ was updated. Then, from our induction hypothesis and Case 1, we know that $Q_h^{j+1}(s,a) \geq Q_h^{j,\star}(s,a) + b_\Delta$. Since $Q_h(s,a)$ has not been updated from episode $j+1$ to episode k, we know that $Q_h^{k+1}(s,a) = Q_h^k(s,a) = \cdots = Q_h^{j+1}(s,a) \geq Q_h^{j,\star}(s,a) + b_\Delta \geq Q_h^{k,\star}(s,a)$, where the last inequality holds because of Lemma 1.

A union bound over all time steps completes our proof. □

B.3. Proof of Lemma 7

Proof. It holds that

$$\sum_{k=1}^{K} \frac{1}{\check{n}_h^k} \sum_{i=1}^{\check{n}_h^k} \delta_{h+1}^{\check{l}_{h,i}^k} = \sum_{k=1}^{K} \sum_{j=1}^{K} \frac{1}{\check{n}_h^k} \delta_{h+1}^{j} \sum_{i=1}^{\check{n}_h^k} \mathbb{I}\left[\check{l}_{h,i}^k = j\right] = \sum_{j=1}^{K} \delta_{h+1}^{j} \sum_{k=1}^{K} \frac{1}{\check{n}_h^k} \sum_{i=1}^{\check{n}_h^k} \mathbb{I}\left[\check{l}_{h,i}^k = j\right]. \tag{12}$$

For a fixed episode j, notice that $\sum_{i=1}^{\check{n}_h^k} \mathbb{I}[\check{l}_{h,i}^k = j] \leq 1$, and that $\sum_{i=1}^{\check{n}_h^k} \mathbb{I}[\check{l}_{h,i}^k = j] = 1$ happens if and only if $(s_h^k, a_h^k) = (s_h^j, a_h^j)$ and (j,h) lies in the previous stage of (k,h) with respect to the triple (s_h^k, a_h^k, h). Let $\mathcal{K} \overset{\text{def}}{=} \{k \in [K] : \sum_{i=1}^{\check{n}_h^k} \mathbb{I}[\check{l}_{h,i}^k = j] = 1\}$; then, we know that every element $k \in \mathcal{K}$ has the same value of $\check{n}_h^k$, i.e., there exists an integer $N_j > 0$, such that $\check{n}_h^k = N_j, \forall k \in \mathcal{K}$. Further, by our definition of the stages, we know that $|\mathcal{K}| \leq (1 + \frac{1}{H})N_j$, because the current stage is at most $(1 + \frac{1}{H})$ times longer than the previous stage. Therefore, for every j, we know that

$$\sum_{k=1}^{K} \frac{1}{\check{n}_h^k} \sum_{i=1}^{\check{n}_h^k} \mathbb{I}\left[\check{l}_{h,i}^k = j\right] \leq 1 + \frac{1}{H}. \tag{13}$$

Combining (12) and (13) completes the proof of $\sum_{k=1}^{K} \frac{1}{\check{n}_h^k} \sum_{i=1}^{\check{n}_h^k} \delta_{h+1}^{\check{l}_{h,i}^k} \leq (1 + \frac{1}{H}) \sum_{k=1}^{K} \delta_{h+1}^k$. □

B.4. Proof of Proposition 2

In the following, we will bound each term in Λ_{h+1}^k separately in a series of lemmas.

Lemma 4. *With probability* 1, *we have that*

$$\sum_{h=1}^{H} \sum_{k=1}^{K} (1 + \frac{1}{H})^{h-1} (3b_h^k + 5b_\Delta) \leq O(\sqrt{SAKH^5\iota} + KH\Delta_r^{(1)} + KH^2\Delta_p^{(1)}).$$

Proof. First, by the definition of b_Δ, it is easy to see that $\sum_{h=1}^{H} \sum_{k=1}^{K} (1 + \frac{1}{H})^{h-1} 5b_\Delta \leq \sum_{h=1}^{H} \sum_{k=1}^{K} O(\Delta_r^{(1)} + H\Delta_p^{(1)}) \leq O(KH\Delta_r^{(1)} + KH^2\Delta_p^{(1)})$. Recall our definition that $e_1 = H$ and $e_{i+1} = \left\lfloor (1 + \frac{1}{H}) e_i \right\rfloor, i \geq 1$. For a fixed $h \in [H]$, since $H^2 \geq 1$,

$$\begin{aligned}
&\sum_{k=1}^{K} (1 + \frac{1}{H})^{h-1} 3b_h^k \leq \sum_{k=1}^{K} (1 + \frac{1}{H})^{h-1} 12 \sqrt{\frac{H^2}{\check{N}_h^k(s_h^k, a_h^k)} \iota} \\
=&12H\sqrt{\iota} \sum_{s,a} \sum_{j \geq 1} (1 + \frac{1}{H})^{h-1} \sqrt{\frac{1}{e_j}} \sum_{k=1}^{K} \mathbb{I}\left[(s_h^k, a_h^k) = (s,a), \check{N}_h^k(s_h^k, a_h^k) = e_j\right] \\
=&12H\sqrt{\iota} \sum_{s,a} \sum_{j \geq 1} (1 + \frac{1}{H})^{h-1} w(s,a,j) \sqrt{\frac{1}{e_j}},
\end{aligned}$$

where $w(s,a,j) \overset{\text{def}}{=} \sum_{k=1}^{K} \mathbb{I}\left[(s_h^k, a_h^k) = (s,a), \check{N}_h^k(s_h^k, a_h^k) = e_j\right]$, and $w(s,a) \overset{\text{def}}{=} \sum_{j\geq 1} w(s,a,j)$. We then know that $\sum_{s,a} w(s,a) = K$. For a fixed (s,a), let us now find an upper bound of j, denoted as J. Since each stage is $(1+\frac{1}{H})$ times longer than the previous stage, we know for $1 \leq j \leq J$, $w(s,a,j) = \sum_{k=1}^{K} \mathbb{I}\left[(s_h^k, a_h^k) = (s,a), \check{N}_h^k(s_h^k, a_h^k) = e_j\right] = \left\lfloor (1+\frac{1}{H}) e_j \right\rfloor$. From $\sum_{j=1}^{J} w(s,a,j) = w(s,a)$, we get $e_J \leq (1+\frac{1}{H})^{J-1} \leq \frac{10}{1+\frac{1}{H}} \frac{w(s,a)}{H}$. Therefore,

$$\sum_{j\geq 1} (1+\frac{1}{H})^{h-1} w(s,a,j) \sqrt{\frac{1}{e_j}} \leq O\left(\sum_{j=1}^{J} \sqrt{e_j}\right) \leq O\left(\sqrt{w(s,a)H}\right).$$

Finally, by the Cauchy-Schwartz inequality, we have

$$\sum_{h=1}^{H}\sum_{k=1}^{K} (1+\frac{1}{H})^{h-1} 3 b_h^k = O\left(H^2\sqrt{\iota} \sum_{s,a}\sum_{j\geq 1} w(s,a,j)\sqrt{\frac{1}{e_j}}\right) \leq \sqrt{SAKH^5\iota}.$$

Combining the bounds for b_h^k and b_Δ completes the proof. □

Lemma 5. *With probability at least* $1-\delta$*, it holds that*

$$\sum_{h=1}^{H}\sum_{k=1}^{K} (1+\frac{1}{H})^{h-1} \phi_{h+1}^k \leq O(\sqrt{KH^3\iota} + KH\Delta_r^{(1)} + KH^2\Delta_p^{(1)}).$$

Proof. We have that

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K} (1+\frac{1}{H})^{h-1} \phi_{h+1}^k \\
=&\sum_{h=1}^{H}\sum_{k=1}^{K} (1+\frac{1}{H})^{h-1} \frac{1}{\check{n}} \sum_{i=1}^{\check{n}} \left(P_h^k - \mathbf{e}_{s_{h+1}^k}\right)\left(V_{h+1}^{\check{l}_i,\star} - V_{h+1}^{k,\pi}\right)(s_h^k, a_h^k) \\
=&\sum_{h=1}^{H}\sum_{k=1}^{K} (1+\frac{1}{H})^{h-1} \frac{1}{\check{n}} \sum_{i=1}^{\check{n}} \left(P_h^k - \mathbf{e}_{s_{h+1}^k}\right)\left(V_{h+1}^{\check{l}_i,\star} - V_{h+1}^{k,\star} + V_{h+1}^{k,\star} - V_{h+1}^{k,\pi}\right)(s_h^k, a_h^k) \\
\leq&\sum_{h=1}^{H}\sum_{k=1}^{K} (1+\frac{1}{H})^{h-1} 2b_\Delta + \sum_{h=1}^{H}\sum_{k=1}^{K} (1+\frac{1}{H})^{h-1} \left(P_h^k - \mathbf{e}_{s_{h+1}^k}\right)\left(V_{h+1}^{k,\star} - V_{h+1}^{k,\pi}\right)(s_h^k, a_h^k),
\end{aligned}$$

where the last inequality follows from Lemma 1 and the definition of b_Δ. From the proof of Lemma 4, we know that the first term can be bounded as

$$\sum_{h=1}^{H}\sum_{k=1}^{K} (1+\frac{1}{H})^{h-1} 2b_\Delta \leq O(KH\Delta_r^{(1)} + KH^2\Delta_p^{(1)}).$$

Further, the second term is bounded by the Azuma-Hoeffding inequality as

$$\sum_{h=1}^{H}\sum_{k=1}^{K} (1+\frac{1}{H})^{h-1} \left(P_h^k - \mathbf{e}_{s_{h+1}^k}\right)\left(V_{h+1}^{k,\star} - V_{h+1}^{k,\pi}\right)(s_h^k, a_h^k) \leq O(\sqrt{KH^3\iota}).$$

Combining the two terms completes the proof. □

Lemma 6. *With probability at least $1-(KH+1)\delta$, it holds that*

$$\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\xi_{h+1}^{k} \leq O(\sqrt{SAKH^3\iota}+KH^2\Delta_p^{(1)}).$$

Proof. We have that

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\xi_{h+1}^{k}\\
=&\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\frac{1}{\check{n}}\sum_{i=1}^{\check{n}}\left(P_h^k-\mathbf{e}_{s_{h+1}^{\check{l}_i}}\right)\left(V_{h+1}^{\check{l}_i}-V_{h+1}^{\check{l}_i,\star}\right)(s_h^k,a_h^k)\\
=&\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\frac{1}{\check{n}}\sum_{i=1}^{\check{n}}\left(P_h^k-P_h^{\check{l}_i}+P_h^{\check{l}_i}-\mathbf{e}_{s_{h+1}^{\check{l}_i}}\right)\left(V_{h+1}^{\check{l}_i}-V_{h+1}^{\check{l}_i,\star}\right)(s_h^k,a_h^k)\\
\leq&O(KH^2\Delta_p^{(1)})+\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\frac{1}{\check{n}}\sum_{i=1}^{\check{n}}\left(P_h^{\check{l}_i}-\mathbf{e}_{s_{h+1}^{\check{l}_i}}\right)\left(V_{h+1}^{\check{l}_i}-V_{h+1}^{\check{l}_i,\star}\right)(s_h^k,a_h^k), \qquad (14)
\end{aligned}$$

where the last step is by the fact that $V_{h+1}^{\check{l}_i}(s_h^k,a_h^k)\geq V_{h+1}^{\check{l}_i,\star}(s_h^k,a_h^k)$ from Lemma 2, and then by Hölder's inequality and the triangle inequality. The following proof is analogous to the proof of Lemma 15 in Zhang et al. (2020). For completeness we reproduce it here. We have

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\frac{1}{\check{n}}\sum_{i=1}^{\check{n}}\left(P_h^{\check{l}_i}-\mathbf{e}_{s_{h+1}^{\check{l}_i}}\right)\left(V_{h+1}^{\check{l}_i}-V_{h+1}^{\check{l}_i,\star}\right)(s_h^k,a_h^k)\\
=&\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{j=1}^{K}(1+\frac{1}{H})^{h-1}\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\mathbb{I}\left[\check{l}_{h,i}^k=j\right]\left(P_h^j-\mathbf{e}_{s_{h+1}^j}\right)\left(V_{h+1}^j-V_{h+1}^{j,\star}\right)(s_h^k,a_h^k)\\
=&\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{j=1}^{K}(1+\frac{1}{H})^{h-1}\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\mathbb{I}\left[\check{l}_{h,i}^k=j\right]\left(P_h^j-\mathbf{e}_{s_{h+1}^j}\right)\left(V_{h+1}^j-V_{h+1}^{j,\star}\right)(s_h^j,a_h^j), \qquad (15)
\end{aligned}$$

where (15) holds because $\check{l}_{h,i}^k(s_h^k,a_h^k)=j$ if and only if j is in the previous stage of k and $(s_h^k,a_h^k)=(s_h^j,a_h^j)$. For simplicity of notations, we define $\theta_{h+1}^k \overset{\text{def}}{=} (1+\frac{1}{H})^{h-1}\sum_{j=1}^K\frac{1}{\check{n}_h^j}\sum_{i=1}^{\check{n}_h^j}\mathbb{I}\left[\check{l}_{h,i}^j=k\right]$. Then we further have (note that we have swapped the notation of j and k)

$$(15)=\sum_{h=1}^{H}\sum_{k=1}^{K}\theta_{h+1}^k\left(P_h^k-\mathbf{e}_{s_{h+1}^k}\right)\left(V_{h+1}^k-V_{h+1}^{k,\star}\right)(s_h^k,a_h^k).$$

For $(k,h)\in[K]\times[H]$, let x_h^k denote the number of occurrences of the triple (s_h^k,a_h^k,h) in the current stage. Define $\tilde{\theta}_{h+1}^k \overset{\text{def}}{=} (1+\frac{1}{H})^{h-1}\frac{\lfloor(1+\frac{1}{H})x_h^k\rfloor}{x_h^k}\leq 3$. Define $\mathcal{K}\overset{\text{def}}{=}\{(k,h):\theta_{h+1}^k=\tilde{\theta}_{h+1}^k\}$, and $\bar{\mathcal{K}}\overset{\text{def}}{=}\{(k,h)\in[K]\times[H]:\theta_{h+1}^k\neq\tilde{\theta}_{h+1}^k\}$. Then, we have that

$$\begin{aligned}
(15)=&\sum_{h=1}^{H}\sum_{k=1}^{K}\tilde{\theta}_{h+1}^k\left(P_h^k-\mathbf{e}_{s_{h+1}^k}\right)\left(V_{h+1}^k-V_{h+1}^{k,\star}\right)(s_h^k,a_h^k)\\
&+\sum_{h=1}^{H}\sum_{k=1}^{K}(\theta_{h+1}^k-\tilde{\theta}_{h+1}^k)\left(P_h^k-\mathbf{e}_{s_{h+1}^k}\right)\left(V_{h+1}^k-V_{h+1}^{k,\star}\right)(s_h^k,a_h^k).
\end{aligned}$$

Since $\tilde{\theta}_{h+1}^{k}$ is independent of s_{h+1}^{k}, by the Azuma-Hoeffding inequality, it holds with probability at least $1-\delta$ that

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\tilde{\theta}_{h+1}^{k}\left(P_h^k-\mathbf{e}_{s_{h+1}^k}\right)\left(V_{h+1}^k-V_{h+1}^{k,\star}\right)(s_h^k,a_h^k)\le O(\sqrt{KH^3\iota}). \tag{16}$$

It is easy to see that if k is in a stage that is before the second last stage of the triple (s_h^k, a_h^k, h), then $(k,h)\in\mathcal{K}$. For a triple (s,a,h), define $\mathcal{K}_h^{\perp}(s,a)\overset{\text{def}}{=}\{k\in[K]:$ k is in the second last stage of the triple $(s,a,h),\ (s_h^k,a_h^k)=(s,a)\}$. We have that

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K}(\theta_{h+1}^k-\tilde{\theta}_{h+1}^k)\left(P_h^k-\mathbf{e}_{s_{h+1}^k}\right)\left(V_{h+1}^k-V_{h+1}^{k,\star}\right)(s_h^k,a_h^k)\\
=&\sum_{s,a,h}\sum_{k:(k,h)\in\bar{\mathcal{K}}}\mathbb{I}\left[(s_h^k,a_h^k)=(s,a)\right](\theta_{h+1}^k-\tilde{\theta}_{h+1}^k)\left(P_h^k-\mathbf{e}_{s_{h+1}^k}\right)\left(V_{h+1}^k-V_{h+1}^{k,\star}\right)(s,a)\\
=&\sum_{s,a,h}(\theta_{h+1}(s,a)-\tilde{\theta}_{h+1}(s,a))\sum_{k\in\mathcal{K}_h^{\perp}(s,a)}\left(P_h^k-\mathbf{e}_{s_{h+1}^k}\right)\left(V_{h+1}^k-V_{h+1}^{k,\star}\right)(s,a), \qquad (17)
\end{aligned}$$

where for a fixed triple (s,a,h), we have defined $\theta_{h+1}(s,a)\overset{\text{def}}{=}\theta_{h+1}^k$, for any $k\in\mathcal{K}_h^{\perp}(s,a)$. Note that $\theta_{h+1}(s,a)$ is well-defined, because $\theta_{h+1}^{k_1}=\theta_{h+1}^{k_2},\forall k_1,k_2\in\mathcal{K}_h^{\perp}(s,a)$. Similarly, let $\tilde{\theta}_{h+1}(s,a)\overset{\text{def}}{=}\tilde{\theta}_{h+1}^k$ for any $k\in\mathcal{K}_h^{\perp}(s,a)$, and $\tilde{\theta}_{h+1}(s,a)$ is also well-defined. By the Azuma-Hoeffding inequality and a union bound, it holds with probability at least $1-KH\delta$ that

$$\begin{aligned}
(17)\le&\sum_{s,a,h}O\left(\sqrt{H^2\left|\mathcal{K}_h^{\perp}(s,a)\right|\iota}\right)\\
=&\sum_{s,a,h}O\left(\sqrt{H^2\check{N}_h^{K+1}(s,a)\iota}\right)\\
\le&O\left(\sqrt{SAH^3\iota\sum_{s,a,h}\check{N}_h^{K+1}(s,a)}\right) \qquad (18)\\
\le&O\left(\sqrt{SAKH^3\iota}\right) \qquad (19)
\end{aligned}$$

where $\check{N}_h^{K+1}(s,a)$ is defined to be the total number of visitations to the triple (s,a,h) over the entire K episodes. (18) is by the Cauchy-Schwartz inequality. (19) holds because by the way stages are defined, for each triple (s,a,h), the length of its last two stages is at most an $O(1/H)$ fraction of the total number of visitations.

Combining (14), (16) and (19) completes the proof. □

B.5. Proof of Lemma 3

Proof. This proof follows a similar structure as the proof of Lemma 2. It should be clear from the way we update $Q_h(s,a)$ that $Q_h^k(s,a)$ is monotonically decreasing in k. We now prove $Q_h^{k,\star}(s,a)-2(H-h+1)b_\Delta\le Q_h^{k+1}(s,a)$ for all s,a,h,k by induction on k. First, it holds for $k=1$ by our initialization of $Q_h(s,a)$. For $k\ge 2$, now suppose $Q_h^{j,\star}(s,a)-2(H-h+1)b_\Delta\le Q_h^{j+1}(s,a)\le Q_h^j(s,a)$ for all s,a,h and $1\le j\le k$. For a fixed triple (s,a,h), we consider the following two cases.

Case 1: $Q_h(s,a)$ is updated in episode k. Then with probability at least $1-2\delta$

$$Q_h^{k+1}(s,a)=\frac{\check{r}_h(s,a)}{\check{N}_h^k(s,a)}+\frac{\check{v}_h(s,a)}{\check{N}_h^k(s,a)}+b_h^k$$

$$\geq \frac{\check{r}_h(s,a)}{\check{n}} + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}} V_{h+1}^{\check{l}_i,\star}(s_{h+1}^{\check{l}_i}) - 2(H-h)b_\Delta + \sqrt{\frac{H^2}{\check{n}}\iota} + \sqrt{\frac{\iota}{\check{n}}} \tag{20}$$

$$\geq \frac{\check{r}_h(s,a)}{\check{n}} + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}} P_h^{\check{l}_i} V_{h+1}^{\check{l}_i,\star}(s,a) + \sqrt{\frac{\iota}{\check{n}}} - 2(H-h)b_\Delta \tag{21}$$

$$= \frac{\check{r}_h(s,a)}{\check{n}} + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}} \left(Q_h^{\check{l}_i,\star}(s,a) - r_h^{\check{l}_i}(s,a)\right) + \sqrt{\frac{\iota}{\check{n}}} - 2(H-h)b_\Delta \tag{22}$$

$$\geq Q_h^{k,\star}(s,a) - b_\Delta - 2(H-h)b_\Delta. \tag{23}$$

Inequality (20) is by the induction hypothesis that $Q_{h+1}^{\check{l}_i}(s_{h+1}^{\check{l}_i}, a) \geq Q_{h+1}^{\check{l}_i,\star}(s_{h+1}^{\check{l}_i}, a) - 2(H-h)b_\Delta, \forall a \in \mathcal{A}$, and hence $V_{h+1}^{\check{l}_i}(s_{h+1}^{\check{l}_i}) \geq V_{h+1}^{\check{l}_i,\star}(s_{h+1}^{\check{l}_i}) - 2(H-h)b_\Delta$. Inequality (21) follows from the Azuma-Hoeffding inequality. (22) uses the Bellman optimality equation. Inequality (23) is by the Hoeffding's inequality that $\frac{1}{\check{n}}\left(\sum_{i=1}^{\check{n}} r_h^{\check{l}_i}(s,a) - \check{r}_h(s,a)\right) \leq \sqrt{\frac{\iota}{\check{n}}}$ with high probability, and by Lemma 1 that $Q_h^{\check{l}_i,\star}(s,a) \geq Q_h^{k,\star}(s,a) - b_\Delta$. According to the monotonicity of $Q_h^k(s,a)$, we know that $Q_h^{k,\star}(s,a) - 2(H-h+1)b_\Delta \leq Q_h^{k+1}(s,a) \leq Q_h^k(s,a)$. In fact, we have proved the stronger statement $Q_h^{k+1}(s,a) \geq Q_h^{k,\star}(s,a) - b_\Delta - 2(H-h)b_\Delta$ that will be useful in Case 2 below.

Case 2: $Q_h(s,a)$ is not updated in episode k. Then there are two possibilities:

1. If $Q_h(s,a)$ has never been updated from episode 1 to episode k: It is easy to see that $Q_h^{k+1}(s,a) = Q_h^k(s,a) = \cdots = Q_h^1(s,a) = H - h + 1 \geq Q_h^{k,\star}(s,a) - 2(H-h+1)b_\Delta$ holds.

2. If $Q_h(s,a)$ has been updated at least once from episode 1 to episode k: Let j be the index of the latest episode that $Q_h(s,a)$ was updated. Then, from our induction hypothesis and Case 1, we know that $Q_h^{j+1}(s,a) \geq Q_h^{j,\star}(s,a) - b_\Delta - 2(H-h)b_\Delta$. Since $Q_h(s,a)$ has not been updated from episode $j+1$ to episode k, we know that $Q_h^{k+1}(s,a) = Q_h^k(s,a) = \cdots = Q_h^{j+1}(s,a) \geq Q_h^{j,\star}(s,a) - b_\Delta - 2(H-h)b_\Delta \geq Q_h^{k,\star}(s,a) - 2(H-h+1)b_\Delta$, where the last inequality holds because of Lemma 1.

A union bound over all time steps completes our proof. □

C. Proof of Theorem 1

We introduce a few terms to facilitate the analysis. Denote by s_h^k and a_h^k respectively the state and action taken at step h of episode k. Let $N_h^k(s,a), \check{N}_h^k(s,a), Q_h^k(s,a)$ and $V_h^k(s)$ denote, respectively, the values of $N_h(s,a), \check{N}_h(s,a), Q_h(s,a)$ and $V_h(s)$ at the *beginning* of the k-th episode in Algorithm 1. Further, for the triple (s_h^k, a_h^k, h), let n_h^k be the total number of episodes that this triple has been visited prior to the current stage, and let $l_{h,i}^k$ denote the index of the episode that this triple was visited the i-th time among the total n_h^k times. Similarly, let $\check{n}_h^k$ denote the number of visits to the triple (s_h^k, a_h^k, h) in the stage right before the current stage, and let $\check{l}_{h,i}^k$ be the i-th episode among the $\check{n}_h^k$ episodes right before the current stage. For simplicity, we use l_i and $\check{l}_i$ to denote $l_{h,i}^k$ and $\check{l}_{h,i}^k$, and $\check{n}$ to denote $\check{n}_h^k$, when h and k are clear from the context. We also use $\check{r}_h(s,a)$ and $\check{v}_h(s,a)$ to denote the values of $\check{r}_h(s_h^k, a_h^k)$ and $\check{v}_h(s_h^k, a_h^k)$ when updating the $Q_h(s_h^k, a_h^k)$ value in Line 12 of Algorithm 1.

We now proceed to analyze the dynamic regret in one epoch, and at the very end of this section, we will see how to combine the dynamic regret over all the epochs to prove Theorem 1. The following analysis will be conditioned on the successful event of Lemma 2.

The dynamic regret of Algorithm 1 in epoch $d = 1$ can hence be expressed as

$$\begin{aligned}\mathcal{R}^{(d)}(\pi, K) &= \sum_{k=1}^{K} \left(V_1^{k,*}\left(s_1^k\right) - V_1^{k,\pi}\left(s_1^k\right) \right) \\ &\leq \sum_{k=1}^{K} \left(V_1^{k}\left(s_1^k\right) - V_1^{k,\pi}\left(s_1^k\right) \right). \end{aligned} \tag{24}$$

From the update rules of the value functions in Algorithm 1, we have

$$\begin{aligned} V_h^k(s_h^k) &\leq \mathbb{I}\left[n_h^k = 0\right] H + \frac{\check{r}_h(s_h^k, a_h^k)}{\check{N}_h^k(s_h^k, a_h^k)} + \frac{\check{v}_h(s_h^k, a_h^k)}{\check{N}_h^k(s_h^k, a_h^k)} + b_h^k + 2b_\Delta \\ &= \mathbb{I}\left[n_h^k = 0\right] H + \frac{\check{r}_h(s_h^k, a_h^k)}{\check{N}_h^k(s_h^k, a_h^k)} + \frac{1}{\check{n}} \sum_{i=1}^{\check{n}} V_{h+1}^{\check{l}_i}(s_{h+1}^{\check{l}_i}) + b_h^k + 2b_\Delta. \end{aligned}$$

For ease of exposition, we define the following terms:

$$\delta_h^k \overset{\text{def}}{=} V_h^k(s_h^k) - V_h^{k,\star}(s_h^k),\ \zeta_h^k \overset{\text{def}}{=} V_h^k(s_h^k) - V_h^{k,\pi}(s_h^k). \tag{25}$$

We further define $\tilde{r}_h^k(s_h^k, a_h^k) \overset{\text{def}}{=} \frac{\check{r}_h(s_h^k, a_h^k)}{\check{N}_h^k(s_h^k, a_h^k)} - r_h^k(s_h^k, a_h^k)$. Then by the Hoeffding's inequality, it holds with high probability that

$$\begin{aligned} \tilde{r}_h^k(s_h^k, a_h^k) &\leq \frac{1}{\check{n}} \sum_{i=1}^{\check{n}} r_h^{\check{l}_i}(s_h^k, a_h^k) + \sqrt{\frac{\iota}{\check{n}}} - r_h^k(s_h^k, a_h^k) \\ &\leq b_h^k + b_\Delta. \end{aligned} \tag{26}$$

By the Bellman equation $V_h^{k,\pi}(s_h^k) = Q_h^{k,\pi}(s_h^k, \pi(s_h^k)) = r_h^k(s_h^k, a_h^k) + P_h^k V_{h+1}^{k,\pi}(s_h^k, a_h^k)$, we have

$$\begin{aligned} \zeta_h^k &\leq \mathbb{I}\left[n_h^k = 0\right] H + \frac{1}{\check{n}} \sum_{i=1}^{\check{n}} V_{h+1}^{\check{l}_i}(s_{h+1}^{\check{l}_i}) + b_h^k + 2b_\Delta + \tilde{r}_h^k(s_h^k, a_h^k) - P_h^k V_{h+1}^{k,\pi}(s_h^k, a_h^k) \\ &\leq \mathbb{I}\left[n_h^k = 0\right] H + \frac{1}{\check{n}} \sum_{i=1}^{\check{n}} P_h^{\check{l}_i} V_{h+1}^{\check{l}_i}(s_h^k, a_h^k) - P_h^k V_{h+1}^{k,\pi}(s_h^k, a_h^k) + 3b_h^k + 3b_\Delta \qquad (27) \\ &= \mathbb{I}\left[n_h^k = 0\right] H + \underbrace{\frac{1}{\check{n}} \sum_{i=1}^{\check{n}} \left(P_h^{\check{l}_i} - P_h^k\right) V_{h+1}^{\check{l}_i}(s_h^k, a_h^k)}_{①} + \underbrace{\frac{1}{\check{n}} \sum_{i=1}^{\check{n}} P_h^k \left(V_{h+1}^{\check{l}_i} - V_{h+1}^{\check{l}_i,\star}\right)(s_h^k, a_h^k)}_{②} + 3b_h^k + 3b_\Delta \\ &\quad + \underbrace{\frac{1}{\check{n}} \sum_{i=1}^{\check{n}} P_h^k \left(V_{h+1}^{\check{l}_i,\star} - V_{h+1}^{k,\pi}\right)(s_h^k, a_h^k)}_{③}, \qquad (28) \end{aligned}$$

where (27) is by the Azuma-Hoeffding inequality and by (26). In the following, we bound each term in (28) separately. First, by Hölder's inequality, we have

$$① \leq \frac{1}{\check{n}} \sum_{i=1}^{\check{n}} \Delta_p^{(1)}(H - h) \leq b_\Delta. \tag{29}$$

Let $\mathbf{e}_j$ denote a standard basis vector of proper dimensions that has a 1 at the j-th entry and 0s at the others, in the form of $(0, \ldots, 0, 1, 0, \ldots, 0)$. Recall the definition of δ_h^k in (25), and we have

$$\textcircled{2} = \underbrace{\frac{1}{\check{n}} \sum_{i=1}^{\check{n}} \left(P_h^k - \mathbf{e}_{s_{h+1}^{\check{l}_i}} \right) \left(V_{h+1}^{\check{l}_i} - V_{h+1}^{\check{l}_i,\star} \right) (s_h^k, a_h^k)}_{\xi_{h+1}^k} + \frac{1}{\check{n}} \sum_{i=1}^{\check{n}} \delta_{h+1}^{\check{l}_i} = \xi_{h+1}^k + \frac{1}{\check{n}} \sum_{i=1}^{\check{n}} \delta_{h+1}^{\check{l}_i}. \tag{30}$$

Finally, recalling the definition of ζ_h^k in (25), we have that

$$\begin{aligned}
\textcircled{3} =& \frac{1}{\check{n}} \sum_{i=1}^{\check{n}} \left(V_{h+1}^{\check{l}_i,\star}(s_{h+1}^k) - V_{h+1}^{k,\pi}(s_{h+1}^k) \right) \\
& + \underbrace{\frac{1}{\check{n}} \sum_{i=1}^{\check{n}} \left(P_h^k - \mathbf{e}_{s_{h+1}^k} \right) \left(V_{h+1}^{\check{l}_i,\star} - V_{h+1}^{k,\pi} \right) (s_h^k, a_h^k)}_{\phi_{h+1}^k} \\
=& \frac{1}{\check{n}} \sum_{i=1}^{\check{n}} \left(V_{h+1}^{\check{l}_i,\star}(s_{h+1}^k) - V_{h+1}^{k,\star}(s_{h+1}^k) \right) + \zeta_{h+1}^k - \delta_{h+1}^k + \phi_{h+1}^k \\
\leq& b_\Delta + \zeta_{h+1}^k - \delta_{h+1}^k + \phi_{h+1}^k
\end{aligned} \tag{31}$$

where inequality (31) is by Lemma 1. Combining (28), (29), (30), and (31) leads to

$$\zeta_h^k \leq \mathbb{I}\left[n_h^k = 0 \right] H + \frac{1}{\check{n}} \sum_{i=1}^{\check{n}} \delta_{h+1}^{\check{l}_i} + \xi_{h+1}^k + \zeta_{h+1}^k - \delta_{h+1}^k + \phi_{h+1}^k + 3b_h^k + 5b_\Delta. \tag{32}$$

To find an upper bound of $\sum_{k=1}^K \zeta_h^k$, we proceed to upper bound each term on the RHS of (32) separately. First, notice that $\sum_{k=1}^K \mathbb{I}\left[n_h^k = 0\right] \leq SAH$, because each fixed triple (s, a, h) contributes at most 1 to $\sum_{k=1}^K \mathbb{I}\left[n_h^k = 0\right]$. The second term in (32) can be upper bounded by the following lemma:

Lemma 7. $\sum_{k=1}^K \frac{1}{\check{n}_h^k} \sum_{i=1}^{\check{n}_h^k} \delta_{h+1}^{\check{l}_{h,i}^k} \leq (1 + \frac{1}{H}) \sum_{k=1}^K \delta_{h+1}^k$.

Combining (32) and Lemma 7, we now have that

$$\begin{aligned}
\sum_{k=1}^K \zeta_h^k \leq& SAH^2 + \frac{1}{H} \sum_{k=1}^K \delta_{h+1}^k + \sum_{k=1}^K \left(\xi_{h+1}^k + \zeta_{h+1}^k + \phi_{h+1}^k + 3b_h^k + 5b_\Delta \right) \\
\leq& SAH^2 + (1 + \frac{1}{H}) \sum_{k=1}^K \zeta_{h+1}^k + \sum_{k=1}^K \underbrace{\left(\xi_{h+1}^k + \phi_{h+1}^k + 3b_h^k + 5b_\Delta \right)}_{\Lambda_{h+1}^k},
\end{aligned} \tag{33}$$

where in (33) we have used the fact that $\delta_{h+1}^k \leq \zeta_{h+1}^k$, which in turn is due to the optimality that $V_h^{k,\star}(s_h^k) \geq V_h^{k,\pi}(s_h^k)$. Notice that we have ζ_h^k on the LHS of (33) and ζ_{h+1}^k on the RHS. By iterating (33) over $h = H, H-1, \ldots, 1$, we conclude that

$$\sum_{k=1}^K \zeta_1^k \leq O\left(SAH^3 + \sum_{h=1}^H \sum_{k=1}^K (1 + \frac{1}{H})^{h-1} \Lambda_{h+1}^k \right). \tag{34}$$

We bound $\sum_{h=1}^H \sum_{k=1}^K (1 + \frac{1}{H})^{h-1} \Lambda_{h+1}^k$ in the proposition below. Its proof relies on a series of lemmas in Appendix B that upper bound each term in Λ_{h+1}^k separately.

Proposition 2. *With probability at least $1-(KH+2)\delta$, it holds that*

$$\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\Lambda_{h+1}^{k}\leq\widetilde{O}(\sqrt{SAKH^5}+KH\Delta_r^{(1)}+KH^2\Delta_p^{(1)}).$$

Now we are ready to prove Theorem 1.

Proof. (of Theorem 1) By (24) and (34), and by replacing δ with $\frac{\delta}{KH+2}$ in Proposition 2, we know that the dynamic regret in epoch $d=1$ can be upper bounded with probability at least $1-\delta$ by:

$$\mathcal{R}^{(d)}(\pi,K)\leq\widetilde{O}(SAH^3+\sqrt{SAKH^5}+KH\Delta_r^{(1)}+KH^2\Delta_p^{(1)}),$$

and this holds for every epoch $d\in[D]$. Suppose $T=\Omega(SA\Delta H^2)$; summing up the dynamic regret over all the D epochs gives us an upper bound of $\widetilde{O}(D\sqrt{SAKH^5}+\sum_{d=1}^{D}KH\Delta_r^{(d)}+\sum_{d=1}^{D}KH^2\Delta_p^{(d)})$. Recall the definition that $\sum_{d=1}^{D}\Delta_r^{(d)}\leq\Delta_r$, $\sum_{d=1}^{D}\Delta_p^{(d)}\leq\Delta_p$, $\Delta=\Delta_r+\Delta_p$, and that $K=\Theta(\frac{T}{DH})$. By setting $D=S^{-\frac{1}{3}}A^{-\frac{1}{3}}\Delta^{\frac{2}{3}}H^{-\frac{2}{3}}T^{\frac{1}{3}}$, the dynamic regret over the entire T steps is bounded by $\mathcal{R}(\pi,M)\leq\widetilde{O}(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}H^{\frac{5}{3}}T^{\frac{2}{3}})$, which completes the proof. □

D. Proof Sketch of Theorem 2

Proof sketch. We only outline the difference with respect to the proof of Theorem 1 in the main text. The reader should have no difficulty recovering the complete proof by following the same routine as Theorem 1. Specifically, it suffices to investigate the steps that are involved with Lemma 2.

The dynamic regret of the new algorithm in epoch $d=1$ now can be expressed as

$$\mathcal{R}^{(d)}(\pi,K)=\sum_{k=1}^{K}\left(V_1^{k,*}\left(s_1^k\right)-V_1^{k,\pi}\left(s_1^k\right)\right)\leq\sum_{k=1}^{K}\left(V_1^{k}\left(s_1^k\right)-V_1^{k,\pi}\left(s_1^k\right)\right)+2KHb_{\Delta},\tag{35}$$

where we applied the results of Lemma 3 instead of Lemma 2. The reader should bear in mind that from the new update rules of the value functions, we now have

$$V_h^k(s_h^k)\leq\mathbb{I}\Big[n_h^k=0\Big]H+\frac{\check{r}_h(s_h^k,a_h^k)}{\check{N}_h^k(s_h^k,a_h^k)}+\frac{\check{v}_h(s_h^k,a_h^k)}{\check{N}_h^k(s_h^k,a_h^k)}+b_h^k,\tag{36}$$

where the RHS no longer has the additional bonus term b_Δ. If we define ζ_h^k, ξ_{h+1}^k, and ϕ_{h+1}^k in the same way as before, the reader can easily verify that all the derivations until Equation (34) still holds, although the value of Λ_{h+1}^k should be re-defined as $\Lambda_{h+1}^k\overset{\text{def}}{=}\xi_{h+1}^k+\phi_{h+1}^k+3b_h^k+3b_\Delta$ due to the new upper bound in (36) that is independent of b_Δ. Proposition 2 also follows analogously though some additional attention should be paid to the proof of Lemma 6 where the results of Lemma 2 have been utilized. Finally, we obtain the dynamic regret upper bound in epoch $d=1$ as follows:

$$\mathcal{R}^{(d)}(\pi,K)\leq\widetilde{O}\left(SAH^3+\sqrt{SAKH^5}+KH\Delta_r^{(1)}+KH^2\Delta_p^{(1)}\right)+2KHb_\Delta,$$

where the additional term $2KHb_\Delta$ comes from (35). From our definition of b_Δ, we can easily see that $2KHb_\Delta\leq O(KH\Delta_r^{(1)}+KH^2\Delta_p^{(1)})$. Therefore, we can conclude that the dynamic regret upper bound in one epoch remains the same order, which leaves the dynamic regret over the entire horizon also unchanged. □

Algorithm 2: RestartQ-UCB (Freedman)

```
for epoch $d \leftarrow 1$ to $D$ do
    Initialize: $V_h(s) \leftarrow H-h+1, Q_h(s,a) \leftarrow H-h+1, N_h(s,a) \leftarrow 0, \check{N}_h(s,a) \leftarrow 0,$
    $\check{r}_h(s,a) \leftarrow 0, \check{\mu}_h(s,a) \leftarrow 0, \check{v}_h(s,a) \leftarrow 0, \check{\sigma}_h(s,a) \leftarrow 0, \mu_h^{\text{ref}}(s,a) \leftarrow 0, \sigma_h^{\text{ref}}(s,a) \leftarrow 0, V_h^{\text{ref}}(s) \leftarrow H$, for all
    $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$;
    for episode $k \leftarrow (d-1)K+1$ to $\min\{dK, M\}$ do
        observe $s_1^k$;
        for step $h \leftarrow 1$ to $H$ do
            Take action $a_h^k \leftarrow \arg\max_a Q_h(s_h^k, a)$, receive $R_h^k(s_h^k, a_h^k)$, and observe $s_{h+1}^k$;
            $\check{r} \leftarrow \check{r} + R_h^k(s_h^k, a_h^k), \check{v} \leftarrow \check{v} + V_{h+1}(s_{h+1}^k)$;
            $\check{\mu} \leftarrow \check{\mu} + V_{h+1}(s_{h+1}^k) - V_{h+1}^{\text{ref}}(s_{h+1}^k), \check{\sigma} \leftarrow \check{\sigma} + \big(V_{h+1}(s_{h+1}^k) - V_{h+1}^{\text{ref}}(s_{h+1}^k)\big)^2$;
            $\mu^{\text{ref}} \leftarrow \mu^{\text{ref}} + V_{h+1}^{\text{ref}}(s_{h+1}^k), \sigma^{\text{ref}} \leftarrow \sigma^{\text{ref}} + (V_{h+1}^{\text{ref}}(s_{h+1}^k))^2$;
            $n \leftarrow n+1, \check{n} \leftarrow \check{n}+1$;
            if $n \in \mathcal{L}$   // Reaching the end of the stage
             then
                $b_h^k \leftarrow \sqrt{\frac{H^2}{\check{n}}\iota} + \sqrt{\frac{1}{\check{n}}\iota},\ b_\Delta \leftarrow \Delta_r^{(d)} + H\Delta_p^{(d)}$;
                $\underline{b}_h^k \leftarrow 2\sqrt{\frac{\sigma^{\text{ref}}/n - (\mu^{\text{ref}}/n)^2}{n}\iota} + 2\sqrt{\frac{\check{\sigma}/\check{n} - (\check{\mu}/\check{n})^2}{\check{n}}\iota} + 5(\frac{H\iota}{n} + \frac{H\iota}{\check{n}} + \frac{H\iota^{3/4}}{n^{3/4}} + \frac{H\iota^{3/4}}{\check{n}^{3/4}}) + \sqrt{\frac{1}{\check{n}}\iota}$;
                $Q_h(s_h^k, a_h^k) \leftarrow \min\left\{\frac{\check{r}}{\check{n}} + \frac{\check{v}}{\check{n}} + b_h^k + 2b_\Delta, \frac{\check{r}}{\check{n}} + \frac{\mu^{\text{ref}}}{n} + \frac{\check{\mu}}{\check{n}} + 2\underline{b}_h^k + 4b_\Delta, Q_h(s_h^k, a_h^k)\right\}$;
                $V_h(s_h^k) \leftarrow \max_a Q_h(s_h^k, a)$;
                $\check{N}_h(s_h^k, a_h^k), \check{r}_h(s_h^k, a_h^k), \check{v}_h(s_h^k, a_h^k), \check{\mu}_h(s_h^k, a_h^k), \check{\sigma}_h(s_h^k, a_h^k) \leftarrow 0$;
            if $\sum_a N_h(s_h^k, a) = \Omega(SAH^6\iota)$// Learn the reference value
             then
                $V_h^{\text{ref}}(s_h^k) \leftarrow V_h(s_h^k)$;
```

E. Algorithm: RestartQ-UCB (Freedman)

The algorithm Restarted Q-Learning with Freedman Upper Confidence Bounds (RestartQ-UCB Freedman) is presented in Algorithm 2. For ease of exposition, we use $\check{r}$, $\check{\mu}$, $\check{v}$, $\check{\sigma}$, μ^{ref}, σ^{ref}, $\check{n}$, and n to denote $\check{r}_h(s,a)$, $\check{\mu}_h(s_h^k, a_h^k)$, $\check{v}_h(s_h^k, a_h^k)$, $\check{\sigma}_h(s_h^k, a_h^k)$, $\mu_h^{\text{ref}}(s_h^k, a_h^k)$, $\sigma_h^{\text{ref}}(s_h^k, a_h^k)$, $\check{N}_h(s_h^k, a_h^k)$, and $N_h(s_h^k, a_h^k)$ respectively, when the values of (s,a,h,k) are clear from the context.

Compared with Algorithm 1, there are two major improvements in Algorithm 2. The first one is to replace the Hoeffding-based bonus term b_h^k with a tighter term $\underline{b}_h^k$. The latter term takes into account the second moment information of the random variables, which allows sharper tail bounds that rely on second moments to come into use (in our case, the Freedman's inequality). The second improvement is a variance reduction technique, or more specifically, the reference-advantage decomposition as coined in Zhang et al. (2020). The intuition is to first learn a reference value function V^{ref} that serves as a roughly accurate estimate of the optimal value function $V^\star$. The goal of learning the optimal value function $V^\star = V^{\text{ref}} + (V^* - V_{\text{ref}})$ can hence be decomposed into estimating two terms V^{ref} and $V^* - V_{\text{ref}}$. The reference value V^{ref} is a fixed term, and can be accurately estimated using a large number of samples (in Algorithm 2, we estimate V^{ref} only when we have $cSAH^6\iota$ samples for a large constant c). The advantage term $V^* - V^{\text{ref}}$ can also be accurately estimated due to the reduced variance.

F. Proof of Theorem 3

Similar to the proof of Theorem 1, we start with the dynamic regret in one epoch, and then extend to all epochs in the end. The proof follows the same routine as in the proof of Theorem 1. Given that a rigorous analysis on the Freedman-based bonus with variance reduction is present in Zhang et al. (2020), one should not find it difficult to extend our Hoeffding-based algorithm to Algorithm 2. Therefore, rather than providing a complete proof of Theorem 3, in the following, we sketch the differences and highlight the additional analysis needed that is not covered by the proof of Theorem 1 and Zhang et al. (2020).

To facilitate the analysis, first recall a few notations $N_h^k, \check{N}_h^k, Q_h^k(s,a), V_h^k(s), n_h^k, l_{h,i}^k, \check{n}_h^k, \check{l}_{h,i}^k, l_i$ and $\check{l}_i$ that we have defined in Section 4. In addition, when (h,k) is clear from the context, we drop the time indices and simply use $\check{\mu}, \check{\sigma}, \mu^{\text{ref}}, \sigma^{\text{ref}}$ to denote their corresponding values in the computation of the $Q_h(s_h^k, a_h^k)$ value in Line 15 of Algorithm 2.

We start with the following lemma, which is an analogue of Lemma 2 but requires a more careful treatment of variations accumulated in μ^{ref} and $\check{\mu}_h$. It states that the optimistic $Q_h^k(s,a)$ is an upper bound of the optimal $Q_h^{k,\star}(s,a)$ with high probability.

Lemma 8. *(Freedman) For $\delta \in (0,1)$, with probability at least $1 - 2KH\delta$, it holds that $Q_h^{k,\star}(s,a) \leq Q_h^{k+1}(s,a) \leq Q_h^k(s,a), \forall (s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$.*

Proof. It should be clear from the way we update $Q_h(s,a)$ that $Q_h^k(s,a)$ is monotonically decreasing in k. We now prove $Q_h^{k,\star}(s,a) \leq Q_h^{k+1}(s,a)$ for all s,a,h,k by induction on k. First, it holds for $k=1$ by our initialization of $Q_h(s,a)$. For $k \geq 2$, now suppose $Q_h^{j,\star}(s,a) \leq Q_h^j(s,a)$ for all s,a,h and $1 \leq j \leq k$. For a fixed triple (s,a,h), we consider the following two cases.

Case 1: $Q_h(s,a)$ is updated in episode k. Notice that it suffices to analyze the case where $Q_h(s,a)$ is updated using $\underline{b}_h^k$, because the other case of b_h^k would be exactly the same as in Lemma 2. With probability at least $1-\delta$,

$$
\begin{aligned}
Q_h^{k+1}(s,a) =& \frac{\check{r}_h(s,a)}{\check{N}_h^k(s,a)} + \frac{\mu^{\text{ref}}(s,a)}{N_h^k(s,a)} + \frac{\check{\mu}_h(s,a)}{\check{N}_h^k(s,a)} + 2\underline{b}_h^k + 4b_\Delta \\
=& \frac{\check{r}_h(s,a)}{\check{n}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(V_{h+1}^{\text{ref},l_i}(s_{h+1}^{l_i}) - P_h^{l_i}V_{h+1}^{\text{ref},l_i}(s,a)\right)}_{\chi_1} \\
&+ \underbrace{\frac{1}{\check{n}}\sum_{i=1}^{\check{n}}\left[\left(V_{h+1}^{\check{l}_i}(s_{h+1}^{\check{l}_i}) - V_{h+1}^{\text{ref},\check{l}_i}(s_{h+1}^{\check{l}_i})\right) - \left(P_h^{\check{l}_i}V_{h+1}^{\check{l}_i} - P_h^{\check{l}_i}V_{h+1}^{\text{ref},\check{l}_i}\right)(s,a)\right]}_{\chi_2} \\
&+ \underbrace{\frac{1}{n}\sum_{i=1}^{n}P_h^{l_i}V_{h+1}^{\text{ref},l_i} + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}}\left(P_h^{\check{l}_i}V_{h+1}^{\check{l}_i} - P_h^{\check{l}_i}V_{h+1}^{\text{ref},\check{l}_i}\right)(s,a)}_{\chi_3} + 2\underline{b}_h^k + 4b_\Delta
\end{aligned}
\tag{37}
$$

In the following, we will bound each term in (37) separately. First, we have that

$$
\begin{aligned}
\chi_3 + 2b_\Delta =& \frac{1}{n}\sum_{i=1}^{n}\left(P_h^{l_i}V_{h+1}^{\text{ref},l_i} - P_h^kV_{h+1}^{\text{ref},l_i}\right)(s,a) + b_\Delta & (38)\\
&- \frac{1}{\check{n}}\sum_{i=1}^{\check{n}}\left(P_h^{\check{l}_i}V_{h+1}^{\text{ref},\check{l}_i} - P_h^kV_{h+1}^{\text{ref},\check{l}_i}\right)(s,a) + b_\Delta & (39)\\
&+ \frac{1}{n}\sum_{i=1}^{n}P_h^kV_{h+1}^{\text{ref},l_i}(s,a) - \frac{1}{\check{n}}\sum_{i=1}^{\check{n}}P_h^kV_{h+1}^{\text{ref},\check{l}_i}(s,a) + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}}P_h^{\check{l}_i}V_{h+1}^{\check{l}_i}(s,a) & (40)\\
\geq& \frac{1}{\check{n}}\sum_{i=1}^{\check{n}}P_h^{\check{l}_i}V_{h+1}^{\check{l}_i}(s,a), & (41)
\end{aligned}
$$

where (38)$\geq$ 0 and (39)$\geq$ 0 by Hölder's inequality and the definition of b_Δ. In (40), we have that $\frac{1}{n}\sum_{i=1}^{n} P_h^k V_{h+1}^{\mathrm{ref},l_i}(s,a) - \frac{1}{\check{n}}\sum_{i=1}^{\check{n}} P_h^k V_{h+1}^{\mathrm{ref},\check{l}_i}(s,a) \geq 0$, because $V_{h+1}^{\mathrm{ref},k}(s)$ is non-increasing in k.

Following a similar procedure as in Lemma 10, Lemma 12, and Lemma 13 in Zhang et al. (2020), we can further bound $|\chi_1|$ and $|\chi_2|$ as follows:

$$|\chi_1| \leq 2\sqrt{\frac{\nu^{\mathrm{ref}}\iota}{n}} + \frac{5H\iota^{\frac{3}{4}}}{n^{\frac{3}{4}}} + \frac{2\sqrt{\iota}}{Tn} + \frac{2H\iota}{n}, \tag{42}$$

$$|\chi_2| \leq 2\sqrt{\frac{\check{\nu}\iota}{\check{n}}} + \frac{5H\iota^{\frac{3}{4}}}{\check{n}^{\frac{3}{4}}} + \frac{2\sqrt{\iota}}{T\check{n}} + \frac{2H\iota}{\check{n}}, \tag{43}$$

where $\nu^{\mathrm{ref}} \overset{\mathrm{def}}{=} \frac{\sigma^{\mathrm{ref}}}{n} - \left(\frac{\mu^{\mathrm{ref}}}{n}\right)^2$ and $\check{\nu} \overset{\mathrm{def}}{=} \frac{\check{\sigma}}{\check{n}} - \left(\frac{\check{\mu}}{\check{n}}\right)^2$. These are the steps where Freedman's inequality Freedman (1975) come into use, and we omit these steps since they are essentially the same as the derivations in Zhang et al. (2020). We can see from (42), (43), and the definition of $\underline{b}_h^k$ that $|\chi_1| + |\chi_2| \leq \underline{b}_h^k$.

Substituting the results on χ_1, χ_2 and χ_3 back to (37), it holds that with probability at least $1-\delta$,

$$\begin{aligned}
Q_h^{k+1}(s,a) =& \frac{\check{r}_h(s,a)}{\check{n}} + \chi_1 + \chi_2 + \chi_3 + 2\underline{b}_h^k + 4b_\Delta \\
\geq& \frac{\check{r}_h(s,a)}{\check{n}} + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}} P_h^{\check{l}_i} V_{h+1}^{\check{l}_i}(s,a) + \underline{b}_h^k + 2b_\Delta && (44) \\
\geq& \frac{\check{r}_h(s,a)}{\check{n}} + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}} P_h^{\check{l}_i} V_{h+1}^{\check{l}_i,\star}(s,a) + \underline{b}_h^k + 2b_\Delta && (45) \\
=& \frac{\check{r}_h(s,a)}{\check{n}} + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}} \left(Q_h^{\check{l}_i,\star}(s,a) - r_h^{\check{l}_i}(s,a)\right) + \underline{b}_h^k + 2b_\Delta \\
\geq& \frac{1}{\check{n}}\sum_{i=1}^{\check{n}} Q_h^{\check{l}_i,\star}(s,a) + 2b_\Delta \geq Q_h^{k,\star}(s,a) + b_\Delta, && (46)
\end{aligned}$$

where in (44) we used (41), (42), (43), and the definition of $\underline{b}_h^k$ in Algorithm 2. (45) is by the induction hypothesis that $Q_{h+1}^{\check{l}_i}(s_{h+1}^{\check{l}_i},a) \geq Q_{h+1}^{\check{l}_i,\star}(s_{h+1}^{\check{l}_i},a), \forall a \in \mathcal{A}, 1 \leq \check{l}_i \leq k$. The second to last inequality holds due to the Hofdding's inequality that $\frac{1}{\check{n}}\left(\sum_{i=1}^{\check{n}} r_h^{\check{l}_i}(s,a) - \check{r}_h(s,a)\right) \leq \sqrt{\frac{\iota}{\check{n}}} \leq \underline{b}_h^k$ with high probability. Finally, the last inequality follows from Lemma 1.

According to the monotonicity of $Q_h^k(s,a)$, we can conclude from (46) that $Q_h^{k,\star}(s,a) \leq Q_h^{k+1}(s,a) \leq Q_h^k(s,a)$. In fact, we have proved the stronger statement $Q_h^{k+1}(s,a) \geq Q_h^{k,\star}(s,a) + b_\Delta$ that will be useful in Case 2 below.

Case 2: $Q_h(s,a)$ is not updated in episode k. Then, there are two possibilities:

1. If $Q_h(s,a)$ has never been updated from episode 1 to episode k: It is easy to see that $Q_h^{k+1}(s,a) = Q_h^k(s,a) = \cdots = Q_h^1(s,a) = H - h + 1 \geq Q_h^{k,\star}(s,a)$ holds.

2. If $Q_h(s,a)$ has been updated at least once from episode 1 to episode k: Let j be the index of the latest episode that $Q_h(s,a)$ was updated. Then, from our induction hypothesis and Case 1, we know that $Q_h^{j+1}(s,a) \geq Q_h^{j,\star}(s,a) + b_\Delta$. Since $Q_h(s,a)$ has not been updated from episode $j+1$ to episode k, we know that $Q_h^{k+1}(s,a) = Q_h^k(s,a) = \cdots = Q_h^{j+1}(s,a) \geq Q_h^{j,\star}(s,a) + b_\Delta \geq Q_h^{k,\star}(s,a)$, where the last inequality holds because of Lemma 1.

A union bound over all time steps completes our proof. □

Conditional on the successful event of Lemma 8, the dynamic regret of Algorithm 2 in epoch $d = 1$ can hence be expressed as

$$\mathcal{R}^{(d)}(\pi, K) = \sum_{k=1}^{K} \left(V_1^{k,*}\left(s_1^k\right) - V_1^{k,\pi}\left(s_1^k\right) \right) \leq \sum_{k=1}^{K} \left(V_1^{k}\left(s_1^k\right) - V_1^{k,\pi}\left(s_1^k\right) \right). \tag{47}$$

From the update rules of the value functions in Algorithm 2, we have

$$\begin{aligned} V_h^k(s_h^k) \leq & \mathbb{I}\left[n_h^k = 0\right] H + \frac{\check{r}_h(s_h^k, a_h^k)}{\check{n}} + \frac{\mu_h^{\text{ref},k}}{n} + \frac{\check{\mu}_h^k}{\check{n}} + 2\underline{b}_h^k + 4b_\Delta \\ = & \mathbb{I}\left[n_h^k = 0\right] H + \frac{\check{r}_h(s_h^k, a_h^k)}{\check{n}} + \frac{1}{n}\sum_{i=1}^{n} V_{h+1}^{\text{ref},l_i}(s_{h+1}^{l_i}) + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}} (V_{h+1}^{\check{l}_i}(s_{h+1}^{\check{l}_i}) - V_{h+1}^{\text{ref},\check{l}_i}(s_{h+1}^{\check{l}_i})) + 2\underline{b}_h^k + 4b_\Delta. \end{aligned}$$

If we again define $\zeta_h^k \overset{\text{def}}{=} V_h^k(s_h^k) - V_h^{k,\pi}(s_h^k)$, we can follow a similar routine as in the proof of Theorem 1 (details can be found in Zhang et al. (2020)) and obtain

$$\sum_{k=1}^{K} \zeta_1^k \leq O\left(SAH^3 + \sum_{h=1}^{H}\sum_{k=1}^{K} (1 + \frac{1}{H})^{h-1} \Lambda_{h+1}^k \right),$$

where $\Lambda_{h+1}^k \overset{\text{def}}{=} \psi_{h+1}^k + \xi_{h+1}^k + \phi_{h+1}^k + 4\underline{b}_h^k + 8b_\Delta$ with the following definitions:

$$\begin{aligned} \psi_{h+1}^k &\overset{\text{def}}{=} \frac{1}{n_h^k} \sum_{i=1}^{n_h^k} \left(P_h^k V_{h+1}^{\text{ref},l_i} - P_h^k V_{h+1}^{\text{ref},K+1} \right)(s_h^k, a_h^k), \\ \xi_{h+1}^k &\overset{\text{def}}{=} \frac{1}{\check{n}_h^k} \sum_{i=1}^{\check{n}_h^k} \left(P_h^k - \mathbf{e}_{s_{h+1}^{\check{l}_i}} \right)\left(V_{h+1}^{\check{l}_i} - V_{h+1}^{\check{l}_i,\star} \right)(s_h^k, a_h^k), \\ \phi_{h+1}^k &\overset{\text{def}}{=} \left(P_h^k - \mathbf{e}_{s_{h+1}^k} \right)\left(V_{h+1}^{\check{l}_i,\star} - V_{h+1}^{k,\pi} \right)(s_h^k, a_h^k). \end{aligned}$$

An upper bound on the first four terms in Λ_{h+1}^k is derived in the proof of Lemma 7 in Zhang et al. (2020) (There is an extra term of $\sqrt{\frac{1}{\check{n}}\iota}$ in our defnition of $\underline{b}_h^k$ compared to theirs, but it does not affect the leading term in the upper bound). By further recalling the definition of b_Δ, we can obtain the following lemma.

Lemma 9. *(Lemma 7 in Zhang et al. (2020)) With probability at least $(1 - O(H^2T^4\delta))$, it holds that*

$$\sum_{h=1}^{H}\sum_{k=1}^{K} (1 + \frac{1}{H})^{h-1} \Lambda_{h+1}^k = O\left(\sqrt{SAH^2T\iota} + H\sqrt{T\iota}\log(T) + S^2A^{\frac{3}{2}}H^8T^{\frac{1}{4}}\iota + KH\Delta_r^{(1)} + KH^2\Delta_p^{(1)} \right).$$

Combined with (47) and the definition of ζ_h^k, we obtain the dynamic regret bound in a single epoch:

$$\mathcal{R}^{(d)}(\pi, K) = O\left(\sqrt{SAH^2T\iota} + H\sqrt{T\iota}\log(T) + S^2A^{\frac{3}{2}}H^8T^{\frac{1}{4}}\iota + KH\Delta_r^{(1)} + KH^2\Delta_p^{(1)} \right), \forall d \in [D].$$

Finally, suppose T is greater than a polynomial of S, A, Δ and H, $\sqrt{SAH^2T\iota}$ would be the leading term of the dynamic regret in a single epoch. In this case, summing up the dynamic regret over all the D epochs gives us an upper bound of $\widetilde{O}\left(D\sqrt{SAH^2T} + \sum_{d=1}^{D} KH\Delta_r^{(d)} + \sum_{d=1}^{D} KH^2\Delta_p^{(d)} \right)$. Recall that $\sum_{d=1}^{D} \Delta_r^{(d)} \leq \Delta_r$,

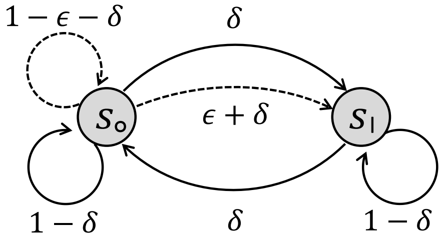

Figure 2: The "JAO MDP" constructed in Jaksch et al. (2010). Dashed lines denote transitions related to the good action $a^\star$.

$\sum_{d=1}^{D}\Delta_p^{(d)} \le \Delta_p$, $\Delta = \Delta_r + \Delta_p$, and that $K = \Theta(\frac{T}{DH})$. By setting $D = S^{-\frac{1}{3}}A^{-\frac{1}{3}}\Delta^{\frac{2}{3}}T^{\frac{1}{3}}$, the dynamic regret over the entire T steps is bounded by

$$\mathcal{R}(\pi, M) \le \widetilde{O}\left(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}HT^{\frac{2}{3}}\right).$$

This completes the proof of Theorem 3.

G. Proof of Theorem 4

The proof of our lower bound relies on the construction of a "hard instance" of non-stationary MDPs. The instance we construct is essentially a switching-MDP: an MDP with piecewise constant dynamics on each *segment* of the horizon, and its dynamics experience an abrupt change at the beginning of each new segment. More specifically, we divide the horizon T into L segments[6], where each segment has $T_0 \overset{\text{def}}{=} \lfloor \frac{T}{L} \rfloor$ steps and contains $M_0 \overset{\text{def}}{=} \lfloor \frac{M}{L} \rfloor$ episodes, each episode having a length of H. Within each such segment, the system dynamics of the MDP do not vary, and we construct the dynamics for each segment in a way such that the instance is a hard instance of stationary MDPs on its own. The MDP within each segment is essentially similar to the hard instances constructed in stationary RL problems (Osband & Van Roy, 2016; Jin et al., 2018). Between two consecutive segments, the dynamics of the MDP change abruptly, and we let the dynamics vary in a way such that no information learned from previous interactions with the MDP can be used in the new segment. In this sense, the agent needs to learn a new hard stationary MDP in each segment. Finally, optimizing the value of L and the variation magnitude between consecutive segments (subject to the constraints of the total variation budget) leads to our lower bound.

We start with a simplified episodic setting where the transition kernels and reward functions are held constant within each episode, i.e., $P_1^m = \cdots = P_h^m = \ldots P_H^m$ and $r_1^m = \cdots = r_h^m = \ldots r_H^m, \forall m \in [M]$. This is a popular but less challenging episodic setting, and its stationary counterpart has been studied in Azar et al. (2017). We further require that when the environment varies due to the non-stationarity, all steps in one episode should vary simultaneously in the same way. This simplified setting is easier to analyze, and its analysis conveniently leads to a lower bound for the un-discounted setting as a side result along the way. Later we will show how the analysis can be naturally extended to the more general setting we introduced in Section 2, using techniques that have also been utilized in Jin et al. (2018). For simplicity of notations, we temporarily drop the h indices and use P^m and r^m to denote the transition kernel and reward function whenever there is no ambiguity.

Consider a two-state MDP as depicted in Figure 2. This MDP was initially proposed in Jaksch et al. (2010) as a hard instance of stationary MDPs, and following Jin et al. (2018) we will refer to this construction as the "JAO MDP". This MDP has 2 states $\mathcal{S} = \{s_\circ, s_\shortmid\}$ and SA actions $\mathcal{A} = \{1, 2, \ldots, SA\}$. The reward does not depend on actions: state $s_\shortmid$ always gives reward 1 whatever action is taken, and state $s_\circ$ always gives reward 0. Any action taken at state $s_\shortmid$ takes the agent to state $s_\circ$ with probability δ, and to state $s_\shortmid$ with probability $1-\delta$. At state $s_\circ$, for all but a single "good" action $a^\star$, the agent is taken to state $s_\shortmid$ with probability δ, and for the good

[6] The definition of segments is irrelevant to, and should not be confused with, the notion of epochs we previously defined.

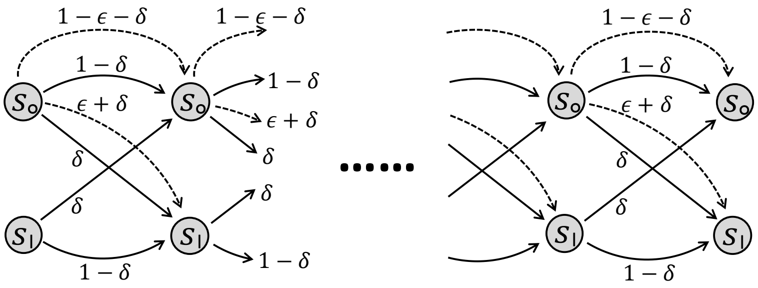

Figure 3: A chain with H copies of JAO MDPs correlated in time. At the end of an episode, the state should deterministically transition from any state in the last copy to the $s_\circ$ state in the first copy of the chain, the arrows of which are not shown in the figure. Also, the $s_{\shortmid}$ state in the first copy is actually never reached and hence is redundant.

action $a^\star$, the agent is taken to state $s_{\shortmid}$ with probability $\delta + \varepsilon$ for some $0 < \varepsilon < \delta$. The exact values of δ and ε will be chosen later. Note that this is not an MDP with S states and A actions as we desire, but the extension to an MDP with S states and A actions is routine (Jaksch et al., 2010), and is hence omitted here.

To apply the JAO MDP to the simplified episodic setting, we "concatenate" H copies of exactly the same JAO MDP into a chain as depicted in Figure 3, denoting the H steps in an episode. The initial state of this MDP is the $s_\circ$ state in the first copy of the chain, and after each episode the state is "reset" to the initial state. In the following, we first show that the constructed MDP is a hard instance of stationary MDPs, without worrying about the evolution of the system dynamics. The techniques that we will be using are essentially the same as in the proofs of the lower bound in the multi-armed bandit problem (Auer et al., 2002) or the reinforcement learning problem in the un-discounted setting (Jaksch et al., 2010).

The good action $a^\star$ is chosen uniformly at random from the action space $\mathcal{A}$, and we use $\mathbb{E}_\star[\cdot]$ to denote the expectation with respect to the random choice of $a^\star$. We write $\mathbb{E}_a[\cdot]$ for the expectation conditioned on action a being the good action $a^\star$. Finally, we use $\mathbb{E}_{\text{unif}}[\cdot]$ to denote the expectation when there is no good action in the MDP, i.e., every action in $\mathcal{A}$ takes the agent from state $s_\circ$ to $s_{\shortmid}$ with probability δ. Define the probability notations $\mathbb{P}_\star(\cdot), \mathbb{P}_a(\cdot)$, and $\mathbb{P}_{\text{unif}}(\cdot)$ analogously.

Consider running a reinforcement learning algorithm on the constructed MDP for T_0 steps, where $T_0 = M_0 H$. It has been shown in Auer et al. (2002) and Jaksch et al. (2010) that it is sufficient to consider deterministic policies. Therefore, we assume that the algorithm maps deterministically from a sequence of observations to an action a_t at time t. Define the random variables $N_{\shortmid}, N_\circ$ and $N_\circ^\star$ to be the total number of visits to state $s_{\shortmid}$, the total number of visits to $s_\circ$, and the total number of times that $a^\star$ is taken at state $s_\circ$, respectively. Let s_t denote the state observed at time t, and a_t the action taken at time t. When there is no chance of ambiguity, we sometimes also use s_h^m to denote the state at step h of episode m, which should be interpreted as the state s_t observed at time $t = (m-1) \times H + h$. The notation a_h^m is used analogously. Since $s_\circ$ is assumed to be the initial state, we have that

$$
\begin{aligned}
&\mathbb{E}_a[N_{\shortmid}] = \sum_{t=1}^{T_0} \mathbb{P}_a(s_t = s_{\shortmid}) = \sum_{m=1}^{M_0} \sum_{h=2}^{H} \mathbb{P}_a(s_h^m = s_{\shortmid}) \\
&= \sum_{m=1}^{M_0} \sum_{h=2}^{H} \big(\mathbb{P}_a(s_{h-1}^m = s_\circ) \cdot \mathbb{P}_a(s_h^m = s_{\shortmid} \mid s_{h-1}^m = s_\circ) + \mathbb{P}_a(s_{h-1}^m = s_{\shortmid}) \cdot \mathbb{P}_a(s_h^m = s_{\shortmid} \mid s_{h-1}^m = s_{\shortmid}) \big) \\
&= \sum_{m=1}^{M_0} \sum_{h=2}^{H} \big(\delta \mathbb{P}_a(s_{h-1}^m = s_\circ, a_h^m \neq a^\star) + (\delta + \varepsilon) \mathbb{P}_a(s_{h-1}^m = s_\circ, a_h^m = a^\star) + (1-\delta) \mathbb{P}_a(s_{h-1}^m = s_{\shortmid}) \big) \\
&\leq \delta \mathbb{E}_a[N_\circ - N_\circ^\star] + (\delta + \varepsilon) \mathbb{E}_a[N_\circ^\star] + (1-\delta) \mathbb{E}_a[N_{\shortmid}],
\end{aligned}
$$

and rearranging the last inequality gives us $\mathbb{E}_a[N_{\shortmid}] \le \mathbb{E}_a[N_\circ - N_\circ^\star] + (1+\frac{\varepsilon}{\delta})\mathbb{E}_a[N_\circ^\star]$.

For this proof only, define the random variable $W(T_0)$ to be the total reward of the algorithm over the horizon T_0, and define $G(T_0)$ to be the (static) regret with respect to the optimal policy. Since for any algorithm, the probability of staying in state $s_\circ$ under $\mathbb{P}_a(\cdot)$ is no larger than under $\mathbb{P}_{\text{unif}}(\cdot)$, it follows that

$$\begin{aligned}
&\mathbb{E}_a[W(T_0)] \le \mathbb{E}_a[N_{\shortmid}] \le \mathbb{E}_a[N_\circ - N_\circ^\star] + (1+\frac{\varepsilon}{\delta})\mathbb{E}_a[N_\circ^\star] \\
=&\mathbb{E}_a[N_\circ] + \frac{\varepsilon}{\delta}\mathbb{E}_a[N_\circ^\star] \le \mathbb{E}_{\text{unif}}[N_\circ] + \frac{\varepsilon}{\delta}\mathbb{E}_a[N_\circ^\star] \\
=&T_0 - \mathbb{E}_{\text{unif}}[N_{\shortmid}] + \frac{\varepsilon}{\delta}\mathbb{E}_a[N_\circ^\star]. \qquad (48)
\end{aligned}$$

Let $\tau_{\circ\shortmid}^m$ denote the first step that the state transits from state $s_\circ$ to $s_{\shortmid}$ in the m-th episode, then

$$\begin{aligned}
\mathbb{E}_{\text{unif}}[N_{\shortmid}] =& \sum_{m=1}^{M_0}\sum_{h=1}^{H}\mathbb{P}_{\text{unif}}(\tau_{\circ\shortmid}^m = h)\mathbb{E}_{\text{unif}}[N_{\shortmid} \mid \tau_{\circ\shortmid}^m = h] = \sum_{m=1}^{M_0}\sum_{h=1}^{H}(1-\delta)^{h-1}\delta\mathbb{E}_{\text{unif}}[N_{\shortmid} \mid \tau_{\circ\shortmid}^m = h] \\
\ge& \sum_{m=1}^{M_0}\sum_{h=1}^{H}(1-\delta)^{h-1}\delta\frac{H-h}{2} = \sum_{m=1}^{M_0}\left(\frac{H}{2} - \frac{1}{2\delta} + \frac{(1-\delta)^H}{2\delta}\right) \\
\ge& \frac{T_0}{2} - \frac{M_0}{2\delta}. \qquad (49)
\end{aligned}$$

Since the algorithm is a deterministic mapping from the observation sequence to an action, the random variable $N_\circ^\star$ is also a function of the observations up to time T. In addition, since the immediate reward only depends on the current state, $N_\circ^\star$ can further be considered as a function of just the state sequence up to T. Therefore, the following lemma from Jaksch et al. (2010), which in turn was adapted from Lemma A.1 in Auer et al. (2002), also applies in our setting.

Lemma 10. *(Lemma 13 in Jaksch et al. (2010)) For any finite constant B, let $f : \{s_\circ, s_{\shortmid}\}^{T_0+1} \to [0, B]$ be any function defined on the state sequence $\mathbf{s} \in \{s_\circ, s_{\shortmid}\}^{T_0+1}$. Then, for any $0 < \delta \le \frac{1}{2}$, any $0 < \varepsilon \le 1 - 2\delta$, and any $a \in \mathcal{A}$, it holds that*

$$\mathbb{E}_a[f(\mathbf{s})] \le \mathbb{E}_{\textit{unif}}[f(\mathbf{s})] + \frac{B}{2}\cdot\frac{\varepsilon}{\sqrt{\delta}}\sqrt{2\mathbb{E}_{\textit{unif}}[N_\circ^\star]}.$$

Since $N_\circ^\star$ itself is a function from the state sequence to $[0, T_0]$, we can apply Lemma 10 and arrive at

$$\mathbb{E}_a[N_\circ^\star] \le \mathbb{E}_{\text{unif}}\left[N_\circ^\star\right] + \frac{T_0}{2}\cdot\frac{\varepsilon}{\sqrt{\delta}}\sqrt{2\mathbb{E}_{\text{unif}}\left[N_\circ^\star\right]}. \qquad (50)$$

From (49), we have that $\sum_{a=1}^{SA}\mathbb{E}_{\text{unif}}\left[N_\circ^\star\right] = T_0 - \mathbb{E}_{\text{unif}}\left[N_{\shortmid}\right] \le \frac{T_0}{2} + \frac{M_0}{2\delta}$. By the Cauchy-Schwarz inequality, we further have that $\sum_{a=1}^{SA}\sqrt{2\mathbb{E}_{\text{unif}}\left[N_\circ^\star\right]} \le \sqrt{SA(T_0 + \frac{M_0}{\delta})}$. Therefore, from (50), we obtain

$$\sum_{a=1}^{SA}\mathbb{E}_a[N_\circ^\star] \le \frac{T_0}{2} + \frac{M_0}{2\delta} + \frac{T_0}{2}\cdot\frac{\varepsilon}{\sqrt{\delta}}\sqrt{SA(T_0 + \frac{M_0}{\delta})}.$$

Together with (48) and (49), it holds that

$$\begin{aligned}
\mathbb{E}_\star[W(T_0)] \le& \frac{1}{SA}\sum_{a=1}^{SA}\mathbb{E}_a[W(T_0)] \\
&\le \frac{T_0}{2} + \frac{M_0}{2\delta} + \frac{\varepsilon}{\delta}\frac{1}{SA}\left(\frac{T_0}{2} + \frac{M_0}{2\delta} + \frac{T_0}{2}\cdot\frac{\varepsilon}{\sqrt{\delta}}\sqrt{SA(T_0 + \frac{M_0}{\delta})}\right). \qquad (51)
\end{aligned}$$

G.1. The Un-discounted Setting

Let us now momentarily deviate from the episodic setting and consider the un-discounted setting (with $M_0 = 1$). This is the case of the JAO MDP in Figure 2 where there is not reset. We could calculate the stationary distribution and find that the optimal average reward for the JAO MDP is $\frac{\delta+\varepsilon}{2\delta+\varepsilon}$. It is also easy to calculate that the diameter of the JAO MDP is $D = \frac{1}{\delta}$. Therefore, the expected (static) regret with respect to the randomness of a^* can be lower bounded by

$$\begin{aligned}\mathbb{E}_\star[G(T_0)] =& \frac{\delta+\varepsilon}{2\delta+\varepsilon}T_0 - \mathbb{E}_\star[W(T_0)] \\ \geq& \frac{\varepsilon T_0}{4\delta+2\varepsilon} - \frac{D}{2} - \frac{\varepsilon D(T_0+D)}{2SA} - \frac{\varepsilon^2 T_0 D\sqrt{D}}{2\sqrt{SA}}(\sqrt{T_0}+\sqrt{D}).\end{aligned}$$

By assuming $T_0 \geq DSA$ (which in turn suggests $D \leq \sqrt{\frac{T_0 D}{SA}}$) and setting $\varepsilon = c\sqrt{\frac{SA}{T_0 D}}$ for $c = \frac{3}{40}$, we further have that

$$\begin{aligned}\mathbb{E}_\star[G(T_0)] \geq& \left(\frac{c}{6} - \frac{c}{2SA} - \frac{cD}{2SAT_0} - \frac{c^2}{2} - \frac{c^2}{2}\sqrt{\frac{D}{T_0}}\right)\sqrt{SAT_0D} - \frac{D}{2} \\ \geq& \left(\frac{3}{20}c - c^2 - \frac{1}{200}\right)\sqrt{SAT_0D} = \frac{1}{1600}\sqrt{SAT_0D}.\end{aligned}$$

It is easy to verify that our choice of δ and ε satisfies our assumption that $0 < \varepsilon < \delta$. So far, we have recovered the (static) regret lower bound of $\Omega(\sqrt{SAT_0D})$ in the un-discounted setting, which was originally proved in Jaksch et al. (2010).

Based on this result, let us now incorporate the non-stationarity of the MDP and derive a lower bound for the dynamic regret $\mathcal{R}(T)$. Recall that we are constructing the non-stationary environment as a switching-MDP. For each segment of length T_0, the environment is held constant, and the regret lower bound for each segment is $\Omega(\sqrt{SAT_0D})$. At the beginning of each new segment, we uniformly sample a new action a^* at random from the action space $\mathcal{A}$ to be the good action for the new segment. In this case, the learning algorithm cannot use the information it learned during its previous interactions with the environment, even if it knows the switching structure of the environment. Therefore, the algorithm needs to learn a new (static) MDP in each segment, which leads to a dynamic regret lower bound of $\Omega(L\sqrt{SAT_0D}) = \Omega(\sqrt{SATLD})$, where let us recall that L is the number of segments. Every time the good action a^* varies, it will cause a variation of magnitude 2ε in the transition kernel. The constraint of the overall variation budget requires that $2\varepsilon L = \frac{3}{20}\sqrt{\frac{SA}{T_0D}}L \leq \Delta$, which in turn requires $L \leq 4\Delta^{\frac{2}{3}}T^{\frac{1}{3}}D^{\frac{1}{3}}S^{-\frac{1}{3}}A^{-\frac{1}{3}}$. Finally, by assigning the largest possible value to L subject to the variation budget, we obtain a dynamic regret lower bound of $\Omega\left(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}D^{\frac{2}{3}}T^{\frac{2}{3}}\right)$. This completes the proof of Proposition 1.

G.2. The Episodic Settings

Now let us go back to our simplified episodic setting, as depicted in Figure 3. One major difference with the previous un-discounted setting is that we might not have time to mix between $s_\circ$ and $s_|$ in H steps. (Note that we only need to reach the stationary distribution over the $(s_\circ, s_|)$ pair in each step h, rather than the stationary distribution over the entire MDP. In fact, the latter case is never possible because the entire MDP is not aperiodic.) It can be shown that the optimal policy on this MDP has a mixing time of $\Theta\left(\frac{1}{\delta}\right)$ (Jin et al., 2018), and hence we can choose δ to be slightly larger than $\Theta(\frac{1}{H})$ to guarantee sufficient time to mix. All the analysis up to inequality (51) carries over to the episodic setting, and essentially we can set δ to be $\Theta\left(\frac{1}{H}\right)$ to get a (static) regret lower bound of $\Omega(\sqrt{SAT_0H})$ in each segment. Another difference with the previous setting lies in the usage of the variation budget. Since we require that all the steps in the same episode should vary simultaneously, it now takes a variation budget of $2\varepsilon H$ each time we switch to a new action a^* at the beginning of a new segment. Therefore, the overall variation budget now puts a constraint of $2\varepsilon HL \leq O(\Delta)$ on the magnitude of each switch. Again, by choosing $\varepsilon = \Theta\left(\sqrt{\frac{SA}{T_0 H}}\right)$ and optimizing over possible values of L subject to the budget constraint, we obtain a dynamic regret lower bound of $\Omega\left(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}H^{\frac{1}{3}}T^{\frac{2}{3}}\right)$ in the simplified episodic setting.

Finally, we consider the standard episodic setting as introduced in Section 2. In this setting, we essentially will be concatenating H distinct JAO MDPs, each with an independent good action a^*, into a chain like Figure 3. The transition kernels in these JAO MDPs are also allowed to vary asynchronously in each step h, although our construction of the lower bound does not make use of this property. As argued similarly in Jin et al. (2018), the number of observations for each specific JAO MDP is only T_0/H, instead of T_0. Therefore, we can assign a slightly larger value to ε and the learning algorithm would still not be able to identify the good action given the fewer observations. Setting $\delta = \Theta\left(\frac{1}{H}\right)$ and $\varepsilon = \Theta\left(\sqrt{\frac{SA}{T_0}}\right)$ leads to a (static) regret lower bound of $\Omega(H\sqrt{SAT_0})$ in the stationary RL problem. Again, the transition kernels in all the H JAO MDPs vary simultaneously at the beginning of each new segment. By optimizing L subject to the overall budget constraint $2\varepsilon HL \leq O(\Delta)$, we obtain a dynamic regret lower bound of $\Omega\left(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}H^{\frac{2}{3}}T^{\frac{2}{3}}\right)$ in the episodic setting. This completes our proof of Theorem 4.

H. Supplementary Proofs for Section 7

H.1. Proof of Theorem 5

Proof. The proof follows easily from the following observation: From the perspective of agent 1, the environment is non-stationary due to the fact that agent 2 is changing its policy over time. Since the switching cost of agent 2 is upper bounded by $O(T^\beta)$, by the definition of Δ_r and Δ_p in Section 2, we know that the variation of the environment from the perspective of agent 1 is upper bounded by $O(T^\beta)$. Substituting the value of Δ with $O(T^\beta)$ in Theorem 2 or Theorem 3 leads to the desired result. □

H.2. Proof of Theorem 6

Proof. From the (λ, μ)-smoothness of the MDP, it follows that

$$\lambda \cdot V_h^{(\pi^{1\star},\pi^{2\star})}(s) - \mu \cdot V_h^{(\pi^1,\pi^2)}(s) \leq V_h^{(\pi^{1\star},\pi^2)}(s), \forall s \in \mathcal{S}, h \in [H].$$

Therefore, it holds that

$$\begin{aligned}
&\sum_{m=1}^{M}\left(\lambda \cdot V_1^{(\pi^{1\star},\pi^{2\star})}(s_1^m) - (1+\mu)\cdot V_1^{(\pi^1,\pi^2)}(s_1^m)\right)\\
\leq&\sum_{m=1}^{M}\left(V_1^{(\pi^{1\star},\pi^2)}(s_1^m) - V_1^{(\pi^1,\pi^2)}(s_1^m)\right)\\
\leq&\sum_{m=1}^{M}\left(\sup_{\pi^{1\star}} V_1^{(\pi^{1\star},\pi^2)}(s_1^m) - V_1^{(\pi^1,\pi^2)}(s_1^m)\right)\\
=&\mathcal{R}^{\pi^2}(\pi^1, M) = \widetilde{O}(T^{\frac{\beta+2}{3}}),
\end{aligned}$$

where the last step follows from Theorem 5. Rearranging the terms leads to the desired result. □

I. Simulations Setup

We compare RestartQ-UCB with three baseline algorithms: LSVI-UCB-Restart (Zhou et al., 2020), Q-Learning UCB, and Epsilon-Greedy (Watkins, 1989). LSVI-UCB-Restart is a state-of-the-art non-stationary RL algorithm that combines optimistic least-squares value iteration with periodic restarts. It is originally designed for non-stationary RL in linear MDPs, but in our simulations we reduce it to the tabular case by setting the feature map to be essentially an identity mapping, i.e., the feature dimension is set to be $d = S \times A$. Q-Learning UCB is simply our RestartQ-UCB algorithm with no restart. It is a Q-learning based algorithm that uses upper confidence bounds to guide the exploration. Epsilon-Greedy is also a Q-learning based algorithm with restarts. Compared with RestartQ-UCB, Epsilon-Greedy does not employ a UCB-based bonus term to explicitly force exploration. Instead, it takes the greedy action according to the estimated Q function with a high probability $1 - \varepsilon$, and explores an action from the action set uniformly at random with probability ε.

We evaluate the cumulative rewards of the four algorithms on a variant of a reinforcement learning task named Bidirectional Diabolical Combination Lock (Agarwal et al., 2020; Misra et al., 2020). This task is designed to be particularly difficult for *exploration*. At the beginning of each episode, the agent starts at a fixed state. According to its first action, the agent transitions to one of the two paths, or "combination locks", each of length H. Each path is a chain of H states, where the state at the endpoint of each path gives a high reward. At each step on the path, there is only one "correct" action that leads the agent to the next state on the path, while the other $A - 1$ actions lead it to a sinking state that yields a small per-step reward of $\frac{1}{8H}$ ever since. Since we are considering a non-deterministic MDP, each intended transition "succeeds" with probability 0.98; that is, even if the agent takes the correct action at a certain step, there is still a 0.02 probability that it will end in the sinking state. The agent obtains a 0 reward when taking a correct action, and gets a $\frac{1}{8H}$ reward at the step when it transitions to the sinking state. Finally, the endpoint state of one path gives a reward of 1, while the other endpoint only gives a reward of 0.25. As argued in Agarwal et al. (2020), the following properties make this task especially challenging: First, it has sparse high rewards, and uniform exploration only has a A^{-H} probability of reaching a high reward endpoint. Second, it has dense low rewards, and a locally optimal policy will lead to the sinking state quickly. Third, there is no indication which path has the globally optimal reward, and the agent must remember to still visit the other one. Interested readers can refer to Section 5.1 of Agarwal et al. (2020) for detailed descriptions of the task.

We introduce two types of non-stationarity to the Bidirectional Diabolical Combination Lock task, namely *abrupt* variations and *gradual* variations. For abrupt variations, we periodically switch the two high-reward endpoints: One high-reward endpoint gives a reward of 1 at the beginning, and abruptly changes to a reward of 0.25 after a certain number of episodes, and then switches back to the reward of 1 after the same number of episodes. The other high-reward endpoint goes the other way around. For gradual changes, we gradually vary the transition probability at the starting state: At the first episode, one action leads to the first path with 0.98 probability, and to the second path with 0.02 probability. We linearly decrease its probability of leading to the first path, and increase its probability to the second path. As a result, at the last episode, this action would lead to the first path with 0.02 probability, and to the second path with 0.98 probability instead. The same is true for the other actions.

For simplicity, we use Hoeffding-based bonus terms in the simulations for RestartQ-UCB. We set $M = 5000, H = 5, S = 10$, and $A = 2$. For abrupt variations, we switch the two high-reward endpoints after every 1000 episodes. The hyper-parameters for each algorithm are optimized individually. For RestartQ-UCB, LSVI-UCB-Restart, and Epsilon-Greedy, we restart the algorithms after every 1000 episodes both for abrupt variations and gradual variations. This is the same frequency as the abrupt variation of the environment (because the restart frequency is optimized as a hyper-parameter), although it turns out that other restart frequencies lead to very similar results. For Epsilon-Greedy, we set the exploration probability to be $\varepsilon = 0.05$. All results are averaged over 30 runs on a laptop with an Intel Core i5-9300H CPU and 16 GB memory.