A. Proof of Theorem 1

To start, we list some technical lemmas. Then we prove that our estimation of the Q-function is optimistic with high probability, *i.e.*, $\tilde{Q}_h^k(s,a) \geq Q_h^*(s,a)$. We begin by proving Lemma 1 and cite a lemma in (Zhang et al., 2020a).

Lemma 2 (Bennet's inequality.)**.** *Let $Z, Z_1, ...Z_n$ be i.i.d. random variables with values in $[0,1]$ and let $\delta > 0$. Define $\mathbb{V}Z = \mathbb{E}[(Z - \mathbb{E}Z)^2]$. Then we have*

$$\mathbb{P}[|\mathbb{E}[Z] - \frac{1}{n}\sum_{i=1}^{n} Z_i| > \sqrt{\frac{2\mathbb{V}Z\ln(2/\delta)}{n}} + \frac{\ln(2/\delta)}{n}] \leq \delta.$$

Lemma 3 (Lemma 10 in (Zhang et al., 2020a).)**.** *Let $(M_n)_{n\geq 0}$ be a martingale such that $M_0 = 0$ and $|M_n - M_{n-1}| \leq c$ for some $c > 0$ and any $n \geq 1$. Let $\mathrm{Var}_n = \sum_{k=1}^{n} \mathbb{E}[(M_k - M_{k-1})^2 | \mathcal{F}_{k-1}]$ for $n \geq 0$, where $\mathcal{F}_k = \sigma(M_1, ..., M_k)$. Then for any positive integer n, and any $\epsilon, \delta > 0$, we have that*

$$\mathbb{P}[|M_n| \geq 2\sqrt{2}\sqrt{\mathrm{Var}_n \ln(1/\delta)} + 2\sqrt{\epsilon\ln(1/\delta)} + 2c\ln(1/\delta)] \leq 2(\log_2(\frac{nc^2}{\epsilon}) + 1)\delta.$$

A.1. Proof of Lemma 1

Proof. Recall that $Q_h(s,a) - \hat{Q}_h(s,a) = (\tilde{r}(s,a) - \hat{r}(s,a)) + (\tilde{P}_{s,a} - \hat{P}_{s,a})V_{h+1} + (\tilde{b}_h(s,a) - \hat{b}_h(s,a))$, where

$$\begin{aligned}
\hat{b}_h =& c_1\sqrt{\frac{\mathbb{V}(\hat{P}, V_{h+1})\iota}{\max\{\hat{n}(s,a),1\}}} + c_2\sqrt{\frac{\hat{r}(s,a)\iota}{\max\{\hat{n}(s,a),1\}}} + c_3\frac{\iota}{\max\{\hat{n}(s,a),1\}},\\
\tilde{b}_h =& c_1\min\left\{\sqrt{\frac{\mathbb{V}(\tilde{P}, V_{h+1})\iota}{|\tilde{n}(s,a) - C|}}, 1\right\} + c_2\min\left\{\sqrt{\frac{\tilde{r}\iota}{|\tilde{n}(s,a) - C|}}, 1\right\} + c_3\min\left\{\frac{\iota}{|\tilde{n}(s,a) - C|}, 1\right\}\\
&+ 2\min\left\{\frac{2C}{|\tilde{n}(s,a) - C|}, 1\right\} + (c_1 + c_2)\min\left\{\frac{\sqrt{C\iota}}{|\tilde{n}(s,a) - C|}, 1\right\}.
\end{aligned}$$

We prove the next lemma that characterize the the difference between biased and unbiased estimators,

Lemma 4. *For any vector $V \in R^S$, $V(s) \in [0,1]$ for any $s \in S$, it holds that*

$$||\tilde{P}_{s,a} - \hat{P}_{s,a}||_1 \leq 2\min\{\frac{C}{|\tilde{n}(s,a) - C|}, 1\}, \tag{4}$$

$$|\mathbb{V}(\tilde{P}_{s,a}, V) - \mathbb{V}(\hat{P}_{s,a}, V)| \leq 6\min\{\frac{C}{|\tilde{n}(s,a) - C|}, 1\}, \tag{5}$$

$$|\tilde{r} - \hat{r}| \leq \min\{\frac{C}{|\tilde{n} - C|}, 1\}. \tag{6}$$

For (4), we calculate that

$$\begin{aligned}
||\tilde{P}_{s,a} - \hat{P}_{s,a}||_1 =& \sum_{s'} |\tilde{P}_{s,a,s'} - \hat{P}_{s,a,s'}|\\
=& \sum_{s'} |\frac{\tilde{n}(s,a,s')}{\tilde{n}(s,a)} - \frac{\hat{n}(s,a,s')}{\hat{n}(s,a)}|\\
\leq& \sum_{s'} |\frac{\tilde{n}(s,a,s')}{\tilde{n}(s,a)} - \frac{\hat{n}(s,a,s')}{\tilde{n}(s,a) - C}|\\
\leq& \sum_{s'} \frac{\tilde{n}(s,a)|\tilde{n}(s,a,s') - \hat{n}(s,a,s')|}{\tilde{n}(s,a)|\tilde{n}(s,a) - C|} + \frac{C\tilde{n}(s,a,s')}{\tilde{n}(s,a)|\tilde{n}(s,a) - C|}\\
\leq& \frac{2C}{|\tilde{n}(s,a) - C|}.
\end{aligned}$$

Here the second inequality holds because $\sum_{s'}|\hat{n}(s,a,s')-\tilde{n}(s,a,s')|\leq C$ for any s,a. Note that $||\tilde{P}_{s,a}-\hat{P}_{s,a}||_1\leq 2$, we have (4) holds. For (5),

$$\begin{aligned}
&|\mathbb{V}(\tilde{P}_{s,a},V)-\mathbb{V}(\hat{P}_{s,a},V)|\\
\leq&|\tilde{P}_{s,a}(V)^2-\hat{P}_{s,a}(V)^2|+|(\tilde{P}_{s,a}V)^2-(\hat{P}_{s,a}V)^2|\\
\leq&||\tilde{P}_{s,a}-\hat{P}_{s,a}||_1+||\tilde{P}_{s,a}-\hat{P}_{s,a}||_1(\tilde{P}V+\hat{P}V)\\
\leq&3||\tilde{P}_{s,a}-\hat{P}_{s,a}||_1
\end{aligned}$$

Hence by (4), we finished the proof of (5). (6) comes straight from the definition of $\tilde{r}$ and $\hat{r}$. Finally, by plugging (4), (5) and (6) into $Q_h-\hat{Q}_h$ and using $(\tilde{P}_{s,a}-\hat{P}_{s.a})V_{h+1}\leq||\tilde{P}_{s,a}-\hat{P}_{s.a}||_1$, we finished the proof of Lemma 1. □

Recall that the Lemma 1 in (Zhang et al., 2020a) states that $\hat{Q}_h^k(s,a)\geq Q_h^*(s,a)$ holds with probability $1-2SA(\log_2(KH)+1)\delta$, as a result, $Q_h^k(s,a)\geq Q_h^*(s,a)$, which implies that our estimation of Q-function is optimistic. We denote by $\mathcal{E}_1$ the event that $Q_h^k(s,a)\geq Q_h^*(s,a)$ for any (k,h,s,a).

A.2. Bounding Bellman Error

We begin with the following lemma that bounds the Bellman error induced by the Q-function.

Lemma 5. *With probability* $1-3S^2AH(\log_2(KH)+1)\delta$, *for any* $1\leq k\leq K$, $1\leq h\leq H$ *and* (s,a), *it holds that*

$$\begin{aligned}
&Q_h^k(s,a)-r(s,a)-P_{s,a}V_{h+1}^k\\
\leq&\min\{2\tilde{b}_h^k(s,a)+\frac{2C}{\tilde{n}(s,a)+C}+\sqrt{\frac{2\mathbb{V}(P_{s,a},V_{h+1}^*)\iota}{\hat{n}^k(s,a)}}+\sqrt{\frac{2S\mathbb{V}(P_{s,a},V_{h+1}^k-V_{h+1}^*)\iota}{\hat{n}^k(s,a)}}+\frac{2S\iota}{3\hat{n}^k(s,a)},1\}.
\end{aligned}\tag{7}$$

Proof. Under event $\mathcal{E}_1$, we have that with probability $1-SAH(\log_2(KH)+1)\delta$, for all $(s,a,k,h)\in\mathcal{S}\times\mathcal{A}\times[K]\times[H]$,

$$\begin{aligned}
&Q_h^k(s,a)-r(s,a)-P_{s,a}V_{h+1}^k\\
\leq&\tilde{r}_h^k(s,a)-r(s,a)+\tilde{b}_h^k(s,a)+(\tilde{P}_{s,a}-\hat{P}_{s,a})V_{h+1}^k+(\hat{P}_{s,a}-P_{s,a})(V_{h+1}^k-V_{h+1}^*)+(\hat{P}_{s,a}-P_{s,a})V_{h+1}^*\\
\leq&2\tilde{b}_h^k(s,a)+(\tilde{P}_{s,a}-\hat{P}_{s,a})V_{h+1}^k+(\hat{P}_{s,a}-P_{s,a})(V_{h+1}^k-V_{h+1}^*)+(\hat{P}_{s,a}-P_{s,a})V_{h+1}^*.
\end{aligned}\tag{8}$$

By Bennet's inequality (Lemma 2) we have that for each s',

$$\mathbb{P}[|\hat{P}_{s,a,s'}-P_{s,a,s'}|>\sqrt{\frac{2P_{s,a,s'}\iota}{\hat{n}^k(s,a)}}+\frac{\iota}{3\hat{n}^k(s,a)}]\leq\delta.$$

So with probability at least $1-S\delta$, we have,

$$\begin{aligned}
(\hat{P}_{s,a}-P_{s,a})(V_{h+1}^k-V_{h+1}^*)=&\sum_{s'}(\hat{P}_{s,a,s'}^k-P_{s,a,s'})(V_{h+1}^k(s')-V_{h+1}^*(s')-P_{s,a}(V_{h+1}^k-V_{h+1}^*))\\
\leq&\sum_{s'}\sqrt{\frac{2P_{s,a,s'}\iota}{\hat{n}^k(s,a)}}|V_{h+1}^k(s')-V_{h+1}^*(s')-P_{s,a}(V_{h+1}^k-V_{h+1}^*)|+\frac{S\iota}{3\hat{n}^k(s,a)}\\
\leq&\sqrt{\frac{\hat{n}^k(s,a)}{2S\mathbb{V}(P_{s,a},V_{h+1}^k-V_{h+1}^*)}}+\frac{S\iota}{3\hat{n}^k(s,a)},
\end{aligned}\tag{9}$$

where the first equality holds because $\sum_{s'}\hat{P}_{s,a,s'}^k=\sum_{s'}P_{s,a,s'}=1$ and the last inequality holds by Cauchy-Schwartz inequality. By Bennet's inequality again (Lemma 2), we obtain that

$$\mathbb{P}[|(\hat{P}_{s,a}^k-P_{s,a})V_{h+1}^*|>\sqrt{\frac{2\mathbb{V}(P_{s,a},V_{h+1}^*)\iota}{\hat{n}^k(s,a)}}+\frac{\iota}{3\hat{n}^k(s,a)}]\leq\delta.\tag{10}$$

On the other hand, we have that

$$\begin{aligned}(\tilde{P}-\hat{P})V_{h+1}^k =& \sum_{s'}(\frac{\tilde{n}(s,a,s')}{\tilde{n}(s,a)}-\frac{\hat{n}(s,a,s')}{\hat{n}(s,a)})V_{h+1}^k(s')\\ \le& \sum_{s'}|\frac{\tilde{n}(s,a,s')}{\tilde{n}(s,a)}-\frac{\hat{n}(s,a,s')}{\tilde{n}(s,a)+C}|\\ \le& \sum_{s'}\frac{\tilde{n}(s,a)|\tilde{n}(s,a,s')-\hat{n}(s,a,s')|}{\tilde{n}(s,a)(\tilde{n}(s,a)+C)}+\frac{C\tilde{n}(s,a,s')}{\tilde{n}(s,a)(\tilde{n}(s,a)+C)}\\ \le& \frac{2C}{\tilde{n}(s,a)+C},\end{aligned} \tag{11}$$

where the first inequality holds because $|\hat{n}(s,a)-\tilde{n}(s,a)|\le C$, and the last inequality holds because $\sum_{s'}|\tilde{n}(s,a,s')-\hat{n}(s,a,s')|\le C$. Combining (8), (9), (10) and (11) and via a union bound over k,h,s,a, we conclude that (7) holds with probability $1-3S^2AH(\log_2(KH)+1)\delta$. □

In the rest of this section, we let $\beta_h^k(s,a)$ be the shorthand of RHS of (7), *i.e.*,

$$\beta_h^k(s,a):=\min\{2\tilde{b}_h^k(s,a)+\frac{2C}{\tilde{n}(s,a)+C}+\sqrt{\frac{2\mathbb{V}(P_{s,a},V_{h+1}^*)\iota}{\hat{n}^k(s,a)}}+\sqrt{\frac{2S\mathbb{V}(P_{s,a},V_{h+1}^k-V_{h+1}^*)\iota}{\hat{n}^k(s,a)}}+\frac{2S\iota}{3\hat{n}^k(s,a)},1\}.$$

A.3. Regret Analysis

Let $\mathcal{K}$ be the set of indice of episodes in which no update is triggered. By the update rule, it is easy to see that $|\mathcal{K}^C|\le SA(\log_2(KH)+1)$. Denote $h_0(k)$ to be the first time an update is triggered in the k-th episode if there is and otherwise $H+1$. Define $\mathcal{X}_0=\{(k,h_0(k))|k\in\mathcal{K}^C\}$ and $\mathcal{X}=\{(k,h)|k\in\mathcal{K}^C,h_0(k)+1\le h\le H\}$. We further define $\bar{V}_h^k(s_h^k,a_h^k)=\mathbb{I}[(k,h)\notin\mathcal{X}]\tilde{V}_h^k(s_h^k,a_h^k)$. We also set $\bar{\beta}_h^k(s_h^k,a_h^k)=\mathbb{I}[(k,h)\notin\mathcal{X}]\beta_h^k(s_h^k,a_h^k)$ and $\bar{r}_h^k=\mathbb{I}[(k,h)\notin\mathcal{X}]r(s_h^k,a_h^k)$. By Lemma 5, we have that with probability $1-3S^2AH(\log_2(KH)+1)\delta$,

$$\bar{V}_h^k(s_h^k,a_h^k)\le\bar{r}_h^k+\bar{\beta}_h^k(s_h^k,a_h^k)+P_{s,a}\bar{V}_{h+1}^k,$$

for any $(k,h)\notin\mathcal{X}_0$ and

$$\bar{V}_h^k(s_h^k,a_h^k)\le\bar{r}_h^k+\bar{\beta}_h^k(s_h^k,a_h^k)+P_{s,a}\bar{V}_{h+1}^k+1,$$

for any $(k,h)\in\mathcal{X}_0$.

By Lemma 5, with probability at least $1-5S^2AH(\log_2(KH)+1)\delta$, it holds that

$$\begin{aligned}\mathrm{Regret}(K):=&\sum_{k=1}^K(V_1^*(s_1^k)-\tilde{V}_1^{\pi^k}(s_1^k))\\ \le&\sum_{k=1}^K(V_1^k(s_1^k)-\tilde{V}_1^{\pi^k}(s_1^k))\\ =&\sum_{k=1}^K(\bar{V}_1^k(s_1^k)-\tilde{V}_1^{\pi^k}(s_1^k))\\ =&\sum_{k=1}^K(\bar{V}_1^k(s_1^k)-\sum_{h=1}^H\bar{r}_h^k)+\sum_{k=1}^K(\sum_{h=1}^H\bar{r}_h^k-\tilde{V}_1^{\pi^k}(s_1^k))\\ =&\sum_{k=1}^K\sum_{h=1}^H(P_{s_h^k,a_h^k}\bar{V}_{h+1}^k-\bar{V}_{h+1}^k(s_{h+1}^k))+\sum_{k=1}^K\sum_{h=1}^H(\bar{V}_h^k(s_h^k)-\bar{r}_h^k-P_{s_h^k,a_h^k}\bar{V}_{h+1}^k)+\sum_{k=1}^K(\sum_{h=1}^H\bar{r}_h^k-\tilde{V}_1^{\pi^k}(s_1^k))\\ \le&\sum_{k=1}^K\sum_{h=1}^H(P_{s_h^k,a_h^k}\bar{V}_{h+1}^k-\bar{V}_{h+1}^k(s_{h+1}^k))+\sum_{k=1}^K\sum_{h=1}^H\bar{\beta}_h^k(s_h^k,a_h^k)+\sum_{k=1}^K(\sum_{h=1}^H\bar{r}_h^k-\tilde{V}_1^{\pi^k}(s_1^k))+|\mathcal{K}^C|.\end{aligned} \tag{12}$$

Here the first inequality is due to the optimism of Q-function, and the last inequality holds by Lemma 5. Define $M_1 = \sum_{k=1}^K \sum_{h=1}^H (P_{s_h^k,a_h^k} \bar{V}_{h+1}^k - \bar{V}_{h+1}^k(s_{h+1}^k))$, $M_2 = \sum_{k=1}^K \sum_{h=1}^H \bar{\beta}_h^k(s_h^k, a_h^k)$ and $M_3 = \sum_{k=1}^K (\sum_{h=1}^H \bar{r}_h^k - \tilde{V}_1^{\pi^k}(s_1^k))$.

Bounding M_1: For the M_1 term, we note that it can be viewed as a martingale. Hence by using a variance dependent concentraion inequality (3), we have that

$$\mathbb{P}[|M_1| > 2\sqrt{2\sum_{k=1}^K \sum_{h=1}^H \mathbb{V}(P_{s_h^k,a_h^k}, \bar{V}_{h+1}^k)\iota + 6\iota}] \le 2(\log_2(KH)+1)\delta.$$

So in order to bound M_1, it suffices to bound $M_4 := \sum_{k=1}^K \sum_{h=1}^H \mathbb{V}(P_{s_h^k,a_h^k}, \bar{V}_{h+1}^k)$. We will deal with this term later.

Bounding M_3: For the M_3 term, we have

$$\begin{aligned} M_3 =& \sum_{k=1}^K (\sum_{h=1}^H \bar{r}_h^k - \tilde{V}_1^{\pi^k}(s_1^k)) \\ =& \sum_{k=1}^K \sum_{h=1}^H (\bar{r}_h^k - r_h^k) + \sum_{k=1}^K (\sum_{h=1}^H r_h^k - \tilde{V}_1^{\pi^k}(s_1^k)) \\ \le& \sum_{k=1}^K \sum_{h=1}^H (r(s_h^k, a_h^k) - r_h^k) + \sum_{k=1}^K (\sum_{h=1}^H r_h^k - \tilde{V}_1^{\pi^k}(s_1^k)). \end{aligned} \tag{13}$$

For the first term $\sum_{k=1}^K \sum_{h=1}^H (r(s_h^k, a_h^k) - r_h^k)$ in (13), by (3), we obtain that

$$\mathbb{P}[|\sum_{k=1}^K \sum_{h=1}^H (r(s_h^k, a_h^k) - r_h^k)| > 2\sqrt{2\sum_{k=1}^K \sum_{h=1}^H \mathrm{Var}(s,a)\iota + 6\iota}] \le 2(\log_2(KH)+1)\delta,$$

where $\mathrm{Var}(s,a) := \mathbb{E}[(R(s,a) - \mathbb{E}[R(s,a)])^2]$. Moreover, since the random variable $R(s,a) \in [0,1]$, we have

$$\sum_{k=1}^K \sum_{h=1}^H \mathrm{Var}(s,a) \le \sum_{k=1}^K \sum_{h=1}^H r(s,a) \le \sum_{k=1}^K \sum_{h=1}^H (r(s_h^k, a_h^k) - r_h^k) + K.$$

So we have $\mathbb{P}[|\sum_{k=1}^K \sum_{h=1}^H (r(s_h^k, a_h^k) - r_h^k)| > 2\sqrt{2(\sum_{k=1}^K \sum_{h=1}^H (r(s_h^k, a_h^k) - r_h^k) + K)\iota + 6\iota}] \le 2(\log_2(KH)+1)\delta$, which implies that with probability $1 - 2(\log_2(KH)+1)\delta$ we have

$$\sum_{k=1}^K \sum_{h=1}^H (r(s_h^k, a_h^k) - r_h^k) \le 6\sqrt{K\iota} + 21\iota.\iota \tag{14}$$

For the second term in (13), because $\sum_{h=1}^H r_h^k - \tilde{V}_1^{\pi^k}(s_1^k)$, $k \in [K]$ is a martingale. Hence by Azuma's inequality, we have with probability $1-\delta$ it holds that

$$|\sum_{k=1}^K (\sum_{h=1}^H r_h^k - \tilde{V}_1^{\pi^k}(s_1^k))| \le \sqrt{2K\iota}. \tag{15}$$

Combining (14) and (15), we have that $\mathbb{P}[|M_3| > 8\sqrt{K\iota} + 6\iota] \le 2(\log_2(KH)+1)\delta$.

Bounding M_2: For the M_2 term, recall that

$$\begin{aligned} \beta_h^k(s,a) =& O(\min\{\sqrt{\frac{\mathbb{V}(\hat{P}_{s,a}, V_{h+1}^k)\iota}{|\tilde{n}^k(s,a) - C|}}, 1\} + \min\{\sqrt{\frac{\hat{r}_h^k(s,a)\iota}{|\tilde{n}^k(s,a) - C|}}, 1\} + \min\{\frac{\iota}{|\tilde{n}^k(s,a) - C|}, 1\} + \min\{\frac{2C}{|\tilde{n}^k(s,a) - C|}, 1\} \\ &+ \min\{\frac{\sqrt{C\iota}}{|\tilde{n}^k(s,a) - C|}, 1\} + \frac{2C}{\tilde{n}^k(s,a) + C} + \sqrt{\frac{\mathbb{V}(P_{s,a}, V_{h+1}^*)}{\hat{n}^k(s,a)}} + \sqrt{\frac{S\mathbb{V}(P_{s,a}, V_{h+1}^k - V_{h+1}^*)\iota}{\hat{n}^k(s,a)}} + \frac{S\iota}{\hat{n}^k(s,a)}). \end{aligned}$$

First note that

$$O(\min\{\frac{\iota}{|\tilde{n}^k(s,a)-C|},1\}+\min\{\frac{2C}{|\tilde{n}^k(s,a)-C|},1\}+\min\{\frac{\sqrt{C\iota}}{|\tilde{n}^k(s,a)-C|},1\}+\frac{2C}{\tilde{n}^k(s,a)+C})=O(\min\{\frac{C\iota}{|\tilde{n}^k(s,a)-C|},1\}).$$

By Lemma 2, we have

$$\mathbb{P}[\hat{P}_{s,a,s'}>\frac{3}{2}P_{s,a,s'}+\frac{4\iota}{3\hat{n}^k(s,a)}]\leq\mathbb{P}[\hat{P}_{s,a,s'}-P_{s,a,s'}>\sqrt{\frac{2P_{s,a,s'}\iota}{\hat{n}^k(s,a)}}+\frac{\iota}{3\hat{n}^k(s,a)}]\leq\delta,$$

which implies that, with probability $1-2S^2AH(\log_2(KH)+1)\delta$, it holds that

$$\begin{aligned}\mathbb{V}(\hat{P}_{s,a,s'},V^k_{h+1})=&\sum_{s'}\hat{P}_{s,a,s'}(V^k_{h+1}(s')-\hat{P}^k_{s,a}V^k_{h+1})^2\\ \leq&\sum_{s'}\hat{P}^k_{s,a,s'}(V^k_{h+1}(s')-P_{s,a}V^k_{h+1})^2\\ \leq&\sum_{s'}(\frac{3}{2}P_{s,a,s'}+\frac{4\iota}{3\hat{n}^k(s,a)})(V^k_{h+1}(s')-P_{s,a}V^k_{h+1})^2\\ \leq&\frac{3}{2}\mathbb{V}(P_{s,a},V^k_{h+1})+\frac{4S\iota}{3\hat{n}^k(s,a)}.\end{aligned}$$

Note that $\mathbb{V}(P,X+Y)\leq 2(\mathbb{V}(P,X)+\mathbb{V}(P,Y))$ for any P,X,Y, so we conclude that

$$\begin{aligned}\beta^k_h(s,a)\leq O(&\min\{\sqrt{\frac{\mathbb{V}(P_{s,a},V^k_{h+1})\iota}{|\tilde{n}^k(s,a)-C|}},1\}+\min\{\sqrt{\frac{\hat{r}^k_h(s,a)\iota}{|\tilde{n}^k(s,a)-C|}},1\}+\sqrt{\frac{S\mathbb{V}(P_{s,a},V^k_{h+1}-V^*_{h+1})\iota}{\hat{n}^k(s,a)}}\\ &+\min\{\frac{C\iota}{|\tilde{n}^k(s,a)-C|},1\}+\frac{S\iota}{\hat{n}^k(s,a)}))\end{aligned}$$

According to the update rule, despite those episodes in which an update is triggered, the number of visit of (s,a) between the i-th update of $\hat{P}_{s,a}$ and the $i+1$-th update of $\hat{P}_{s,a}$ do not exceeds 2^{i-1}, *i.e.*, for any (s,a) and any $i\geq 3$, we have

$$\sum_{k=1}^K\sum_{h=1}^H\mathbb{I}[(s^k_h,a^k_h)=(s,a),\tilde{n}^k(s,a)-C=2^{i-1}]\mathbb{I}[(k,h)\notin\mathcal{X}]\leq 2^{i-1}. \tag{16}$$

Let $l=\max\{i|2^{i-1}+C\leq KH\}$. We calculate that

$$\begin{aligned}&\sum_{k=1}^K\sum_{h=1}^H\min\{\frac{C\iota}{|\tilde{n}^k(s,a)-C|},1\}\mathbb{I}[(k,h)\notin\mathcal{X}]\\ =&\sum_{k=1}^K\sum_{h=1}^H\sum_{s,a}^l(\sum_{i=3}\mathbb{I}[(s^k_h,a^k_h)=(s,a),\tilde{n}^k(s,a)-C=2^{i-1}]\min\{\frac{C\iota}{|\tilde{n}^k(s,a)-C|},1\}\mathbb{I}[(k,h)\notin\mathcal{X}]\\ &+\mathbb{I}[(s^k_h,a^k_h)=(s,a),\tilde{n}^k(s,a)-C<4])\\ \leq&\sum_{k=1}^K\sum_{h=1}^H\sum_{s,a}^l\sum_{i=3}^l(\mathbb{I}[(s^k_h,a^k_h)=(s,a),\tilde{n}^k(s,a)-C=2^{i-1}]\mathbb{I}[(k,h)\notin\mathcal{X}]\frac{C\iota}{2^{i-1}})+(C+4)\\ \leq&\sum_{s,a}Cl\iota+(C+4)SA \qquad (17)\\ =&\tilde{O}(CSA). \qquad (18)\end{aligned}$$

Here (17) is by using (16). Let $\omega=\{\omega^k_h\geq 0|1\leq h\leq H,1\leq k\leq K\}$ be a group of non-negative weights such that

$w_h^k \in [0,1]$ for any k,h and $w_h^k = 0$ if $(k,h) \in \mathcal{X}$. We prove the following useful inequality:

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} \min\{\sqrt{\frac{\omega_h^k \iota}{\tilde{n}^k(s,a) - C}}, 1\} \\
\leq &\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s,a}\sum_{i=3}^{l} \mathbb{I}[(s_h^k, a_h^k) = (s,a), \tilde{n}^k(s,a) = 2^{i-1}]\sqrt{\frac{\omega_h^k \iota}{2^{i-1}}} + CSA + 8SA(\log_2(KH)+4)\iota \\
= &\sum_{s,a}\sum_{i=3}^{l} \frac{1}{\sqrt{2^{i-1}}} \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{I}[(s_h^k, a_h^k) = (s,a), \tilde{n}^k(s,a) = 2^{i-1}]\sqrt{\omega_h^k \iota} + CSA + 8SA(\log_2(KH)+4)\iota \\
\leq &\sum_{s,a}\sum_{i=3}^{l} \sqrt{\frac{\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{I}[(s_h^k, a_h^k) = (s,a), \tilde{n}^k(s,a) - C = 2^{i-1}]\iota}{2^{i-1}}} \sqrt{(\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{I}[(s_h^k, a_h^k) = (s,a), \tilde{n}^k(s,a) - C = 2^{i-1}]\omega_h^k)} \qquad (19) \\
&+ CSA + 8SA(\log_2(KH)+4)\iota \\
\leq &\sqrt{SAl\iota \sum_{k=1}^{K}\sum_{h=1}^{H} \omega_h^k} + CSA + 8SA(\log_2(KH)+4)\iota. \qquad (20)
\end{aligned}$$

Here (19) is by Cauchy-Schwarz inequality and (20) is due to (16). By using the same technique (or see (29) in (Zhang et al., 2020a)), we can prove that

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \sqrt{\frac{\omega_h^k}{\hat{n}^k(s_h^k, a_h^k)}} \leq \sqrt{SAl \sum_{k=1}^{K}\sum_{h=1}^{H} \omega_h^k} + 8SA(\log_2(KH)+4). \qquad (21)$$

Let $I(k,h)$ be the shorthand of $\mathbb{I}[(k,h) \notin \mathcal{X}]$. By plugging in respectively $\omega_h^k = I(k,h)\hat{r}_h^k(s_h^k, a_h^k), I(k,h)\mathbb{V}(P_{s_h^k, a_h^k}, V_{h+1}^k)$ into (20) and $\omega_h^k = I(k,h)\mathbb{V}(V_{h+1}^k - V_{h+1}^*)$ into (21), we obtain that

$$\begin{aligned}
M_2 = &\sum_{k=1}^{K}\sum_{h=1}^{H} \bar{\beta}_h^k(s_h^k, a_h^k) \\
= &\sum_{k=1}^{K}\sum_{h=1}^{H} I(k,h)\beta_h^k(s_h^k, a_h^k) \\
\leq &O(\sqrt{SAl\iota \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{V}(P_{s_h^k, a_h^k}, V_{h+1}^k) I(k,h)} + \sqrt{S^2Al\iota \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{V}(P_{s_h^k, a_h^k}, V_{h+1}^k - V_{h+1}^*) I(k,h)} \\
&+ O(\sqrt{SAl\iota \sum_{k=1}^{K}\sum_{h=1}^{H} \hat{r}_h^k(s_h^k, a_h^k) I(k,h)} + \tilde{O}(CSA + S^2A) \qquad (22)
\end{aligned}$$

We define $M_5 := \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{V}(P_{s_h^k, a_h^k}, V_{h+1}^k - V_{h+1}^*) I(k, h+1)$. We state the following Lemmas in (Zhang et al., 2020a) to complete the proof of Theorem 1, which we omit the detailed proof.

Lemma 6 (Lemma 5 in (Zhang et al., 2020a)). $\sum_{k=1}^{K}\sum_{h=1}^{H} \hat{r}_h^k(s_h^k, a_h^k) I(k,h) \leq 2\sum_{k=1}^{K}\sum_{h=1}^{H} r_h^k + 4SA \leq 2K + 4SA$.

Lemma 7 (Lemma 6 in (Zhang et al., 2020a)). *With probability* $1 - 2(\log_2(KH)+1)\log_2(KH)\delta$, *it holds that*

$$M_4 \leq 2M_2 + 2|\mathcal{K}^C| + 2K + \max\{46\iota, 8\sqrt{(M_2 + |\mathcal{K}^C| + K)\iota} + 6\iota\}.$$

Lemma 8 (Lemma 7 in (Zhang et al., 2020a)). *With probability* $1 - 2(\log_2(KH)+1)\log_2(KH)\delta$, *it holds that*

$$M_5 \leq 2\max\{M_2, 1\} + 2|\mathcal{K}^C| + \max\{46\iota, 8\sqrt{(M_2 + |\mathcal{K}^C|)\iota} + 6\iota\}.$$

Combining the above lemmas with (22), we have that with probability $1-(6S^2AH(\log_2(KH)+1)+6(\log_2(KH)+1)\log_2(H))\delta$,

$$\begin{aligned}
M_2 \leq & O(\sqrt{SAl\iota(M_4+|\mathcal{K}^C|)}+\sqrt{S^2Al\iota(M_5+|\mathcal{K}^C|)}+\sqrt{SAl\iota K}+CSA+S^2A\iota\log_2(KH)),\\
M_4 \leq & 2M_2+2|\mathcal{K}^C|+\max\{46\iota, 8\sqrt{(M_2+2K)\iota}+6\iota\},\\
M_5 \leq & 2\max\{M_2,1\}+2|\mathcal{K}^C|+\max\{46\iota,\sqrt{M_2\iota}+6\iota\}.
\end{aligned}$$

Which implies that

$$\begin{aligned}
M_2 \leq & O(\sqrt{SAKl\iota}+CSA+\sqrt{S^2Al\sqrt{M_2}\iota^{3/2}}+\sqrt{SAl\iota K}+S^2A\iota\log_2(KH)) && (23)\\
\leq & O(\sqrt{SAKl\iota}+CSA+S^2A\iota\log_2(KH)). && (24)
\end{aligned}$$

Recalling (12) and (23), we conclude that, with probability $1-(10S^2AH(\log_2(KH)+2)+6(\log_2(KH)+1)\log_2(KH)+1)\delta$,

$$\begin{aligned}
\mathrm{Regret}(K) \leq & M_1+M_2+M_3\\
\leq & O(\sqrt{SAKl\iota}+S^2A\iota\log_2(KH)+CSA+\sqrt{K\iota})\\
= & O(\sqrt{SAKl\iota}+S^2A\iota\log_2(KH)+CSA).
\end{aligned}$$

Hence by rescaling δ, we finish the proof of Theorem 1.

B. Proof of Theorem 3

Proof. First, since we run Algorithm 4 to learn a $\frac{\mu_{min}}{2}$-policy cover, from Theorem 4, this requires

$$H(N'_{\mathrm{g}}+N'_{\phi}+N'_{\mathrm{p}})=\tilde{\Omega}\left(\frac{M^4A^4H^2\log|\mathcal{G}|}{\mu_{\min}^4\gamma^2}+\frac{HMA}{\mu_{\min}}+\frac{M^2AH^2}{\mu_{\min}^3}\right)$$

trajectories.

Next, we prove that in phase 2, Algorithm 2 learns a decoding function with ϵ_{f} decoding error, where ϵ_{f} is to be defined later. Formally, we prove that the following condition holds with high probability:

Condition 1 (Bijection between learned and true states)**.** *There exists $\epsilon_{\mathrm{f}}<\frac{1}{2}$ such that there is a bijective mapping $\alpha_h:\widehat{\mathcal{S}}_h\to\mathcal{S}_h$ for which*

$$\mathbb{P}_{x\sim q(\cdot|\alpha_h(\hat{s}))}\left[\hat{f}_h(x)=\hat{s}\right]\geq 1-\epsilon_{\mathrm{f}}.$$

In other words, this condition states that every estimated latent state $\hat{s}$ roughly corresponds to a true latent state $\alpha_h(\hat{s})$, when we use the decoding function $\hat{f}_h$. This is because all but an ϵ_{f} fraction of contexts drawn from $\alpha_h(\hat{s})$ are decoded to their true latent state, and for each latent state s, there is a distinct estimated state $\alpha_h^{-1}(s)$ as the map α_h is a bijection. For simplicity, we define $\mathbf{p}(s,a)\in\mathbb{R}^M$ to be the forward transition distribution over $\mathcal{S}_h$ for $s\in\mathcal{S}_{h-1}$ and $a\in\mathcal{A}$. We abuse notation to similarly use $\mathbf{p}(\hat{s},a)\in\mathbb{R}^M$ to be the vector $\{\mathbb{P}(s\mid\hat{s},a)\}_{s\in\mathcal{S}_h}$ of conditional probabilities $\mathcal{S}_h$ for $\hat{s}\in\widehat{\mathcal{S}}_{h-1}$ and $a\in\mathcal{A}$. Note that unlike $s\in\mathcal{S}_{h-1},\hat{s}\in\widehat{\mathcal{S}}_{h-1}$ is not a Markovian state and hence the conditional probability vector $\mathbf{p}(\hat{s},a)$ depends on the specific distribution over $\widehat{\mathcal{S}}_{h-1}\times\mathcal{A}$. In the following we will use $\mathbf{p}^{\nu}(\hat{s},a)$ to emphasize this dependency where ν is the distribution, where ν is a distribution over $\widehat{\mathcal{S}}_{h-1}\times\mathcal{A}$.

In the proof, we often compare two vectors indexed by $\mathcal{S}_h$ and $\widehat{\mathcal{S}}_h$. We will assume the order of the indices of these two vectors are matched according to α_h.

Establishing Condition 1. In order to establish the condition, we need to show that our decoding function $\hat{f}_h$ predicts the underlying latent state correctly almost always. We do this in two steps. Since the functions $\hat{f}_h$ are derived based on $\hat{\mathrm{g}}_h$ and $\widehat{\phi}_h$, we analyze the properties of these two objects in the following two lemmas in (Du et al., 2019). In order to state the first lemma, we need some additional notation. Note that η_h and $\hat{f}_{h-1}$ induce a distribution over $\mathcal{S}_{h-1}\times\widehat{\mathcal{S}}_{h-1}\times\mathcal{A}\times\mathcal{S}_h$. We denote

this distribution as ν_h. With this distribution, we define the conditional backward probability $\hat{\mathbf{b}}_{\nu_h} : \mathcal{S}_h \to \triangle\left(\widehat{\mathcal{S}}_{h-1} \times \mathcal{A}\right)$ as

$$\hat{b}_{\nu_h}\left(\hat{s}, a \mid s_1'\right) = \frac{p_{h-1}^{\nu_h}\left(s_1' \mid \hat{s}, a\right) \mathbb{P}^{\nu_h}(\hat{s}, a)}{\sum_{\hat{s}_1, a_1} p_{h-1}^{\nu_h}\left(s_1' \mid \hat{s}_1, a_1\right) \mathbb{P}^{\nu_h}\left(\hat{s}, a_1\right)} \tag{25}$$

Recall that $\mathbf{p}_{h-1}^{\nu_h}$ above refers to the distribution over s_1' according the transition dynamics, when $\hat{s}, a$ are induced by ν_h. With this notation, we have the following lemma.

Lemma 9 (Lemma G.2 in (Du et al., 2019)). *Assume $\epsilon_{\mathrm{f}} \leq \frac{\mu_{\min}^3 \gamma}{100 M^4 A^3}$. Then the distributions $\hat{b}_{\nu_h}\left(\hat{s}, a \mid s'\right)$ are well separated for any pair $s_1', s_2' \in \mathcal{S}_h$:*

$$\left\|\hat{\mathbf{b}}_{\nu_h}\left(s_1'\right) - \hat{\mathbf{b}}_{\nu_h}\left(s_2'\right)\right\|_1 \geq \frac{\mu_{\min}\gamma}{3MA} \tag{26}$$

Furthermore, if $N_{\mathrm{g}} = \Omega\left(\frac{M^3 A^3}{\epsilon_{\mathrm{f}} \mu_{\min}^3 \gamma^2} \log\left(\frac{|\mathcal{G}|H}{\delta}\right)\right)$, with probability at least $1 - \delta/H$, for every $s' \in \mathcal{S}_h, \hat{\mathrm{g}}_h$ satisfies

$$\mathbb{P}_{x' \sim q(\cdot|s')}\left[\left\|\hat{\mathbf{g}}_h\left(x'\right) - \hat{\mathbf{b}}_{\nu_h}\left(s'\right)\right\|_1 \geq \frac{\gamma\mu_{\min}}{100MA}\right] \leq \epsilon_{\mathrm{f}} \tag{27}$$

Proof. See (Du et al., 2019) for details. □

The first part of Lemma 9 tell us that the latent states at level h are well separated if we embed them using $\phi\left(s'\right) = \hat{\mathbf{b}}_{\nu_h}\left(s'\right)$ as the state embedding. The second part guarantees that our regression procedure estimates this representation accurately. Together, these assertions imply that any two contexts from the same latent state (up to an ϵ_{f} fraction) are close to each other, while contexts from two different latent states are well-separated. Formally, with probability at least $1 - \frac{\delta}{H}$ over the N_{g} training data:

1. For any $s' \in \mathcal{S}_h$ and $x_1', x_2' \sim q\left(\cdot \mid s'\right)$, we have with probability at least $1 - 2\epsilon_{\mathrm{f}}$ over the emission process

$$\left\|\hat{\mathrm{g}}_h\left(x_1'\right) - \hat{\mathrm{g}}_h\left(x_2'\right)\right\|_1 \leq \frac{\mu_{\min}\gamma}{50MA}$$

2. For any $s_1', s_2' \in \mathcal{S}_h$ such that $s_1' \neq s_2', x_1' \sim q\left(\cdot \mid s_1'\right)$ and $x_2' \sim q\left(\cdot \mid s_2'\right)$, we have with probability at least $1 - 2\epsilon_{\mathrm{f}}$ over the emission process

$$\left\|\hat{\mathrm{g}}_h\left(x_1'\right) - \hat{\mathrm{g}}_h\left(x_2'\right)\right\|_1 \geq \frac{\mu_{\min}\gamma}{4MA}$$

In other words, the mapping of contexts, as performed through the functions $\hat{\mathrm{g}}_h$ should be easy to cluster with each cluster roughly corresponding to a true latent state. Our next lemma guarantees that with enough samples for clustering, this is indeed the case.

Lemma 10 (Lemma G.3 in (Du et al., 2019)). *If $N_\phi = \Theta\left(\frac{MA}{\mu_{\min}} \log\left(\frac{MH}{\delta}\right)\right)$ and $\epsilon_{\mathrm{f}} \leq \frac{\delta}{100 H N_\phi}$ we have with probability at least $1 - \frac{\delta}{H}$, (1) for every $s' \in \mathcal{S}_h$, there exists at least one point $\mathbf{z} \in \mathcal{Z}$ such that $\mathbf{z} = \hat{\mathrm{g}}_h\left(x'\right)$ with $x' \sim q\left(\cdot \mid s'\right)$ and $\left\|\hat{\mathrm{g}}_h\left(x'\right) - \hat{\mathbf{b}}_{\nu_h}\left(s'\right)\right\|_1 \leq \frac{\mu_{\min}\gamma}{100MA}$ and (2) for every $\mathbf{z} = \hat{g}_h\left(x'\right) \in \mathcal{Z}$ with $x' \sim q\left(\cdot \mid s'\right), \left\|\hat{\mathrm{g}}_h\left(x'\right) - \hat{\mathbf{b}}_{\nu_h}\left(s'\right)\right\|_1 \leq \frac{\mu_{\min}\gamma}{100MA}$*

Proof. See (Du et al., 2019) for details. □

Based on Lemma 9 and 10, we can establish that Condition 1 holds with high probability. Note that Condition 1 consists of two parts. The first part states that there exists a bijective map $\alpha_h : \widehat{\mathcal{S}}_h \to \mathcal{S}_h$. The second part states that the decoding error is small. To prove the first part, we explicitly construct the map α_h and show it is bijective. We define $\alpha_h : \widehat{\mathcal{S}}_h \to \mathcal{S}_h$ as

$$\alpha_h\left(\hat{s}'\right) = \operatorname*{argmin}_{s \in \mathcal{S}_h} \left\|\phi\left(s'\right) - \widehat{\phi}\left(\hat{s}'\right)\right\|_1 \tag{28}$$

First observe that for any $\hat{s}' \in \widehat{\mathcal{S}}_h$, by the second conclusion of Lemma 10 we know there exists $s' \in \mathcal{S}_h$ such that

$$\left\|\widehat{\phi}\left(\hat{s}'\right) - \phi\left(s'\right)\right\| \leq \frac{\gamma\mu_{\min}}{100MA}$$

This also implies for any $s'' \neq s'$

$$\left\|\widehat{\phi}\left(\hat{s}'\right)-\phi\left(s''\right)\right\| \geq \left\|\phi\left(s''\right)-\phi\left(s''\right)\right\| - \left\|\widehat{\phi}\left(\hat{s}'\right)-\phi\left(s'\right)\right\| \geq \frac{\gamma\mu_{\min}}{4MA}$$

Therefore we know $\alpha_h\left(\hat{s}'\right)=s'$, *i.e.*, α_h always maps the learned state to the correct original state.

We now prove α_h is injective, *i.e.*, $\alpha\left(\hat{s}'\right) \neq \alpha_h\left(\hat{s}''\right)$ for $\hat{s}' \neq \hat{s}'' \in \widehat{\mathcal{S}}_h$. Suppose there are $\hat{s}', \hat{s}'' \in \widehat{\mathcal{S}}_h$ such that $\alpha_h\left(\hat{s}'\right)=\alpha_h\left(\hat{s}''\right)=s'$ for some $s' \in \mathcal{S}_h$. Then using the second conclusion of Lemma 10, we know

$$\left\|\widehat{\phi}\left(\hat{s}'\right)-\widehat{\phi}\left(\hat{s}''\right)\right\|_1 \leq \left\|\widehat{\phi}\left(\hat{s}'\right)-\phi\left(s'\right)\right\|_1 + \left\|\phi\left(s'\right)-\widehat{\phi}\left(\hat{s}''\right)\right\|_1 \leq \frac{\gamma\mu_{\min}}{50MA} \tag{29}$$

However, we know by Algorithm 2, every $\hat{s}' \neq \hat{s}'' \in \widehat{\mathcal{S}}_h$ must satisfy

$$\left\|\widehat{\phi}\left(\hat{s}'\right)-\widehat{\phi}\left(\hat{s}''\right)\right\|_1 > \tau = \frac{\gamma\mu_{\min}}{30MA}$$

This leads to a contradiction and thus α_h is injective.

Next we prove α_h is surjective, *i.e.*, for every $s' \in \mathcal{S}_h$, there exists $\hat{s}' \in \widehat{\mathcal{S}}_h$ such that $\alpha_h\left(\hat{s}'\right)=s'$. The first conclusion in Lemma G. 3 guarantees that for each latent state $s' \in \mathcal{S}_h$, there exists $\mathbf{z}=\hat{\mathrm{g}}\left(x'\right) \in \mathcal{Z}$ with $x' \sim q\left(\cdot \mid s'\right)$. The second conclusion of Lemma 10 guarantees that

$$\left\|\mathbf{z}-\phi\left(s'\right)\right\|_1 \leq \frac{\gamma\mu_{\min}}{100MA}$$

Now we first assert that all points in a cluster are emitted from the same latent state by combining Equation (10), the second part of Lemma G. 3 and our setting of τ. Now the second part of Lemma G. 3 implies that there exists $\hat{s}' \in \widehat{\mathcal{S}}_h$ such that $\left\|\mathrm{z}-\widehat{\phi}\left(\hat{s}'\right)\right\|_1 \leq \frac{\mu_{\min}\gamma}{50MA}$, since z and $\widehat{\phi}\left(\hat{s}'\right)$ correspond to $\hat{\mathrm{g}}$ evaluated on two different contexts in the same cluster. Therefore we have

$$\left\|\boldsymbol{\phi}\left(s'\right)-\widehat{\boldsymbol{\phi}}\left(\hat{s}'\right)\right\|_1 \leq \left\|\boldsymbol{\phi}\left(s'\right)-\mathbf{z}\right\|_1 + \left\|\mathbf{z}-\widehat{\boldsymbol{\phi}}\left(\hat{s}'\right)\right\|_1 \leq \frac{\mu_{\min}\gamma}{30MA}$$

Now we can show that $\alpha_h\left(\hat{s}'\right)=s'$. To do this, we show that $\widehat{\phi}\left(\hat{s}'\right)$ is closer to $\phi\left(s'\right)$ than the embedding of any state in $\mathcal{S}_h$. Using the second conclusion of Lemma G. 3 and Equation 10 we know for any $s'' \neq s'$

$$\left\|\widehat{\phi}\left(\hat{s}'\right)-\phi\left(s''\right)\right\|_1 \geq \left\|\phi\left(s'\right)-\phi\left(s''\right)\right\|_1 - \left\|\widehat{\phi}\left(\hat{s}'\right)-\phi\left(s'\right)\right\|_1 \geq \frac{\gamma\mu_{\min}}{4MA}$$

We know $s'=\operatorname{argmin}_{s_1 \in \mathcal{S}_h}\left\|\widehat{\phi}\left(s_1\right)-\widehat{\phi}\left(\hat{s}'\right)\right\|_1$. Therefore, by the definition of α_h we know $\alpha_h\left(\hat{s}'\right)=s$. Now we have finished the proof of the first part of Condition G.1.

For the second part of Condition G. 1, note for any $s' \in \mathcal{S}_h$ and $x' \sim q\left(\cdot \mid s'\right)$, by Lemma G.2, we know with probability at least $1-\epsilon_{\mathrm{f}}$ over the emission process we have

$$\left\|\hat{\mathrm{g}}_h\left(x'\right)-\phi\left(s'\right)\right\|_1 \leq \frac{\gamma\mu_{\min}}{100MA} \tag{30}$$

For $\hat{s}'=\alpha_h^{-1}\left(s'\right)$, we have

$$\left\|\hat{\mathrm{g}}_h\left(x'\right)-\widehat{\phi}\left(\hat{s}'\right)\right\|_1 \leq \left\|\hat{\mathrm{g}}_h\left(x'\right)-\phi\left(s'\right)\right\|_1 + \left\|\phi\left(s'\right)-\widehat{\phi}\left(\hat{s}'\right)\right\|_1 \leq \frac{\gamma\mu_{\min}}{50MA}$$

On the other hand, for $\hat{s}'' \in \widehat{\mathcal{S}}_h$ with $\hat{s}'' \neq \alpha_h^{-1}\left(s'\right)$, we have

$$\left\|\hat{\mathrm{g}}_h\left(x'\right)-\widehat{\phi}\left(\hat{s}''\right)\right\|_1 \geq -\left\|\hat{\mathrm{g}}_h\left(x'\right)-\phi\left(s'\right)\right\|_1 + \left\|\phi\left(s'\right)-\phi\left(\alpha_h\left(\hat{s}''\right)\right)\right\|_1 - \left\|\phi\left(\alpha_h\left(\hat{s}''\right)\right)-\widehat{\phi}\left(\hat{s}''\right)\right\|_1 \geq \frac{\gamma\mu_{\min}}{4MA}$$

Therefore we have with probability at least $1-\epsilon_{\mathrm{f}}$

$$\hat{f}_h\left(x'\right)=\operatorname*{argmin}_{\hat{s}' \in \widehat{\mathcal{S}}_h}\left\|\widehat{\boldsymbol{\phi}}\left(\hat{s}'\right)-\hat{\mathrm{g}}_h\left(x'\right)\right\|_1=\alpha_h^{-1}\left(s'\right)$$

which is equivalent to the second part of Condition G.1.

Combine the analysis above, we have the following lemma.

Lemma 11. *Assume* $N_\phi = \Theta\left(\frac{MA}{\mu_{\min}}\log\left(\frac{MH}{\delta}\right)\right)$. *For any* $\epsilon_{\mathrm{f}} < \min\left\{\frac{\mu_{\min}^3\gamma}{100M^4A^3}, \frac{\delta}{100HN_\phi}\right\}$, *set* $N_{\mathrm{g}} = \Omega\left(\frac{M^3A^3}{\epsilon_{\mathrm{f}}\mu_{\min}^3\gamma^2}\log\left(\frac{|\mathcal{G}|H}{\delta}\right)\right)$, *then Algorithm 2 learns a decoding function* $\hat{f}$ *such that there is a bijective mapping* $\alpha_h : \widehat{\mathcal{S}}_h \to \mathcal{S}_h$ *for which*

$$\mathbb{P}_{x\sim q(\cdot|\alpha_h(\hat{s}))}\left[\hat{f}_h(x) = \hat{s}\right] \geq 1 - \epsilon_{\mathrm{f}}.$$

Finally, we prove that with an ϵ_{f}-error decoding function, with high probability there will be at most $2\epsilon_{\mathrm{f}}T' + \sqrt{2T'\ln\frac{\delta}{2}}$ states in phase 3 where the agent makes a mistake. Here T' is the total number of steps in phase 3. This allows us to call CR-MVP with $C = 2\epsilon_{\mathrm{f}}T' + \sqrt{2T'\ln\frac{\delta}{2}}$.

Lemma 12. *Assume that the decoding function has* ϵ_f *decoding error, that is,* $\mathbb{P}\left(f(s) \neq s\right) \leq \epsilon_f$. *Define* $\tilde{N}(s,a,s')$ *as the counter without error, and* $\hat{N}(s,a,s')$ *as the counter with error. we have*

$$\mathbb{P}\left(\sum_{s,a,s'}\left|\hat{N}(s,a,s') - \tilde{N}(s,a,s')\right| \leq 2\epsilon_f T' + \sqrt{2T'\ln\frac{2}{\delta}}\right) \leq \frac{\delta}{2} \tag{31}$$

Proof. This is a simple corollary of Chernoff-Hoeffding inequality. □

Combine the three parts of the algorithm together, we can bound the total regret as

$$\begin{aligned}
&H(N'_{\mathrm{g}} + N'_\phi + N'_{\mathrm{p}}) + H(N_{\mathrm{g}} + N_\phi) + \tilde{O}\left(\sqrt{HMAK} + H^2M^2A + HMA\left(\epsilon_{\mathrm{f}}T' + \sqrt{T'\ln\frac{\delta}{2}}\right)\right)\\
=&\tilde{O}\left(\frac{M^4A^4H^2\log|\mathcal{G}|}{\mu_{\min}^4\gamma^2} + \frac{HMA}{\mu_{\min}} + \frac{M^2AH^2}{\mu_{\min}^3} + \frac{HMA}{\mu_{\min}}\log\left(\frac{MH}{\delta}\right) + \right.\\
&\left.\frac{HM^3A^3}{\epsilon_{\mathrm{f}}\mu_{\min}^3\gamma^2}\log\left(\frac{|\mathcal{G}|H}{\delta}\right) + \sqrt{HMAK} + H^2M^2A + HMA\left(\epsilon_{\mathrm{f}}T' + \sqrt{T'\ln\frac{\delta}{2}}\right)\right)
\end{aligned}$$

Finally, by setting $\epsilon_{\mathrm{f}} = \sqrt{\frac{M^2A^2}{\mu_{\min}^3\gamma^2HK}\log\left(\frac{|\mathcal{G}|H}{\delta}\right)}$, the regret is upper bounded by $\tilde{O}\left(\frac{H^{3/2}M^2A^2\sqrt{K}}{\mu_{min}^{3/2}\gamma} + \mathrm{poly}(H, M, A, \mu_{min}^{-1}, \gamma^{-1})\right)$, which finishes the proof.

□

C. Proof of Lower Bounds

C.1. Proof of Theorem 2

Proof. First we prove the lower bound in the multi-armed bandit case, *i.e.* when there is only one state and A actions (called A arms in previous literature). Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be two bandit instances. In $\mathcal{E}_1$, the reward of arm 1 is $Ber(\frac{1}{2})$, and the reward of the other $A-1$ arms is $Ber(0)$. There is no adversarial corruption in $\mathcal{E}_1$. We denote $\mathrm{Regret}\left(\mathcal{A}, \mathcal{E}_1, T\right)$ as the regret $\mathcal{A}$ incurs under environment $\mathcal{E}_1$ after T steps.

Let $T_a(n)$ denote the number of times arm a is chosen in the first n steps. Under environment $\mathcal{E}_1$. The agent incurs regret only when it makes a suboptimal action, which can only happen when it pulls arm $a \neq 1$. So the regret under $\mathcal{E}_1$ can be represented as

$$\mathrm{Regret}(\mathcal{A}, \mathcal{E}_1, T) = \frac{1}{2}\sum_{i=2}^{A}\mathbb{E}[T_i(T)] \tag{32}$$

Now we consider two cases: **Case 1:** $\forall i \geq 2, \mathbb{P}_{\mathcal{A},\mathcal{E}_1}\left(T_i(T) > \frac{C}{2}\right) \geq \frac{1}{2}$. Then by Lemma 13, we have

$$\text{Regret}(\mathcal{A}, \mathcal{E}_1, T) = \frac{1}{2}\sum_{i=2}^{A} \mathbb{E}[T_{s_1,i}(T)] \tag{33}$$

$$\geq \frac{C(A-1)}{16}\mathbb{P}\left(\sum_{i=2}^{A} T_{s_1,i}(T) \geq \frac{C(A-1)}{8}\right) \tag{34}$$

$$\geq \frac{C(A-1)}{48} \tag{35}$$

Case 2: $\exists i \geq 2$, **such that** $\mathbb{P}_{\mathcal{A},\mathcal{E}_1}\left(T_i(T) \leq \frac{C}{2}\right) \geq \frac{1}{2}$**.** In this case we define another environment $\mathcal{E}_2$ identical to $\mathcal{E}_1$ except that the reward of arm i is $Ber(1)$. And for the first $\frac{C}{2}$ times the agent pulls arm i, the adversary will corrupt the reward to be $Ber(0)$. In such environment, the corruption level is C and the agent incurs regret at least $\frac{1}{2}$ each time the agent takes action $a \neq i$.

Now since the rewards $\mathcal{E}_1$ and $\mathcal{E}_2$ generate are the same before $T_i(T)$ exceeds $\frac{C}{2}$, the agent will pull arms identical to that in $\mathcal{E}_1$ when it runs algorithm $\mathcal{A}$, until it pulls arm i for more than $\frac{C}{2}$ times. So $\mathbb{P}_{\mathcal{A},\mathcal{E}_2}\left(T_i(T) \leq \frac{C}{2}\right) \geq \frac{1}{2}$. Then the regret under $\mathcal{E}_2$ is

$$\text{Regret}(\mathcal{A}, \mathcal{E}_2, T) \geq \frac{1}{2}\sum_{j\neq i} \mathbb{E}[T_j(T)] \tag{36}$$

$$\geq \frac{1}{2}\left(T - \frac{C}{2}\right)\mathbb{P}_{\mathcal{A},\mathcal{E}_1}\left(\sum_{j\neq i} T_j(T) \geq T - \frac{C}{2}\right) \tag{37}$$

$$= \frac{1}{2}\left(T - \frac{C}{2}\right)\mathbb{P}_{\mathcal{A},\mathcal{E}_1}\left(T_i(T) \leq \frac{C}{2}\right) \tag{38}$$

$$\geq \frac{1}{4}\left(T - \frac{C}{2}\right) \tag{39}$$

Since $T \geq 2AC$, in case 2 the regret is at least $O(AC)$, which completes the proof.

Now we consider the MDP case. Without loss of generosity, assume that $S = A^H$, where H is a positive integer. We construct the following MDP with H horizon: In each episode the MDP starts from a fixed state s_1. And for any $h \in [H]$, in step h, each state s_h can be represented by a sequence of actions with length $h-1$. And when the environment is in state $s_h = (a_1, ..., a_{h-1})$ and the agent takes action a_h, it will transit to state $s_{h+1} = (a_1, ..., a_{h-1}, a_h)$. In other words, the transition of this MDP can be represented by an A-nary tree, which is shown in Figure 1. The reward for the first $H-1$ episodes are all 0, so the agent receives a reward signal only when it reaches a leaf node in step H.

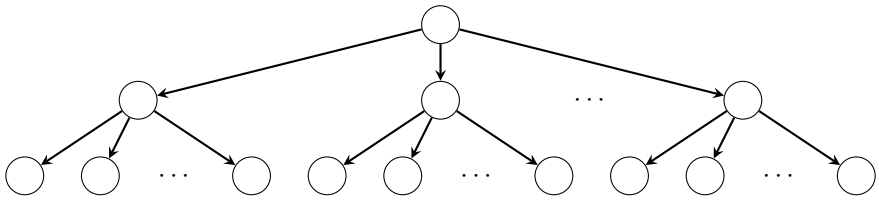

Figure 1. An A-nary tree with three layers.

The MDP is equivalent to a multi-armed bandit problem with SA arms. So we can use the result on multi-armed bandits and obtain a $\Omega(CSA)$ lower bound on the regret.

□

Lemma 13. *Assume X_i are nonnegative random variables and $\mathbb{P}(X_i \geq C) \geq \frac{1}{2}$ for all $i \in [n]$. Then $\mathbb{P}(\sum_{i=1}^{n} X_i \geq \frac{nC}{4}) \geq \frac{1}{3}$.*

Proof. Define $Y_i = \min\{C, X_i\}$, then $\mathbb{P}(\sum_{i=1}^n X_i \geq \frac{nC}{4}) \geq \mathbb{P}(\sum_{i=1}^n Y_i \geq \frac{nC}{4})$ Denote $z = \mathbb{P}(\sum_{i=1}^n X_i \geq \frac{nC}{4})$, then

$$\frac{nC}{2} \leq \mathbb{E}(\sum_{i=1}^n Y_i) \leq \mathbb{P}(\sum_{i=1}^n Y_i \geq \frac{nC}{4})nC + \mathbb{P}(\sum_{i=1}^n Y_i < \frac{nC}{4})\frac{nC}{4} \tag{40}$$

This yields

$$\frac{nC}{2} \leq znC + (1-z)\frac{nC}{4} \Rightarrow z \geq \frac{1}{3} \tag{41}$$

which finishes the proof. □

C.2. Proof of Proposition 1

Proof. As in the proof of Theorem 3, we define $\mathcal{E}_1$ as a A-arm bandit instance without corruptions, and the reward vector of the A arms is $(\frac{1}{2}, 0, ..., 0)$. Consider running algorithm $\mathcal{A}$ on $\mathcal{E}_1$. Suppose $i = \text{argmin}_{j>1} \mathbb{E}_1[T_i]$, where T_i is the number of times arm i is pulled and $\mathbb{E}_1$ denotes taking expectation in $\mathcal{E}_1$. We define environment $\mathcal{E}_2$ as follows: let the corruption level C be $[2\mathbb{E}[T_i]] + 1$, and the reward vector is $(\frac{1}{2}, 0, ..., 0, 1, 0, ...0)$, where the i-th element is 1. Moreover, the adversary create a false 0 reward of the i-th arm whenever the i-th arm is pulled less than C times.

First we consider the regret incurred in $\mathcal{E}_1$. Note that by the definition of T_i, the algorithm will incur at least $O(\mathbb{E}[T_i])$ regret in $\mathcal{E}_1$, since arm i is suboptimal in $\mathcal{E}_1$ and is pulled for at least $O(\mathbb{E}[T_i])$ times. In $\mathcal{E}_1$ the corruption level is 0, so the regret is upper bounded by $O(\sqrt{K})$. This implies that

$$O(\sqrt{K}) \geq \text{Regret}(\mathcal{A}, \mathcal{E}_1, K) \geq \Omega(\mathbb{E}[T_i]) = \Omega(C). \tag{42}$$

On the other hand, consider running the algorithm $\mathcal{A}$ on $\mathcal{E}_2$. Since the agent will receive a reward of 0 for the first C times it pulls arm i, the agent will act just as it were in $\mathcal{E}_1$, unless it pulls arm i for more than C times. Note that the probability that $T_i \leq C$ is at least $\frac{1}{2}$ by the definition of T_i and C, which means that with high probability the algorithm will pull arm i for more than C times. As a result, the algorithm will incur regret at least $\frac{1}{2}(K - C) = \Omega(K)$ with probability $\frac{1}{2}$, since arm i is the optimal arm in $\mathcal{E}_2$. This implies that

$$O(\sqrt{K} + K^\alpha C^\beta) \geq \text{Regret}(\mathcal{A}, \mathcal{E}_2, K) \geq \Omega(K). \tag{43}$$

Hence by combining (42) and (43) we have,

$$K^{\frac{1}{2}} \geq C \geq K^{\frac{1-\alpha}{\beta}},$$

which is equal to $\alpha + \frac{\beta}{2} \geq 1$. □

D. Detailed description of PCID algorithm

Here we present the original PCID algorithm in (Du et al., 2019) for convenience.

The following theorem shows that with a polynomial sample complexity, PCID returns an ϵ -policy cover of all latent states with high probablilty.

Theorem 4 (Theorem 4.1 in (Du et al., 2019)). *Fix any $\epsilon = O\left(\frac{\mu_{\min}^3 \gamma}{M^4 A^3 H}\right)$ and a failure probability $\delta > 0$. Set $N_{\text{g}} = \tilde{\Omega}\left(\frac{M^4 A^4 H \log|\mathcal{G}|}{\epsilon \mu_{\min}^3 \gamma^2}\right), N_\phi = \tilde{\Theta}\left(\frac{MA}{\mu_{\min}}\right), N_{\text{p}} = \tilde{\Omega}\left(\frac{M^2 A H^2}{\mu_{\min} \epsilon^2}\right)$, $\tau = \frac{\gamma}{30MA}$. Then with probability at least $1 - \delta$, Algorithm 4 returns an ϵ-policy cover of $\mathcal{S}$, with size at most MH.*

Algorithm 4 PCID

Input: N_{g} : sample size for learning context embeddings, N_{ϕ} : sample size for learning state embeddings, N_{p} : sample size for estimating transition probabilities, $\tau > 0$: a clustering threshold for learning latent states
Output: policy cover $\Pi = \Pi_1 \cup \cdots \cup \Pi_{H+1}$
Let $\widehat{\mathcal{S}}_1 = \{s_1\}$.
Let $\hat{f}_1(x) = s_1$ for all $x \in \mathcal{X}$
Let $\Pi_1 = \{\pi_0\}$ where π_0 is the trivial 0 -step policy. Initialize $\hat{p}$ to an empty mapping.
for $h = 2, ..., H+1$ **do**
  Let $\eta_h = U(\Pi_{h-1}) \odot U(\mathcal{A})$
  Execute η_h for N_{g} times. $D_{\mathrm{g}} = \left\{\hat{s}^i_{h-1}, a^i_{h-1}, x^i_h\right\}_{i=1}^{N_{\mathrm{g}}}$ for $\hat{s}_{h-1} = \hat{f}_{h-1}(x_{h-1})$
  Learn $\hat{\mathrm{g}}_h$ by calling ERM oracle on input D_{g} :

$$\hat{\mathrm{g}}_h = \mathrm{argmin}_{\mathrm{g} \in \mathcal{G}} \sum_{(\hat{s}, a, x') \in D_{\mathrm{g}}} \left\| \mathrm{g}(x') - \mathbf{e}_{(\hat{s}, a)} \right\|^2$$

  Execute η_h for N_{ϕ} times. $\mathcal{Z} = \left\{\widehat{\mathbf{z}}_i = \hat{\mathbf{g}}_h\left(x^i_h\right)\right\}_{i=1}^{N_{\phi}}$
  Learn $\widehat{\mathcal{S}}_h$ and the state embedding map $\widehat{\phi}_h : \widehat{\mathcal{S}}_h \to \mathcal{Z}$ by clustering $\mathcal{Z}$ with threshold τ (see Algorithm 3).
  Define $\hat{f}_h(x') = \mathrm{argmin}_{\hat{s} \in \widehat{\mathcal{S}}_h} \left\| \widehat{\phi}(\hat{s}) - \hat{\mathrm{g}}_h(x') \right\|_1$
  Execute η_h for N_{p} times. $D_{\mathrm{p}} = \left\{\hat{s}^i_{h-1}, a^i_{h-1}, \hat{s}^i_h\right\}_{i=1}^{N_{\mathrm{p}}}$ for $\hat{s}_{h-1} = \hat{f}_{h-1}(x_{h-1}), \hat{s}_h = \hat{f}_h(x_h)$
  Define $\hat{p}(\hat{s}_h \mid \hat{s}_{h-1}, a_{h-1})$ equal to empirical conditional probabilities in D_{p}.
end for
for $\hat{s}' \in \widehat{\mathcal{S}}_h$ **do**
  Run Algorithm 5 with inputs $\hat{p}$ and $\hat{s}'$ to obtain $(h-1)$ -step policy $\psi_{\hat{s}'} : \widehat{\mathcal{S}}_{[h-1]} \to \mathcal{A}$.
  Set $\pi_{\hat{s}'}(x_\ell) = \psi_{\hat{s}'}\left(\hat{f}_\ell(x_\ell)\right), \ell \in [h-1], x_\ell \in \mathcal{X}_\ell$
end for
Let $\Pi_h = (\pi_{\hat{s}})_{\hat{s} \in \widehat{\mathcal{S}}_h}$.

Algorithm 5 Dynamic Programming for Reaching a State

Input: target state $\hat{s}^* \in \widehat{\mathcal{S}}_h$, transition probabilities $\hat{p}(\hat{s}' \mid \hat{s}, a)$ for all $\hat{s} \in \widehat{\mathcal{S}}_\ell, a \in \mathcal{A}, \hat{s}' \in \widehat{\mathcal{S}}_{\ell+1}, \ell \in [h-1]$
Output: policy $\psi : \widehat{\mathcal{S}}_{[h-1]} \to \mathcal{A}$ maximizing $\hat{\mathbb{P}}^{\psi}(\hat{s}^*)$.
Let $v(\hat{s}^*) = 1$ and let $v(\hat{s}) = 0$ for all other $\hat{s} \in \widehat{\mathcal{S}}_h$.
for $l = h-1, h-2, ..., 1$ **do**
  for $\hat{s} \in \widehat{\mathcal{S}}_\ell$ **do**
    $\psi(\hat{s}) = \max_{a \in \mathcal{A}} \left[\sum_{\hat{s}' \in \widehat{\mathcal{S}}_{\ell+1}} v(\hat{s}') \hat{p}(\hat{s}' \mid \hat{s}, a)\right]$
    $v(\hat{s}) = \sum_{\hat{s}' \in \mathcal{S}_{\ell+1}} v(\hat{s}') \hat{p}(\hat{s}' \mid \hat{s}, a = \psi(\hat{s}))$
  end for
end for