A. Organization of appendix

In Appendix B, we give detailed proofs for Theorem 1, which is the result for the non-cheated setting. In Appendix C, we describe the algorithm omitted in the main paper for the cheated setting as well as its proofs. Then in Appendix D, we give detailed proofs for Theorem 3 and 4, which are the results for the reward-free exploration sub-algorithm. Finally, in Appendix E, we give a justification on why efficient reward-free exploration methods proposed in Kaufmann et al. (2020) and Ménard et al. (2020) are difficult to be used as sub-algorithms here.

B. Regret Analysis for Theorem 1 (the non-cheated case)

B.1. Notations

We use E_m to denote the m-th epoch. Because the epoch will be restarted when there is an unfinished ESTALL as shown in line 14 and 15, each E_m can be decomposed into one or more sub-epochs, denoted as $E_m^1, E_m^2, \ldots, E_m^{\Gamma_m}$, each with length N_m. In the last sub-epoch, either all the ESTALL are finished or the whole algorithm ends.

For convenience, we also define the following notations

- $\mathring{\pi} = \text{argmax}_{\pi \in \Pi_{1/T}} V_*^\pi$, $\mathring{V} = V_*^{\mathring{\pi}}$ and $\mathring{\Delta}_\pi = \mathring{V} - V_*^\pi$,
- $\pi_*^m = \text{argmax}_{\pi \in \Pi_{1/T}} \{\hat{r}_m(\pi) - \frac{1}{16}\hat{\Delta}_\pi^{m-1}\}$
- $\tilde{n}_j^{m,k}$ be the real number of times that policy set Π_j^m interacting with environment inside E_m^k
- $\rho_m = \sum_{s=1}^m \frac{8\lambda_1\lambda_2(HC_s^p + C_s^r)}{16^{m-s}N_s}$
- $\mathring{\Delta}_j^m = \max_{\pi \in \Pi_j^m} \mathring{\Delta}_\pi$.

B.2. High Probability Events

We define the following events and show that these events occur with high probability.

Definition 1. *Define an event $\mathcal{E}_{overall}$ which implies that the actual length of all sub-algorithms is closed to their scheduled time*

$$\mathcal{E}_{overall} := \left\{ \forall m, \forall k \in [\Gamma_m], \forall j \in [S_m] : \tilde{n}_j^{m,k} \in [\frac{1}{2}n_j^m, \frac{3}{2}n_j^m] \right\} \tag{3}$$

Definition 2. *Define an event $\mathcal{E}_{est}$, which implies that, for all the completed sub-epochs, we can estimated all the policy uniformly at the end of epoch*

$$\mathcal{E}_{est} := \left\{ \forall m, \pi : |\hat{r}_m(\pi) - V_*^\pi| \leq 2\lambda_1\lambda_2 \frac{2(HC_{m,k}^p + C_{m,k}^r)}{N_m} + \frac{1}{16}\hat{\Delta}_\pi^{m-1} \right\}$$

Definition 3. *Define an event $\mathcal{E}_{unfinished}$, which implies that, for all sub-epochs with unfinished sub-algorithm, we always have large corruption as long as $\mathcal{E}_{overall}$ holds,*

$$\mathcal{E}_{unfinished} := \left\{ \forall m, \forall k \in [\Gamma_m] : C_{m,k}^p \geq \frac{1}{4}\sqrt{\frac{\ln(10T|\Pi_{1/T}|/\delta_{overall})}{\lambda_1\lambda_2} N_m} \right\} \textit{ and } \mathcal{E}_{overall}$$

Now we are going to prove that Prob$[\mathcal{E}_{overall} \cap \mathcal{E}_{est} \cap \mathcal{E}_{unfinished}] \geq 1 - \delta_{overall}$. We first show that with high probability, $\mathcal{E}_{overall}$ holds,

Lemma 1 (High Probability for $\mathcal{E}_{overall}$)**.** *Prob* $[\mathcal{E}_{overall}] \geq 1 - \delta_{overall}/4$

Proof. For any fixed E_m^k and Π_j^m, we use a Chernoff-Hoeffding bound on the r.v. $\tilde{n}_j^{m,k}$. The expected value is $\mathbb{E}[\tilde{n}_j^m] = n_j^m \geq \lambda_2 = 12\log(8T/\delta_{overall})$, so

$$\text{Prob}\left[|\tilde{n}_j^{m,k} - n_j^m| \geq \frac{1}{2}n_j^m\right] \leq 2\exp\left(-(\frac{1}{4}n_j^m)/3\right) \leq \delta_{overall}/4T\log(T)$$

Because of the possible failure of a sub-algorithm, there will be at most T sub-epochs and $\log(T)$ sub-policy sets. So by taking the union bound over all the sub-epochs and sub-policy sets, we get the target result □

Next, we are going to show with high probability we have $\mathcal{E}_{overall} \cap \mathcal{E}_{est}$. But before we actually prove those, we will first prove the following lemma that gives an estimation on the total amount of corruptions that will be included in each sub-algorithm.

Lemma 2. *For any fixed sub-epoch E_m^k and any fixed Π_j^m, we have*

$$Prob\left[\sum_{t\in E_m^k} c_t^p \mathbf{1}\{\pi_t \in \Pi_j^m\} \geq \frac{2n_j^m}{N_m}C_{m,k}^p + H\ln 4/\delta \text{ and } \sum_{t\in E_m^k} c_t^r \mathbf{1}\{\pi_t \in \Pi_j^m\} \geq \frac{2n_j^m}{N_m}C_{m,k}^r + H\ln 4/\delta\right] \leq \frac{\delta}{4}$$

Proof. It follows a very similar proof of Eqn.3 in (Gupta et al., 2019). Let $Y_j^t = \mathbf{1}\{\pi_t \in \Pi_j^m\}$ and $B_j^m = \sum_{t\in E_m} Y_j^t c_t^{rp}$. Notice that Y_j^t is an independent Bernoulli variable with mean q_j^t. Consider the sequence of r.v.s $X_1, \ldots, X_{N_m}$ defined by $X_{t-T_m^s+1} = (Y_j^t - q_j^t)c_t^{rp}$ for $t \in E_m$. Then it is a martingale difference sequence with predictable quadratic variation $Var = q_j^m \sum_{t\in E_m} c_t^{rp}$. Then by applying the freedman inequality we get that, with probability at least $1-\delta$,

$$B_j^m \leq q_j^m \sum_{t\in E_m^k} c_t^{rp} + (Var/H + H\ln 4/\delta) \leq 2q_j^m \sum_{t\in E_m} c_t^{rp} + H\ln 4/\delta$$

By replacing $q_j^m = n_j^m/N_m$ and $\sum_{t\in E_m^k} c_t^{rp} \leq C_{m,k}^{rp}$ into that, we have $B_j^m \leq \frac{2n_j^m}{N_m}C_{m,k}^{rp} + H\ln 4/\delta$ □

We now continue proving our claim:

Lemma 3 (High Probability for $\mathcal{E}_{est}$)**.** *Prob* $[\mathcal{E}_{est}] \geq 1 - \delta_{overall}/4$

Proof. For any fixed m, j, suppose the EstAll_j^m is completed. From Lemma 2, we know that, with high probability $1-\delta_j^m/4$, there will be at most $\left(\frac{2n_j^m}{N_m}C_{m,k}^{rp} + H\ln(4/\delta_j^m)\right)$ amount of corruptions included in the sub-algorithm EstAll_j^m. Then by Theorem 4 , we have that, with probability as least $1-\delta_j^m$, for all $\pi \in \Pi_j^m$

$$\begin{aligned}
\left|\hat{r}_m(\pi) - V_*^\pi\right| &\leq 7\epsilon_{est}^j + \frac{n_j^m}{F_j^m}\left(\frac{2(HC_{m,k}^p + C_{m,k}^r)}{N_m}\right) + \frac{H\ln(4/\delta_j^m)}{F_j^m} \\
&\leq 7\epsilon_{est}^j + 2\lambda_1\lambda_2\left(\frac{2(HC_{m,k}^p + C_{m,k}^r)}{N_m}\right) + \epsilon_{est}^j \\
&\leq \frac{1}{16}\epsilon_j + 2\lambda_1\lambda_2\left(\frac{2(HC_{m,k}^p + C_{m,k}^r)}{N_m}\right)
\end{aligned}$$

Now by taking the union bound over at most $\log T$ epochs and at most $\log T$ sub-algorithms for each epoch, as well as replacing the value of $\mathring{\Delta}_j^m$, we have that, with probability at least $1-\delta_{overall}/4$, for all m, j and all $\pi \in \Pi_j^m$

$$|\hat{r}_m(\pi) - V_*^\pi| \leq \epsilon_j/16 + 2\lambda_1\lambda_2\frac{2(HC_{m,k}^p + C_{m,k}^r)}{N_m}$$

By the definition of $\hat{\Delta}_\pi^m$ and Π_j^m, this can also be written as, for all m and all $\pi \in \Pi$, with probability at least $1-\delta_{overall}/4$,

$$|\hat{r}_m(\pi) - V_*^\pi| \leq \hat{\Delta}_\pi^m/16 + 2\lambda_1\lambda_2\frac{2(HC_{m,k}^p + C_{m,k}^r)}{N_m}$$

□

Lemma 4 (High Probability for $\mathcal{E}_{unfinished}$). *Prob* $[\mathcal{E}_{unfinished}] \geq 1 - \delta_{overall}/4$

Proof. Given $\mathcal{E}_{overall}$, all the $\textsc{EstAll}_j^{m,k}$ will have more than $\tilde{n}_j^m \geq \lambda_1 F_j^m \geq 6|\mathcal{S}||\mathcal{A}|F_j^m \log(H|\mathcal{S}||\mathcal{A}|)$ number of interactions with the environment. Then by Theorem 3 , we know that since $\textsc{EstAll}_j^{m,k}$ is unfinished, then with probability at least $1-\delta_j^m$, we will have more than $\frac{\epsilon_{est}^j}{2|\mathcal{S}||\mathcal{A}|H^2}F_j^m$ amount of corruptions being included in any fixed $\textsc{EstAll}_j^{m,k}$.

Next by Lemma 2, we know that with probability at least $1-\delta_j^m/4$,

$$\frac{2n_j^m}{N_m}C_{m,k}^p + H\ln(4/\delta_j^m) \geq \frac{\epsilon_{est}^j}{2|\mathcal{S}||\mathcal{A}|H^2}F_j^m$$

By replacing the values of $2n_j^m$, F_j^m and ϵ_{est}^j, we have for any fixed $\textsc{EstAll}_j^{m,k}$,

$$2\lambda_1\lambda_2\left(\frac{2C_{m,k}^p}{N_m}\right) \geq \epsilon_{est}^j\left(\frac{1}{2|\mathcal{S}||\mathcal{A}|H^2} - \frac{\epsilon_j}{96|\mathcal{S}||\mathcal{A}|H^2}\right) \geq \frac{1}{4|\mathcal{S}||\mathcal{A}|H^2}\epsilon_{est}^j$$

Rearranging the inequality we get

$$C_{m,k}^p \geq \frac{1}{16|\mathcal{S}||\mathcal{A}|H^2}\frac{N_m}{\lambda_1\lambda_2}\epsilon_{est}^j \geq \frac{N_m\epsilon_{est}^m}{16|\mathcal{S}||\mathcal{A}|H^2\lambda_1\lambda_2} \geq \frac{1}{4}\sqrt{\frac{\ln(10T|\Pi_{1/T}|/\delta_{overall})}{\lambda_1\lambda_2}N_m}$$

where the third inequality comes from the fact that $\epsilon_{est}^m \geq 4H^2|\mathcal{S}||\mathcal{A}|\sqrt{\frac{\lambda_1\lambda_2\log(10T|\Pi_{1/T}|/\delta_{overall})}{N_m}}$, which is an rearrangement from the inequality in Lemma 5.

Finally, we know there are at most T number of sub-epochs. So by taking the union bound over all the sub-epochs and over all the sub-policy set Π_j^m inside each sub-epoch E_m^k, we get the target result. □

In what follows we assume events $\mathcal{E}_{overall}.\mathcal{E}_{est}$ and $\mathcal{E}_{unfinished}$ hold, since they do so with probability at least $1-\delta_{est}$.

B.3. Auxiliary Lemmas

Lemma 5. *The length of N_m of epoch m satisfies*

$$16 * 128^2\lambda_1\lambda_2|\mathcal{S}|^2H^4|\mathcal{A}|^2\ln(10T|\Pi_{1/T}|/\delta_{overall})/(\epsilon_m)^2 \leq N_m \leq 64 * 128^2\lambda_1\lambda_2|\mathcal{S}|^2H^4|\mathcal{A}|^2 10T\log(2/\delta_{overall})/(\epsilon_m)^2$$

Sometimes we will use the following

$$16\lambda_1\lambda_2|\mathcal{S}|^2H^4|\mathcal{A}|^2\ln(10T|\Pi_{1/T}|/\delta_{overall})/(\epsilon_{est}^m)^2 \leq N_m \leq 64\lambda_1\lambda_2|\mathcal{S}|^2H^4|\mathcal{A}|^2 10T\log(2/\delta_{overall})/(\epsilon_{est}^m)^2$$

Proof. Because $\hat{r}_*^m - \hat{r}_m(\pi_*^m) \leq 0$, so it has $\hat{\Delta}_{\pi_*^m}^m = \epsilon_m$. This immediately implies the lower bound as

$$N_m \geq \min_{j\in S_m} n_j^m \geq 16 * 128^2\lambda_1\lambda_2|\mathcal{S}|^2H^4|\mathcal{A}|^2\ln(10T|\Pi_{1/T}|/\delta_{overall})/(\epsilon_m)^2$$

We get the upper bound from the fact that

$$N_m = \sum_{j\in S_m} n_j^m \leq 64 * 128^2\lambda_1\lambda_2|\mathcal{S}|^2H^4|\mathcal{A}|^2\ln(10T|\Pi_{1/T}|/\delta_{overall})/(\epsilon_m)^2$$

□

B.4. Lemmas related to completed sub-algorithm

In the case that all the sub-algorithms are completed, the proof steps are the very similar to the ones in (Gupta et al., 2019). Here we restate and refined related lemmas.

Lemma 6 (similar to Lemma 5 (Gupta et al., 2019))**.** *Suppose that $\mathcal{E}_{est}$ occurs. Then for all epochs m,*

$$-2\lambda_1\lambda_2\frac{2(HC_m^p+C_m^r)}{N_m}-\frac{2}{16}\hat{\Delta}_{\mathring{\pi}}^{m-1}\le \hat{r}_*^m-\mathring{V}\le 2\lambda_1\lambda_2\frac{2(HC_m^p+C_m^r)}{N_m}.$$

Proof. For the upper bound, by the definition of $\hat{r}_*^m$ and the occurrence of $\mathcal{E}_{est}$, we have

$$\begin{aligned}
\hat{r}_*^m &= \hat{r}_m(\pi_*^m)-\frac{1}{16}\hat{\Delta}_{\pi_*^m}^{m-1}\\
&\le V_*^{\pi_*^m}+2\lambda_1\lambda_2 H\frac{2(HC_m^p+C_m^r)}{N_m}+\frac{1}{16}\hat{\Delta}_{\pi_*^m}^{m-1}-\frac{1}{16}\hat{\Delta}_{\pi_*^m}^{m-1}\\
&\le \mathring{V}+2\lambda_1\lambda_2\frac{2(HC_m^p+C_m^r)}{N_m}+\frac{1}{16}\hat{\Delta}_{\pi_*^m}^{m-1}-\frac{1}{16}\hat{\Delta}_{\pi_*^m}^{m-1}=\mathring{V}+2\lambda_1\lambda_2\frac{2(HC_m^p+C_m^r)}{N_m}.
\end{aligned}$$

For the lower bound, we have

$$\hat{r}_*^m\ge \hat{r}_m(\mathring{\pi})-\frac{1}{16}\hat{\Delta}_{\mathring{\pi}}^{m-1}\ge \mathring{V}-2\lambda_1\lambda_2\frac{2(HC_m^p+C_m^r)}{N_m}-2\frac{1}{16}\hat{\Delta}_{\mathring{\pi}}^{m-1}$$

□

Lemma 7 (similar to Lemma 6 (Gupta et al., 2019))**.** *Suppose that $\mathcal{E}_{est}$ occurs. Then for all epoch m and all policies π*

$$\hat{\Delta}_\pi^m\le 2\left(\mathring{\Delta}_\pi+2^{-m}+\sum_{s=1}^m\frac{8\lambda_1\lambda_2(HC_s^p+C_s^r)}{16^{m-s}N_s}\right)$$

Proof. The proof is by induction on m. For $m=1$, the claim is trivially true because $\hat{\Delta}_\pi^1\le 2*2^{-1}=1$. Next, suppose that the claim holds for $m-1$. Using Lemma 6 and the definition of $\mathcal{E}_{est}$, we write

$$\begin{aligned}
\hat{r}_*^m-\hat{r}_m(\pi)&=(\hat{r}_*^m-\mathring{V})+(\mathring{V}-V_*^\pi)+(V_*^\pi-\hat{r}_m(\pi))\\
&\le 2\lambda_1\lambda_2\frac{2(HC_m^p+C_m^r)}{N_m}+\mathring{\Delta}_\pi+2\lambda_1\lambda_2\frac{2(HC_m^p+C_m^r)}{N_m}+\frac{1}{16}\hat{\Delta}_\pi^{m-1}
\end{aligned}$$

Now using the induction hypothesis, we have

$$\begin{aligned}
\hat{r}_*^m-\hat{r}_m(\pi)&\le \mathring{\Delta}_\pi+2\lambda_1\lambda_2\frac{4(HC_m^p+C_m^r)}{N_m}+\frac{1}{16}\left(2\mathring{\Delta}_\pi+2*2^{-(m-1)}+\sum_{s=1}^{m-1}\frac{8\lambda_1\lambda_2(HC_m^p+C_m^r)}{16^{m-1-s}N_s}\right)\\
&\le 2\mathring{\Delta}_\pi+2*2^{-m}+\sum_{s=1}^{m}\frac{8\lambda_1\lambda_2(HC_s^p+C_s^r)}{16^{m-s}N_s}
\end{aligned}$$

Now by the definition of $\hat{\Delta}_\pi^m$, if $\hat{r}_*^m-\hat{r}_m(\pi)\le 2^{-m}$, then we directly have $\hat{\Delta}_\pi^m<2^{-m}$. Otherwise if $\hat{r}_*^m-\hat{r}_m(\pi)>2^{-m}$, then $\hat{\Delta}_\pi^m<\hat{r}_*^m-\hat{r}_m(\pi)$ □

Lemma 8 (similar to Lemma 7 (Gupta et al., 2019))**.** *Suppose that $\mathcal{E}_{est}$ occurs. Then for all epochs m and all policies π*

$$\hat{\Delta}_\pi^m\ge\frac{1}{4}\mathring{\Delta}_\pi-3\sum_{s=1}^m\frac{8\lambda_1\lambda_2(HC_s^p+C_s^r)}{16^{m-s}N_s}-\frac{3}{8}2^{-m}:=\frac{1}{4}\mathring{\Delta}_\pi-3\rho_m-\frac{3}{8}2^{-m}$$

Proof.

$$\begin{aligned}
\hat{\Delta}_\pi^m &\geq \frac{1}{2}(\hat{r}_*^m - \hat{r}_m(\pi)) \\
&\geq \left(\frac{\mathring{V}}{2} - \lambda_1\lambda_2\frac{2(HC_m^p + C_m^r)}{N_m} - \frac{1}{16}\Delta_{\mathring{\pi}}^{m-1}\right) - \left(\frac{V_*^\pi}{2} + \lambda_1\lambda_2\frac{2(HC_m^p + C_m^r)}{N_m} + \frac{1}{32}\mathring{\Delta}_\pi^{m-1}\right) \\
&= \frac{\mathring{\Delta}_\pi}{2} - \lambda_1\lambda_2\frac{4C_m}{N_m} - \frac{3}{32}\hat{\Delta}_{\mathring{\pi}}^{m-1} \\
&\geq \frac{\mathring{\Delta}_\pi}{2} - \lambda_1\lambda_2\frac{4C_m}{N_m} - \frac{6}{32}\left(\mathring{\Delta}_\pi + 2^{-(m-1)} + \sum_{s=1}^{m-1}\frac{8\lambda_1\lambda_2(HC_s^p + C_s^r)}{16^{m-s}N_s}\right) \\
&\geq \frac{1}{4}\mathring{\Delta}_\pi - 3\underbrace{\sum_{s=1}^{m}\frac{8\lambda_1\lambda_2(HC_s^p + C_s^r)}{16^{m-s}N_s}}_{\rho_m} - \frac{3}{8}2^{-m}
\end{aligned}$$

The first inequality is by the definition of $\hat{\Delta}_\pi^m$. The first term of the second inequality comes from Lemma 6 and the second term of the second inequality comes from the occurrence of $\mathcal{E}_{est}$. And the third inequality comes from Lemma 7. □

Corollary 1. *Suppose that $\mathcal{E}_{est}$ occurs. Then for all epoch m and all policies π.*

$$\epsilon_j \geq \frac{1}{4}\mathring{\Delta}_j^m - 3\rho_{m-1} - \frac{3}{8}2^{-(m-1)}$$

Proof. The above lemma 8 holds for all $\pi \in \Pi_j^m$ including the one leads to Δ_j^m. Furthermore, we have $\epsilon_j = \hat{\Delta}_\pi^{m-1}$. Therefore, we get the target result. □

B.5. Lemmas related to unfinished sub-algorithms

Now we will show that, if the sub-algorithm is unfinished, then the number of repeated sub-epochs can be upper bounded in terms of corruption.

Lemma 9. *If $\mathcal{E}_{unfinished}$ occurs, then we have*

$$\Gamma_m - 1 \leq C_m^p\epsilon_m/(H^2|\mathcal{S}||\mathcal{A}|\ln(10T|\Pi_{1/T}|/\delta_{overall}) \leq C_m^p/(H^2|\mathcal{S}||\mathcal{A}|\ln(10T|\Pi_{1/T}|/\delta_{overall})$$

Proof. Condition on $\mathcal{E}_{unfinished}$, we have

$$\begin{aligned}
N_m &\leq \frac{16\lambda_1\lambda_2}{\ln(10T|\Pi_{1/T}|/\delta_{overall})}\min_{k\in[\Gamma_m-1]}(C_{m,k}^p)^2 \\
&\leq \frac{16\lambda_1\lambda_2}{\ln(10T|\Pi_{1/T}|/\delta_{overall})}(\frac{C_m^p - C_{m,\Gamma_m}^p}{\Gamma_m - 1})^2 \\
&\leq \frac{16\lambda_1\lambda_2}{\ln(10T|\Pi_{1/T}|/\delta_{overall})}(\frac{C_m^p}{\Gamma_m - 1})^2
\end{aligned}$$

Also from Lemma 5, we know a lower bound on N_m. Therefore we have

$$16 * 128^2\lambda_1\lambda_2|\mathcal{S}|^2H^4|\mathcal{A}|^2\ln(10T|\Pi_{1/T}|/\delta_{overall})/(\epsilon_m)^2 \leq \frac{16\lambda_1\lambda_2}{\ln(10T|\Pi_{1/T}|/\delta_{overall})}(\frac{C_m^p}{\Gamma_m - 1})^2$$

Rearranging the above inequality we get

$$\Gamma_m - 1 \leq C_m^p\epsilon_m/(128H^2|\mathcal{S}||\mathcal{A}|\ln(10T|\Pi_{1/T}|/\delta_{overall})$$

□

B.6. Proof for main theorem

Proof. Assume $\mathcal{E}_{overall}$, $\mathcal{E}_{est}$ and $\mathcal{E}_{unfinished}$ occur. Now we decompose the regret into

$$\begin{aligned}
\text{Reg} &= \sum_{m=1}^{M} \sum_{\pi\in\Pi} \sum_{k=1}^{\Gamma_m} \sum_{t\in E_m^k} (\mathring{V} - V_*^{\pi})\mathbf{1}\{\pi_t = \pi\} + T(V^* - \mathring{V}) \\
&\leq \sum_{m=1}^{M} \sum_{j\in S_m} \sum_{k=1}^{\Gamma_m} \mathring{\Delta}_j^m \tilde{n}_j^{m,k} + \mathcal{O}(H) \\
&\leq \underbrace{\frac{3}{2}\sum_{m=1}^{M} \sum_{j\in S_m} \mathring{\Delta}_j^m n_j^{m,\Gamma_m}}_{\text{NON-REPEAT TERM}} + \underbrace{\frac{3}{2}\sum_{m=1}^{M} \sum_{k=1}^{\Gamma_m-1} \sum_{j\in S_m} \mathring{\Delta}_j^m n_j^{m,k}}_{\text{REPEAT TERM}} + \mathcal{O}(H)
\end{aligned}$$

where the last inequality comes from event $\mathcal{E}_{overall}$. For convenience, denote $R_j^{m,k} = \mathring{\Delta}_j^m n_j^{m,k}$, $\beta = 512\sqrt{\lambda_1\lambda_2 \ln(10T|\Pi_{1/T}|/\delta_{overall})}|\mathcal{S}||\mathcal{A}|H^2$ and we know by definition that $\epsilon_j \leq \beta\sqrt{1/n_j^m}$.

We first give upper bounds on term $R_j^{m,k}$ for any fixed m, k. Notice that when the algorithm goes to epoch m, it suggests that all the sub-algorithms ran before m are completed. Therefore, we will use lemmas stated in Section B.4 for the following proof.

Case 1: $\rho_{m-1} < \mathring{\Delta}_j^m/64$. In this case, if $\mathring{\Delta}_j^m/2 \geq 2^{-(m-1)}$, given $\mathcal{E}_{est}$, we can use Corollary 1 to get

$$\epsilon_j \geq \frac{1}{4}\mathring{\Delta}_j^m - 3\rho_{m-1} - \frac{3}{8}2^{-(m-1)} \geq \left(\frac{1}{4} - \frac{3}{64} - \frac{3}{16}\right)\mathring{\Delta}_j^m = \frac{\mathring{\Delta}_j^m}{64}$$

If $\mathring{\Delta}_j^m/2 < 2^{-(m-1)}$, then $\epsilon_j \geq \frac{\mathring{\Delta}_j^m}{64}$ trivially holds.

In turn, we have $n_j^m \leq \beta/\epsilon_j^2$ according to the definition of n_j^m, from which follows

$$R_i^{m,k} \leq 64\beta\sqrt{n_j^m}$$

This can be also be written as

$$R_i^{m,k} \leq \mathring{\Delta}_j^m \beta/\epsilon_j^2 \leq 64^2\mathring{\Delta}_j^m\beta/(\mathring{\Delta}_j^m)^2 = 64^2\beta/\mathring{\Delta}_j^m \leq 64^2\beta/\min_{\pi\in\Pi_{1/T}}\mathring{\Delta}_\pi$$

Case 2: $\rho_{m-1} \geq \mathring{\Delta}_j^m/64$. We again use the upper bound of $n_j^m \leq \beta^2/\epsilon_m^2$

$$R_i^{m,k} \leq 96\beta^2\rho_{m-1}/\epsilon_m^2 = 96\beta^2\rho_{m-1}2^{2m}$$

By combining these two cases, we have

$$R_j^{m,k} \leq 64\beta\min\left\{\sqrt{n_j^m}, \frac{64}{\min_{\pi\in\Pi_{1/T}}\mathring{\Delta}_\pi}\right\} + 96\beta^2\rho_{m-1}/\epsilon_m^2$$

Secondly, we deal with the NON-REPEAT TERM. By summing $R_j^{m,k}$ over all policy sets for $k = \Gamma_m$, we get

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{j\in S_m}\mathring{\Delta}_j^m n_j^{m,\Gamma_m} \\
&\le 64\beta\sum_{m=1}^{M}\min\left\{\sqrt{\log T N_m}, \frac{64\log T}{\min_{\pi\in\Pi_{1/T}}\mathring{\Delta}_\pi}\right\} + 96\beta^2(\log T)\sum_{m=1}^{M}\rho_{m-1}2^{2m} \\
&\le 64\beta(\log T)\min\left\{\sqrt{T}, \frac{64\log T}{\min_{\pi\in\Pi_{1/T}}\mathring{\Delta}_\pi}\right\} + 96\beta^2(\log T)\sum_{m=1}^{M}\rho_{m-1}2^{2m} \\
&\le \tilde{\mathcal{O}}\left(|\mathcal{S}|^2|\mathcal{A}|^{3/2}H^2\min\{H^{1/2},|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/2}\}\ln(1/\delta_{overall})\min\left\{\sqrt{T},\frac{1}{\min_{\pi\in\Pi_{1/T}}\mathring{\Delta}_\pi}\right\}\right) \\
&\quad + \tilde{\mathcal{O}}\left(|\mathcal{S}||\mathcal{A}|\ln(1/\delta_{overall})(HC^p + C^r)\right) \\
&= \tilde{\mathcal{O}}\left(|\mathcal{S}|^2|\mathcal{A}|^{3/2}H^2\min\{H^{1/2},|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/2}\}\ln(1/\delta_{overall})\min\left\{\sqrt{T},\frac{1}{\min_{\pi\in\Pi}\Delta_\pi}\right\}\right) \\
&\quad + \tilde{\mathcal{O}}\left(|\mathcal{S}||\mathcal{A}|\ln(1/\delta_{overall})(HC^p + C^r)\right)
\end{aligned}$$

The last equation comes from the fact that $\Pi_{1/T}$ is $1/T$-net of policy and $\sqrt{T} > \frac{1}{\min_{\pi\in\Pi_{1/T}}\mathring{\Delta}_\pi}$ when $\min_{\pi\in\Pi_{1/T}}\mathring{\Delta}_\pi < o(\sqrt{1/T})$.

Here the result of $\sum_{m=1}^{M}\rho_{m-1}2^{2m}$ comes from the following,

$$\begin{aligned}
\sum_{m=1}^{M}\beta^2\rho_{m-1}/\epsilon_m^2 &= \sum_{m=1}^{M}\beta^2\sum_{s=1}^{m-1}4^m\frac{8\lambda_1\lambda_2(HC_s^p + C_s^r)}{16^{m-1-s}N_s} \\
&= 8\lambda_1\lambda_2\beta^2\sum_{s=1}^{M}(HC_s^p + C_s^r)\sum_{m=s}^{M}4^m\frac{1}{16^{m-1-s}N_s} \\
&\le 8\lambda_1\lambda_2\beta^2\sum_{s=1}^{M}(HC_s^p + C_s^r)\sum_{m=s}^{M}4^m\frac{4^{-s}}{16^{m-1-s}\beta^2} \\
&= 32\lambda_1\lambda_2\sum_{s=1}^{M}(HC_s^p + C_s^r)\sum_{m=s}^{M}\frac{4^{m-1-s}}{16^{m-1-s}} \\
&= \tilde{\mathcal{O}}\left(|\mathcal{S}||\mathcal{A}|\ln(1/\delta_{overall})(HC^p + C^r)\right)
\end{aligned}$$

where the first equality use changing order of summation techniques and the second inequality comes from the lower bound of N_s in Lemma 5.

Thirdly, we consider the REPEAT TERM. From the previous analysis, we have

$$\sum_{m=1}^{M}\sum_{k=1}^{\Gamma_m-1}\sum_{j\in S_m}\mathring{\Delta}_j^m n_j^{m,k} \le 64\beta\sum_{m=1}^{M}\sum_{k=1}^{\Gamma_m-1}\sqrt{(\log T)N_m} + \sum_{m=1}^{M}(\Gamma_{m'} - 1)96\beta^2(\log T)\rho_{m-1}2^{2m}$$

First, given $\mathcal{E}_{unfinished}$, we can bound the first term by

$$64\beta\sum_{m=1}^{M}\sum_{k=1}^{\Gamma_m-1}\sqrt{\log T}C_{m,k}^p\frac{16\sqrt{\lambda_1\lambda_2}}{\sqrt{\ln(10T|\Pi_{1/T}|/\delta_{overall})}} \le \tilde{\mathcal{O}}\left(H^2|\mathcal{S}|^2|\mathcal{A}|^2\ln(1/\delta_{overall})C^p\right)$$

Then, by Lemma 9, we can bound the first term by bounding the $\Gamma_m - 1$ as below

$$\begin{aligned}
&\beta^2(\log T)\sum_{m=1}^{M}(\Gamma_m - 1)\rho_{m-1}2^{2m} \\
&\leq \beta^2(\log T)\sum_{m=1}^{M}\frac{C^p_m}{H^2|\mathcal{S}||\mathcal{A}|\ln(10T|\Pi_{1/T}|/\delta_{overall})}\rho_{m-1}2^{2m} \\
&\leq \frac{\log T}{H^2|\mathcal{S}||\mathcal{A}|\ln(10T|\Pi_{1/T}|/\delta_{overall})}\left(\sum_{m'=1}^{M}C^p_m\right)\left(\sum_{m=1}^{M}\beta^2\sum_{m'\in M}\rho_{m-1}2^{2m}\right) \\
&\leq \frac{C^p(\log T)^2}{H^2|\mathcal{S}||\mathcal{A}|\ln(10T|\Pi_{1/T}|/\delta_{overall})}\left(\beta^2\sum_{m=1}^{M}\rho_{m-1}2^{2m}\right) \\
&\leq \tilde{O}\left(\frac{1}{H^2}C^p(HC^p + C^r)\right)
\end{aligned}$$

Combing all the upper bounds, we get the final result. □

B.7. Relationship between PolicyGapComlexity and the GapCompelxity in Simchowitz & Jamieson (2019)

In the main paper, we assume a single starting states. Here, in order to make a comparison, we remove this assumption and assume a starting distribution over all states. As stated in the **Related Work** section, the most common GapComplexity used in reinforcement learning is in the following form. Note that to aid the exposition, we omit other states and actions dependency below.

$$\begin{aligned}
\text{gap}_h(s,a) &= V^*_h(s) - Q^*_h(s,a), \\
\text{GapComplexity} &= \frac{1}{\min_{s,a,h}\text{gap}_h(s,a)}
\end{aligned}$$

To get an intuition about its relation to policy gap Δ_π, consider the optimal policy π^* and the second optimal policy π'. If there is a tie, we just arbitrarily choose two policies with closest behavior. Define

$$\mathcal{H}_{identical} = \{h | \forall h' \in [0, h-1], \forall s \in \mathcal{S}_{h'}, \pi^*(s) = \pi'(s)\}$$

where $\mathcal{S}_h = \{s \in \mathcal{S} | \max_{\pi\in\Pi}\text{Prob}\left(\pi \text{ visits } s \text{ at } h\right) > 0\}$ and $\mathcal{S}_0 = \emptyset$. So $\mathcal{H}_{identical}$ is a collection of steps, before which, the optimal policy π^* and the second optimal policy π' are unidentifiable. Note that $h = 1$ is always included in $\mathcal{H}_{identical}$. Now we have

$$\begin{aligned}
\Delta_{\pi'} &= V^* - V^{\pi'}_* \\
&= \max_{h\in\mathcal{H}_{identical}}\sum_{s\in\mathcal{S}}\text{Prob}\left(\pi^* \text{ visits } s \text{ at } h\right)\left(V^*_h(s) - Q^{\pi'}_{*,h}(s,\pi'(s))\right) \\
&\geq \max_{h\in\mathcal{H}_{identical}}\sum_{s\in\mathcal{S}}\text{Prob}\left(\pi^* \text{ visits } s \text{ at } h\right)\left(V^*_h(s) - Q^*_h(s,\pi'(s))\right) \\
&\geq \min_{s,a,h}\text{gap}_h(s,a)\max_{h\in\mathcal{H}_{identical}}\sum_{s\in\mathcal{S}}\text{Prob}\left(\pi^* \text{ visits } s \text{ at } h\right)\mathbf{1}\{\pi^*(s) \neq \pi'(s)\}
\end{aligned}$$

It is easy to see that $\max_{h\in\mathcal{H}_{identical}}\sum_{s\in\mathcal{S}}\text{Prob}\left(\pi^* \text{ visits } s \text{ at } h\right)\mathbf{1}\{\pi^*(s) \neq \pi'(s)\}$ is positive due to the definition of $\mathcal{H}_{identical}$.

Recall the the PolicyGapComplexity is defined as $\frac{1}{\Delta_{\pi'}}$, so we have

$$\begin{aligned}
\text{PolicyGapComplexity} &\leq \frac{1}{\max_{h\in\mathcal{H}_{identical}}\sum_{s\in\mathcal{S}}\text{Prob}\left(\pi^* \text{ visits } s \text{ at } h\right)\mathbf{1}\{\pi^*(s) \neq \pi'(s)\}}\frac{1}{\min_{s,a,h}\text{gap}_h(s,a)} \\
&\leq \frac{\text{GapComplexity}}{\max_{h\in\mathcal{H}_{identical}}\sum_{s\in\mathcal{S}}\text{Prob}\left(\pi^* \text{ visits } s \text{ at } h\right)\mathbf{1}\{\pi^*(s) \neq \pi'(s)\}}
\end{aligned}$$

Therefore, with respect to the gap term, the PolicyGapComplexity and the GapComplexity are close when $\max_{h\in\mathcal{H}_{identical}} \sum_{s\in\mathcal{S}} \text{Prob}\left(\pi^* \text{ visits } s \text{ at } h\right) \mathbf{1}\{\pi^*(s) \neq \pi'(s)\}$ is large.

Because step $h = 1$ is always included in $\mathcal{H}_{identical}$, so one nontrivial case satisfying the above condition is that the starting states are uniformly chosen from some subset of states. It is easy to see that the single starting states is also one of the special cases. Besides, there are also many other cases satisfying the above condition, for example, a MDP that starts from various states and always concentrates on some states with equal chances in later steps included in $\mathcal{H}_{identical}$.

Finally, whether the PolicyGapComplexity-dependent bound can also get some refined dependency on $|\mathcal{S}|, |\mathcal{A}|, H$ like the GapComplexity-dependent bound in Xu et al. (2021) in some special cases remains further investigation.

C. Meta-algorithm and Results for cheated Adversary

Algorithm 5 BRUTE-FORCE-POLICY-ELIMINATION-RL

```
1: Input: time horizon $T$, confidence $\delta_{overall}$
2: Construct a $1/T$-net for non-stationary policies, denoted as $\Pi_{1/T}$.
3: Initialize $S_1 = 0, \Pi^1 = \Pi$. And for $j \in \log T$, initialize $\epsilon_j = 2^{-j}.\epsilon_{sim}^j = \epsilon_j/128$
4: Set $\lambda_1 = 6|\mathcal{S}||\mathcal{A}|log(H^2|\mathcal{S}||\mathcal{A}|/\epsilon_{sim})$ and $\lambda_2 = 12\ln(8T/\delta_{overall})$
5: for epoch $m = 1, 2, \ldots$ do
6:     Set $\delta^m = \delta_{overall}/(5T)$
7:     Set $F^m = \frac{8|\mathcal{S}|^2 H^4 |\mathcal{A}|^2 \ln(2|\Pi^m|/\delta^m)}{(\epsilon_{sim}^m)^2}$
8:     Set $N_m = 2\lambda_1\lambda_2 F^m$ and $T_m^s = T_{m-1}^s + N_{m-1}$
9:     Initialize a sub-algorithm ESTALL$^m$ = EstAll$(\epsilon_{sim}^m, \Pi^m, \delta^m, F^m)$
10:    for $t = T_m^s, T_m^s + 1, \ldots, T_m^s + N_m - 1$ do
11:        Play the policy according the awaiting ESTALL$^m$.CONTINUE. Then continue running ESTALL$^m$ until the next
           ROLLOUT is met. (If no more ROLLOUT needed, then just uniformly play one )
12:    end for
13:    if ESTALL$^m$ is unfinished then
14:        Set $T_m^s = T_m^s + N_m$ and repeat the whole process from line 9.    ▷ So each repeat is a sub-epoch.
15:    else
16:        Obtain $\hat{r}_m(\pi)$ for all $\pi$.
17:    end if
18:    Update the active policy set
           $$\Pi^{m+1} \leftarrow \{\pi | \max_{\pi' \in \Pi^m} \hat{r}_m(\pi') - \hat{r}_m(\pi) \leq 8\lambda_1\lambda_2 H^2 \sqrt{|\mathcal{S}||\mathcal{A}| \ln(10T|\Pi_{1/T}|/\delta_{overall})T}/N_m + \frac{1}{8}\epsilon_m\}$$
19: end for
```

Theorem 5. *The regret is upper bounded by*

$$\begin{aligned}\text{Reg} \leq\ & \tilde{\mathcal{O}}\left(|\mathcal{S}|^2|\mathcal{A}|^{3/2}H^2 \min\{\sqrt{H}, \sqrt{|\mathcal{S}||\mathcal{A}|}\} \ln(1/\delta_{overall})\sqrt{T}\right) \\ &+ \tilde{\mathcal{O}}\left(\frac{(C^r)^2}{H^3|\mathcal{S}||\mathcal{A}|} + H|\mathcal{S}||\mathcal{A}|(C^p)^2\right)\end{aligned}$$

Remark In Section 2.2 in (Bogunovic et al., 2020), they proved that in order to get $\tilde{\mathcal{O}}(\sqrt{HT})$, the corruption terms can go as low as $\tilde{\Omega}(\frac{C^2}{\log C})$ for the linear bandits. Therefore, we conjecture that $\tilde{\mathcal{O}}((C^r + C^p)^2)$ term is also unavoidable in our setting.

C.1. Regret Analysis for Theorem 5

For convenience, we rearrange this upper bound a little bit. So now our target is to show the follows.

$$\begin{aligned}\text{Reg} &\leq \tilde{\mathcal{O}}\left(|\mathcal{S}|^2|\mathcal{A}|^{3/2}H^2 \min\{\sqrt{H}, \sqrt{|\mathcal{S}||\mathcal{A}|}\} \ln(1/\delta_{overall})\sqrt{T}\right) \\ &+ \tilde{\mathcal{O}}\left(\frac{(HC^p + C^r)^2}{H^3|\mathcal{S}||\mathcal{A}|\ln(|\Pi_{1/T}|)} + H\frac{\ln(1/\delta_{overall})}{\ln(|\Pi_{1/T}|/\delta_{overall})}|\mathcal{S}||\mathcal{A}|(C^p)^2\right)\end{aligned}$$

We only need to consider the case that $C^r + HC^p \leq H^2\sqrt{|\mathcal{S}||\mathcal{A}|\ln(|\Pi_{1/T}|)T}$, otherwise we will get a trivial linear regret.

It easy to see that the following events sill holds with at least $1 - \delta_{overall}$ probability,

$$\begin{aligned}&\mathcal{E}_{overall} := \left\{\forall m, \forall k \in [\Gamma_m] : \tilde{n}^{m,k} \in [\frac{1}{2}n^m, \frac{3}{2}n^m]\right\} \\ &\mathcal{E}_{est} := \left\{\forall m, \pi \in \Pi^m : |\hat{r}^m(\pi) - V_*^\pi| \leq 2\lambda_1\lambda_2\frac{2(HC^p_{m,\Gamma_m} + C^r_{m,\Gamma_m})}{N_m} + \frac{1}{16}\epsilon_m\right\} \\ &\mathcal{E}_{unfinished} := \left\{\forall m, \forall k \in [\Gamma_m] : C^p_{m,k} \geq \frac{1}{4}\sqrt{\frac{\ln(10T|\Pi|/\delta_{overall})}{\lambda_1\lambda_2}N_m}\right\} \text{ and } \mathcal{E}_{overall}\end{aligned}$$

Notice here we will permanently eliminate a policy instead of maintaining different subset of policies, therefore, in $\mathcal{E}_{est}$, all the active policies have same levels of estimation. Next we show that given the above events, we will never eliminate the best policy from the active policy set Π^{m+1}.

Again we use the following notations $\mathring{\pi} = \text{argmax}_{\pi \in \Pi_{1/T}} V_*^\pi$, $\mathring{V} = V_*^{\mathring{\pi}}$ and $\mathring{\Delta}_\pi = \mathring{V} - V_*^\pi$.

Lemma 10. *For any epoch m, we always have $\mathring{\pi} \in \Pi^m$.*

Proof. Given $\mathcal{E}_{est}$, let $\hat{\pi}_m = \text{argmax}_{\pi' \in \Pi^m}\hat{r}_m(\pi')$, we know that

$$\begin{aligned}\hat{r}_m(\hat{\pi}_m) - \hat{r}_m(\mathring{\pi}) &\leq V_*^{\hat{\pi}_m} - \mathring{V} + 4\lambda_1\lambda_2\frac{2(HC^p_{m,\Gamma_m} + C^r_{m,\Gamma_m})}{N_m} + \frac{1}{8}\epsilon_m \\ &\leq 4\lambda_1\lambda_2\frac{2(HC^p_{m,\Gamma_m} + C^r_{m,\Gamma_m})}{N_m} + \frac{1}{8}\epsilon_m \\ &\leq 8\lambda_1\lambda_2H^2\sqrt{|\mathcal{S}||\mathcal{A}|\ln(|\Pi_{1/T}|)T}/N_m + \frac{1}{8}\epsilon_m *\end{aligned}$$

where the last inequality comes from the assumption that $C^r + HC^p \leq H^2\sqrt{|\mathcal{S}||\mathcal{A}|\ln(|\Pi_{1/T}|)T}$. Now by the elimination condition in Line 18 , we can get our target result. □

Then we can upper bounded $\max_{\pi \in \Pi^m} \Delta_\pi$ as follows

Lemma 11. *For any active policy set Π^m, we have*

$$\max_{\pi \in \Pi^m} \Delta_\pi \leq \tilde{\mathcal{O}}\left(|\mathcal{S}|^2|\mathcal{A}|^{3/2}H^{3/2}(\frac{1}{\sqrt{N_m}} + \frac{\sqrt{HT}}{N_m})\right)$$

Proof. Let $\pi' = \text{argmax}_{\pi \in \Pi^{m+1}} \Delta_\pi$

$$
\begin{aligned}
\mathring{\Delta}_{\pi'} &\leq \mathring{V} - V_*^{\pi'} \\
&\leq \hat{r}_m(\mathring{\pi}) - \hat{r}_m(\pi') + 4\lambda_1\lambda_2 \frac{2(HC^p_{m,\Gamma_m} + C^r_{m,\Gamma_m})}{N_m} + \frac{1}{8}\epsilon_m \\
&\leq 8\lambda_1\lambda_2 H^2 \sqrt{|\mathcal{S}||\mathcal{A}| \ln(|\Pi_{1/T}|)T}/N_m + \frac{1}{4}\epsilon_{m+1} \\
&= \tilde{\mathcal{O}}\left(|\mathcal{S}||\mathcal{A}| \ln(1/\delta_{overall}) H^2 \sqrt{|\mathcal{S}||\mathcal{A}| \ln(|\Pi_{1/T}|)} \frac{\sqrt{T}}{N_{m+1}} + |\mathcal{S}|^{3/2}|\mathcal{A}|^{3/2} H^2 \sqrt{\frac{\ln(1/\delta_{overall}) \ln(10T|\Pi_{1/T}|/\delta_{overall})}{N_{m+1}}} \right) \\
&\leq \tilde{\mathcal{O}}\left(|\mathcal{S}|^{3/2}|\mathcal{A}|^{3/2} H^2 \ln(1/\delta_{overall}) \sqrt{\ln(|\Pi_{1/T}|)} (\sqrt{T} + \sqrt{\frac{1}{N_{m+1}}}) \right) \\
&\leq \tilde{\mathcal{O}}\left(|\mathcal{S}|^2|\mathcal{A}|^{3/2} H^2 \min\{\sqrt{H}, \sqrt{|\mathcal{S}||\mathcal{A}|}\} \ln(1/\delta_{overall}) \left(\sqrt{T} + \sqrt{\frac{1}{N_{m+1}}} \right) \right)
\end{aligned}
$$

Here the second inequality comes from Lemma 10. The third inequality comes from the elimination condition in Line 18 and the assumption that the assumption that $C^r + HC^p \leq H^2\sqrt{|\mathcal{S}||\mathcal{A}| \ln(|\Pi_{1/T}|)T}$. Replace the value of ϵ_m in the term of N_m we get the target result. □

Now given $\mathcal{E}_{overall}$, we again have regret that

$$
\text{Reg} \leq \underbrace{\frac{3}{2}\sum_{m=1}^{M} (\max_{\pi \in \Pi^m} \Delta_\pi) N_m}_{\text{NON-REPEAT TERM}} + \underbrace{\sum_{m=1}^{M}\sum_{k=1}^{\Gamma_m - 1} N_m}_{\text{REPEAT TERM}}
$$

First, we deal with the NON-REPEAT TERM. By applying Lemma 11, we have

$$
\begin{aligned}
\sum_{m=1}^{M} (\max_{\pi \in \Pi^m} \Delta_\pi) N &\leq \sum_{m=1}^{M} \tilde{\mathcal{O}}\left(|\mathcal{S}|^2|\mathcal{A}|^{3/2} H^2 \min\{\sqrt{H}, \sqrt{|\mathcal{S}||\mathcal{A}|}\} \ln(1/\delta_{overall}) \left(\sqrt{T} + \sqrt{\frac{1}{N_{m+1}}} \right) \right) \\
&\leq \tilde{\mathcal{O}}\left(|\mathcal{S}|^2|\mathcal{A}|^{3/2} H^2 \min\{\sqrt{H}, \sqrt{|\mathcal{S}||\mathcal{A}|}\} \ln(1/\delta_{overall}) \sqrt{T} \right)
\end{aligned}
$$

Next, we deal with the REPEAT TERM. By $\mathcal{E}_{unfinished}$, we have

$$
\begin{aligned}
\sum_{m=1}^{M} (\max_{\pi \in \Pi^m} \Delta_\pi) N_m \leq H \sum_{m=1}^{M}\sum_{k=1}^{\Gamma_m - 1} N^m &\leq H|\mathcal{A}||\mathcal{S}| \frac{\ln(1/\delta_{overall})}{\ln(10T|\Pi_{1/T}|/\delta_{overall})} \sum_{m=1}^{M}\sum_{k=1}^{\Gamma_m - 1} (C^p_{m,k})^2 \\
&\leq H|\mathcal{A}||\mathcal{S}|(C^p)^2
\end{aligned}
$$

D. Analysis for EstAll Sub-algorithm

D.1. Preliminaries

We define the set of episodes that the learner interacts with environment as $\mathcal{I}_{est}$ and the total corruption included these episodes as $C^{r(p)}_{est} = \sum_{t \in \mathcal{I}_{est}} c^{r(p)}_t$.

D.2. Key results

Theorem 6 (Sample complexity restated here)**.** *Suppose* $F \geq \frac{8|\mathcal{S}|^2 H^4 |\mathcal{A}|^2 \ln(2|\Pi|/\delta_{est})}{\epsilon^2_{est}}$ *and* $\tau \geq 6$*. Under the corruption assumption* $C^p_{est} \leq \frac{\epsilon_{est} F}{2|\mathcal{S}||\mathcal{A}|H^2}$*, with probability at least* $1 - \delta_{est}$*, the algorithm interacts with environment at most*

$$
|\mathcal{S}||\mathcal{A}|F\tau log(H^2|\mathcal{S}||\mathcal{A}|/\epsilon_{est})
$$

times. Note, if the algorithm interacts with environment more than the above number of times, then with probability at least $1 - \delta_{est}$*,* $C^p_{est} > \frac{\epsilon_{est} F}{2|\mathcal{S}||\mathcal{A}|H^2}$

Proof. By Lemma 14, we know that with probability at least $1 - \delta_{est}$, for any fixed state-action pair (s, a), Line 7 in Algorithm 2 will fail at most $\log_2(H^2|\mathcal{S}||\mathcal{A}|/\epsilon_{est})$ times by doubling from $\frac{\epsilon_{est}}{H|\mathcal{S}||\mathcal{A}|}$ to H. So the maximum number of policies that will be added into policy set $\Pi_{\mathcal{D}}$ is at most $\log_2(H^2|\mathcal{S}||\mathcal{A}|/\epsilon_{est})|\mathcal{S}||\mathcal{A}|$. Now because for each policy added into $\Pi_{\mathcal{D}}$, we will greedily sample $F\tau$ times according to Algorithm 4, so the total interaction time is at most $\log_2(H^2|\mathcal{S}||\mathcal{A}|/\epsilon_{est})|\mathcal{S}||\mathcal{A}|F\tau$ times. □

Theorem 7 (Estimation correctness restated here)**.** *Suppose* $F \geq \frac{8|\mathcal{S}|^2 H^4 |\mathcal{A}|^2 \ln(2|\Pi|/\delta_{est})}{\epsilon_{est}^2}$ *and* $\tau \geq 6$*. Then for all* $\pi \in \Pi$*, with probability at least* $1 - \delta_{est}$*,*

$$\left|\hat{r}(\pi) - V^\pi(s_1)\right| \leq (1+\tau)\epsilon_{est} + (HC^p_{est} + C^r_{est})/F$$

Proof. By definition, $\hat{r}(\pi) = \frac{1}{F}\sum_{i=1}^F r(z_i^\pi)$ and $\{r(z_i^\pi)\}_{i=1}^F$ is a sequence of independent random variables. We denote its expected value $\mathbb{E}[r(z_i^\pi)]$ as $\{V_i^\pi\}_{i=1}^F$. Here V_i is not a real existing value function but an "average value function" whose rewards and transition functions are the average of rewards and transition functions generated by the MDPs under different times (so some are corrupted). Now we can use Hoeffding's inequality to bound $\left|\hat{r}(\pi) - \frac{1}{F}\sum_{i=1}^F V_i^\pi\right|$.

For those $\pi \in \Pi_{\mathcal{D}}$,

$$\text{Prob}\left[\left|\hat{r}(\pi) - \frac{1}{F}\sum_{i=1}^F V_i^\pi\right| \leq \epsilon_{est}\right] \geq 1 - 2\exp(-2F\epsilon_{est}^2/H^2) \geq 1 - \delta_{est}/2|\Pi|$$

For those $\pi \notin \Pi_{\mathcal{D}}$, if none of then are failed, we again have

$$\text{Prob}\left[\left|\hat{r}(\pi) - \frac{1}{F}\sum_{i=1}^F V_i^\pi\right| \leq \epsilon_{est}\right] \geq 1 - \delta_{est}/2|\Pi|$$

Then because at each (s, a), the policy *fails* at most $\epsilon_{est}\tau F/H|\mathcal{S}||\mathcal{A}|$, there will be at most $\tau\epsilon_{est}F/H$ trajectories with *Fails*. Each failed trajectory will cause at most H rewards, therefore,

$$\text{Prob}\left[\left|\hat{r}(\pi) - \frac{1}{F}\sum_{i=1}^F V_i^\pi\right| \leq (1+\tau)\epsilon_{est}\right] \geq 1 - \delta_{est}/2|\Pi|$$

Now we can decompose our target result into,

$$\left|\hat{r}(\pi) - V^\pi\right| \leq \left|\hat{r}(\pi) - \frac{1}{F}\sum_{i=1}^F V_i^\pi\right| + \left|\frac{1}{F}\sum_{i=1}^F V_i^\pi - V^\pi\right|$$

The first term can be upper bounded by the previous results. The second term can be upper bounded by lemma 16. Finally, by taking a union bound over all policies in Π, we get our target result. □

D.3. Detailed Analysis

D.4. Notations

For convenience, we write F instead of F_{est} in this section.

D.4.1. Main Lemmas

Claim 1 For any fixed π, each of the trajectories in $\{z_i^\pi\}_{i\in[F]}$ is independent to each other due to the property of MDP.

Definition 4. *Define* $f^\pi(s, a)$ *as the random variable which is the total number of times a trajectory induced by* π *visits* (s, a) *with respect to the underlying MDP* $\mathcal{M}$ *and then define its expectation as*

$$\mathbb{E}[f^\pi(s, a)] = \mu^\pi(s, a)$$

For any policy set Π, we define the following $\mu^{\Pi}_{\max}$

$$\mu^{\Pi}_{\max}(s,a) = \max_{\pi \in \Pi} \mu^{\pi}(s,a).$$

This can be leveraged to compute a lower bound on the expected number of times of visiting (s,a) after rolling out each π in Π once.

Lemma 12. *Under the assumption of $C^p_{est} \leq \frac{\epsilon_{est} F}{2|\mathcal{S}||\mathcal{A}|H^2}$. For any fixed policy π, let $\Pi_{\mathcal{D}}$ be an exploration set of policies before simulating π. Then when $\mu^{\pi}(s,a) \in \left[\frac{\epsilon_{est}}{|\mathcal{S}||\mathcal{A}|H}, 2\mu^{\Pi_{\mathcal{D}}}_{\max}(s,a)\right]$, $\mu^{\Pi_{\mathcal{D}}}_{\max}(s,a) \geq \frac{\epsilon_{est}}{|\mathcal{S}||\mathcal{A}|H}$,$F \geq \frac{8|\mathcal{S}|^2 H^4 |\mathcal{A}|^2 \ln(2|\Pi|/\delta_{est})}{\epsilon^2_{est}}$ and $\tau \geq 6$, we have with probability at least $1 - \frac{\delta_{est}}{|\Pi|}$*

$$\sum_{i=1}^{F} \underbrace{|\{(s,a) \text{ or } Fail(s,a,i) \text{ included in } z^{\pi}_i\}|}_{\text{total number of times } z^{\pi}_i \text{ visited } (s,a)} < |\mathcal{D}_{s,a}| + \frac{\tau \epsilon_{est}}{|\mathcal{S}||\mathcal{A}|H} F$$

Proof. First, we are going to get the high probability lower bound on $|\mathcal{D}_{s,a}|$. Denote $\sum_{h=1}^{H} \mathbf{1}\{\pi''$ visits (s,a) at layer h during the rollout $j\}$ as X_j, where $\pi'' = \text{argmax}_{\pi \in \Pi^{\mathcal{D}}} \mu^{\pi}(s,a)$. We have

$$|\mathcal{D}_{s,a}| = \sum_{j=1}^{F\tau} \sum_{\pi' \in \Pi_{\mathcal{D}}} \sum_{h=1}^{H} \mathbf{1}\{\pi' \text{ visit } (s,a) \text{ at layer h during the rollout } j\} \geq \sum_{j=1}^{F\tau} X_j.$$

Note that $\{X_j\}$ is a sequence of independent random variable with each $X_j \in [0,H]$. We denote $\mathbb{E}[X_j]$ as $\mu^{\pi''}_{j,rollout}(s,a)$. From the corruption assumption $C^p_{est} \leq \frac{\epsilon_{est} F}{2|\mathcal{S}||\mathcal{A}|H^2}$ and by corollary 15, we have

$$|\frac{1}{F\tau} \sum_{j=1}^{F\tau} \mu^{\pi''}_{j,rollout}(s,a) - \mu^{\Pi_{\mathcal{D}}}_{\max}(s,a)| \leq \frac{HC^p_{est}}{F\tau} \leq \frac{\epsilon_{est}}{2|\mathcal{S}||\mathcal{A}|H} \tag{4}$$

which, combined with $\mu^{\Pi_{\mathcal{D}}}_{\max}(s,a) \geq \frac{\epsilon_{est}}{|\mathcal{S}||\mathcal{A}|H}$, also leads to

$$\frac{1}{F\tau} \sum_{j=1}^{F\tau} \mu^{\pi''}_{j,rollout}(s,a) \geq \frac{\epsilon_{est}}{2|\mathcal{S}||\mathcal{A}|H}$$

Then by using the Hoeffding's inequality, we get

$$\text{Prob}\left[\sum_{j}^{F\tau} X_j \leq \frac{1}{2} \sum_{j=1}^{F\tau} \mu^{\pi''}_{j,rollout}(s,a)\right] \leq \exp\left(-\frac{2F^2\tau^2}{F\tau H^2} (\frac{\epsilon_{est}}{4|\mathcal{S}||\mathcal{A}|H})^2\right) \leq \frac{\delta_{est}}{2|\Pi|}$$

Therefore, we get that with probability at least $1 - \frac{\delta_{est}}{2|\Pi|}$, $|D_{s,a}| > \frac{1}{2} \sum_{j=1}^{F\tau} \mu^{\pi}_j(s,a)$

Second, we are going to get the high probability upper bound on $\sum_{i=1}^{F} |\{(s,a)$ or $Fail(s,a,i)$ included in $z^{\pi}_i\}|$. Denote $|\{(s,a)$ or $Fail(s,a,i)$ included in $z^{\pi}_i\}|$ as $Y_i \in [0,H]$ and its expectation $\mathbb{E}[Y_i] = \mu^{\pi}_{i,sim}(s,a)$. By Claim 1, we know that each trajectory in $\{z^{\pi}_i\}_{i \in [F]}$ is independent to each other. Again from the corruption assumption $C^p_{est} \leq \frac{\epsilon_{est} F}{2|\mathcal{S}||\mathcal{A}|H^2}$ and by corollary 15, we have

$$|\frac{1}{F} \sum_{i=1}^{F} \mu^{\pi}_{i,sim}(s,a) - \mu^{\pi}(s,a)| \leq \frac{HC^p_{est}}{F} \leq \frac{\epsilon_{est}}{2|\mathcal{S}||\mathcal{A}|H} \tag{5}$$

which, combined with $\mu^{\pi}(s,a) \geq \frac{\epsilon_{est}}{|\mathcal{S}||\mathcal{A}|H}$, also leads to

$$\frac{1}{F} \sum_{j=1}^{F} \mu^{\pi}_{i,sim}(s,a) \geq \frac{\epsilon_{est}}{2|\mathcal{S}||\mathcal{A}|H}$$

So by using the hoeffding inequality again, we get that with probability at least $1-\frac{\delta_{est}}{2|\Pi|}$,

$$\sum_{i=1}^{F}|\{(s,a) \text{ or } Fail(s,a,i) \text{ included in } z_i^{\pi}\}| < \frac{3}{2}\sum_{i=1}^{F}\mu_{i,sim}^{\pi}(s,a)$$

Finally, combine the high probability upper bound and lower bound, we have that with probability at least $1-\frac{\delta_{est}}{|\Pi|}$

$$\begin{aligned}
&\sum_{i=1}^{F}|\{(s,a) \text{ or } Fail(s,a,i) \text{ included in } z_i^{\pi}\}| - |\mathcal{D}_{s,a}| \\
&< \frac{3}{2}\sum_{i=1}^{F}\mu_{i,sim}^{\pi}(s,a) - \frac{1}{2}\sum_{j=1}^{F\tau}\mu_{j,rollout}^{\pi''}(s,a) \\
&\le \frac{3}{2}F\mu^{\pi}(s,a) - \frac{1}{2}F\tau\mu_{max}^{\Pi_{\mathcal{D}}}(s,a) + \frac{\epsilon_{est}}{2|\mathcal{S}||\mathcal{A}|H}(F+F\tau) \\
&\le \frac{\epsilon_{est}}{2|\mathcal{S}||\mathcal{A}|H}(\frac{3}{2}F+\frac{1}{2}F\tau) < \frac{\epsilon_{est}}{|\mathcal{S}||\mathcal{A}|H}F\tau
\end{aligned}$$

where the second inequality comes from eq. 4, 5 and the last inequality comes form the the assumption $\mu^{\pi}(s,a) < 2\mu_{\max}^{\Pi_{\mathcal{D}}}(s,a), \tau \ge 6$. □

Lemma 13. *Under the assumption of* $C_{est}^{p} \le \frac{\epsilon_{est}F}{2|\mathcal{S}||\mathcal{A}|H^2}$ *. For any fixed policy* π*, let* $\Pi_{\mathcal{D}}$ *be an exploration set of policies before simulating* π*. Then when* $\mu^{\pi}(s,a) < \frac{\epsilon_{est}}{|\mathcal{S}||\mathcal{A}|H}$*,*$F \ge \frac{8|\mathcal{S}|^2H^4|\mathcal{A}|^2\ln(2|\Pi|/\delta_{est})}{\epsilon_{est}^2}$ *and* $\tau \ge 6$*, we have with probability at least* $1-\frac{\delta_{est}}{|\Pi|}$

$$\sum_{i=1}^{F}|\{(s,a) \textit{ or } Fail(s,a,i) \textit{ included in } z_i^{\pi}\}| < |\mathcal{D}_{s,a}| + \frac{\epsilon_{est}}{|\mathcal{S}||\mathcal{A}|H}F\tau$$

Proof. We just need to show that under this condition, $\sum_{i=1}^{F}|\{(s,a) \text{ or } Fail(s,a,i) \text{ included in } z_i^{\pi}\}| < \frac{\tau\epsilon_{est}}{|\mathcal{S}||\mathcal{A}|H}F$. To show this, we use the same method and notation used in the proof of Lemma 12 and get that with probability at least $1-\frac{\delta_{est}}{2|\Pi|}$,

$$\begin{aligned}
&\sum_{i=1}^{F}|\{(s,a) \text{ or } Fail(s,a,i) \text{ included in } z_i^{\pi}\}| \\
&\le \frac{3}{2}F\mu^{\pi}(s,a) + \frac{\epsilon_{est}}{2|\mathcal{S}||\mathcal{A}|H}F < \frac{2\epsilon_{est}}{|\mathcal{S}||\mathcal{A}|H}F < \frac{\tau\epsilon_{est}}{|\mathcal{S}||\mathcal{A}|H}F
\end{aligned}$$

□

Lemma 14. *Let* $\Pi_{\mathcal{D}}$ *be the set of policies maintained before executing line 9 and let* $\hat{\Pi}_{\mathcal{D}}$ *be the set of policies maintained after executing. Let* (s,a) *be the state action pair where the Fail occurs. Then we have, with probability at least* $1-\delta_{est}$*,*

$$\mu_{\max}^{\hat{\Pi}_{\mathcal{D}}}(s,a) \ge \max\{2\mu_{\max}^{\Pi_{\mathcal{D}}}(s,a), \frac{\epsilon_{est}}{|\mathcal{S}||\mathcal{A}|H}\}$$

Proof. If $\mu_{max}^{\Pi_{\mathcal{D}}} < \frac{\epsilon_{est}}{|\mathcal{S}||\mathcal{A}|H}$, by Lemma 13, we know that with probability at least $1-\frac{\delta_{est}}{|\Pi|}$, we always have $\mu_{max}^{\hat{\Pi}_{\mathcal{D}}} \ge \frac{\epsilon_{est}}{|\mathcal{S}||\mathcal{A}|H}$. Otherwise, if we already have $\mu_{max}^{\Pi_{\mathcal{D}}} \ge \frac{\epsilon_{est}}{|\mathcal{S}||\mathcal{A}|H}$, then by Lemma 12, we know that with probability at $1-\frac{\delta_{est}}{|\Pi|}$, $\mu_{max}^{\hat{\Pi}_{\mathcal{D}}} \ge 2\mu_{max}^{\Pi_{\mathcal{D}}}$. Finally, we take the union bound over all policies in Π to get the target result. □

D.4.2. Auxiliary Lemma

Definition 5. *Define $q_P^\pi(s,h)$ as the probability that policy π will visit s at step h given the underlying transition probability P. Also define $V_M^\pi(s_1)$ as the value function that policy π will induce given the underlaying MDP M.*

The change of the visiting probability and the value function for any fixed π can be upper bounded in terms of the change of transition functions and expected rewards. Here we consider the most general case that the transition function and the expected rewards is non-stationary between each layers. We want to remark that, although our underlying MDP is stationary by assumption, our corruptions is allowed to be non-stationary. Also our algorithm will simulate a trajectory by the sample collected from different times. Therefore, we prove the following lemma for the non-stationary case.

Lemma 15 (Corruption Effects on Visiting Probability). *For any step h',*

$$\begin{aligned}
&\sum_{s\in\mathcal{S}} |q_{P_1}^\pi(s,h') - q_{P_2}^\pi(s,h')| \\
&\leq \min\{1, \sum_{h=2}^{h'-1} \sup_{s\in\mathcal{S},a\in\mathcal{A}} \|P_1(\cdot|s,a,h) - P_2(\cdot|s,a,h)\|_1 + \sup_{a\in\mathcal{A}} \|P_1(\cdot|s_0,a,1) - P_2(\cdot|s_0,a,1)\|_1\}
\end{aligned}$$

Proof. We prove this by induction. First, we can easily get the base case that

$$\sum_{s\in\mathcal{S}} |q_{P_1}^\pi(s,2) - q_{P_2}^\pi(s,2)| \leq \sup_{a\in\mathcal{A}} \|P_1(\cdot|s_0,a) - P_2(\cdot|s_0,a)\|_1\}.$$

Then by assuming that, for any step $h' \geq 3$,

$$\begin{aligned}
&\sum_{s\in\mathcal{S}} |q_{P_1}^\pi(s,h') - q_{P_2}^\pi(s,h')| \\
&\leq \sum_{h=2}^{h'-1} \sup_{s\in\mathcal{S},a\in\mathcal{A}} \|P_1(\cdot|s,a,h) - P_2(\cdot|s,a,h)\|_1 + \sup_{a\in\mathcal{A}} \|P_1(\cdot|s_0,a,1) - P_2(\cdot|s_0,a,1)\|_1,
\end{aligned}$$

we have that, for any step $h'+1$,

$$\begin{aligned}
&\sum_{s\in\mathcal{S}} |q_{P_1}^\pi(s,h'+1) - q_{P_2}^\pi(s,h'+1)| \\
&\leq \sum_{s\in\mathcal{S}} | \sum_{s'\in\mathcal{S}} \left(q_{P_1}^\pi(s',h') - q_{P_2}^\pi(s',h')\right) P_1(s|s',\pi_{h'}(s'),h')| \\
&\quad + \sum_{s\in\mathcal{S}} | \sum_{s'\in\mathcal{S}} q_{P_2}^\pi(s',h') \left(P_1(s|s',\pi_{h'}(s',h',h') - P_2(s|s',\pi_{h'}(s'),h')\right) | \\
&\leq \sum_{s'\in\mathcal{S}} \left|q_{P_1}^\pi(s',h') - q_{P_2}^\pi(s',h')\right| \sum_{s\in\mathcal{S}} P_1(s|s',\pi_{h'}(s')) + \sum_{s'\in\mathcal{S}} q_{P_2}^\pi(s',h') \sum_{s\in\mathcal{S}} |P_1(s|s',\pi(s',h') - P_2(s|s',\pi_{h'}(s'))| \\
&\leq \sum_{s'\in\mathcal{S}} \left|q_{P_1}^\pi(s',h') - q_{P_2}^\pi(s'.h')\right| + \sup_{s'\in\mathcal{S}} \sum_{s\in\mathcal{S}} |P_1(s|s',\pi_{h'}(s'),h') - P_2(s|s',\pi_{h'}(s'),h')| \\
&\leq \sum_{h=2}^{h'} \sup_{s\in\mathcal{S},a\in\mathcal{A}} \|P_1(\cdot|s,a) - P_2(\cdot|s,a)\|_1 + \sup_{a\in\mathcal{A}} \|P_1(\cdot|s_0,a,h') - P_2(\cdot|s_0,a,h')\|_1
\end{aligned}$$

□

Lemma 16 (Corruption effects on value function).

$$\begin{aligned}
|V^{M_1,\pi} - V^{M_2,\pi}| \leq H \sum_{h=2}^{H} \sup_{s'\in\mathcal{S}} \|P_1(\cdot|s',\pi(s'),h) - P_2(\cdot|s',\pi(s'),h)\|_1 + \sum_{h=2}^{H} \sup_{s\in\mathcal{S}} |\mu_1(s,\pi(s),) - \mu_2(s,\pi(s),h)| \\
+ \|P_1(\cdot|s_0,\pi(s_0),1) - P_2(\cdot|s_0,\pi(s_0),1)\|_1 + |\mu_1(s_0,\pi(s_0),1) - \mu_2(s_0,\pi(s_0),1)|
\end{aligned}$$

Proof. For convenience, when I write $\sum_{h=1}^H \sum_{s\in\mathcal{S}}$ in the following, I actually mean $\sum_{h=2}^H \sum_{s\in\mathcal{S}} + \sum_{s=s_0}$.

$$
\begin{aligned}
&|V^{M_1,\pi}(s_0) - V^{M_2,\pi}(s_0)| \\
&\leq |\sum_{h=1}^H \sum_{s\in\mathcal{S}} \left(q_{P_1}^\pi(s,h) - q_{P_2}^\pi(s,h)\right) \mu_1(s,\pi(s),h)| + |\sum_{h=1}^H \sum_{s\in\mathcal{S}} q_{P_2}^\pi(s,h)\left(\mu_1(s,\pi(s),h) - \mu_2(s,\pi(s),h)\right)\ | \\
&\leq |\sum_{h=1}^H \sup_{s\in\mathcal{S}} \mu_1(s,\pi_1(s)) \sum_{s\in\mathcal{S}} \left(q_{P_1}^\pi(s,h) - q_{P_2}^\pi(s,h)\right)| + \sum_{h=1}^H \sup_{s\in\mathcal{S}} |\mu_1(s,\pi(s),h) - \mu_2(s,\pi(s),h)| \\
&\leq \left(\sum_{h=1}^H \sup_{s\in\mathcal{S}} \mu_1(s,\pi_1(s))\right)\left(\sum_{h=1}^H \sup_{s\in\mathcal{S}} \|P_1(\cdot|s,\pi(s),h) - P_2(\cdot|s,\pi(s),h)\|_1\right) \\
&\quad + \sum_{h=1}^H \sup_{s\in\mathcal{S}} |\mu_1(s,\pi(s),h) - \mu_2(s,\pi(s),h)| \\
&\leq H \sum_{h=1}^H \sup_{s\in\mathcal{S}_h} \|P_1(\cdot|s,\pi(s),h) - P_2(\cdot|s,\pi(s),h)\|_1 + \sum_{h=1}^H \sup_{s\in\mathcal{S}} |\mu_1(s,\pi(s),h) - \mu_2(s,\pi(s),h)|
\end{aligned}
$$

Here the third inequality comes from Lemma 15 and the last inequality comes from the assumption on the reward function. □

E. Discussion on Reward-free Exploration Algorithm under Corruptions

In the Related Work section, we mentioned that algorithms proposed in Kaufmann et al. (2020) and Ménard et al. (2020) can *efficiently* achieve uniform ϵ-close estimations for all the polices with near-optimal sample complexity in the no-corruption setting. Their main idea is to construct a computable estimator of Q-value estimation error for all the state-action pairs and greedily play the action that maximize such estimator at every step until all the state-action pairs have sufficiently small Q-value estimation errors. So a natural question to ask is,

Can we replace the ESTALL with this type of efficient algorithms ?

To be specific, firstly, in the non-corrupted setting, we want to find an efficient algorithm that can guarantee uniform estimations on all the policies in any given policy set Π by only implementing polices inside Π. Secondly, we also want this algorithm has corruption robustness at least not worse than the ESTALL.

For the first target, we can easily define an estimator $W_t(\pi) = \sum_{h=1}^H \sum_{s\in\mathcal{S}} \frac{\hat{p}_{t,h}^\pi(s)}{n_h^t(s,\pi(s))}$, where $n_h^t(s,\pi(s))$ is the empirical number of times state-action-step pair $(s,\pi(s),h)$ has been visited before time $t+1$ and $\hat{p}_{t,h}^\pi(s)$ is the empirical probability that the policy π reach state s at h before time $t+1$. Suppose we have an efficient oracle that can calculate the following in the polynomial times,

$$\text{argmax}_{\pi\in\Pi} W_t(\pi)$$

Then we can find an oracle-efficient algorithm by greedily sampling $\pi_{t+1} = \text{argmax}_{\pi\in\Pi} W_t(\pi)$ until all the $W_t(\pi)$ are small enough.

Unfortunately, in the presence of corruptions, we find it is hard to get a good robustness. Roughly speaking, suppose the rewards are fixed, then the estimation error $\hat{V}^\pi$ for any policy π is upper bounded by

$$|V^\pi - \hat{V}^\pi| \leq \min_{t\in\mathcal{I}} C_\mathcal{I}^p W_t(\pi_{t+1}) + \sqrt{W_t(\pi_{t+1})}$$

where $\mathcal{I}$ represents the whole time period this algorithm is running. Then from our perspective, when $|\mathcal{I}| = o(1/\epsilon^2)$, we can only guarantee $\min_{t\in\mathcal{I}} W_t(\pi_{t+1}) \leq \tilde{\mathcal{O}}\left(poly(|\mathcal{S}||\mathcal{A}|H(\epsilon^2 + C_\mathcal{I}^p\epsilon^2)\right)$, which gives

$$|V^\pi - \hat{V}^\pi| \leq \tilde{\mathcal{O}}\left(poly(|\mathcal{S}||\mathcal{A}|H)((C_\mathcal{I}^p)^2\epsilon^2 + \sqrt{C_\mathcal{I}^p}\epsilon)\right)$$

Note that EstAll gives $\tilde{\mathcal{O}}\left(poly(|\mathcal{S}||\mathcal{A}|H((C^p_{\mathcal{I}})^2\epsilon^2+\epsilon))\right)$-close estimations when $C^p_{\mathcal{I}} \leq 1/\epsilon$. Therefore, plug-in this algorithm instead of EstAll in BARBAR-RL will give worse dependence in T.

Whether we can find a better estimator in this type of reward-free sub-algorithms or whether we can find another proper meta-algorithm for this type of sub-algorithms remains open.