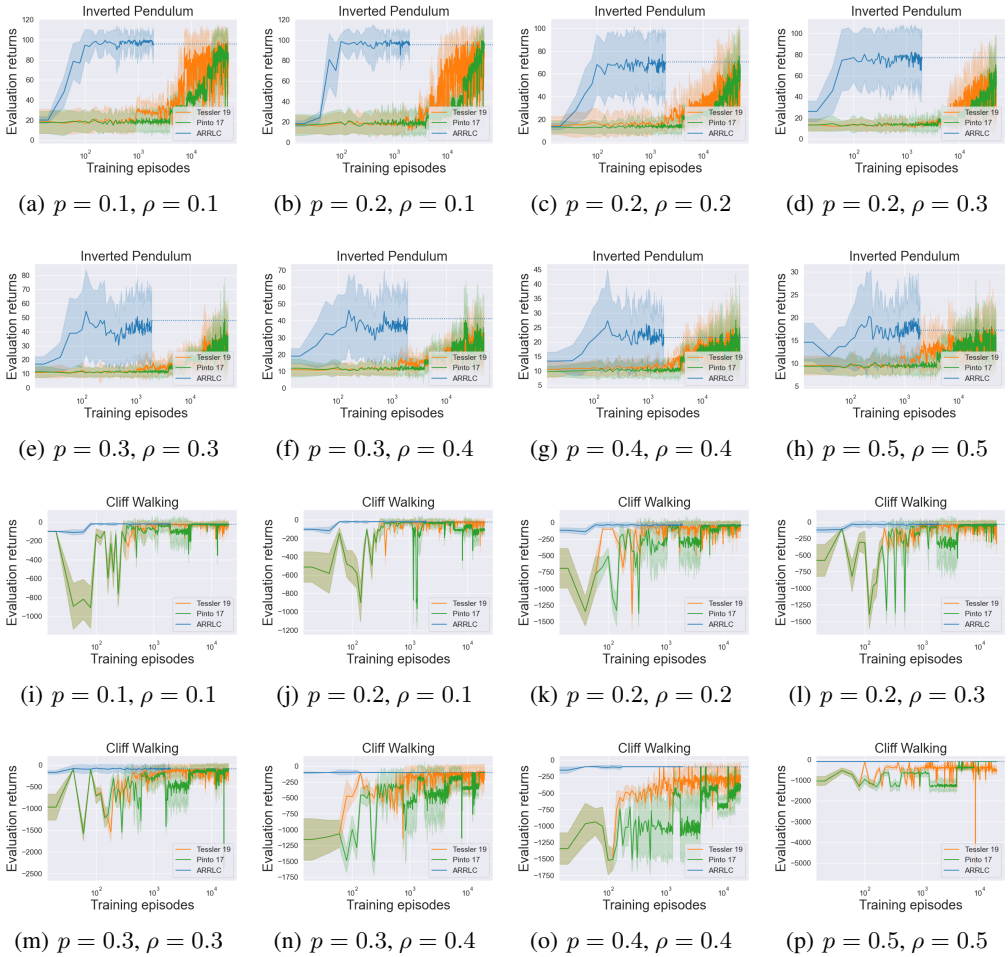
A Additional Numerical Results

A.1 Ablation study and more comparisons

In Section 7, we compared our algorithm with the robust TD algorithm in (Klima et al., 2019). Here, we compare our algorithm with the algorithms in (Pinto et al., 2017; Tessler et al., 2019). The method in (Tessler et al., 2019) requires an MDP solver to solve the optimal adversarial policy when the agent policy is given and the optimal agent policy when the adversarial policy is given. The white-box MDP solver requires knowledge of the underline MDP so that there is no learning curve and sample complexity discussion in (Tessler et al., 2019). Thus, we implement the algorithms in (Pinto et al., 2017; Tessler et al., 2019) with a Q-learning MDP solver, and compared the final evaluation rewards and the learning curve. In addition, we implement the ablation study by setting different ρ and p. In our experiments, the policy is learned in a clean environment, and is then tested on the perturbed environment. ρ is the parameter in algorithm when learning the robust policy. ρ can be considered as the agent's guess about the probability of a disturbance occurring. However, p is the probability that the perturb happens in the perturbed environment. In the perturbed environment, with probability p, the action is perturbed by an adversarial action.



(a) $p = 0.1, \rho = 0.1$ (b) $p = 0.2, \rho = 0.1$ (c) $p = 0.2, \rho = 0.2$ (d) $p = 0.2, \rho = 0.3$

(e) $p = 0.3, \rho = 0.3$ (f) $p = 0.3, \rho = 0.4$ (g) $p = 0.4, \rho = 0.4$ (h) $p = 0.5, \rho = 0.5$

(i) $p = 0.1, \rho = 0.1$ (j) $p = 0.2, \rho = 0.1$ (k) $p = 0.2, \rho = 0.2$ (l) $p = 0.2, \rho = 0.3$

(m) $p = 0.3, \rho = 0.3$ (n) $p = 0.3, \rho = 0.4$ (o) $p = 0.4, \rho = 0.4$ (p) $p = 0.5, \rho = 0.5$

Figure 6: ARRLC v.s. RARL v.s. PR-PI

In Figure 6, we show the learning curves under different p and ρ. It can be seen that our ARRLC algorithm converges faster than the other algorithms. This demonstrates the efficiency of our ARRLC algorithm to learn optimal policy under policy execution uncertainty.

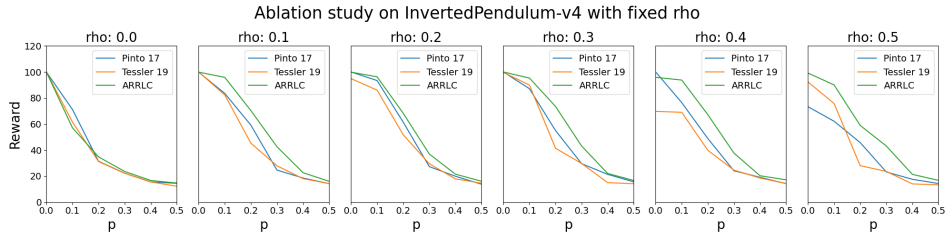


Figure 7: Ablation study on InvertedPendulum-v4 with fixed ρ.

In Figure 7, given the agents trained with fixed rho, we test the agents in different disturbed environments with different p. In Figure 8, we compared the different agents trained with different rho. The x-axis is the different choice of ρ or p. The y-axis is the final evaluation rewards.

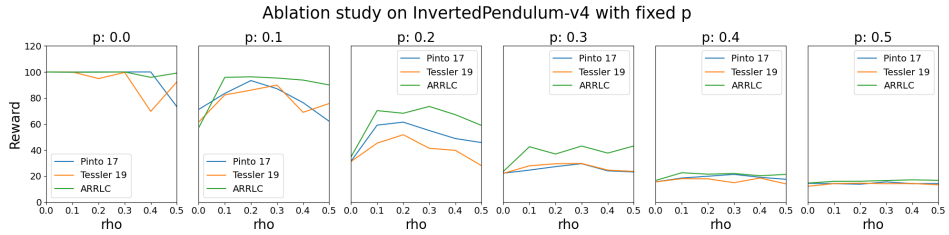


Figure 8: Ablation study on InvertedPendulum-v4 with fixed ρ.

The theoretical guarantee on sample complexity and regret of our algorithm relies on the assumption of known uncertainty parameter. However, in the experimental results shown in 7, the parameter can mismatch with the true disturb probability. In the main paper Figure 9, we test the mismatch of the uncertainty parameter ρ and true uncertainty probability p. We trained the agent with $\rho = 0.2$, but we use $p = 0.1$ in the test. The proposed robust algorithm still outperforms the non-robust algorithm.

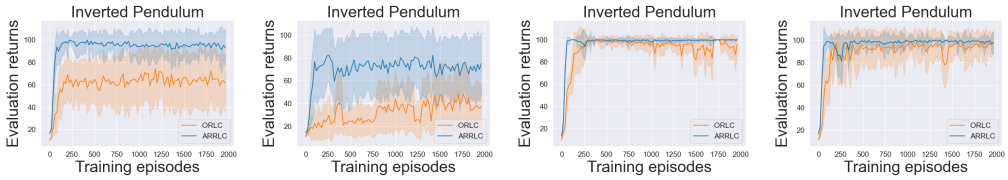
A.2 Robustness to different adversary

In this section we considered different adversary policies include both the fixed policy in the main page and a random adversary policy. After the agent takes an action, with probability p, the random adversary will uniformly randomly choose an adversary action to replace the agent's action. In Figure 9 and Figure 10, "fix" represents that the actions are perturbed by a fixed adversarial policy during the testing, "random" represents that the actions are randomly perturbed during the testing, p is the action perturbation probability.

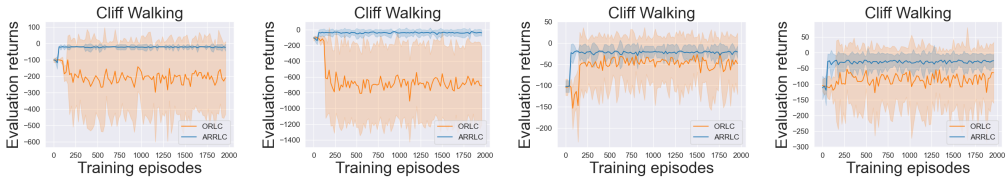
Since we do not know whether the fixed policy or the random policy is the strongest adversary policy against the agent, a more direct comparison is to use the learned worst-case policy in different algorithms to do a cross-comparison. We used the learned worst-case policies to disturb the different robust agents. We report the final evaluation rewards in Table 1. We trained our method in 2000 episodes and the approaches of Pinto et al. (2017); Tessler et al. (2019) in 30000 episodes. We set that $p = \rho = 0.2$. The ARRLC agent performs the best against three different adversaries and the ARRLC adversary impacts the most on three different agents.

Table 1: Final rewards under cross-comparison between ARRLC, PR-PI and RAPL

	ARRLC adversary	RAPL adversary	PR-PI adversary
ARRLC agent	**72.536**	**81.736**	**89.824**
RAPL agent	49.936	72.216	70.6
PR-PI agent	52.788	63.784	86.648

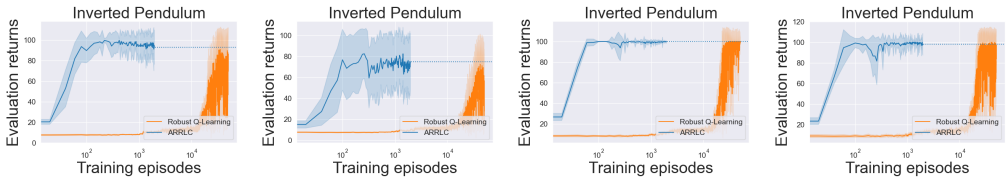


(a) p=0.1, fix (b) p=0.2, fix (c) p=0.1, random (d) p=0.2, random

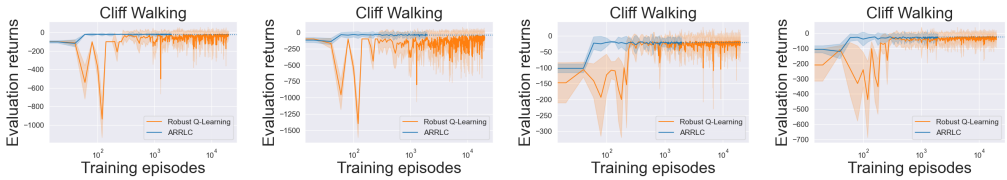


(e) p=0.1, fix (f) p=0.2, fix (g) p=0.1, random (h) p=0.2, random

Figure 9: ARRLC v.s. ORLC.



(a) p=0.1, fix (b) p=0.2, fix (c) p=0.1, random (d) p=0.2, random



(e) p=0.1, fix (f) p=0.2, fix (g) p=0.1, random (h) p=0.2, random

Figure 10: ARRLC v.s. Robust TD

B PROOF OF PROPOSITION 1

The uncertainty set of the policy execution has the form in:

$$\Pi^{\rho}(\pi) := \{\widetilde{\pi} | \forall s, \widetilde{\pi}_h(\cdot|s) = (1-\rho)\pi(\cdot|s) + \rho\pi'_h(\cdot|s), \pi'_h(\cdot|s) \in \Delta_{\mathcal{A}}\}. \tag{12}$$

We define

$$\begin{aligned} C_h^{\pi,\pi',\rho}(s) &:= \mathbb{E}\left[\sum_{h'=h}^{H} R_{h'}(s_{h'}, a_{h'}) | s_h = s, a_{h'} \sim \widetilde{\pi}_{h'}(\cdot|s_{h'})\right] \\ D_h^{\pi,\pi',\rho}(s,a) &:= \mathbb{E}\left[\sum_{h'=h}^{H} R_{h'}(s_{h'}, a_{h'}) | s_h = s, a_h = a, a_{h'} \sim \widetilde{\pi}_{h'}(\cdot|s_{h'})\right]. \end{aligned}$$

Robust Bellman Equation First we prove the action robust Bellman equation holds for any policy π, state s action a and step h. From the definition of the robust value function in (1), we have $V_{H+1}^{\pi}(s) = 0, \forall s \in \mathcal{S}$.

We prove the robust Bellman equation by building a policy π^{-}. Here, policy π^{-} is the optimal adversarial policy towards the policy π.

At step H, we set $\pi_H^-(s) = \arg\min_{a\in\mathcal{A}} R_H(s,a)$. We have

$$\begin{aligned} V_H^{\pi}(s) &= \min_{\pi'} C_H^{\pi,\pi',\rho}(s) \\ &= (1-\rho)[\mathbb{D}_{\pi_H} R_H](s) + \rho \min_{\pi'}[\mathbb{D}_{\pi'_H} R_H](s) \\ &= (1-\rho)[\mathbb{D}_{\pi_H} Q_H^{\pi}](s) + \rho \min_{a\in\mathcal{A}} Q_H^{\pi}(s,a) = C_H^{\pi,\pi^-,\rho}(s), \end{aligned} \tag{13}$$

as $V_{H+1} = 0$.

The robust Bellman equation holds at step H and $\min_{\pi'} \sum_s w(s) C_H^{\pi,\pi',\rho}(s) = \sum_s w(s) \min_{\pi'} C_H^{\pi,\pi',\rho}(s) = \sum_s w(s) C_H^{\pi,\pi^-,\rho}(s)$ for any state s and any weighted function $w : \mathcal{S} \to \Delta_{\mathcal{S}}$.

Suppose the robust Bellman equation holds at step $h+1$ and $\min_{\pi'} \sum_s w(s) C_{h+1}^{\pi,\pi',\rho}(s) = \sum_s w(s) \min_{\pi'} C_{h+1}^{\pi,\pi',\rho}(s) = \sum_s w(s) C_{h+1}^{\pi,\pi^-,\rho}(s)$ for any state s and any weighted function $w : \mathcal{S} \to \Delta_{\mathcal{S}}$.

Now we prove the robust Bellman equation holds at step h. From the definition of the robust Q-function in (2) and the form of uncertainty set, we have

$$\begin{aligned} Q_h^{\pi}(s,a) &= \min_{\widetilde{\pi}\in\Pi(\pi)} \mathbb{E}\left[\sum_{h'=h}^{H} R_{h'}(s_{h'}, a_{h'}) | s_h = s, a_h = a, a_{h'} \sim \widetilde{\pi}_{h'}(\cdot|s_{h'})\right] \\ &= \min_{\pi'} D_h^{\pi,\pi',\rho}(s,a) \\ &= R_h(s,a) + \min_{\pi'} \mathbb{E}_{s'\sim P_h(\cdot|s,a)} C_{h+1}^{\pi,\pi',\rho}(s) \\ &= R_h(s,a) + \mathbb{E}_{s'\sim P_h(\cdot|s,a)} \min_{\pi'} C_{h+1}^{\pi,\pi',\rho}(s) \\ &= R_h(s,a) + [P_h V_{h+1}^{\pi}](s,a). \end{aligned} \tag{14}$$

We also have that $Q_h^{\pi}(s,a) = D_h^{\pi,\pi^-,\rho}(s,a)$.

Recall that a (stochastic) Markov policy is a set of H maps $\pi := \{\pi_h : \mathcal{S} \to \Delta_{\mathcal{A}}\}_{h\in[H]}$. From the definition of the robust value function in (1) and the form of uncertainty set, we have

$$\begin{aligned} V_h^{\pi}(s) &= \min_{\widetilde{\pi}\in\Pi(\pi)} \mathbb{E}\left[\sum_{h'=h}^{H} R_{h'}(s_{h'}, a_{h'}) | s_h = s, a_{h'} \sim \widetilde{\pi}_{h'}(\cdot|s_{h'})\right] \\ &= \min_{\pi'} C_h^{\pi,\pi',\rho}(s) \\ &= \min_{\pi'_h} \min_{\{\pi'_{h'}\}_{h'=h+1}^{H}} C_h^{\pi,\pi',\rho}(s) \\ &\geq (1-\rho) \min_{\{\pi'_{h'}\}_{h'=h+1}^{H}} \mathbb{E}_{a\sim\pi_h(\cdot|s)} D_h^{\pi,\pi',\rho}(s,a) + \rho \min_{\pi'_h} \min_{\{\pi'_{h'}\}_{h'=h+1}^{H}} \mathbb{E}_{a\sim\pi'_h(\cdot|s)} D_h^{\pi,\pi',\rho}(s,a) \\ &\geq (1-\rho) \mathbb{E}_{a\sim\pi_h(\cdot|s)} \min_{\{\pi'_{h'}\}_{h'=h+1}^{H}} D_h^{\pi,\pi',\rho}(s,a) + \rho \min_{\pi'_h} \mathbb{E}_{a\sim\pi'_h(\cdot|s)} \min_{\{\pi'_{h'}\}_{h'=h+1}^{H}} D_h^{\pi,\pi',\rho}(s,a) \\ &= (1-\rho)[\mathbb{D}_{\pi_h} Q_h^{\pi}](s) + \rho \min_{a\in\mathcal{A}} Q_h^{\pi}(s,a). \end{aligned} \tag{15}$$

We set $\pi_h^-(s) = \arg\min_{a\in\mathcal{A}} Q_h^{\pi}(s,a) = \arg\min_{a\in\mathcal{A}} D_h^{\pi,\pi^-,\rho}(s,a)$.

At step h, we have

$$\begin{aligned} V_h^{\pi}(s) &\leq C_h^{\pi,\pi^-,\rho}(s) \\ &= (1-\rho)[\mathbb{D}_{\pi_h} D_h^{\pi,\pi^-,\rho}](s) + \rho \min_{a\in\mathcal{A}} D_h^{\pi,\pi^-,\rho}(s,a) \\ &= (1-\rho)[\mathbb{D}_{\pi_h} Q_h^{\pi}](s) + \rho \min_{a\in\mathcal{A}} Q_h^{\pi}(s,a), \end{aligned} \tag{16}$$

where the last equation comes from the robust Bellman equation at step $h+1$ and

$$D_h^{\pi,\pi^-,\rho}(s,a) = R_h(s,a) + [P_h C_{h+1}^{\pi,\pi^-,\rho}](s,a) = R_h(s,a) + [P_h V_{h+1}^{\pi}](s,a).$$

Thus, the robust Bellman equation holds at step h.

Then, we prove the commutability of the expectation and the minimization operations at step h. For any weighted function w, we have $\min_{\pi'} \sum_s w(s) C_h^{\pi,\pi',\rho}(s) \geq \sum_s w(s) \min_{\pi'} C_h^{\pi,\pi',\rho}(s)$. Then, $\min_{\pi'} \sum_s w(s) C_h^{\pi,\pi',\rho}(s) \leq \sum_s w(s) C_h^{\pi,\pi^-,\rho}(s) = \sum_s w(s) \min_{\pi'} C_h^{\pi,\pi',\rho}(s)$.

By induction on $h = H, \cdots, 1$, we prove the robust Bellman equation.

Perfect Duality and Robust Bellman Optimality Equation We now prove that the perfect duality holds and can be solved by the optimal robust Bellman equation.

The control problem in the LHS of (4) is equivalent to

$$\max_{\pi} \min_{\widetilde{\pi} \in \Pi^{\rho}(\pi)} \mathbb{E}\left[\sum_{h'=h}^{H} R_{h'}(s_{h'}, a_{h'}) | s_h = s, a_{h'} \sim \widetilde{\pi}_{h'}(\cdot|s_{h'})\right] = \max_{\pi} \min_{\pi'} C_h^{\pi,\pi',\rho}(s). \tag{17}$$

The control problem in the RHS of (4) is equivalent to

$$\min_{\widetilde{\pi} \in \Pi^{\rho}(\pi)} \max_{\pi} \mathbb{E}\left[\sum_{h'=h}^{H} R_{h'}(s_{h'}, a_{h'}) | s_h = s, a_{h'} \sim \widetilde{\pi}_{h'}(\cdot|s_{h'})\right] = \min_{\pi'} \max_{\pi} C_h^{\pi,\pi',\rho}(s). \tag{18}$$

For step H, we have $C_H^{\pi,\pi',\rho}(s) = [\mathbb{D}_{((1-\rho)\pi+\rho\pi')_H} R_H](s) = (1-\rho)[\mathbb{D}_{\pi_H} R_H](s) + \rho[\mathbb{D}_{\pi'_H} R_H](s)$. Thus, we have

$$\begin{aligned} \max_{\pi} \min_{\pi'} C_H^{\pi,\pi',\rho}(s) =& (1-\rho) \max_{\pi} [\mathbb{D}_{\pi_H} R_H](s) + \rho \min_{\pi'} [\mathbb{D}_{\pi'_H} R_H](s) \\ =& (1-\rho) \max_{a \in \mathcal{A}} R_H(s,a) + \rho \min_{b \in \mathcal{A}} R_H(s,b), \end{aligned} \tag{19}$$

and

$$\begin{aligned} \min_{\pi'} \max_{\pi} C_H^{\pi,\pi',\rho}(s) =& (1-\rho) \max_{\pi} [\mathbb{D}_{\pi_H} R_H](s) + \rho \min_{\pi'} [\mathbb{D}_{\pi'_H} R_H](s) \\ =& (1-\rho) \max_{a \in \mathcal{A}} R_H(s,a) + \rho \min_{b \in \mathcal{A}} R_H(s,b). \end{aligned} \tag{20}$$

At step H, the perfect duality holds for all s and there always exists an optimal robust policy $\pi_H^*(s) = \arg\max_{a \in \mathcal{A}} Q_H^*(s,a) = \arg\max_{a \in \mathcal{A}} R_H(s,a)$ and its corresponding optimal adversarial policy $\pi_H^-(s) = \arg\min_{a \in \mathcal{A}} R_H(s,a)$ which are deterministic. The action robust Bellman optimality equation holds at step H for any stats s and action a.

In addition, $\max_{\pi} \min_{\pi'} \sum_s w(s) C_H^{\pi,\pi',\rho}(s) = \sum_s w(s) \max_{\pi} \min_{\pi'} C_H^{\pi,\pi',\rho}(s)$ for any weighted function $w : \mathcal{S} \to \Delta_{\mathcal{S}}$. This can be shown as

$$\begin{aligned} & \max_{\pi} \min_{\pi'} \sum_{s \in \mathcal{S}} w(s) C_H^{\pi,\pi',\rho}(s) \\ =& (1-\rho) \max_{\pi} \sum_{s \in \mathcal{S}} w(s) [\mathbb{D}_{\pi_H} R_H](s) + \rho \min_{\pi'} \sum_{s \in \mathcal{S}} w(s) [\mathbb{D}_{\pi'_H} R_H](s) \\ =& (1-\rho) \sum_{s \in \mathcal{S}} w(s) \max_{a \in \mathcal{A}} R_H(s,a) + \rho \sum_{s \in \mathcal{S}} w(s) \min_{b \in \mathcal{A}} R_H(s,b). \end{aligned} \tag{21}$$

Suppose that at steps from $h+1$ to H, the perfect duality holds for any s, the action robust Bellman optimality equation holds for any state s and action a, there always exists an optimal robust policy $\pi_{h'}^* = \arg\max_{a \in \mathcal{A}} Q_{h'}^*(s,a)$ and its corresponding optimal adversarial policy $\pi_{h'}^-(s) = \arg\min_{a \in \mathcal{A}} Q_{h'}^*(s,a)$, $\forall h' \geq h+1$, which is deterministic, and $\max_{\pi} \min_{\pi'} \sum_s w(s) C_{h'}^{\pi,\pi',\rho}(s) = \sum_s w(s) \max_{\pi} \min_{\pi'} C_{h'}^{\pi,\pi',\rho}(s)$ for any state s, any weighted function $w : \mathcal{S} \to \Delta_{\mathcal{S}}$ and any

$h' \geq h+1$. We have $V^*_{h'}(s) = V^{\pi^*}_{h'}(s) = C^{\pi^*,\pi^-,\rho}_{h'}(s)$ and $Q^*_{h'}(s,a) = Q^{\pi^*}_{h'}(s,a) = D^{\pi^*,\pi^-,\rho}_{h'}(s,a)$ for any state s and any $h' \geq h+1$.

We first prove that the robust Bellman optimality equation holds at step h.

We have

$$\begin{aligned} Q^*_h(s,a) &= \max_{\pi} \min_{\pi'} D^{\pi,\pi',\rho}_h(s,a) \\ &= \max_{\pi} \min_{\pi'} (R_h(s,a) + [P_h C^{\pi,\pi',\rho}_{h+1}](s,a)) \\ &= R_h(s,a) + [P_h(\max_{\pi} \min_{\pi'} C^{\pi,\pi',\rho}_{h+1})](s,a) \\ &= R_h(s,a) + [P_h V^*_{h+1}](s,a). \end{aligned} \tag{22}$$

and also $Q^*_h(s,a) = Q^{\pi^*}_h(s,a) = D^{\pi^*,\pi^-,\rho}_h(s,a)$.

From the robust Bellman equation, we have

$$\begin{aligned} \max_{\pi} V^{\pi}_h(s) =& \max_{\pi} \left((1-\rho)[\mathbb{D}_{\pi_h} Q^{\pi}_h](s) + \rho \min_{a\in\mathcal{A}} Q^{\pi}_h(s,a) \right) \\ \leq& (1-\rho) \max_{\pi_h} \max_{\{\pi_h\}^H_{h'=h+1}} [\mathbb{D}_{\pi_h} Q^{\pi}_h](s) + \rho \max_{\{\pi_h\}^H_{h'=h+1}} \min_{a\in\mathcal{A}} Q^{\pi}_h(s,a) \\ \leq& (1-\rho) \max_{\pi_h} \max_{\{\pi_h\}^H_{h'=h+1}} [\mathbb{D}_{\pi_h} Q^{\pi}_h](s) + \rho \min_{a\in\mathcal{A}} \max_{\{\pi_h\}^H_{h'=h+1}} Q^{\pi}_h(s,a) \\ \leq& (1-\rho) \max_{\pi_h} [\mathbb{D}_{\pi_h} Q^*_h](s) + \rho \min_{a\in\mathcal{A}} Q^*_h(s,a) \\ =& (1-\rho) \max_{a\in\mathcal{A}} Q^*_h(s,a) + \rho \min_{a\in\mathcal{A}} Q^*_h(s,a). \end{aligned} \tag{23}$$

We set $\pi^*_h(s) = \max_{a\in\mathcal{A}} Q^*_h(s,a)$. According to the robust bellman equation, we have

$$\begin{aligned} \max_{\pi} V^{\pi}_h(s) \geq V^{\pi^*}_h(s) &= (1-\rho)[\mathbb{D}_{\pi^*_h} Q^{\pi^*}_h](s) + \rho \min_{a\in\mathcal{A}} Q^{\pi^*}_h(s,a) \\ &= (1-\rho) \max_{a\in\mathcal{A}} Q^{\pi^*}_h(s,a) + \rho \min_{a\in\mathcal{A}} Q^{\pi^*}_h(s,a) \\ &= (1-\rho) \max_{a\in\mathcal{A}} Q^*_h(s,a) + \rho \min_{a\in\mathcal{A}} Q^*_h(s,a). \end{aligned} \tag{24}$$

Thus, the robust Bellman optimality equation holds at step h. There always exists an optimal robust policy $\pi^*_h = \arg\max_{a\in\mathcal{A}} Q^*_h(s,a)$ and its corresponding optimal adversarial policy $\pi^-_h(s) = \arg\min_{a\in\mathcal{A}} Q^*_h(s,a)$ that is deterministic so that $C^{\pi^*,\pi^-,\rho}_h(s) = V^*_h(s)$.

Then, we prove the commutability of the expectation, the minimization and the maximization operations at step h.

In the proof of robust Bellman equation, we have shown that

$$\min_{\pi'} \sum_s w(s) C^{\pi,\pi',\rho}_h(s) = \sum_s w(s) \min_{\pi'} C^{\pi,\pi',\rho}_h(s)$$

for any policy π and any weighted function w. Hence

$$\max_{\pi} \min_{\pi'} \sum_s w(s) C^{\pi,\pi',\rho}_h(s) \sum_s = \max_{\pi} \sum_s w(s) \min_{\pi'} C^{\pi,\pi',\rho}_h(s).$$

First, we have

$$\max_{\pi} \sum_s w(s) \min_{\pi'} C^{\pi,\pi',\rho}_h(s) \leq \sum_s w(s) \max_{\pi} \min_{\pi'} C^{\pi,\pi',\rho}_h(s).$$

Then, we can show

$$\begin{aligned} \max_{\pi} \sum_s w(s) \min_{\pi'} C^{\pi,\pi',\rho}_h(s) &\geq \sum_s w(s) \min_{\pi'} C^{\pi^*,\pi',\rho}_h(s) \\ &= \sum_s w(s) C^{\pi^*,\pi^-,\rho}_h(s) \\ &= \sum_s w(s) \max_{\pi} \min_{\pi'} C^{\pi,\pi',\rho}_h(s). \end{aligned} \tag{25}$$

In summary,

$$\max_{\pi}\min_{\pi'}\sum_{s} w(s)C_h^{\pi,\pi',\rho}(s)\sum_{s} = w(s)\max_{\pi}\min_{\pi'}C_h^{\pi,\pi',\rho}(s).$$

We can show the perfect duality at step h by

$$\max_{\pi}\min_{\pi'}C_h^{\pi,\pi',\rho}(s) = C_h^{\pi^*,\pi^-,\rho}(s) = \max_{\pi}C_h^{\pi,\pi^-,\rho}(s) \geq \min_{\pi'}\max_{\pi}C_h^{\pi,\pi',\rho}(s). \tag{26}$$

By induction on $h = H, \cdots, 1$, we prove Proposition 1.

C Proof for Action Robust Reinforcement Learning with Certificates

In this section, we prove Theorem 1. Recall that we use $\overline{Q}_h^k$,$\overline{V}_h^k$,$\underline{Q}_h^k$,$\underline{V}_h^k$, N_h^k, $\hat{P}_h^k$,$\hat{r}_h^k$ and θ_h^k to denote the values of $\overline{Q}_h$,$\overline{V}_h$,$\underline{Q}_h$,$\underline{V}_h$, $\max\{N_h, 1\}$, $\hat{P}_h$, r_h and θ_h at the beginning of the k-th episode in Algorithm 1.

C.1 Proof of monotonicity

C.1.1 Proof of Lemma 1

When $N_h^k(s,a) \leq 1$, (8), (9) and (7) hold trivially by the bound of the rewards and value functions.

For every $h \in [H]$ the empiric Bernstein inequality combined with a union bound argument, to take into account that $N_h^k(s,a) > 1$ is a random number, leads to the following inequality w.p. $1 - SAH\delta$ (see Theorem 4 in (Maurer & Pontil, 2009))

$$\left|(\hat{P}_h^k - P_h)V_{h+1}^*(s,a)\right| \leq \sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}V_{h+1}^*(s,a)\iota}{N_h^k(s,a)}} + \frac{7H\iota}{3(N_h^k(s,a))}, \tag{27}$$

and

$$\left|(\hat{P}_h^k - P_h)V_{h+1}^{\overline{\pi}^k}(s,a)\right| \leq \sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}V_{h+1}^{\overline{\pi}^k}(s,a)\iota}{N_h^k(s,a)}} + \frac{7H\iota}{3(N_h^k(s,a))}. \tag{28}$$

Similarly, with Azuma's inequality, w.p. $1 - SAH\delta$

$$\left|\hat{r}_h^k(s,a) - R_h(s,a)\right| \leq \sqrt{\frac{2Var(r_h^k(s,a))\iota}{N_h^k(s,a)}} + \frac{7\iota}{3(N_h^k(s,a))} \leq \sqrt{\frac{2\hat{r}_h^k(s,a)\iota}{N_h^k(s,a)}} + \frac{7\iota}{3(N_h^k(s,a))}, \tag{29}$$

where $Var(r_h^k(s,a))$ is the empirical variance of $R_h(s,a)$ computed by the $N_h^k(s,a)$ samples and $Var(r_h^k(s,a)) \leq \hat{r}_h^k(s,a)$.

C.1.2 Proof of Lemma 2

We first prove that $\overline{Q}_h^k(s,a) \geq Q_h^*(s,a)$ for all $(s,a,h,k) \in S \times A \times [H] \times [K]$, by backward induction conditioned on the event $\mathcal{E}^R \cap \mathcal{E}^{PV}$. Firstly, the conclusion holds for $h = H+1$ because $\overline{V}_{H+1}(s) = \underline{V}_{H+1}(s) = 0$ and $\overline{Q}_{H+1}(s,a) = \underline{Q}_{H+1}(s,a) = 0$ for all s and a. For $h \in [H]$,

assuming the conclusion holds for $h+1$, by Algorithm 1, we have

$$\begin{aligned}
&\hat{r}_h^k(s,a) + \hat{P}_h^k \overline{V}_{h+1}(s,a) + \theta_h^k(s,a) - Q_h^*(s,a) \\
=&\hat{r}_h^k(s,a) + \hat{P}_h^k \overline{V}_{h+1}(s,a) + \theta_h^k(s,a) - R_h(s,a) - P_h V_{h+1}^*(s,a) \\
=&\hat{r}_h^k(s,a) - R_h(s,a) + \hat{P}_h^k \left(\overline{V}_{h+1} - V_{h+1}^*\right)(s,a) + (\hat{P}_h^k - P_h)V_{h+1}^*(s,a) + \theta_h^k(s,a) \\
\geq&(\hat{P}_h^k - P_h)V_{h+1}^*(s,a) + \sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s,a)\iota}{N_h^k(s,a)}} + \frac{\hat{P}_h^k\left(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k\right)(s,a)}{H} + \frac{8H^2\iota}{N_h^k(s,a)} \\
\geq&\sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s,a)\iota}{N_h^k(s,a)}} + \frac{\hat{P}_h^k\left(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k\right)(s,a)}{H} + \frac{8H^2\iota}{N_h^k(s,a)} - \sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}V_{h+1}^*(s,a)\iota}{N_h^k(s,a)}},
\end{aligned} \tag{30}$$

where the first inequality comes from event $\mathcal{E}^R$, $\overline{V}_{h+1}(s) \geq V_{h+1}^*(s)$ and the definition of $\theta_h^k(s,a)$ and the last inequality from event $\mathcal{E}^{PV}$. By the relation of V-values in the step $(h+1)$,

$$\begin{aligned}
&\left|\mathbb{V}_{\hat{P}_h^k}\left(\frac{\overline{V}_{h+1}^k + \underline{V}_{h+1}^k}{2}\right)(s,a) - \mathbb{V}_{\hat{P}_h^k}V_{h+1}^*(s,a)\right| \\
\leq&\left|[\hat{P}_h^k(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2]^2 - (\hat{P}_h^k V_{h+1}^*)^2\right|(s,a) + \left|\hat{P}_h^k[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2]^2 - \hat{P}_h^k(V_{h+1}^*)^2\right|(s,a) \\
\leq&4H\hat{P}_h^k\left|(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2 - V_{h+1}^*\right|(s,a) \\
\leq&2H\hat{P}_h^k\left(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k\right)(s,a)
\end{aligned} \tag{31}$$

and

$$\begin{aligned}
&\sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}V_{h+1}^*(s,a)\iota}{N_h^k(s,a)}} \\
\leq&\sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s,a)\iota + 4H\hat{P}_h^k\left(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k\right)(s,a)\iota}{N_h^k(s,a)}} \\
\leq&\sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s,a)\iota}{N_h^k(s,a)}} + \sqrt{\frac{4H\hat{P}_h^k\left(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k\right)(s,a)\iota}{N_h^k(s,a)}} \\
\leq&\sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s,a)\iota}{N_h^k(s,a)}} + \frac{\hat{P}_h^k\left(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k\right)(s,a)}{H} + \frac{8H^2\iota}{N_h^k(s,a)}.
\end{aligned} \tag{32}$$

Plugging (32) back into (30), we have $\hat{r}_h^k(s,a) + \hat{P}_h^k\overline{V}_{h+1}(s,a) + \theta_h^k(s,a) \geq Q_h^*(s,a)$. Thus, $\overline{Q}_h^k(s,a) = \min\{H-h+1, \hat{r}_h^k(s,a) + \hat{P}_h^k\overline{V}_{h+1}^k(s,a) + \theta_h^k(s,a)\} \geq Q_h^*(s,a)$.

From the definition of $\overline{V}_h^k(s)$ and $\overline{\pi}_h^k$, we have

$$\begin{aligned}
\overline{V}_h^k(s) =&(1-\rho)\overline{Q}_h^k(s,\overline{\pi}_h^k(s)) + \rho\overline{Q}_h^k(s,\underline{\pi}_h^k(s)) \\
\geq&(1-\rho)\overline{Q}_h^k(s,\pi_h^*(s)) + \rho Q_h^*(s,\underline{\pi}_h^k(s)) \\
\geq&(1-\rho)Q_h^*(s,\pi_h^*(s)) + \rho\min_{a\in\mathcal{A}}Q_h^*(s,a) = V_h^*(s).
\end{aligned} \tag{33}$$

Similarly, we can prove that $\underline{Q}_h^k(s,a) \leq Q_h^{\overline{\pi}^k}(s,a)$ and $\underline{V}_h^k(s) \leq V_h^{\overline{\pi}^k}(s)$.

$$
\begin{aligned}
&\hat{r}_h^k(s,a) + \hat{P}_h^k \underline{V}_{h+1}(s,a) - \theta_h^k(s,a) - Q_h^{\overline{\pi}^k}(s,a) \\
=&\hat{r}_h^k(s,a) + \hat{P}_h^k \underline{V}_{h+1}(s,a) - \theta_h^k(s,a) - R_h(s,a) - P_h V_{h+1}^{\overline{\pi}^k}(s,a) \\
=&\hat{r}_h^k(s,a) - R_h(s,a) + \hat{P}_h^k \left(\underline{V}_{h+1} - V_{h+1}^{\overline{\pi}^k}\right)(s,a) + (\hat{P}_h^k - P_h) V_{h+1}^{\overline{\pi}^k}(s,a) - \theta_h^k(s,a) \\
\leq&(\hat{P}_h^k - P_h) V_{h+1}^{\overline{\pi}^k}(s,a) - \sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s,a)\iota}{N_h^k(s,a)}} \\
&- \frac{\hat{P}_h^k \left(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k\right)(s,a)}{H} - \frac{8H^2\iota}{N_h^k(s,a)} \\
\leq&\sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k} V_{h+1}^{\overline{\pi}^k}(s,a)\iota}{N_h^k(s,a)}} - \sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s,a)\iota}{N_h^k(s,a)}} \\
&- \frac{\hat{P}_h^k \left(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k\right)(s,a)}{H} - \frac{8H^2\iota}{N_h^k(s,a)} \leq 0,
\end{aligned} \tag{34}
$$

and

$$
\begin{aligned}
\underline{V}_h^k(s) =&(1-\rho)\underline{Q}_h^k(s, \overline{\pi}_h^k(s)) + \rho \underline{Q}_h^k(s, \underline{\pi}_h^k(s)) \\
\leq&(1-\rho) Q_h^{\overline{\pi}^k}(s, \overline{\pi}_h^k(s)) + \rho \min_{a \in \mathcal{A}} \underline{Q}_h^k(s,a) \\
\leq&(1-\rho) Q_h^{\overline{\pi}^k}(s, \overline{\pi}_h^k(s)) + \rho \underline{Q}_h^k(s, \arg\min_{a \in \mathcal{A}} Q_h^{\overline{\pi}^k}(s,a)) \\
\leq&(1-\rho) Q_h^{\overline{\pi}^k}(s, \overline{\pi}_h^k(s)) + \rho \min_{a \in \mathcal{A}} Q_h^{\overline{\pi}^k}(s,a) = V_h^{\overline{\pi}^k}(s).
\end{aligned} \tag{35}
$$

C.2 Regret Analysis

C.2.1 Proof of Lemma 3

We consider the event $\mathcal{E}^R \cap \mathcal{E}^{PV}$. The following analysis will be done assuming the successful event $\mathcal{E}^R \cap \mathcal{E}^{PV}$ holds. By Lemma 2, the regret can be bounded by $Regret(K) := \sum_{k=1}^K (V_1^*(s_1^k) - V_1^{\overline{\pi}^k}(s_1^k)) \leq \sum_{k=1}^K (\overline{V}_1^k(s_1^k) - \underline{V}_1^k(s_1^k))$.

By the update steps in Algorithm 1, we have

$$
\begin{aligned}
&\overline{V}_h^k(s_h^k) - \underline{V}_h^k(s_h^k) \\
=&(1-\rho)\overline{Q}_h^k(s_h^k, \overline{\pi}_h^k(s_h^k)) + \rho\overline{Q}_h^k(s_h^k, \underline{\pi}_h^k(s_h^k)) - (1-\rho)\underline{Q}_h^k(s_h^k, \overline{\pi}_h^k(s_h^k)) - \rho\underline{Q}_h^k(s_h^k, \underline{\pi}_h^k(s_h^k)) \\
\leq&[\mathbb{D}_{\widetilde{\pi}_h^k}\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k) + 2\mathbb{D}_{\widetilde{\pi}_h^k}\theta_h(s_h^k) \\
=&[\mathbb{D}_{\widetilde{\pi}_h^k}\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k) - [\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k, a_h^k) + 2\mathbb{D}_{\widetilde{\pi}_h^k}\theta_h(s_h^k) \\
&+ [\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k, a_h^k) \\
=&[\mathbb{D}_{\widetilde{\pi}_h^k}\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k) - [\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k, a_h^k) + 2\mathbb{D}_{\widetilde{\pi}_h^k}\theta_h(s_h^k) \\
&+ [\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k, a_h^k) - c_1 P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k) \\
&+ c_1 P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k) - c_2(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k) + c_2(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k) \\
=&[\mathbb{D}_{\widetilde{\pi}_h^k}\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k) - [\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k, a_h^k) \\
&+ [\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k, a_h^k) - c_1 P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k) \\
&+ c_1 P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k) - c_2(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k) + c_2(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k) \\
&+ 2(1-\rho)\sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s_h^k, \overline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k, \overline{\pi}_h^k(s_h^k))}} + 2(1-\rho)\sqrt{\frac{2\hat{r}_h^k(s_h^k, \overline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k, \overline{\pi}_h^k(s_h^k))}} \\
&+ (1-\rho)\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, \overline{\pi}_h^k(s_h^k))/H + \frac{2(1-\rho)(24H^2+7H+7)\iota}{3N_h^k(s_h^k, \overline{\pi}_h^k(s_h^k)))} \\
&+ 2\rho\sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s_h^k, \underline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k, \underline{\pi}_h^k(s_h^k))}} + 2\rho\sqrt{\frac{2\hat{r}_h^k(s_h^k, \underline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k, \underline{\pi}_h^k(s_h^k))}} \\
&+ \rho\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, \underline{\pi}_h^k(s_h^k))/H + \frac{2\rho(24H^2+7H+7)\iota}{3N_h^k(s_h^k, \underline{\pi}_h^k(s_h^k)))} \\
=&(1+1/H)[\mathbb{D}_{\widetilde{\pi}_h^k}\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k) - (1+1/H)[\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k, a_h^k) \\
&+ \underbrace{(1+1/H)[\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k, a_h^k) - c_1 P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k)}_{(a)} \\
&+ c_1 P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k) - c_2(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k) + c_2(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k) \\
&+ 2(1-\rho)\underbrace{\sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s_h^k, \overline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k, \overline{\pi}_h^k(s_h^k))}}_{(b1)} + 2(1-\rho)\sqrt{\frac{2\hat{r}_h^k(s_h^k, \overline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k, \overline{\pi}_h^k(s_h^k))}} \\
&+ \frac{2(1-\rho)(24H^2+7H+7)\iota}{3N_h^k(s_h^k, \overline{\pi}_h^k(s_h^k)))} + 2\rho\underbrace{\sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s_h^k, \underline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k, \underline{\pi}_h^k(s_h^k))}}_{(b2)} \\
&+ 2\rho\sqrt{\frac{2\hat{r}_h^k(s_h^k, \underline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k, \underline{\pi}_h^k(s_h^k))}} + \frac{2\rho(24H^2+7H+7)\iota}{3N_h^k(s_h^k, \underline{\pi}_h^k(s_h^k)))}.
\end{aligned}
\tag{36}
$$

Bound of the error of the empirical probability estimator (a) By Bennett's inequality, we have that w.p. $1 - S\delta$

$$
|\hat{P}_h^k(s'|s,a) - P_h(s'|s,a)| \leq \sqrt{\frac{2P_h(s'|s,a)\iota}{N_h^k(s,a)}} + \frac{\iota}{3N_h^k(s,a)} \tag{37}
$$

holds for all s, a, h, k, s'.

Thus, we have that

$$\begin{aligned}
&(\hat{P}_h^k - P_h)(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s,a) \\
=&\sum_{s'}(\hat{P}_h^k(s'|s,a) - P_h(s'|s,a))(\overline{V}_{h+1}^k(s') - \underline{V}_{h+1}^k(s')) \\
\leq&\sum_{s'}\sqrt{\frac{2P_h(s'|s,a)\iota}{N_h^k(s,a)}}(\overline{V}_{h+1}^k(s') - \underline{V}_{h+1}^k(s')) + \frac{SH\iota}{3N_h^k(s,a)} \\
\leq&\sum_{s'}\left(\frac{P_h(s'|s,a)\iota}{H} + \frac{H}{2N_h^k(s,a)}\right)\left(\overline{V}_{h+1}^k(s') - \underline{V}_{h+1}^k(s')\right) + \frac{SH\iota}{3N_h^k(s,a)} \\
\leq&P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s,a)/H + \frac{SH^2}{2N_h^k(s,a)} + \frac{SH\iota}{3N_h^k(s,a)} \\
\leq&P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s,a)/H + \frac{SH^2\iota}{N_h^k(s,a)},
\end{aligned} \tag{38}$$

where the second inequality is due to AM-GM inequality.

Bound of the error of the empirical variance estimator (b1) & (b2) Here, we bound $\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s_h^k, a_h^k)$.

Recall that $C_h^{\pi,\pi',\rho}(s) = \mathbb{E}\left[\sum_{h'=h}^H R_{h'}(s_{h'}, a_{h'})|s_h = s, a_{h'} \sim \widetilde{\pi}_{h'}(\cdot|s_{h'})\right]$ in Appendix B. Set π^{k*} here is the optimal policy towards the adversary policy $\underline{\pi}^k$ with $\pi_h^{k*}(s) = \arg\max_\pi C_h^{\pi,\underline{\pi}^k,\rho}(s)$. Similar to the proof in Appendix C.1.2, we can show that $\overline{V}_h^k(s) \geq C_h^{\pi^{k*},\underline{\pi}^k,\rho}(s)$. We also have that $C_h^{\pi^{k*},\underline{\pi}^k,\rho}(s) = \max_\pi C_h^{\pi,\underline{\pi}^k,\rho}(s) \geq C_h^{\overline{\pi}^k,\underline{\pi}^k,\rho}(s) \geq V_h^{\overline{\pi}^k}(s) \geq \underline{V}_h^k(s)$. For any $(s,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$, under event $\mathcal{E}^R \cap \mathcal{E}^{PV}$,

$$\begin{aligned}
&\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s,a) - \mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s,a) \\
=&\hat{P}_h^k[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2]^2(s,a) - [\hat{P}_h^k(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2]^2(s,a) \\
&- P_h(C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho})^2(s,a) + (P_hC_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho})^2(s,a) \\
\leq&[\hat{P}_h^k(\overline{V}_{h+1}^k)^2 - (\hat{P}_h^k\underline{V}_{h+1}^k)^2 - P_h(\underline{V}_{h+1}^k)^2 + (P_h\overline{V}_{h+1}^k)^2](s,a) \\
\leq&|(\hat{P}_h^k - P_h)(\overline{V}_{h+1}^k)^2|(s,a) + |(P_h\underline{V}_{h+1}^k)^2 - (\hat{P}_h^k\underline{V}_{h+1}^k)^2|(s,a) \\
&+ P_h|(\overline{V}_{h+1}^k)^2 - (\underline{V}_{h+1}^k)^2|(s,a) + |(P_h\overline{V}_{h+1}^k)^2 - (P_h\underline{V}_{h+1}^k)^2|(s,a),
\end{aligned} \tag{39}$$

where the first inequality is due $\overline{V}_h^k(s) \geq C_h^{\pi^{k*},\underline{\pi}^k,\rho}(s) \geq \underline{V}_h^k(s)$. The result of (Weissman et al., 2003) combined with a union bound on $N_h^k(s,a) \in [K]$ implies w.p $1-\delta$

$$\|\hat{P}_h^k(\cdot|s,a) - P_h(\cdot|s,a)\|_1 \leq \sqrt{\frac{2S\iota}{N_h^k(s,a)}} \tag{40}$$

holds for all s, a, h, k.

These terms can be bounded separately by

$$\begin{aligned}
&|(\hat{P}_h^k - P_h)(\overline{V}_{h+1}^k)^2|(s,a) \leq H^2\sqrt{\frac{2S\iota}{N_h^k(s,a)}}, \\
&|(P_h\underline{V}_{h+1}^k)^2 - (\hat{P}_h^k\underline{V}_{h+1}^k)^2|(s,a) \leq 2H|(P_h - \hat{P}_h^k)\underline{V}_{h+1}^k| \leq 2H^2\sqrt{\frac{2S\iota}{N_h^k(s,a)}}, \\
&P_h|(\overline{V}_{h+1}^k)^2 - (\underline{V}_{h+1}^k)^2|(s,a) \leq 2HP_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s,a), \\
&|(P_h\overline{V}_{h+1}^k)^2 - (P_h\underline{V}_{h+1}^k)^2|(s,a) \leq 2HP_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s,a),
\end{aligned} \tag{41}$$

where the first two inequality is due to (40). In addition, $3H^2\sqrt{\frac{2S\iota}{N_h^k(s,a)}} \le 1 + \frac{9SH^4\iota}{2N_h^k(s,a)}$. Thus, we have

$$
\begin{aligned}
&(1-\rho)\sqrt{\frac{\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s_h^k,\overline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))}} + \rho\sqrt{\frac{\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s_h^k,\underline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k,\underline{\pi}_h^k(s_h^k))}} \\
\le&(1-\rho)\sqrt{\frac{\mathbb{V}_{P_h} C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\overline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))}} + \rho\sqrt{\frac{\mathbb{V}_{P_h} C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\underline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k,\underline{\pi}_h^k(s_h^k))}} \\
&+ (1-\rho)\sqrt{\frac{4HP_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k,\overline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))}} + \rho\sqrt{\frac{4HP_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k,\underline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k,\underline{\pi}_h^k(s_h^k))}} \\
&+ (1-\rho)\sqrt{\frac{1}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))}} + \rho\sqrt{\frac{1}{N_h^k(s_h^k,\underline{\pi}_h^k(s_h^k))}} + \frac{(1-\rho)\sqrt{9SH^4\iota/2}}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))} + \frac{\rho\sqrt{9SH^4\iota/2}}{N_h^k(s_h^k,\underline{\pi}_h^k(s_h^k))} \\
\le&(1-\rho)\sqrt{\frac{\mathbb{V}_{P_h} C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\overline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))}} + \rho\sqrt{\frac{\mathbb{V}_{P_h} C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\underline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k,\underline{\pi}_h^k(s_h^k))}} \\
&+ (1-\rho)\left(\frac{P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k,\overline{\pi}_h^k(s_h^k))}{2\sqrt{2}H} + \frac{2\sqrt{2}H^2\iota}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))}\right) \\
&+ \rho\left(\frac{P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k,\underline{\pi}_h^k(s_h^k))}{2\sqrt{2}H} + \frac{2\sqrt{2}H^2\iota}{N_h^k(s_h^k,\underline{\pi}_h^k(s_h^k))}\right) \\
&+ (1-\rho)\sqrt{\frac{1}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))}} + \rho\sqrt{\frac{1}{N_h^k(s_h^k,\underline{\pi}_h^k(s_h^k))}} \\
&+ \frac{(1-\rho)\sqrt{9SH^4\iota/2}}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))} + \frac{\rho\sqrt{9SH^4\iota/2}}{N_h^k(s_h^k,\underline{\pi}_h^k(s_h^k))} \\
=&(1-\rho)\sqrt{\frac{\mathbb{V}_{P_h} C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\overline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))}} + \rho\sqrt{\frac{\mathbb{V}_{P_h} C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\underline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k,\underline{\pi}_h^k(s_h^k))}} \\
&+ \frac{\mathbb{D}_{\overline{\pi}_h^k} P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k)}{2\sqrt{2}H} + \frac{2\sqrt{2}(1-\rho)H^2\iota}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))} + \frac{2\sqrt{2}\rho H^2\iota}{N_h^k(s_h^k,\underline{\pi}_h^k(s_h^k))} \\
&+ (1-\rho)\sqrt{\frac{1}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))}} + \rho\sqrt{\frac{1}{N_h^k(s_h^k,\underline{\pi}_h^k(s_h^k))}} \\
&+ \frac{(1-\rho)\sqrt{9SH^4\iota/2}}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))} + \frac{\rho\sqrt{9SH^4\iota/2}}{N_h^k(s_h^k,\underline{\pi}_h^k(s_h^k))},
\end{aligned}
\tag{42}
$$

where the second inequality is due to AM-GM inequality.

Recursing on h Plugging (38) and (42) into (36)and setting $c_1 = 1 + 1/H$ and $c_2 = (1 + 1/H)^3$, we have

$$
\begin{aligned}
&\overline{V}_h^k(s_h^k) - \underline{V}_h^k(s_h^k) \\
\leq&(1+1/H)[\mathbb{D}_{\widetilde{\pi}_h^k}\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k) - (1+1/H)[\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k, a_h^k) \\
&+ (1/H + 1/H^2)P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k) + \frac{(SH + SH^2)\iota}{N_h^k(s_h^k, a_h^k)} \\
&+ c_1 P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k) - c_2(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k) + c_2(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k) \\
&+ 2(1-\rho)\sqrt{\frac{2\hat{r}_h^k(s_h^k, \overline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k, \overline{\pi}_h^k(s_h^k))}} + \frac{2(1-\rho)(24H^2 + 7H + 7)\iota}{3N_h^k(s_h^k, \overline{\pi}_h^k(s_h^k)))} \\
&+ 2\rho\sqrt{\frac{2\hat{r}_h^k(s_h^k, \underline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k, \underline{\pi}_h^k(s_h^k))}} + \frac{2\rho(24H^2 + 7H + 7)\iota}{3N_h^k(s_h^k, \underline{\pi}_h^k(s_h^k)))} \\
&+ (1-\rho)\sqrt{\frac{8\mathbb{V}_{P_h} C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}(s_h^k, \overline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k, \overline{\pi}_h^k(s_h^k))}} + \rho\sqrt{\frac{8\mathbb{V}_{P_h} C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}(s_h^k, \underline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k, \underline{\pi}_h^k(s_h^k))}} \\
&+ \frac{\mathbb{D}_{\widetilde{\pi}_h^k} P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k)}{H} + \frac{8(1-\rho)H^2\iota}{N_h^k(s_h^k, \overline{\pi}_h^k(s_h^k))} + \frac{8\rho H^2\iota}{N_h^k(s_h^k, \underline{\pi}_h^k(s_h^k))} \\
&+ (1-\rho)\sqrt{\frac{8}{N_h^k(s_h^k, \overline{\pi}_h^k(s_h^k))}} + \rho\sqrt{\frac{8}{N_h^k(s_h^k, \underline{\pi}_h^k(s_h^k))}} + \frac{6(1-\rho)\sqrt{SH^4\iota}}{N_h^k(s_h^k, \overline{\pi}_h^k(s_h^k))} + \frac{6\rho\sqrt{SH^4\iota}}{N_h^k(s_h^k, \underline{\pi}_h^k(s_h^k))}.
\end{aligned} \tag{43}
$$

We set $\Theta_h^k(s,a) = \sqrt{\frac{8\mathbb{V}_{P_h} C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}(s,a)\iota}{N_h^k(s,a)}} + \sqrt{\frac{32}{N_h^k(s,a)}} + \frac{46\sqrt{SH^4\iota}}{N_h^k(s,a)}$. Since $r_h^k(s,a) \leq 1$, by organizing the items, we have that

$$
\begin{aligned}
&\overline{V}_h^k(s_h^k) - \underline{V}_h^k(s_h^k) \\
\leq&(1+1/H)[\mathbb{D}_{\widetilde{\pi}_h^k}\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k) - (1+1/H)[\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k, a_h^k) \\
&+ (1/H + 1/H^2)P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k) + \frac{(SH + SH^2)\iota}{N_h^k(s_h^k, a_h^k)} \\
&+ c_1 P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k) - c_2(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k) + c_2(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k) \\
&+ \frac{\mathbb{D}_{\widetilde{\pi}_h^k} P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, \overline{\pi}_h^k(s_h^k))}{H} + \mathbb{D}_{\widetilde{\pi}_h^k}\Theta_h^k(s_h^k) \\
\leq&(1+1/H)[\mathbb{D}_{\widetilde{\pi}_h^k}\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k) - (1+1/H)[\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k, a_h^k) \\
&+ \frac{1}{H}[\mathbb{D}_{\widetilde{\pi}_h^k} P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k) - P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k)] \\
&+ (1 + 3/H + 1/H^2)P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k) - c_2(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k) \\
&+ c_2(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k) + \frac{(SH + SH^2)\iota}{N_h^k(s_h^k, a_h^k)} + \mathbb{D}_{\widetilde{\pi}_h^k}\Theta_h^k(s_h^k) \\
\leq&(1+1/H)[\mathbb{D}_{\widetilde{\pi}_h^k}\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k) - (1+1/H)[\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k, a_h^k) \\
&+ \frac{1}{H}[\mathbb{D}_{\widetilde{\pi}_h^k} P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k) - P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k)] \\
&+ c_2 P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k) - c_2(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k) \\
&+ c_2(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k) + \frac{(SH + SH^2)\iota}{N_h^k(s_h^k, a_h^k)} + \mathbb{D}_{\widetilde{\pi}_h^k}\Theta_h^k(s_h^k).
\end{aligned} \tag{44}
$$

By induction of (36) on $h = 1, \cdots, H$ and $\overline{V}^k_{h+1} = \underline{V}^k_{h+1} = 0$, we have that

$$\begin{aligned} Regret(K) \leq 21 \sum_{k=1}^{K} \sum_{h=1}^{H} (&\mathbb{D}_{\widetilde{\pi}^k_h} \hat{P}^k_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h) - \hat{P}^k_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h, a^k_h) \\ &+ \frac{1}{H} [\mathbb{D}_{\widetilde{\pi}^k_h} P_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h) - P_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h, a^k_h)] \\ &+ P_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h, a^k_h) - (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_{h+1}) \\ &+ \frac{(SH + SH^2)\iota}{N^k_h (s^k_h, a^k_h)} + \mathbb{D}_{\widetilde{\pi}^k_h} \Theta^k_h (s^k_h)). \end{aligned} \tag{45}$$

Here we use $(1 + 1/H)^{3H} < 21$.

C.2.2 Proof of Lemma 4

Recall that $M_1 = \sum_{k=1}^{K} \sum_{h=1}^{H} [\mathbb{D}_{\widetilde{\pi}^k_h} \hat{P}^k_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h) - \hat{P}^k_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h, a^k_h)]$.

Since $\mathbb{E}_{a^k_h \sim \mathbb{D}_{\widetilde{\pi}^k_h}} [\hat{P}^k_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h, a^k_h)] = \mathbb{D}_{\widetilde{\pi}^k_h} \hat{P}^k_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h)$, we have that $\mathbb{D}_{\widetilde{\pi}^k_h} \hat{P}^k_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h) - \hat{P}^k_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h, a^k_h)$ is a martingale difference sequence. By the Azuma-Hoeffding inequality, with probability $1 - \delta$, we have

$$\left| \sum_{k=1}^{K} \sum_{h=1}^{H} [\mathbb{D}_{\widetilde{\pi}^k_h} \hat{P}^k_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h) - \hat{P}^k_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h, a^k_h)] \right| \leq H \sqrt{2HK\iota}. \tag{46}$$

C.2.3 Proof of Lemma 5

Recall that $M_2 = \sum_{k=1}^{K} \sum_{h=1}^{H} \frac{1}{H} [\mathbb{D}_{\widetilde{\pi}^k_h} P_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h) - P_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h, a^k_h)]$.

Since $\mathbb{E}_{a^k_h \sim \mathbb{D}_{\widetilde{\pi}^k_h}} [P_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h, a^k_h)] = \mathbb{D}_{\widetilde{\pi}^k_h} P_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h)$, we have that $\mathbb{D}_{\widetilde{\pi}^k_h} P_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h) - P_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h, a^k_h)$ is a martingale difference sequence. By the Azuma-Hoeffding inequality, with probability $1 - \delta$, we have

$$\left| \sum_{k=1}^{K} \sum_{h=1}^{H} [\mathbb{D}_{\widetilde{\pi}^k_h} P_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h) - P_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h, a^k_h)] \right| \leq H \sqrt{2HK\iota}. \tag{47}$$

C.2.4 Proof of Lemma 6

Recall that $M_3 = \sum_{k=1}^{K} \sum_{h=1}^{H} (P^k_h (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h, a^k_h) - (\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_{h+1}))$.

Let the one-hot vector $\hat{\mathbb{1}}^k_h (\cdot | s^k_h, a^k_h)$ to satisfy that $\hat{\mathbb{1}}^k_h (s^k_{h+1} | s^k_h, a^k_h) = 1$ and $\hat{\mathbb{1}}^k_h (s | s^k_h, a^k_h) = 0$ for $s \neq s^k_{h+1}$. Thus, $[(P^k_h - \hat{\mathbb{1}}^k_h)(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})](s^k_h, a^k_h)$ is a martingale difference sequence. By the Azuma-Hoeffding inequality, with probability $1 - \delta$, we have

$$\left| \sum_{k=1}^{K} \sum_{h=1}^{H} [(P^k_h - \hat{\mathbb{1}}^k_h)(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})](s^k_h, a^k_h) \right| \leq H \sqrt{2HK\iota}. \tag{48}$$

C.2.5 Proof of Lemma 7

We bounded $M_4 = \sum_{k=1}^{K} \sum_{h=1}^{H} [\frac{(SH+SH^2)\iota}{N^k_h (s^k_h, a^k_h)} + \mathbb{D}_{\widetilde{\pi}^k_h} \Theta^k_h (s^k_h)]$ by separately bounding the four items.

Bound $\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{(SH+SH^2)\iota}{N_h^k(s_h^k,a_h^k)}$ We regroup the summands in a different way.

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{(SH+SH^2)\iota}{N_h^k(s_h^k,a_h^k)}=(SH+SH^2)\iota\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{n=1}^{N_h^K(s,a)}\frac{1}{n}\leq(SH+SH^2)SAH\iota^2. \tag{49}$$

Recall that $\Theta_h^k(s,a)=\sqrt{\frac{8\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s,a)\iota}{N_h^k(s,a)}}+\sqrt{\frac{32}{N_h^k(s,a)}}+\frac{46\sqrt{SH^4\iota}}{N_h^k(s,a)}$.

Bound $\sum_{k=1}^{K}\sum_{h=1}^{H}[(1-\rho)\sqrt{\frac{32\iota}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))}}+\rho\sqrt{\frac{32\iota}{N_h^k(s_h^k,\underline{\pi}_h^k(s_h^k))}}]$ We regroup the summands in a different way. For any policy π, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{32\iota}{N_h^k(s_h^k,\pi(s_h^k))}}=\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{n=1}^{N_h^K(s,a)}\sqrt{\frac{32\iota}{n}}\leq 8H\sqrt{SAK\iota}. \tag{50}$$

Bound $\sum_{k=1}^{K}\sum_{h=1}^{H}[(1-\rho)\frac{46SH^2\iota}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))}+\rho\frac{46SH^2\iota}{N_h^k(s_h^k,\underline{\pi}_h^k(s_h^k))}]$ We regroup the summands in a different way. For any policy π, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{46\sqrt{SH^4\iota}}{N_h^k(s_h^k,\pi(s_h^k))}=46\sqrt{SH^4\iota}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{n=1}^{N_h^K(s,a)}\frac{1}{n}\leq 46S^{\frac{3}{2}}AH^3\iota^2. \tag{51}$$

Bound $\sum_{k=1}^{K}\sum_{h=1}^{H}\left[(1-\rho)\sqrt{\frac{8\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\overline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))}}+\rho\sqrt{\frac{8\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\underline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k,\underline{\pi}_h^k(s_h^k))}}\right]$ By Cauchy-Schwarz inequality,

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\overline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))}}\\ \leq&\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\overline{\pi}_h^k(s_h^k))\cdot\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{\iota}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))}}\\ \leq&\sqrt{SAH\iota^2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\overline{\pi}_h^k(s_h^k))}.\end{aligned} \tag{52}$$

Similarly,

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\underline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k,\underline{\pi}_h^k(s_h^k))}}\\ \leq&\sqrt{SAH\iota^2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\underline{\pi}_h^k(s_h^k))}.\end{aligned} \tag{53}$$

By $(1-\rho)a^2+\rho b^2\geq((1-\rho)a+\rho b)^2$,

$$\begin{aligned}&(1-\rho)\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\overline{\pi}_h^k(s_h^k))}+\rho\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\underline{\pi}_h^k(s_h^k))}\\ \leq&\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}[(1-\rho)\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\overline{\pi}_h^k(s_h^k))+\rho\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\underline{\pi}_h^k(s_h^k))]}.\end{aligned} \tag{54}$$

Now we bound the total variance. Let $\mathbb{D}_{\widetilde{\pi}_h^k} P_h(s'|s) = (1-\rho)P_h(s'|s, \overline{\pi}_h^k(s)) + \rho P_h(s'|s, \underline{\pi}_h^k(s))$,

$$[\mathbb{D}_{\widetilde{\pi}_h^k} P_h V_{h+1}](s) = \sum_{s'} [(1-\rho)P_h(s'|s, \overline{\pi}_h^k(s)) + \rho P_h(s'|s, \underline{\pi}_h^k(s))] V_{h+1}(s'), \tag{55}$$

and

$$\begin{aligned} \mathbb{V}_{[\mathbb{D}_{\widetilde{\pi}_h^k} P_h]} V_{h+1}(s) = &\sum_{s'} [(1-\rho)P_h(s'|s, \overline{\pi}_h^k(s)) + \rho P_h(s'|s, \underline{\pi}_h^k(s))][V_{h+1}(s')]^2 \\ &- [\sum_{s'} \left((1-\rho)P_h(s'|s, \overline{\pi}_h^k(s)) + \rho P_h(s'|s, \underline{\pi}_h^k(s))\right) V_{h+1}(s')]^2. \end{aligned} \tag{56}$$

We have that

$$\begin{aligned} &\mathbb{V}_{[\mathbb{D}_{\widetilde{\pi}_h^k} P_h]} C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}(s_h^k) \\ =&\sum_{s'} [(1-\rho)P_h(s'|s_h^k, \overline{\pi}_h^k(s_h^k)) + \rho P_h(s'|s_h^k, \underline{\pi}_h^k(s_h^k))][C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}(s')]^2 \\ &- [\sum_{s'} \left((1-\rho)P_h(s'|s_h^k, \overline{\pi}_h^k(s_h^k)) + \rho P_h(s'|s_h^k, \underline{\pi}_h^k(s_h^k))\right) C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}(s')]^2 \\ \geq&(1-\rho)\mathbb{V}_{P_h} C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}(s_h^k, \overline{\pi}_h^k(s_h^k)) + \rho \mathbb{V}_{P_h} C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}(s_h^k, \underline{\pi}_h^k(s_h^k)) \\ &+ (1-\rho)[P_h C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}(s_h^k, \overline{\pi}_h^k(s_h^k))]^2 + \rho P_h[C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}(s_h^k, \underline{\pi}_h^k(s_h^k))]^2 \\ &- [\sum_{s'} (1-\rho)P_h(s'|s_h^k, \overline{\pi}_h^k(s_h^k)) C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}(s') + \rho P_h(s'|s_h^k, \underline{\pi}_h^k(s_h^k)) C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}(s')]^2 \\ \geq&(1-\rho)\mathbb{V}_{P_h} C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}(s_h^k, \overline{\pi}_h^k(s_h^k)) + \rho \mathbb{V}_{P_h} C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}(s_h^k, \underline{\pi}_h^k(s_h^k)), \end{aligned} \tag{57}$$

where the last inequality is due to $(1-\rho)a^2 + \rho b^2 \geq ((1-\rho)a + \rho b)^2$.

With probability $1-2\delta$, we also have that

$$\begin{aligned} &\sum_{k=1}^K \sum_{h=1}^H \mathbb{V}_{[\mathbb{D}_{\widetilde{\pi}_h^k} P_h]} C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}(s_h^k) \\ =&\sum_{k=1}^K \sum_{h=1}^H \left([\mathbb{D}_{\widetilde{\pi}_h^k} P_h (C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho})^2](s_h^k) - \left([\mathbb{D}_{\widetilde{\pi}_h^k} P_h C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}](s_h^k) \right)^2 \right) \\ =&\sum_{k=1}^K \sum_{h=1}^H \left([\mathbb{D}_{\widetilde{\pi}_h^k} P_h (C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho})^2](s_h^k) - \left(C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}(s_{h+1}^k) \right)^2 \right) \\ &+ \sum_{k=1}^K \sum_{h=1}^H \left(\left(C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}(s_{h+1}^k) \right)^2 - \left([\mathbb{D}_{\widetilde{\pi}_h^k} P_h C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}](s_h^k) \right)^2 \right) \\ \leq&H^2 \sqrt{2HK\iota} + \sum_{k=1}^K \sum_{h=1}^H \left((C_h^{\pi^{k*}, \underline{\pi}^k, \rho}(s_h^k))^2 - \left([\mathbb{D}_{\widetilde{\pi}_h^k} P_h C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}](s_h^k) \right)^2 \right) - \sum_{k=1}^K (C_1^{\pi^{k*}, \underline{\pi}^k, \rho}(s_1^k))^2 \\ \leq&H^2 \sqrt{2HK\iota} + 2H \sum_{k=1}^K \sum_{h=1}^H |C_h^{\pi^{k*}, \underline{\pi}^k, \rho}(s_h^k) - \mathbb{D}_{\widetilde{\pi}_h^k} P_h C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}(s_h^k)| \\ \leq&H^2 \sqrt{2HK\iota} + 2H \sum_{k=1}^K \left(C_1^{\pi^{k*}, \underline{\pi}^k, \rho}(s_1^k) + \sum_{h=1}^H \left(C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}(s_{h+1}^k) - \mathbb{D}_{\widetilde{\pi}_h^k} P_h C_{h+1}^{\pi^{k*}, \underline{\pi}^k, \rho}(s_h^k, a_h^k) \right) \right) \\ \leq&H^2 \sqrt{2HK\iota} + 2H^2 K + 2H^2 \sqrt{2HK\iota} \\ \leq&3H^2 K + 9H^3 \iota/2, \end{aligned} \tag{58}$$

where the first inequality holds with probability $1-\delta$ by Azuma-Hoeffding inequality, the second inequality is due to the bound of V-values, the third inequality is due to Lemma 2 so that

$C_h^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k) \geq \mathbb{D}_{\widetilde{\pi}_h^k} D_h^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k) \geq \mathbb{D}_{\widetilde{\pi}_h^k} P_h C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k)$, the fourth inequality holds with probability $1-\delta$ by Azuma-Hoeffding inequality, and the last inequality holds with $2ab \leq a^2+b^2$.

In summary, with probability at least $1-\delta$, we have $\sum_{k=1}^K \sum_{h=1}^H \mathbb{V}_{P_h} V_{h+1}^{\overline{\pi}^k}(s_h^k, a_h^k) \leq (H^2K + H^3\iota)$.

In summary, $\sum_{k=1}^K \sum_{h=1}^H \mathbb{D}_{\widetilde{\pi}_h^k} \Theta_h^k(s_h^k) \leq 8\sqrt{SAH^2K\iota} + 46S^{\frac{3}{2}}AH^3\iota^2 + \sqrt{24SAH^3K\iota^2 + 36SAH^5\iota^2} \leq 8\sqrt{SAH^2K\iota} + 46S^{\frac{3}{2}}AH^3\iota^2 + \sqrt{24SAH^3K\iota} + 6\sqrt{SAH^5\iota}$.

D MODEL-FREE METHOD

In this section, we develop a model-free algorithm and analyze its theoretical guarantee. We present the proposed Action Robust Q-learning with UCB-Hoeffding (AR-UCBH) algorithm show in Algorithm 2. Here, we highlight the main idea of Algorithm 2. Algorithm 2 follows the same idea of Algorithm 1, which trains the agent in a clean (simulation) environment and learns a policy that performs well when applied to a perturbed environment with probabilistic policy execution uncertainty. To simulate the action perturbation, Algorithm 2 chooses an adversarial action with probability ρ. To learn the agent's optimal policy and the corresponding adversarial policy, Algorithm 2 computes an optimistic estimate $\overline{Q}$ of Q^* and a pessimistic estimate $\underline{Q}$ of $Q^{\overline{\pi}^k}$. Algorithm 2 uses the optimistic estimates to explore the possible optimal policy $\overline{\pi}$ and uses the pessimistic estimates to explore the possible adversarial policy $\underline{\pi}$. The difference is that Algorithm 2 use a model-free method to update Q and V values.

Algorithm 2: Action Robust Q-learning with UCB-Hoeffding (AR-UCBH)

Set $\alpha_t = \frac{H+1}{H+t}$. Initialize $\overline{V}_h(s) = H-h+1$, $\overline{Q}_h(s,a) = H-h+1$, $\underline{V}_h(s) = 0$, $\underline{Q}_h(s,a) = 0$, $\hat{r}_h(s,a)$, $N_h(s,a) = 0$ for any state $s \in \mathcal{S}$, any action $a \in \mathcal{A}$ and any step $h \in [H]$. $\overline{V}_{H+1}(s) = \underline{V}_{H+1}(s) = 0$ and $\overline{Q}_{H+1}(s,a) = \underline{Q}_{H+1}(s,a) = 0$ for all s and a. $\Delta = H$. Initial policy $\overline{\pi}_h^1(a|s)$ and $\underline{\pi}_h^1(a|s) = 1/A$ for any state s, action a and any step $h \in [H]$.
for episode $k = 1, 2, \ldots, K$ **do**
  for step $h = 1, 2, \ldots, H$ **do**
    Observe s_h^k.
    Set $\overline{a}_h^k = \arg\max_a \overline{Q}_h(s_h^k, a)$, $\underline{a}_h^k = \arg\min_a \underline{Q}_h(s_h^k, a)$, $\widetilde{\pi}_h^k(\overline{a}_h^k|s_h^k) = 1-\rho$ and $\widetilde{\pi}_h^k(\underline{a}_h^k|s_h^k) = \rho$.
    Take action $a_h^k \sim \widetilde{\pi}_h^k(\cdot|s_h^k)$.
    Receive reward r_h^k and observe s_{h+1}^k.
    Set $t = N_h(s_h^k, a_h^k) \leftarrow N_h(s_h^k, a_h^k) + 1$; $b_t = \sqrt{H^3\iota/t}$.
    $\overline{Q}_h(s_h^k, a_h^k) \leftarrow (1-\alpha_t)\overline{Q}_h(s_h^k, a_h^k) + \alpha_t(r_h^k + \overline{V}_{h+1}(s_{h+1}^k) + b_t)$,
    $\underline{Q}_h(s_h^k, a_h^k) \leftarrow (1-\alpha_t)\underline{Q}_h(s_h^k, a_h^k) + \alpha_t(r_h^k + \underline{V}_{h+1}(s_{h+1}^k) - b_t)$.
    Set $\overline{\pi}_h^{k+1}(s_h^k) = \arg\max_a \overline{Q}_h(s_h^k, a)$, $\underline{\pi}_h^{k+1}(s_h^k) = \arg\min_a \underline{Q}_h(s_h^{k+1}, a)$.
    $\overline{V}_h(s_h^k) \leftarrow \min\{\overline{V}_h(s_h^k), (1-\rho)\overline{Q}_h(s_h^k, \overline{\pi}_h^{k+1}(s_h^k)) + \rho\overline{Q}_h(s_h^k, \underline{\pi}_h^{k+1}(s_h^k))\}$.
    $\underline{V}_h(s_h^k) \leftarrow \max\{\underline{V}_h(s_h^k), (1-\rho)\underline{Q}_h(s_h^k, \overline{\pi}_h^{k+1}(s_h^k)) + \rho\underline{Q}_h(s_h^k, \underline{\pi}_h^{k+1}(s_h^k))\}$.
    if $\underline{V}_h(s_h^k) > (1-\rho)\underline{Q}_h(s_h^k, \overline{\pi}_h^{k+1}(s_h^k)) + \rho\underline{Q}_h(s_h^k, \underline{\pi}_h^{k+1}(s_h^k))$ **then**
      $\overline{\pi}_h^{k+1} = \overline{\pi}_h^k$.
    end if
  end for
  Output policy $\overline{\pi}^{k+1}$ with certificates $\mathcal{I}_{k+1} = [\underline{V}_1(s_1^k), \overline{V}_1(s_1^k)]$ and $\epsilon_{k+1} = |\mathcal{I}_{k+1}|$.
end for
return $\overline{\pi}^{k+1}$

Here, we highlight the challenges of the model-free planning compared with the model-based planing. In the model-based planning, we performs value iteration and the Q values, V values, agent policy $\overline{\pi}$ and adversarial policy $\underline{\pi}$ are updated on all (s,a). However, in the model-free method, the Q values, V values are updated only on (s_h^k, a_h^k) which are the samples on the trajectories. Compared with

the model-based planning, the model-free planning is slower and less stable. We need to update the output policy carefully. In line 14-16, Algorithm 2 does not update the output policy when the lower bound on the value function of the new policy does not improve. By this, the output policies are stably updated.

We provide the regret and sample complexity bounds of Algorithm 2 in the following:

Theorem 2 *For any $\delta \in (0,1]$, letting $\iota = \log(2SABHK/\delta)$, then with probability at least $1-\delta$, Algorithm 2 achieves:*

- $V_1^*(s_1) - V_1^{\pi^{out}}(s_1) \le \epsilon$, *if the number of episodes* $K \ge \Omega(SAH^5\iota/\epsilon^2 + SAH^2/\epsilon)$.
- $Regret(K) = \sum_{k=1}^{K}(V_1^*(s_1^k) - V_1^{\overline{\pi}^k}(s_1^k)) \le \mathcal{O}(\sqrt{SAH^5K\iota} + SAH^2)$.

The detailed proof is provided in Appendix E.

E PROOF FOR MODEL-FREE ALGORITHM

In this section, we prove Theorem 2. Recall that we use $\overline{Q}_h^k$,$\overline{V}_h^k$,$\underline{Q}_h^k$,$\underline{V}_h^k$ and N_h^k to denote the values of $\overline{Q}_h$,$\overline{V}_h$,$\underline{Q}_h$,$\underline{V}_h$ and $\max\{N_h, 1\}$ at the beginning of the k-th episode.

Property of Learning Rate α_t We refer the readers to the setting of the learning rate $\alpha_t := \frac{H+1}{H+t}$ and the Lemma 4.1 in (Jin et al., 2018). For notational convenience, define $\alpha_t^0 := \prod_{j=1}^{t}(1-\alpha_t)$ and $\alpha_t^i := \alpha_i \prod_{j=i+1}^{t}(1-\alpha_t)$. Here, we introduce some useful properties of α_t^i which were proved in (Jin et al., 2018):
(1) $\sum_{i=1}^{t}\alpha_t^i = 1$ and $\alpha_t^0 = 0$ for $t \ge 1$;
(2) $\sum_{i=1}^{t}\alpha_t^i = 0$ and $\alpha_t^0 = 1$ for $t = 0$;
(3) $\frac{1}{\sqrt{t}} \le \sum_{i=1}^{t}\frac{\alpha_t^i}{\sqrt{t}} \le \frac{2}{\sqrt{t}}$ for every $t \ge 1$;
(4) $\sum_{i=1}^{t}(\alpha_t^i)^2 \le \frac{2H}{t}$ for every $t \ge 1$;
(5) $\sum_{t=i}^{\infty}\alpha_t^i \le (1+\frac{1}{H})$ for every $i \ge 1$.

Recursion on Q As shown in (Jin et al., 2018), at any $(s,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$, let $t = N_h^k(s,a)$ and suppose (s,a) was previously taken by the agent at step h of episodes $k_1, k_2, \ldots, k_t < k$. By the update equations in Algorithm 2 and the definition of α_t^i, we have

$$\begin{aligned}
\overline{Q}_h^k(s,a) &= \alpha_t^0(H-h+1) + \sum_{i=1}^{t}\alpha_t^i\left(r_h^{k_i} + \overline{V}_{h+1}^{k_i}(s_{h+1}^{k_i}) + b_i\right); \\
\underline{Q}_h^k(s,a) &= \sum_{i=1}^{t}\alpha_t^i\left(r_h^{k_i} + \underline{V}_{h+1}^{k_i}(s_{h+1}^{k_i}) - b_i\right).
\end{aligned} \tag{59}$$

Thus,

$$\begin{aligned}
(\overline{Q}_h^k - Q_h^*)(s,a) =& \alpha_t^0(H-h+1) + \sum_{i=1}^{t}\alpha_t^i\left(r_h^{k_i} + \overline{V}_{h+1}^{k_i}(s_{h+1}^{k_i}) + b_i\right) \\
&- \left(\alpha_t^0 Q_h^*(s,a) + \sum_{i=1}^{t}\alpha_t^i\left(R_h(s,a) + P_hV_{h+1}^*(s,a)\right)\right) \\
=& \alpha_t^0(H-h+1-Q_h^*(s,a)) + \sum_{i=1}^{t}\alpha_t^i\left((\overline{V}_{h+1}^{k_i} - V_{h+1}^*)(s_{h+1}^{k_i})\right) \\
&+ \sum_{i=1}^{t}\alpha_t^i\left((r_h^{k_i} - R_h(s,a)) + V_{h+1}^*(s_{h+1}^{k_i}) - P_hV_{h+1}^*(s,a) + b_i\right),
\end{aligned} \tag{60}$$

and similarly

$$\begin{aligned}(\underline{Q}_h^k - Q_h^{\overline{\pi}^k})(s,a) =& \sum_{i=1}^{t} \alpha_t^i \left(r_h^{k_i} + \underline{V}_{h+1}^{k_i}(s_{h+1}^{k_i}) - b_i \right) \\ & - \left(\alpha_t^0 Q_h^{\overline{\pi}^k}(s,a) + \sum_{i=1}^{t} \alpha_t^i \left(R_h(s,a) + P_h V_{h+1}^{\overline{\pi}^k}(s,a) \right) \right) \\ =& -\alpha_t^0 Q_h^{\overline{\pi}^k}(s,a) + \sum_{i=1}^{t} \alpha_t^i \left([P_h(\underline{V}_{h+1}^{k_i} - V_{h+1}^{\overline{\pi}^k})](s,a) \right) \\ & + \sum_{i=1}^{t} \alpha_t^i \left((r_h^{k_i} - R_h(s,a)) + \underline{V}_{h+1}^{k_i}(s_{h+1}^{k_i}) - P_h \underline{V}_{h+1}^{k_i}(s,a) - b_i \right).\end{aligned} \tag{61}$$

In addition, for any $k' \le k$, let $t' = N_h^{k'}(s,a)$. Thus, (s,a) was previously taken by the agent at step h of episodes $k_1, k_2, \ldots, k_{t'} < k'$. We have

$$\begin{aligned}(\underline{Q}_h^{k'} - Q_h^{\overline{\pi}^k})(s,a) =& -\alpha_t^0 Q_h^{\overline{\pi}^k}(s,a) + \sum_{i=1}^{t'} \alpha_{t'}^i \left([P_h(\underline{V}_{h+1}^{k_i} - V_{h+1}^{\overline{\pi}^k})](s,a) \right) \\ & + \sum_{i=1}^{t'} \alpha_{t'}^i \left((r_h^{k_i} - R_h(s,a)) + \underline{V}_{h+1}^{k_i}(s_{h+1}^{k_i}) - P_h \underline{V}_{h+1}^{k_i}(s,a) - b_i \right).\end{aligned} \tag{62}$$

Confidence Bounds By the Azuma-Hoeffding inequality, with probability $1-\delta$, we have that for all s, a, h and $t \le K$,

$$\left| \sum_{i=1}^{t} \alpha_t^i \left((r_h^{k_i} - R_h(s,a)) + \underline{V}_{h+1}^{k_i}(s_{h+1}^{k_i}) - P_h \underline{V}_{h+1}^{k_i}(s,a) \right) \right| \le H \sqrt{\sum_{i=1}^{t} (\alpha_t^i)^2 \iota / 2} \le \sqrt{H^3 \iota / t}. \tag{63}$$

At the same time, with probability $1-\delta$, we have that for all s, a, h and $t \le K$,

$$\left| \sum_{i=1}^{t} \alpha_t^i \left((r_h^{k_i} - R_h(s,a)) + V_{h+1}^*(s_{h+1}^{k_i}) - P_h V_{h+1}^*(s,a) \right) \right| \le \sqrt{H^3 \iota / t}. \tag{64}$$

In addition, we have $\sqrt{H^3\iota/t} \le \sum_{i=1}^{t} \alpha_t^i b_i \le 2\sqrt{H^3\iota/t}$.

Monotonicity Now we prove that $\overline{V}_h^k(s) \ge V_h^*(s) \ge V_h^{\overline{\pi}^k}(s) \ge \underline{V}_h^k(s)$ and $\overline{Q}_h^k(s,a) \ge Q_h^*(s,a) \ge Q_h^{\overline{\pi}^k}(s,a) \ge \underline{Q}_h^k(s,a)$ for all $(s,a,h,k) \in S \times A \times [H] \times [K]$.

At step $H+1$, we have $\overline{V}_{H+1}^k(s) = V_{H+1}^*(s) = V_{H+1}^{\overline{\pi}^k}(s) = \underline{V}_{H+1}^k(s) = 0$ and $\overline{Q}_{H+1}^k(s,a) = Q_{H+1}^*(s,a) = Q_{H+1}^{\overline{\pi}^k}(s,a) = \underline{Q}_{H+1}^k(s,a) = 0$ for all $(s,a,k) \in S \times A \times [K]$.

Consider any step $h \in [H]$ in any episode $k \in [K]$, and suppose that the monotonicity is satisfied for all previous episodes as well as all steps $h' \ge h+1$ in the current episode, which is

$$\begin{aligned}&\overline{V}_{h'}^{k'}(s) \ge V_{h'}^*(s) \ge V_{h'}^{\overline{\pi}^{k'}}(s) \ge \underline{V}_{h'}^{k'}(s) \ \forall (k',h',s) \in [k-1] \times [H+1] \times \mathcal{S}, \\ &\overline{Q}_{h'}^{k'}(s,a) \ge Q_{h'}^*(s,a) \ge Q_{h'}^{\overline{\pi}^{k'}}(s,a) \ge \underline{Q}_{h'}^{k'}(s,a) \ \forall (k',h',s,a) \in [k-1] \times [H+1] \times \mathcal{S} \times \mathcal{A}, \\ &\overline{V}_{h'}^{k}(s) \ge V_{h'}^*(s) \ge V_{h'}^{\overline{\pi}^{k}}(s) \ge \underline{V}_{h'}^{k}(s) \ \forall h' \ge h+1 \text{ and } s \in \mathcal{S}, \\ &\overline{Q}_{h'}^{k}(s,a) \ge Q_{h'}^*(s,a) \ge Q_{h'}^{\overline{\pi}^{k}}(s,a) \ge \underline{Q}_{h'}^{k}(s,a) \ \forall h' \ge h+1 \text{ and } (s,a) \in \mathcal{S} \times \mathcal{A}.\end{aligned} \tag{65}$$

We first show the monotonicity of Q values. We have

$$(\overline{Q}_h^k - Q_h^*)(s,a) \ge \alpha_t^0 (H - h + 1 - Q_h^*(s,a)) + \sum_{i=1}^{t} \alpha_t^i \left((\overline{V}_{h+1}^{k_i} - V_{h+1}^*)(s_{h+1}^{k_i}) \right) \ge 0, \tag{66}$$

and, by to the update rule of $\underline{V}$ values (line 13) in Algorithm 2,

$$\begin{aligned}(\underline{Q}_h^k - Q_h^{\overline{\pi}^k})(s,a) \leq & -\alpha_t^0 Q_h^{\overline{\pi}^k}(s,a) + \sum_{i=1}^{t} \alpha_t^i \left([P_h(\underline{V}_{h+1}^{k_i} - V_{h+1}^{\overline{\pi}^k})](s,a) \right) \\ \leq & -\alpha_t^0 Q_h^{\overline{\pi}^k}(s,a) + \sum_{i=1}^{t} \alpha_t^i \left([P_h(\underline{V}_{h+1}^{k} - V_{h+1}^{\overline{\pi}^k})](s,a) \right) \leq 0.\end{aligned} \tag{67}$$

In addition, for any $k' \leq k$,

$$\begin{aligned}(\underline{Q}_h^{k'} - Q_h^{\overline{\pi}^k})(s,a) \leq & -\alpha_t^0 Q_h^{\overline{\pi}^k}(s,a) + \sum_{i=1}^{t'} \alpha_{t'}^i \left([P_h(\underline{V}_{h+1}^{k_i} - V_{h+1}^{\overline{\pi}^k})](s,a) \right) \\ \leq & -\alpha_t^0 Q_h^{\overline{\pi}^k}(s,a) + \sum_{i=1}^{t'} \alpha_{t'}^i \left([P_h(\underline{V}_{h+1}^{k} - V_{h+1}^{\overline{\pi}^k})](s,a) \right) \leq 0.\end{aligned} \tag{68}$$

Then, we show the monotonicity of V values. We have that

$$\begin{aligned}&(1-\rho)\max_a \overline{Q}_h^k(s,a) + \rho \overline{Q}_h^k(s, \arg\min_a \underline{Q}_h^k(s,a)) \\ \geq &(1-\rho)\max_a \overline{Q}_h^k(s,a) + \rho Q_h^*(s, \arg\min_a \underline{Q}_h^k(s,a)) \\ \geq &(1-\rho)\overline{Q}_h^k(s,\pi_h^*(s)) + \rho \min_{a\in\mathcal{A}} Q_h^*(s,a) \\ \geq &(1-\rho)Q_h^*(s,\pi_h^*(s)) + \rho \min_{a\in\mathcal{A}} Q_h^*(s,a) = V_h^*(s).\end{aligned} \tag{69}$$

By the update rule of $\overline{V}$ values (line 12) in Algorithm 2,

$$\overline{V}_h^k(s) = \min\{\overline{V}_h^{k-1}(s), (1-\rho)\max_a \overline{Q}_h^k(s,a) + \rho \overline{Q}_h^k(s, \arg\min_a \underline{Q}_h^k(s,a))\} \geq V_h^*(s). \tag{70}$$

Here, we need use the update rule of policy $\underline{\pi}$ (line 11-16) in Algorithm 2. Define $\tau(k,h,s) := \max\{k' : k' < k \text{ and } \underline{V}_h^{k'+1}(s) = (1-\rho)\underline{Q}_h^{k'+1}(s, \arg\max_a \overline{Q}_h^{k'+1}(s,a)) + \rho \min_a \underline{Q}_h^{k'+1}(s,a)\}$, which denotes the last episode (before the beginning of the episode k), in which the $\overline{\pi}$ and $\underline{V}$ was updated at (h,s). For notational simplicity, we use τ to denote $\tau(k,h,s)$ here. After the end of episode τ and before the beginning of the episode k, the agent policy $\overline{\pi}$ was not updated and $\underline{V}$ was not updated at (h,s), i.e. $\underline{V}_h^k(s) = \underline{V}_h^{\tau+1}(s) = (1-\rho)\underline{Q}_h^{\tau+1}(s, \overline{\pi}_h^{\tau+1}(s)) + \rho \min_a \underline{Q}_h^{\tau+1}(s,a)$ and $\overline{\pi}_h^k(s) = \overline{\pi}_h^{\tau+1}(s) = \arg\max_a \overline{Q}_h^{\tau+1}(s,a))$. Thus,

$$\begin{aligned}\underline{V}_h^k(s) = &(1-\rho)\underline{Q}_h^{\tau+1}(s, \overline{\pi}_h^{\tau+1}(s)) + \rho \min_a \underline{Q}_h^{\tau+1}(s,a) \\ \leq &(1-\rho)Q_h^{\overline{\pi}^k}(s, \overline{\pi}_h^{\tau+1}(s)) + \rho \min_a \underline{Q}_h^{\tau+1}(s,a) \\ \leq &(1-\rho)Q_h^{\overline{\pi}^k}(s, \overline{\pi}_h^{k}(s)) + \rho \underline{Q}_h^{\tau+1}(s, \arg\min_{a\in\mathcal{A}} Q_h^{\overline{\pi}^k}(s,a)) \\ \leq &(1-\rho)Q_h^{\overline{\pi}^k}(s, \overline{\pi}_h^{k}(s)) + \rho \min_{a\in\mathcal{A}} Q_h^{\overline{\pi}^k}(s,a) = V_h^{\overline{\pi}^k}(s).\end{aligned} \tag{71}$$

By induction from $h = H+1$ to 1 and $k = 1$ to K, we can conclude that $\overline{V}_h^k(s) \geq V_h^*(s) \geq V_h^{\overline{\pi}^k}(s) \geq \underline{V}_h^k(s)$ and $\overline{Q}_h^k(s,a) \geq Q_h^*(s,a) \geq Q_h^{\overline{\pi}^k}(s,a) \geq \underline{Q}_h^k(s,a)$ for all $(s,a,h,k) \in S \times A \times [H] \times [K]$.

Regret Analysis According to the monotonicity, the regret can be bounded by

$$Regret(K) := \sum_{k=1}^{K} (V_1^*(s_1^k) - V_1^{\overline{\pi}^k}(s_1^k)) \leq \sum_{k=1}^{K} (\overline{V}_1^k(s_1^k) - \underline{V}_1^k(s_1^k)). \tag{72}$$

By the update rules in Algorithm 2, we have

$$\begin{aligned}
&\overline{V}_h^k(s_h^k) - \underline{V}_h^k(s_h^k) \\
\leq&(1-\rho)\overline{Q}_h^k(s_h^k, \arg\max_a \overline{Q}_h^k(s_h^k,a)) + \rho\overline{Q}_h^k(s_h^k, \arg\min_a \underline{Q}_h^k(s_h^k,a)) \\
&-(1-\rho)\underline{Q}_h^k(s_h^k, \arg\max_a \overline{Q}_h^k(s_h^k,a)) + \rho\underline{Q}_h^k(s_h^k, \arg\min_a \underline{Q}_h^k(s_h^k,a)) \\
=&(1-\rho)[\overline{Q}_h^k - \underline{Q}_h^k](s_h^k, \overline{a}_h^k) + \rho[\overline{Q}_h^k - \underline{Q}_h^k](s_h^k, \underline{a}_h^k) \\
=&[\overline{Q}_h^k - \underline{Q}_h^k](s_h^k, a_h^k) + [\mathbb{D}_{\widetilde{\pi}_h^k}(\overline{Q}_h^k - \underline{Q}_h^k)](s_h^k) - [\overline{Q}_h^k - \underline{Q}_h^k](s_h^k, a_h^k).
\end{aligned} \tag{73}$$

Set $n_h^k = N_h^k(s_h^k, a_h^k)$ and where $k_i(s_h^k, a_h^k)$ is the episode in which (s_h^k, a_h^k) was taken at step h for the i-th time. For notational simplicity, we set $\phi_h^k = \overline{V}_h^k(s_h^k) - \underline{V}_h^k(s_h^k)$ and $\xi_h^k = [\mathbb{D}_{\widetilde{\pi}_h^k}(\overline{Q}_h^k - \underline{Q}_h^k)](s_h^k) - [\overline{Q}_h^k - \underline{Q}_h^k](s_h^k, a_h^k)$. According to the update rules,

$$\begin{aligned}
\phi_h^k =&\overline{V}_h^k(s_h^k) - \underline{V}_h^k(s_h^k) \\
\leq&\alpha_{n_h^k}^0(H-h+1) + \sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i \left(\overline{V}_{h+1}^{k_i(s_h^k,a_h^k)}(s_{h+1}^{k_i(s_h^k,a_h^k)}) - \underline{V}_{h+1}^{k_i(s_h^k,a_h^k)}(s_{h+1}^{k_i(s_h^k,a_h^k)}) + 2b_i \right) \\
&+ [\mathbb{D}_{\widetilde{\pi}_h^k}(\overline{Q}_h^k - \underline{Q}_h^k)](s_h^k) - [\overline{Q}_h^k - \underline{Q}_h^k](s_h^k, a_h^k) \\
=&\alpha_{n_h^k}^0(H-h+1) + \sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i (\phi_{h+1}^{k_i(s_h^k,a_h^k)} + 2b_i) + \xi_h^k \\
\leq&\alpha_{n_h^k}^0(H-h+1) + \sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i \phi_{h+1}^{k_i(s_h^k,a_h^k)} + \xi_h^k + 4\sqrt{H^3\iota/n_h^k}.
\end{aligned} \tag{74}$$

We add $\overline{V}_h^k(s_h^k) - \underline{V}_h^k(s_h^k)$ over k and regroup the summands in a different way. Note that for any episode k, the term $\sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i \phi_{h+1}^{k_i(s_h^k,a_h^k)}$ takes all the prior episodes $k_i < k$ where (s_h^k, a_h^k) was taken into account. In other words, for any episode k', the term $\phi_{h+1}^{k'}$ appears in the summands at all posterior episodes $k > k'$ where $(s_h^{k'}, a_h^{k'})$ was taken. The first time it appears we have $n_h^k = n_h^{k'} + 1$, and the second time it appears we have $n_h^k = n_h^{k'} + 2$, and so on. Thus, we have

$$\begin{aligned}
&\sum_{k=1}^K (\overline{V}_h^k(s_h^k) - \underline{V}_h^k(s_h^k)) \\
\leq&\sum_{k=1}^K \alpha_{n_h^k}^0(H-h+1) + \sum_{k=1}^K \sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i \phi_{h+1}^{k_i(s_h^k,a_h^k)} + \sum_{k=1}^K \xi_h^k + \sum_{k=1}^K 4\sqrt{H^3\iota/n_h^k} \\
=&\sum_{k=1}^K \alpha_{n_h^k}^0(H-h+1) + \sum_{k'=1}^K \phi_{h+1}^{k'} \sum_{t=n_h^{k'}+1}^{n_h^K} \alpha_t^{n_h^{k'}} + \sum_{k=1}^K \xi_h^k + \sum_{k=1}^K 4\sqrt{H^3\iota/n_h^k} \\
\leq&\sum_{k=1}^K \alpha_{n_h^k}^0(H-h+1) + (1+1/H)\sum_{k=1}^K \phi_{h+1}^k + \sum_{k=1}^K \xi_h^k + \sum_{k=1}^K 4\sqrt{H^3\iota/n_h^k}
\end{aligned} \tag{75}$$

where the final inequality uses the property $\sum_{t=i}^\infty \alpha_t^i \leq (1+\frac{1}{H})$ for every $i \geq 1$.

Taking the induction from $h=1$ to H, we have

$$\begin{aligned}
&\sum_{k=1}^K (\overline{V}_1^k(s_1^k) - \underline{V}_1^k(s_1^k)) \\
\leq&3\sum_{h=1}^H \sum_{k=1}^K \alpha_{n_h^k}^0(H-h+1) + 3\sum_{h=1}^H \sum_{k=1}^K \xi_h^k + \sum_{h=1}^H \sum_{k=1}^K 12\sqrt{H^3\iota/n_h^k}
\end{aligned} \tag{76}$$

where we use the fact that $(1+1/H)^H < 3$ and $\phi_{H+1}^k = 0$ for all k.

We bound the three items separately.

(1) We have $\sum_{h=1}^H \sum_{k=1}^K \alpha_{n_h^k}^0 (H-h+1) = \sum_{h=1}^H \sum_{k=1}^K \mathbb{1}[n_h^k = 0](H-h+1) \leq SAH^2$.

(2) Similar to Lemma 4, by the Azuma-Hoeffding inequality, with probability $1-\delta$, we have $\sum_{h=1}^H \sum_{k=1}^K \xi_h^k \leq H\sqrt{2HK\iota}$.

(3) We have $\sum_{h=1}^H \sum_{k=1}^K 12\sqrt{H^3\iota/n_h^k} = \sum_{h=1}^H \sum_{(s,a)} \sum_{n=1}^{N_h^K(s,a)} \sqrt{H^3\iota/n} \leq H\sqrt{2H^3SAK\iota}$.

In summary,

$$Regret(K) = \sum_{k=1}^K (V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)) \leq \mathcal{O}(\sqrt{SAH^5K\iota} + SAH^2)$$

and

$$\begin{aligned} V_1^*(s_1) - V_1^{\pi^{out}}(s_1) \leq & \overline{V}_1^{K+1}(s_1) - \underline{V}_1^{K+1}(s_1) \\ = & \min_{k\in[K+1]} (\overline{V}_1^k(s_1^k) - \underline{V}_1^k(s_1^k)) \\ \leq & O\left(\frac{\sqrt{SAH^5\iota}}{K} + \frac{SAH^2}{K}\right). \end{aligned} \tag{77}$$