Efficient Action Robust Reinforcement Learning with Probabilistic Policy Execution Uncertainty

Anonymous authors
Paper under double-blind review

Abstract

Robust reinforcement learning (RL) aims to find a policy that optimizes worst-case performance in the face of uncertainties. In this paper, we focus on action robust RL with the probabilistic policy execution uncertainty, in which, instead of always carrying out the action specified by the policy, the agent will take the action specified by the policy with probability $1-\rho$ and an alternative adversarial action with probability ρ. We show the existence of an optimal policy on the action robust MDPs with probabilistic policy execution uncertainty and provide the action robust Bellman optimality equation for its solution. Based on that, we develop Action Robust Reinforcement Learning with Certificates (ARRLC) algorithm that achieves minimax optimal regret and sample complexity. Furthermore, we conduct numerical experiments to validate our approach's robustness, demonstrating that ARRLC outperforms non-robust RL algorithms and converges faster than the other action robust RL algorithms in the presence of action perturbations.

1 Introduction

Reinforcement learning (RL), a framework for solving control-theoretic problem that involves making decisions over time under an unknown environment, has many applications in a variety of scenarios (Zhao et al., 2018; Nazari et al., 2018; O' Kelly et al., 2018; Liu et al., 2020). However, the solutions to standard RL methods are not inherently robust to uncertainties, perturbations, or structural changes in the environment, which are frequently observed in real-world settings. A trustworthy reinforcement learning algorithm should be competent in solving challenging real-world problems with robustness against perturbations and uncertainties. Robust RL aims to improve the worst-case performance of algorithms deterministically or statistically in the face of uncertainties in different MDP components, including observations/states (Zhang et al., 2020a; Sun et al., 2022), actions (Tessler et al., 2019; Klima et al., 2019), transitions (Nilim & El Ghaoui, 2005; Iyengar, 2005; Tamar et al., 2014; Wang & Zou, 2021), and rewards (Huang & Zhu, 2019; Lecarpentier & Rachelson, 2019).

In this paper, we consider action uncertainties, also called policy execution uncertainties, and probabilistic uncertainty set proposed in Tessler et al. (2019). Robust RL against action uncertainties focuses on the discrepancy between the actions generated by the RL agent and the conducted actions. Taking the robot control as an example, such policy execution uncertainty may come from the actuator noise, limited power range, or actuator failures in the real world. Taking the medication advice in healthcare as another example, such policy execution uncertainty may come from the patient's personal behaviors such as drug refusal, forgotten medication, or overdose etc.

To deal with the policy execution uncertainty, robust RL methods (Pinto et al., 2017; Tessler et al., 2019) adopt the adversarial training framework (Goodfellow et al., 2014; Madry et al., 2018) and assume an adversary conducting adversarial attacks to mimic the naturalistic uncertainties. Training with an adversary can be formulated as a zero-sum game between the adversary and the RL agent. However, these interesting works do not provide theoretical guarantee on sample complexity or regret. In this paper, we aim to fill this gap.

The approaches in Pinto et al. (2017) and Tessler et al. (2019) iteratively apply two stages: (i) given a fixed adversary policy, it calculates the agent's optimal policy; and (ii) update the adversary policy against the updated agent's policy. The repetition of stage (i) requires repeatedly solving MDP to find the optimal policy, which is sample inefficient. Motivated by the recent theoretical works on transition probability uncertainty that use the robust dynamic programming method Iyengar (2005) and achieve efficient sample complexity Wang & Zou (2021); Panaganti & Kalathil (2022); Xu et al. (2023), we introduce the action robust Bellman equations and design sample efficient algorithms based on the action robust Bellman equations. Our methods simultaneously update the adversary policy and agent's policy instead of updating one after another is converged. Our major contributions are summarized as follows:

- We show that the robust problem can be solved by the iteration of the action robust Bellman optimality equations. Motivated by this, we design two efficient algorithms.
- We develop a model-based algorithm, Action Robust Reinforcement Learning with Certificates (ARRLC), for episodic action robust MDPs, and show that it achieves minimax order optimal regret and minimax order optimal sample complexity.
- We develop a model-free algorithm for episodic action robust MDPs, and analyze its regret and sample complexity.
- We conduct numerical experiments to validate the robustness of our approach. In our experiments, our robust algorithm achieves a much higher reward than the non-robust RL algorithm when being tested with some action perturbations; and our ARRLC algorithm converges much faster than the robust TD algorithm in (Klima et al., 2019).

2 Related work

We mostly focus on papers that are related to sample complexity bounds for the episodic RL and the two-player zero-sum Markov game, and action robust RL, that are close related to our model. We remark that there are also related settings, e.g., infinite-horizon discounted MDP (Li et al., 2020; He et al., 2021), robust RL with other uncertainties (Iyengar, 2005; Lecarpentier & Rachelson, 2019; Zhang et al., 2020a; Wang & Zou, 2021), robust offline RL (Guo et al., 2022; Shi & Chi, 2022), adversarial training with a generative RL model (Zhou et al., 2021; Yang et al., 2022; Panaganti & Kalathil, 2022; Xu et al., 2023; Shi et al., 2023), adversarial attacks on RL (Zhang et al., 2020b; Liu & Lai, 2021; Sun et al., 2021), etc. Most of robust RL papers considered the robust MDP setting, which have been proposed to handle these problems by letting the transition probability vary in an uncertainty (or ambiguity) set. Table 2 provides the sample complexity comparisons of our result with that of the existing Robust MDP works.

Action robust RL. Pinto et al. (2017) introduces robust adversarial reinforcement learning to address the generalization issues in reinforcement learning by training with a destabilizing adversary that applies disturbance forces to the system. Tessler et al. (2019) introduces two new criteria of robustness for reinforcement learning in the face of action uncertainty. We follow its probabilistic action robust MDP (PR-MDP) in which, instead of the action specified by the policy, an alternative adversarial action is taken with probability ρ. They generalize their policy iteration approach to deep reinforcement learning (DRL) and provide extensive experiments. A similar uncertainty setting was presented in Klima et al. (2019), which extends temporal difference (TD) learning algorithms by a new robust operator and shows that the new algorithms converge to the optimal robust Q-function. However, no theoretical guarantee on sample complexity or regret is provided in these works. We develop a minimax sample efficient algorithm and fill this gap.

Sample complexity bounds for the episodic RL. There is a rich literature on sample complexity guarantees for episodic tabular RL, for example (Kearns & Singh, 2002; Strehl et al., 2006; Auer et al., 2008; Azar et al., 2017; Dann et al., 2017; Jin et al., 2018; Dann et al., 2019; Simchowitz & Jamieson, 2019; Zhang et al., 2020c; 2021). However, these methods can not be directly applied in action robust MDP with small technical changes. Most relevant to our paper is the work about policy certificates (Dann et al., 2019). The algorithm ORLC in Dann et al. (2019) calculate both the upper bound and lower bound of the value functions, and outputs policy certificates that bound the sub-optimality and return of the policy. Our

Algorithm	**Sample complexity**	**Uncertainty**	**Sample from**
(Yang et al., 2022)	$\frac{S^2AH^4(2+\rho)^2}{\rho^2\epsilon^2}$	Transition probability	Generative environment
(Panaganti & Kalathil, 2022)	$\frac{S^2AH^4}{\epsilon^2}$	Transition probability	Generative environment
(Xu et al., 2023)	$\frac{SAH^5}{\epsilon^2}$	Transition probability	Generative environment
(Wang & Zou, 2021)	Asymptotic	Transition probability	Online environment
This work	$\frac{SAH^3}{\epsilon^2}$	Policy execution	Online environment
(Non-robust) Lower bound (Zhang et al., 2020c)	$\frac{SAH^3}{\epsilon^2}$	-	-

Table 1: Sample complexity comparisons of previous results and our results, where S, A are the size of the state space and action space, H is the length of the horizon, K is the number of episodes, ρ is the uncertainty parameter, and ϵ is the level of suboptimality.

proposed ARRLC shares a similar structure with ORLC, but we develop new adversarial trajectory sampling and action robust value iteration method in ARRLC, and new techniques to bound the sum of variances so that our algorithm suits for action robust MDPs.

Sample complexity bounds for the two-player zero-sum Markov game. Training with an adversary can naturally be formulated as a zero-sum game between the adversary and the RL agent. Some sample efficient algorithms for two-player zero-sum Markov game can be used to train the action robust RL agent. The efficient multi-agent RL algorithms (Liu et al., 2021; Jin et al., 2021) can be used to solve the action robust optimal policy but are not minimax optimal. They are a factor of A or H^2 above the minimax lower bound. Our algorithm ARRLC is minimax optimal.

3 Problem formulation

Tabular MDPs. We consider a tabular episodic MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, R)$, where $\mathcal{S}$ is the state space with $|\mathcal{S}| = S$, $\mathcal{A}$ is the action space with $|\mathcal{A}| = A$, $H \in \mathbb{Z}^+$ is the number of steps in each episode, P is the transition matrix so that $P_h(\cdot|s,a)$ represents the probability distribution over states if action a is taken for state s at step $h \in [H]$, and $R_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ represents the reward function at the step h. In this paper, the probability transition functions and the reward functions can be different at different steps.

The agent interacts with the MDP in episodes indexed by k. Each episode k is a trajectory $\{s_1^k, a_1^k, r_1^k, \cdots, s_H^k, a_H^k, r_H^k\}$ of H states $s_h^k \in \mathcal{S}$, actions $a_h^k \in \mathcal{A}$, and rewards $r_h^k \in [0,1]$. At each step $h \in [H]$ of episode k, the agent observes the state s_h^k and chooses an action a_h^k. After receiving the action, the environment generates a random reward $r_h^k \in [0,1]$ derived from a distribution with mean $R_h(s_h^k, a_h^k)$ and next state s_{h+1}^k that is drawn from the distribution $P_h(\cdot|s_h^k, a_h^k)$. For notational simplicity, we assume that the initial states $s_1^k = s_1$ are deterministic in different episode k.

A (stochastic) Markov policy of the agent is a set of H maps $\pi := \{\pi_h : \mathcal{S} \to \Delta_{\mathcal{A}}\}_{h\in[H]}$, where $\Delta_{\mathcal{A}}$ denotes the simplex over $\mathcal{A}$. We use notation $\pi_h(a|s)$ to denote the probability of taking action a in state s under stochastic policy π at step h. A deterministic policy is a policy that maps each state to a particular action. Therefore, when it is clear from the context, we abuse the notation $\pi_h(s)$ for a deterministic policy π to denote the action a which satisfies $\pi_h(a|s) = 1$.

Action robust MDPs. In the action robust case, the policy execution is not accurate and lies in some uncertainty set centered on the agent's policy π. Denote the actual behavior policy by $\widetilde{\pi}$ where $\widetilde{\pi} \in \Pi(\pi)$ and $\Pi(\pi)$ is the uncertainty set of the policy execution. Denote the actual behavior action at episode k and step h by $\widetilde{a}_h^k$ where $\widetilde{a}_h^k \sim \widetilde{\pi}_h^k$. Define the action robust value function of a policy π as the worst-case expected accumulated reward over following any policy in the uncertainty set $\Pi(\pi)$ centered on a fixed policy π:

$$V_h^\pi(s) = \min_{\widetilde{\pi} \in \Pi(\pi)} \mathbb{E}_{\widetilde{\pi}} \left[\sum_{h'=h}^{H} R_{h'}(s_{h'}, a_{h'}) | s_h = s \right]. \tag{1}$$

V_h^π represents the action robust value function of policy π at step h. Similarly, define the action robust Q-function of a policy π:

$$Q_h^\pi(s,a) = \min_{\widetilde{\pi} \in \Pi(\pi)} \mathbb{E}_{\widetilde{\pi}} \left[\sum_{h'=h}^{H} R_{h'}(s_{h'}, a_{h'}) | s_h = s, a_h = a \right]. \tag{2}$$

The goal of action robust RL is to find the optimal robust policy π^* that maximizes the worst-case accumulated reward: $\pi^* = \arg\max_\pi V_1^\pi(s), \forall s \in \mathcal{S}$. We also denote V^{π^*} and Q^{π^*} by V^* and Q^*.

Probabilistic policy execution uncertain set. We follow the setting of the probabilistic action robust MDP (PR-MDP) introduced in (Tessler et al., 2019) to construct the probabilistic policy execution uncertain set. For some $0 \le \rho \le 1$, the policy execution uncertain set is defined as:

$$\Pi^\rho(\pi) := \{\widetilde{\pi} : \forall s, \forall h, \exists \pi_h'(\cdot|s) \in \Delta_{\mathcal{A}} \text{ such that } \widetilde{\pi}_h(\cdot|s) = (1-\rho)\pi_h(\cdot|s) + \rho \pi_h'(\cdot|s)\}. \tag{3}$$

The policy execution uncertain set can also be expressed as $\Pi^\rho(\pi) = (1-\rho)\pi + \rho(\Delta_{\mathcal{A}})^{S \times H}$.

In this setting, an optimal probabilistic robust policy is optimal w.r.t. a scenario in which, with probability at most ρ, an adversary takes control and performs the worst possible action. We call π' as the adversarial policy. For different agent's policy π, the corresponding adversarial policy π' that minimizes the cumulative reward may be different.

Additional notations. We set $\iota = \log(2SAHK/\delta)$ for $\delta > 0$. For simplicity of notation, we treat P as a linear operator such that $[P_h V](s,a) := \mathbb{E}_{s' \sim P_h(\cdot|s,a)} V(s')$, and we define two additional operators $\mathbb{D}$ and $\mathbb{V}$ as follows: $[\mathbb{D}_{\pi_h} Q](s) := \mathbb{E}_{a \sim \pi_h(\cdot|s)} Q(s,a)$ and

$$\begin{aligned}\mathbb{V}_{P_h} V_{h+1}(s,a) &:= \sum_{s'} P_h(s'|s,a) \left(V_{h+1}(s') - [P_h V_{h+1}](s,a)\right)^2 \\ &= [P_h (V_{h+1})^2](s,a) - ([P_h V_{h+1}](s,a))^2.\end{aligned}$$

4 Existence of the optimal robust policy

For the standard tabular MDPs, when the state space, action space, and the horizon are all finite, there always exists an optimal policy. In addition, if the reward functions and the transition probabilities are known to the agent, the optimal policy can be solved by solving the Bellman optimality equation. In the following theorem, we show that the optimal policy also always exists in action robust MDPs and can be solved by the action robust Bellman optimality equation.

Proposition 4.1 *If the uncertainty set of the policy execution has the form in equation 3, the following perfect duality holds for all $s \in \mathcal{S}$ and all $h \in [H]$:*

$$\max_\pi \min_{\widetilde{\pi} \in \Pi^\rho(\pi)} \mathbb{E}_{\widetilde{\pi}} \left[\sum_{h'=h}^{H} R_{h'}(s_{h'}, a_{h'}) | s_h = s \right] = \min_{\widetilde{\pi} \in \Pi^\rho(\pi)} \max_\pi \mathbb{E}_{\widetilde{\pi}} \left[\sum_{h'=h}^{H} R_{h'}(s_{h'}, a_{h'}) | s_h = s \right]. \tag{4}$$

There always exists a deterministic optimal robust policy π^*. The problem can be solved by the iteration of the action robust Bellman optimality equation on $h = H, \cdots, 1$. The action robust Bellman equation and the action robust Bellman optimality equation are:

$$\begin{cases} V_h^\pi(s) = (1-\rho)[\mathbb{D}_{\pi_h} Q_h^\pi](s) + \rho \min_{a \in \mathcal{A}} Q_h^\pi(s,a), \\ Q_h^\pi(s,a) = R_h(s,a) + [P_h V_{h+1}^\pi](s,a), \\ V_{H+1}^\pi(s) = 0, \ \forall s \in \mathcal{S}. \end{cases} \tag{5}$$

$$\begin{cases} V_h^*(s) = (1-\rho)\max_{a\in\mathcal{A}} Q_h^*(s,a) + \rho \min_{b\in\mathcal{A}} Q_h^*(s,b), \\ Q_h^*(s,a) = R_h(s,a) + [P_h V_{h+1}^*](s,a), \\ V_{H+1}^*(s) = 0, \ \forall s \in \mathcal{S}. \end{cases} \tag{6}$$

We define

$$C_h^{\pi,\pi',\rho}(s) := \mathbb{E}_{\widetilde{\pi}}\left[\sum_{h'=h}^{H} R_{h'}(s_{h'}, a_{h'})|s_h = s\right]. \tag{7}$$

Since $\widetilde{\pi}_h(\cdot|s) = (1-\rho)\pi_h(\cdot|s) + \rho\pi'_h(\cdot|s)$, $\min_{\widetilde{\pi}\in\Pi^\rho(\pi)}$ is equivalent to $\min_{\widetilde{\pi}\in\Pi^\rho(\pi)}$ when fixing *pi*. The perfect duality of the control problems in equation 4 is equivalent to $\max_\pi \min_{\pi'} C_h^{\pi,\pi',\rho}(s) = \min_{\pi'} \max_\pi C_h^{\pi,\pi',\rho}(s)$. We provide the detailed proof of the perfect duality and the existence of the optimal policy in Appendix B. Our proposed model-based algorithm in Section 5 and model-free algorithm in Appendix D are based on the action robust Bellman optimality equation. Using the iteration of the proposed action robust Bellman equation to solve the robust problem can simultaneously update the adversary policy and agent policy and avoid inefficient alternating updates.

5 Algorithm and main results

Algorithm 1 ARRLC (**A**ction **R**obust **R**einforcement **L**earning with **C**ertificates)

```
Initialize $\overline{V}_h(s) = H-h+1$, $\overline{Q}_h(s,a) = H-h+1$, $\underline{V}_h(s) = 0$, $\underline{Q}_h(s,a) = 0$, $\hat{r}_h(s,a)$, $N_h(s,a) = 0$ and $N_h(s,a,s') = 0$ 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 any $s$ and $a$. $\Delta = H$.
for episode $k = 1, 2, \ldots, K$ do
  for step $h = 1, 2, \ldots, H$ do
    Observe $s_h^k$.
    Set $\overline{\pi}_h^k(s) = \arg\max_a \overline{Q}_h(s,a)$ , $\underline{\pi}_h^k(s) = \arg\min_a \underline{Q}_h(s,a)$, $\widetilde{\pi}_h^k = (1-\rho)\overline{\pi}_h^k + \rho\underline{\pi}_h^k$.
    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 $N_h(s_h^k, a_h^k) \leftarrow N_h(s_h^k, a_h^k) + 1$, $N_h(s_h^k, a_h^k, s_{h+1}^k) \leftarrow N_h(s_h^k, a_h^k, s_{h+1}^k) + 1$.
    Set $\hat{r}_h^k(s_h^k, a_h^k) \leftarrow \hat{r}_h^k(s_h^k, a_h^k) + (r_h^k - \hat{r}_h^k(s_h^k, a_h^k))/N_h(s_h^k, a_h^k)$.
    Set $\hat{P}_h(\cdot|s_h^k, a_h^k) = N_h(s_h^k, a_h^k, \cdot)/N_h(s_h^k, a_h^k)$.
  end for
  Output policy $\overline{\pi}^k$ with certificates $\mathcal{I}_k = [\underline{V}_1(s_1^k), \overline{V}_1(s_1^k)]$ and $\epsilon_k = |\mathcal{I}_k|$ .
  if $\epsilon_k < \Delta$ then
    $\Delta \leftarrow \epsilon_k$ and $\pi^{out} \leftarrow \overline{\pi}^k$.
  end if
  for step $h = H, H-1, \ldots, 1$ do
    for each $(s,a) \in \mathcal{S} \times \mathcal{A}$ with $N_h(s,a) > 0$ do
      Set $\theta_h(s,a) = \sqrt{\frac{2\mathbb{V}_{\hat{P}_h}[(\overline{V}_{h+1}+\underline{V}_{h+1})/2](s,a)\iota}{N_h(s,a)}} + \sqrt{\frac{2\hat{r}_h(s,a)\iota}{N_h(s,a)}} + \frac{\hat{P}_h\left(\overline{V}_{h+1}-\underline{V}_{h+1}\right)(s,a)}{H} + \frac{(24H^2+7H+7)\iota}{3N_h(s,a)}$,
      $\overline{Q}_h(s,a) \leftarrow \min\{H-h+1, \hat{r}_h(s,a) + \hat{P}_h\overline{V}_{h+1}(s,a) + \theta_h(s,a)\}$,
      $\underline{Q}_h(s,a) \leftarrow \max\{0, \hat{r}_h(s,a) + \hat{P}_h\underline{V}_{h+1}(s,a) - \theta_h(s,a)\}$,
      $\overline{\pi}_h^{k+1}(s) = \arg\max_a \overline{Q}_h(s,a)$ , $\underline{\pi}_h^{k+1}(s) = \arg\min_a \underline{Q}_h(s,a)$,
      $\overline{V}_h(s) \leftarrow (1-\rho)\overline{Q}_h(s, \overline{\pi}_h^{k+1}(s)) + \rho\overline{Q}_h(s, \underline{\pi}_h^{k+1}(s))$,
      $\underline{V}_h(s) \leftarrow (1-\rho)\underline{Q}_h(s, \overline{\pi}_h^{k+1}(s)) + \rho\underline{Q}_h(s, \underline{\pi}_h^{k+1}(s))$.
    end for
  end for
end for
Return $\pi^{out}$
```

In this section, we introduce the proposed **A**ction **R**obust **R**einforcement **L**earning with **C**ertificates (ARRLC) algorithm and provides its theoretical guarantee. The pseudo code is listed in Algorithm 1. Here, we highlight the main idea of our algorithm. Algorithm 1 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 1 chooses an adversarial action with probability ρ. To learn the agent's optimal policy and the corresponding adversarial policy, Algorithm 1 computes an optimistic estimate $\overline{Q}$ of Q^* and a pessimistic estimate $\underline{Q}$ of $Q^{\overline{\pi}^k}$. Algorithm 1 uses the optimistic estimates to explore the possible optimal policy $\overline{\pi}$ and uses the pessimistic estimates to explore the possible adversarial policy $\underline{\pi}$. As shown later in Lemma 6.2, $\overline{V} \geq V^* \geq V^{\overline{\pi}} \geq \underline{V}$ holds with high probabilities. The optimistic and pessimistic estimates $\overline{V}$ and $\underline{V}$ can provide policy certificates, which bounds the cumulative rewards of the return policy $\overline{\pi}^k$ and $\overline{V} - \underline{V}$ bounds the sub-optimality of the return policy $\overline{\pi}^k$ with high probabilities. The policy certificates can give us some insights about the performance of $\overline{\pi}^k$ in the perturbed environment with probabilistic policy execution uncertainty.

5.1 Algorithm description

We now describe the proposed ARRLC algorithm in more details. In each episode, the ARRLC algorithm can be decomposed into two parts.

- Line 3-11 (Sample trajectory and update the model estimate): Simulates the action robust MDP, executes the behavior policy $\widetilde{\pi}$, collects samples, and updates the estimates of the rewards and the transitions.
- Line 16-25 (Adversarial planning from the estimated model): Performs value iteration with bonus to estimate the robust value functions using the empirical estimate of the transition $\hat{P}$, computes a new policy $\overline{\pi}$ that is optimal respect to the estimated robust value functions, and computes a new optimal adversarial policy $\underline{\pi}$ respect to the agent's policy $\overline{\pi}$.

At a high-level, this two-phase policy is standard in the majority of model-based RL algorithms (Azar et al., 2017; Dann et al., 2019). Algorithm 1 shares similar structure with ORLC (Optimistic Reinforcement Learning with Certificates) in Dann et al. (2019) but has some significant differences in line 5-6 and line 18-23. The first main difference is that the ARRLC algorithm simulates the probabilistic policy execution uncertainty by choosing an adversarial action with probability ρ. The adversarial policy and the adversarial actions are computed by the ARRLC algorithm. The second main difference is that the ARRLC algorithm simultaneously plans the agent policy $\overline{\pi}$ and the adversarial policy $\underline{\pi}$ by the action robust Bellman optimality equation.

These two main difference bring two main challenges in the design and analysis of our algorithm.

(1) The ARRLC algorithm simultaneously plans the agent policy and the adversarial policy. However the planned adversarial policy $\underline{\pi}$ is not necessarily the true optimal adversary policy towards the agent policy $\overline{\pi}$ because of the estimation error of the value functions. We carefully design the bonus items and the update role of the value functions so that $\overline{V}_h(s) \geq V_h^*(s) \geq V_h^{\overline{\pi}}(s) \geq \underline{V}_h(s)$ and $\overline{Q}_h(s,a) \geq Q_h^*(s,a) \geq Q_h^{\overline{\pi}}(s,a) \geq \underline{Q}_h(s,a)$ hold for all s and a.

(2) A crucial step in many UCB-type algorithms based on Bernstein inequality is bounding the sum of variance of estimated value function across the planning horizon. The behavior policies in these UCB-type algorithms are deterministic. However, the behavior policy in our ARRLC algorithm is not deterministic due to the simulation of the adversary's behavior. The total variance is the weighted sum of the sum of variance of estimated value function across two trajectories. Even if action $\overline{\pi}(s_h^k)$ or $\underline{\pi}(s_h^k)$ is not sampled at state s_h^k, it counts in the total variance. Thus, the sum of variance is no longer simply the variance of the sum of rewards per episode, and new techniques are introduced. For example, the variance of $\overline{V} + \underline{V}$ can be connected to the variance of $C^{\pi^{k*},\underline{\pi}^k,\rho}$, where π^{k*} is the optimal policy towards the adversary policy $\underline{\pi}^k$ with $\pi_h^{k*}(s) = \arg\max_\pi C_h^{\pi,\underline{\pi}^k,\rho}(s)$. Then the variance of $C^{\pi^{k*},\underline{\pi}^k,\rho}$ can be bounded via recursion on the sampled trajectories.

5.2 Theoretical guarantee

We define the cumulative regret of the output policy $\overline{\pi}^k$ at each episodes k as $Regret(K) := \sum_{k=1}^{K}(V_1^*(s_1^k) - V_1^{\overline{\pi}^k}(s_1^k))$.

Theorem 5.1 *For any $\delta \in (0,1]$, letting $\iota = \log(2SAHK/\delta)$, then with probability at least $1-\delta$, Algorithm 1 achieves:*

- $V_1^*(s_1) - V_1^{\pi^{out}}(s_1) \leq \epsilon$, *if the number of episodes* $K \geq \Omega(SAH^3\iota^2/\epsilon^2 + S^2AH^3\iota^2/\epsilon)$.
- $Regret(K) = \sum_{k=1}^{K}(V_1^*(s_1^k) - V_1^{\overline{\pi}^k}(s_1^k)) \leq \mathcal{O}(\sqrt{SAH^3K\iota} + S^2AH^3\iota^2)$.

For small $\epsilon \leq H/S$, the sample complexity scales as $\mathcal{O}(SAH^3\iota^2/\epsilon^2)$. For the case with a large number of episodes $K \geq S^3AH^3\iota$, the regret scales as $\mathcal{O}(\sqrt{SAH^3K\iota})$. For the standard MDPs, the information-theoretic sample complexity lower bound is $\Omega(SAH^3/\epsilon^2)$ provided in Zhang et al. (2020c) and the regret lower bound is $\Omega(\sqrt{SAH^3K})$ provided in Jin et al. (2018). When $\rho = 0$, the action robust MDPs is equivalent to the standard MDPs. Thus, the information-theoretic sample complexity lower bound and the regret lower bound of the action robust MDPs should have same dependency on S, A, H, K or ϵ. The lower bounds show the optimality of our algorithm up to logarithmic factors.

6 Proof sketch

In this section, we provide sketch of the proof, which will highlight our the main ideas of our proof. First, we will show that $\overline{V}_h(s) \geq V_h^*(s) \geq V_h^{\overline{\pi}}(s) \geq \underline{V}_h(s)$ hold for all s and a. The regret can be bounded by $\overline{V}_1 - \underline{V}_1$ and then be divided by four items, each of which can be bounded separately. The full proof can be found in the appendix contained in the supplementary material.

We first introduce a few notations. 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 in the beginning of the k-th episode in Algorithm 1.

6.1 Proof of monotonicity

We define $\mathcal{E}^R$ to be the event where

$$\left|\hat{r}_h^k(s,a) - R_h(s,a)\right| \leq \sqrt{\frac{2\hat{r}_h^k(s,a)\iota}{N_h^k(s,a)}} + \frac{7\iota}{3(N_h^k(s,a))} \tag{8}$$

holds for all $(s,a,h,k) \in S \times A \times [H] \times [K]$. We also define $\mathcal{E}^{PV}$ to be the event where

$$\left|(\hat{P}_h^k - P_h)V_{h+1}^*(s,a)\right| \leq \sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}V_{h+1}^*(s,a)\iota}{N_h^k(s,a)}} + \frac{7H\iota}{3(N_h^k(s,a))} \tag{9}$$

and

$$\left|(\hat{P}_h^k - P_h)V_{h+1}^{\overline{\pi}^k}(s,a)\right| \leq \sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}V_{h+1}^{\overline{\pi}^k}(s,a)\iota}{N_h^k(s,a)}} + \frac{7H\iota}{3N_h^k(s,a)} \tag{10}$$

hold for all $(s,a,h,k) \in S \times A \times [H] \times [K]$.

Event $\mathcal{E}^R$ means that the estimations of all reward functions stay in certain neighborhood of the true values. Event $\mathcal{E}^{PV}$ represents that the estimation of the value functions at the next step stay in some intervals. The following lemma shows $\mathcal{E}^R$ and $\mathcal{E}^{PV}$ hold with high probability. The analysis will be done assuming the successful event $\mathcal{E}^R \cap \mathcal{E}^{PV}$ holds in the rest of this section.

Lemma 6.1 $\mathbb{P}(\mathcal{E}^R \cap \mathcal{E}^{PV}) \geq 1 - 3\delta$.

Lemma 6.2 *Conditioned on event* $\mathcal{E}^R \cap \mathcal{E}^{PV}$, $\overline{V}_h^k(s) \geq V_h^*(s) \geq V_h^{\overline{\pi}^k}(s) \geq \underline{V}_h^k(s)$ *and* $\overline{Q}_h^k(s,a) \geq Q_h^*(s,a) \geq Q_h^{\overline{\pi}^k}(s,a) \geq \underline{Q}_h^k(s,a)$ *hold for all* $(s,a,h,k) \in S \times A \times [H] \times [K]$.

6.2 Regret analysis

We decompose the regret and analyze the different terms. Set $\Theta_h^k(s,a) = \sqrt{\frac{8\mathbb{V}_{P_h} C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s,a)\iota}{N_h^k(s,a)}} + \sqrt{\frac{32}{N_h^k(s,a)}} + \frac{46\sqrt{SH^4\iota}}{N_h^k(s,a)}$, where π^{k*} is the optimal policy towards the adversary policy $\underline{\pi}^k$ with $\pi_h^{k*}(s) = \arg\max_\pi C_h^{\pi,\underline{\pi}^k,\rho}(s)$. We set

$$M_1 = \sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{D}_{\widetilde{\pi}_h^k}\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k) - \hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k)], \tag{11}$$

$$M_2 = \sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{H}[\mathbb{D}_{\widetilde{\pi}_h^k}P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k) - P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k)] \tag{12}$$

$$M_3 = \sum_{k=1}^{K}\sum_{h=1}^{H}[P_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k) - (\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k)] \tag{13}$$

$$M_4 = \sum_{k=1}^{K}\sum_{h=1}^{H}\left[\frac{(SH + SH^2)\iota}{N_h^k(s_h^k, a_h^k)} + \mathbb{D}_{\widetilde{\pi}_h^k}\Theta_h^k(s_h^k)\right] \tag{14}$$

Here M_1 and M_2 are the cumulative sample error from the random choices of the adversarial policy or agent's policy. M_3 is the cumulative sample error from the randomness of Monte Carlo sampling of the next state. M_4 is the cumulative error from the bonus item θ. Lemma 6.3 shows that the regret can be bounded by these four terms.

Lemma 6.3 *With probability at least* $1 - (S+5)\delta$,

$$\mathit{Regret}(K) \leq \sum_{k=1}^{K}(\overline{V}_1^k(s_1^k) - \underline{V}_1^k(s_1^k)) \leq 21(M_1 + M_2 + M_3 + M_4). \tag{15}$$

We now bound each of these four items separately.

Lemma 6.4 *With probability at least* $1 - \delta$, $|M_1| \leq H\sqrt{2HK\iota}$.

Lemma 6.5 *With probability at least* $1 - \delta$, $|M_2| \leq \sqrt{2HK\iota}$.

Lemma 6.6 *With probability at least* $1 - \delta$, $|M_3| \leq H\sqrt{2HK\iota}$.

Lemma 6.7 *With probability at least* $1-2\delta$, $|M_4| \leq 2S^2AH^3\iota^2 + 8\sqrt{SAH^2K\iota} + 46S^{\frac{3}{2}}AH^3\iota^2 + \sqrt{24SAH^3K\iota} + 6\sqrt{SAH^5}\iota$.

Putting all together. By Lemmas 6.3, 6.4, 6.5, 6.6, and 6.7, we conclude that, with probability $1-(S+10)\delta$,

$$\mathit{Regret}(K) \leq O(\sqrt{SAH^3K\iota} + S^2AH^3\iota^2). \tag{16}$$

By rescaling δ, $\log(\frac{2SAHK}{\delta/(S+10)}) \leq c\iota$ for some constant c and we finish the proof of regret. As $\sum_{k=1}^{K}(\overline{V}_1^k(s_1^k) - \underline{V}_1^k(s_1^k)) \leq O(\sqrt{SAH^3K\iota} + S^2AH^3\iota^2)$, we have that $V_1^*(s_1) - V_1^{\pi^{out}}(s_1) \leq \min_k \overline{V}_1^k(s_1^k) - \underline{V}_1^k(s_1^k) \leq O(\sqrt{\frac{SAH^3}{K}\iota} + \frac{S^2AH^3\iota^2}{K})$ and we finish the proof of sample complexity.

Algorithm 2 Action Robust Q-learning with Hoeffding Confidence Bound (ARQ-H)

```
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 $\pi^{out}$
```

7 Model-free method

In this section, we develop a model-free algorithm, called Action Robust Q-learning with Hoeffding confidence bound (ARQ-H), and analyze its theoretical guarantee. The pseudo code is listed 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 uses a model-free method to update Q and V values.

Here, we highlight the challenges of the model-free method compared with the model-based method. In the model-based planning, we perform 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 and V values are updated only on (s_h^k, a_h^k) which are the samples on the trajectories. The variances of the Q values and V values in model-free method are larger than the model-based method. Compared with the model-based method, the update of the Q values and V values in the model-free method is slower and less stable.

To deal with this challenges, we design a special update rule of the output policy. In line 14-16, Algorithm 2 does not update the output policy until the lower bound on the value function of the new output policy is improved. By this, the output policies are stably improved after every update. The adversary policy is still updated at each episode.

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

Theorem 7.1 *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. The model-free method is more computationally efficient than the model-based method but is less sample efficient.

8 Simulation results

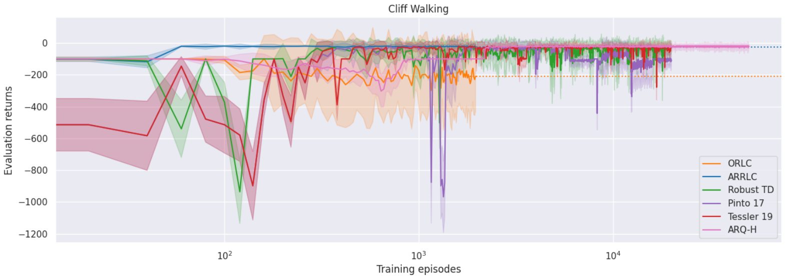
We use OpenAI gym framework (Brockman et al., 2016), and consider two different problems: Cliff Walking, a toy text environment, and Inverted Pendulum, a control environment with the MuJoCo (Todorov et al., 2012) physics simulator. We set $H = 100$. To demonstrate the robustness, the policy is learned in a clean environment, and is then tested in the perturbed environment. Specifically, ρ is the uncertainty parameter used during training. Then, during testing, we set a probability p such that after the agent takes an action, with probability p, the action is chosen by an adversary. The adversary follows a fixed policy. A Monte-Carlo method is used to evaluate the accumulated reward of the learned policy in the perturbed environment. We take the average over 100 trajectories.

Inverted pendulum. The inverted pendulum experiment is a classic control problem in RL. An inverted pendulum is attached by a pivot point to a cart, which is restricted to linear movement in a plane. The cart can be pushed left or right, and the goal is to balance the inverted pendulum on the top of the cart by applying forces on the cart. The observation space of InvertedPendulum-v4 has 4 dimensions: position of the cart along the linear surface, vertical angle of the pole on the cart, linear velocity of the cart, angular velocity of the pole on the cart. We ignore the position of the cart along the linear surface. The vertical angle of the pole on the cart is discretized to 10 intervals between $[-\inf, -0.2, -0.1, -0.05, -0.025, 0, 0.025, 0.05, 0.1, 0.2, \inf]$. The linear velocity of the cart is discretized to 4 intervals between $[-\inf, -0.05, 0, 0.05, \inf]$. The angular velocity of the pole on the cart is discretized to 4 intervals between $[-\inf, -0.05, 0, 0.05, \inf]$. A reward of $+1$ is awarded for each time step that the inverted pendulum stand upright within a certain angle limit. The fixed adversarial policy in the inverted pendulum environment is a force of 0.5 N in the left direction.

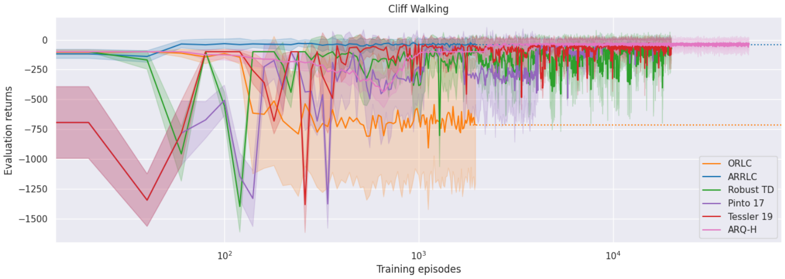
Cliff walking. The cliff walking experiment is a classic scenario proposed in Sutton & Barto (2018). The game starts with the player at location $[3,0]$ of the 4×12 grid world with the goal located at $[3,11]$. A cliff runs along $[3, 1-10]$. If the player moves to a cliff location, it returns to the start location and receives a reward of -100. For every move which does not lead into the cliff, the agent receives a reward of -1. The player makes moves until they reach the goal. The fixed adversarial policy in the cliff walking environment is walking a step to the bottom.

To show the robustness, we compare our algorithm with a non-robust RL algorithm that is ORLC (Optimistic Reinforcement Learning with Certificates) in Dann et al. (2019). It can be seen that overall our ARRLC algorithm achieves a much higher reward than the ORLC algorithm. This demonstrates the robustness of our ARRLC algorithm to policy execution uncertainty. To show the efficiency, we compare our algorithm with the robust TD algorithm in Klima et al. (2019), which can converge to the optimal robust policy but has no theoretical guarantee on sample complexity or regret, the approaches in Pinto et al. (2017) and Tessler et al. (2019), which model the robust problem as a zero-sum game and alternating update the agent policy and adversary policy. In our implementation, Pinto et al. (2017) fixes one policy and updates another for 25 episodes, then alternatively updates another in the next 25 episodes. (Tessler et al., 2019) does not alternate the updating until the current policy is converged. Figure 1 shows the efficiency of our ARRLC algorithm. ARRLC algorithm is more stable than the other algorithms. Compared with our model-free ARQ-H algorithm, the model-based ARRLC algorithm is more stable and more efficient.

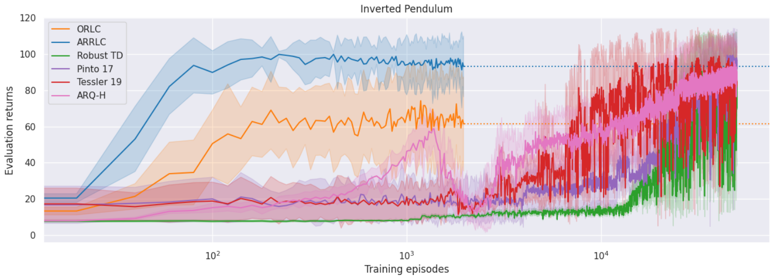
We provide additional experimental results in Appendix A. In these additional experiments, we implement the ablation study by setting different ρ and p and try different adversary policies in the testing environment.



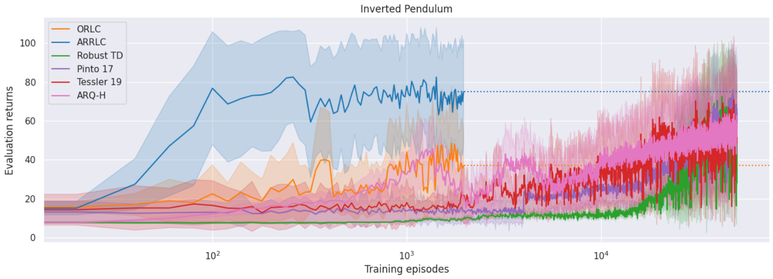
(a) $p = 0.1$



(b) $p = 0.2$



(c) $p = 0.1$



(d) $p = 0.2$

Figure 1: Evaluation returns during training. We plot the accumulated reward of each algorithms under different p and different environment. We set $\rho = 0.2$ for our algorithm, which is the uncertainty parameter used during training.

We also perform cross-comparison experiments in which we use the learned worst-case policies to disturb the different robust agents.

9 Conclusion and Discussion

In this paper, we have developed a novel approach for solving action robust RL problems with probabilistic policy execution uncertainty. We have theoretically proved the sample complexity bound and the regret bound of the algorithms. The upper bound of the sample complexity and the regret of proposed ARRLC algorithm match the lower bound up to logarithmic factors, which shows the minimax optimality of our algorithm. Moreover, we have carried out numerical experiments to validate our algorithm's robustness and efficiency, revealing that ARRLC surpasses non-robust algorithms and converges more rapidly than the robust TD algorithm when faced with action perturbations.

The current theoretical guarantee on the sample complexity and regret of our algorithms are derived for the tabular setting. In the future work, we will explore action robust RL in continuous state or action space. Studying efficient action robust RL with function approximation is also an important direction to pursue. For this purpose, two insights from our work might be useful: (1) The adversary policy and the agent's policy can be simultaneously updated to efficiently sample trajectories; (2) The adversary policies at each episode do not necessarily need be the minimum over the actions, an approximation of the minimum also works. Based on these insights, a policy-gradient method could potentially be designed to handle the continuous action space. We could use policy gradient method, such like PPO, to find an approximation of the adversary policy (the minimum over actions). In addition, similar to (Zhou et al., 2023), considering a scalable uncertainty set is also an interesting direction.

References

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272, 2017.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pp. 1507–1516. PMLR, 2019.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Kaiyang Guo, Shao Yunfeng, and Yanhui Geng. Model-based offline reinforcement learning with pessimism-modulated dynamics belief. *Advances in Neural Information Processing Systems*, 35:449–461, 2022.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for discounted mdps. *Advances in Neural Information Processing Systems*, 34:22288–22300, 2021.

Yunhan Huang and Quanyan Zhu. Deceptive reinforcement learning under adversarial manipulations on cost signals. In *Decision and Game Theory for Security: 10th International Conference, GameSec 2019, Stockholm, Sweden, October 30–November 1, 2019, Proceedings 10*, pp. 217–237. Springer, 2019.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, may 2005.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49:209–232, 2002.

Richard Klima, Daan Bloembergen, Michael Kaisers, and Karl Tuyls. Robust temporal difference learning for critical domains. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, pp. 350–358, 2019.

Erwan Lecarpentier and Emmanuel Rachelson. Non-stationary markov decision processes, a worst-case approach using model-based reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous q-learning: Sharper analysis and variance reduction. *Advances in neural information processing systems*, 33: 7031–7043, 2020.

Guanlin Liu and Lifeng Lai. Provably efficient black-box action poisoning attacks against reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pp. 7001–7010. PMLR, 2021.

Xiao-Yang Liu, Hongyang Yang, Qian Chen, Runjia Zhang, Liuqing Yang, Bowen Xiao, and Christina Dan Wang. Finrl: A deep reinforcement learning library for automated stock trading in quantitative finance. *Deep RL Workshop, NeurIPS 2020*, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

MohammadReza Nazari, Afshin Oroojlooy, Lawrence Snyder, and Martin Takac. Reinforcement learning for solving the vehicle routing problem. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Matthew O' Kelly, Aman Sinha, Hongseok Namkoong, Russ Tedrake, and John C Duchi. Scalable end-to-end autonomous vehicle testing via rare-event simulation. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pp. 9582–9602. PMLR, 2022.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pp. 2817–2826, 2017.

Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *arXiv preprint arXiv:2208.05767*, 2022.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 79903–79917, 2023.

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32, 2019.

Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. Pac model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pp. 881–888, 2006.

Yanchao Sun, Da Huo, and Furong Huang. Vulnerability-aware poisoning mechanism for online rl with unknown dynamics. In *International Conference on Learning Representations*, 2021.

Yanchao Sun, Ruijie Zheng, Yongyuan Liang, and Furong Huang. Who is the strongest enemy? towards optimal and efficient evasion attacks in deep RL. In *International Conference on Learning Representations*, 2022.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction.* MIT press, 2018.

Aviv Tamar, Shie Mannor, and Huan Xu. Scaling up robust mdps using function approximation. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pp. 181–189, 2014.

Chen Tessler, Yonathan Efroni, and Shie Mannor. Action robust reinforcement learning and applications in continuous control. In *International Conference on Machine Learning*, pp. 6215–6224. PMLR, 2019.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34:7193–7206, 2021.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Zaiyan Xu, Kishan Panaganti, and Dileep Kalathil. Improved sample complexity bounds for distributionally robust reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 9728–9754. PMLR, 2023.

Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Toward theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6):3223–3248, 2022.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems*, 33:21024–21037, 2020a.

Xuezhou Zhang, Yuzhe Ma, Adish Singla, and Xiaojin Zhu. Adaptive reward-poisoning attacks against reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 11225–11234, 2020b.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33:15198–15207, 2020c.

Zihan Zhang, Xiangyang Ji, and Simon Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pp. 4528–4531. PMLR, 2021.

Xiangyu Zhao, Long Xia, Liang Zhang, Zhuoye Ding, Dawei Yin, and Jiliang Tang. Deep reinforcement learning for page-wise recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pp. 95–103. ACM, 2018.

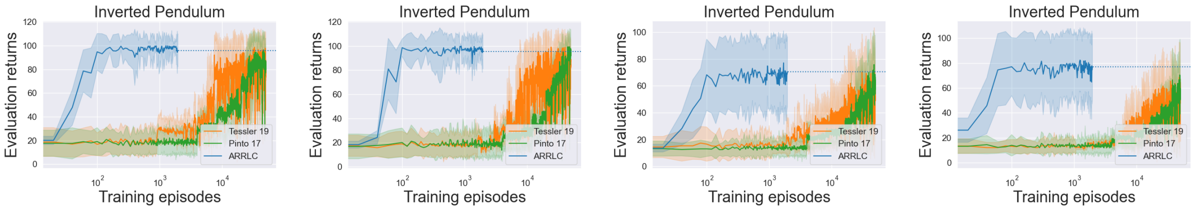
Ruida Zhou, Tao Liu, Min Cheng, Dileep Kalathil, P. R. Kumar, and Chao Tian. Natural actor-critic for robust reinforcement learning with function approximation, 2023.

Zhengqing Zhou, Zhengyuan Zhou, Qinxun Bai, Linhai Qiu, Jose Blanchet, and Peter Glynn. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3331–3339, 13–15 Apr 2021.

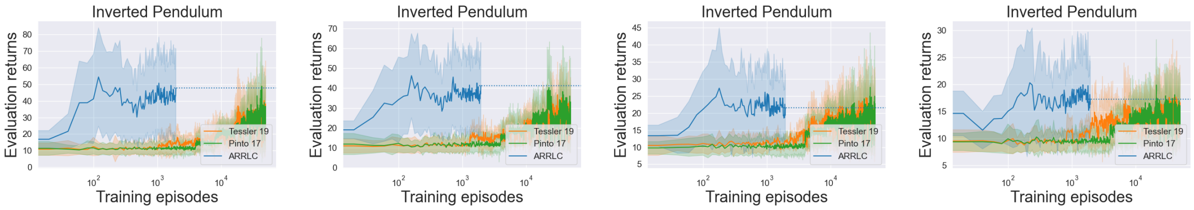
A Additional Numerical Results

A.1 Ablation study and more comparisons

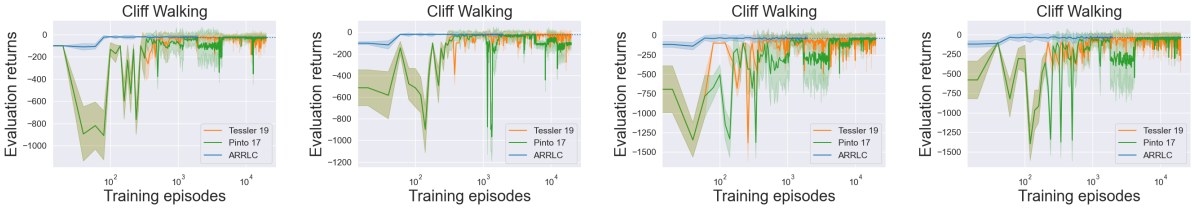
In Section 8, we compared our algorithm with the robust TD algorithm in (Klima et al., 2019). Here, we compare our algorithm with algorithms in (Pinto et al., 2017; Tessler et al., 2019). The method in (Tessler et al., 2019) requires an MDP solver to solve the optimal adversarial policy when the agent 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 in 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.



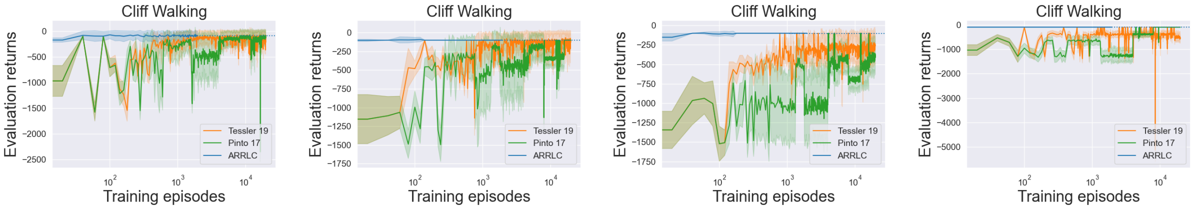
(a) $p = 0.1,\ \rho = 0.1$ (b) $p = 0.2,\ \rho = 0.1$ (c) $p = 0.2,\ \rho = 0.2$ (d) $p = 0.2,\ \rho = 0.3$



(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 2: ARRLC vs. RARL vs. PR-PI

In Figure 2, 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.

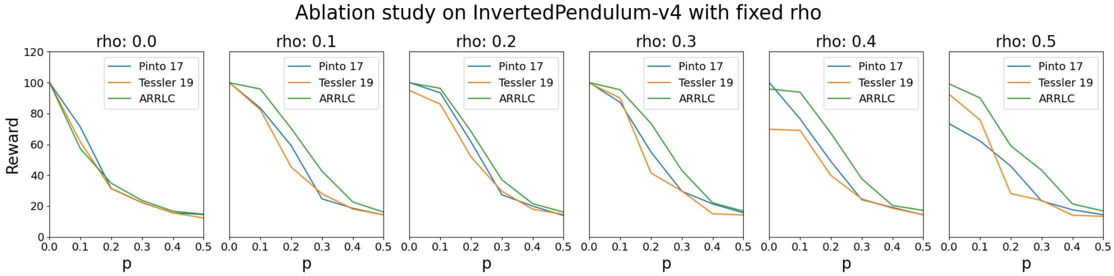


Figure 3: Ablation study on InvertedPendulum-v4 with fixed ρ.

In Figure 3, given the agents trained with fixed ρ (rho), we test the agents in different disturbed environments with different p. In Figure 4, we compare the different agents trained with different ρ. The x-axis is the different choice of ρ or p. The y-axis is the final evaluation rewards.

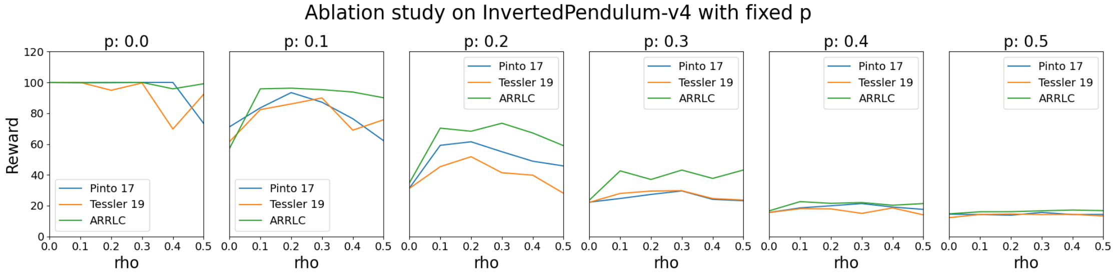


Figure 4: Ablation study on InvertedPendulum-v4 with fixed ρ.

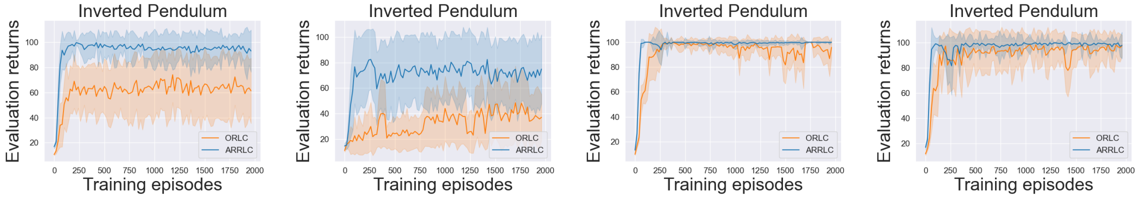
A.2 Robustness to different adversary

In this section, we consider 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 5 and Figure 6, "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.

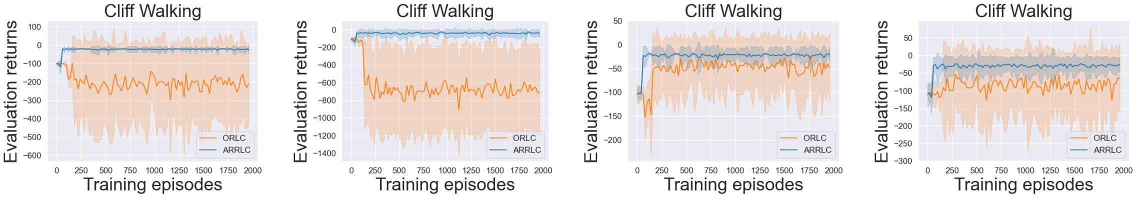
The theoretical guarantee on sample complexity and regret of our algorithm relies on the assumption of known uncertainty parameter. However, in the experimental results shown in Figure 3, the parameter can mismatch with the true disturb probability. In Figure 5, we test the mismatch of the uncertainty parameter ρ and true uncertainty probability p. We train 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.

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 use the learned worst-case policies to disturb the different robust agents. We report the final evaluation rewards in Table 2. We train 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.

We also show the performance of the non-robust approach orlc trained without action-failures, the non-robust approach when trained against the worst adversary learned by ARRLC, and ARRLC against the worst

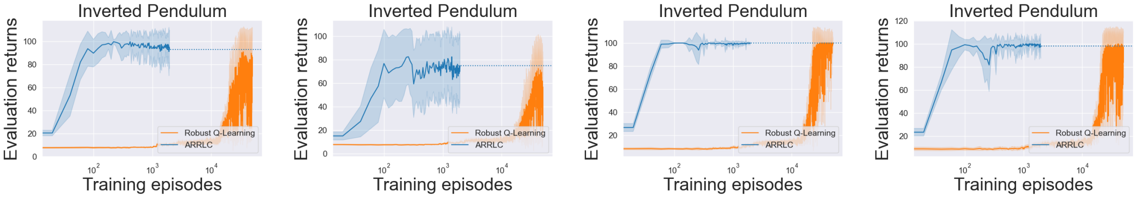


(a) p=0.1, fix (b) p=0.2, fix (c) p=0.1, random (d) p=0.2, random

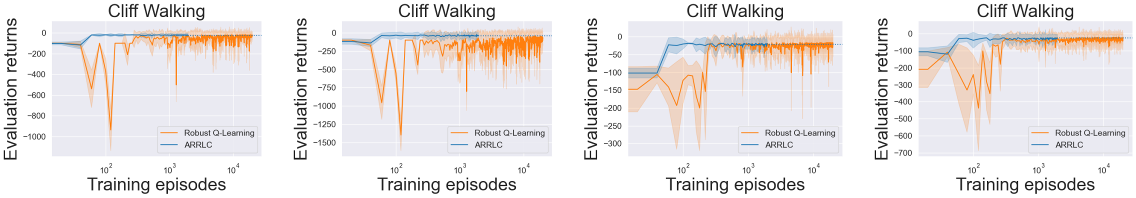


(e) p=0.1, fix (f) p=0.2, fix (g) p=0.1, random (h) p=0.2, random

Figure 5: ARRLC vs. 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 6: ARRLC vs. Robust TD

Table 2: 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

adversary learned by ARRLC. Figure 7 shows how much better ARRLC performs against the adversary than the the non-robust approach, and how far is the performance of ARRLC from the optimal. We set $\rho = 0.2$ and $p = 0.2$ in this experiment.

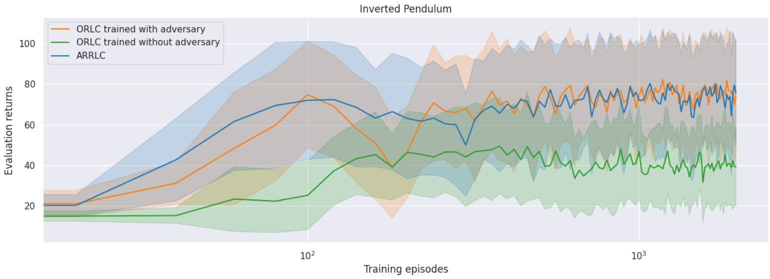


Figure 7: Evaluation returns during training. We compare the performance of the non-robust approach orlc trained without action-failures, the non-robust approach when trained against the worst adversary learned by ARRLC, and ARRLC against the worst adversary learned by ARRLC.

B Proof of Proposition 4.1

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

$$\Pi^{\rho}(\pi) := \{\widetilde{\pi} | \forall s, \widetilde{\pi}_h(\cdot|s) = (1-\rho)\pi(\cdot|s) + \rho\pi'_h(\cdot|s), \pi'_h(\cdot|s) \in \Delta_{\mathcal{A}}\}. \tag{17}$$

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 equation 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

$$\begin{aligned} V_H^{\pi}(s) &= \min_{\pi'} C_H^{\pi,\pi',\rho}(s) \\ &= (1-\rho)[\mathbb{D}_{\pi_H} R_H](s) + \rho \min_{\pi'} [\mathbb{D}_{\pi'_H} R_H](s) \\ &= (1-\rho)[\mathbb{D}_{\pi_H} Q_H^{\pi}](s) + \rho \min_{a \in \mathcal{A}} Q_H^{\pi}(s,a) = C_H^{\pi,\pi^-,\rho}(s), \end{aligned} \tag{18}$$

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 equation 2 and the form of uncertainty set, we have

$$
\begin{aligned}
Q_h^\pi(s,a) =& \min_{\widetilde{\pi} \in \Pi(\pi)} \mathbb{E}\left[\sum_{h'=h}^{H} R_{h'}(s_{h'}, a_{h'}) | s_h = s, a_h = a, a_{h'} \sim \widetilde{\pi}_{h'}(\cdot|s_{h'})\right] \\
=& \min_{\pi'} D_h^{\pi,\pi',\rho}(s,a) \\
=& R_h(s,a) + \min_{\pi'} \mathbb{E}_{s' \sim P_h(\cdot|s,a)} C_{h+1}^{\pi,\pi',\rho}(s) \\
=& R_h(s,a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)} \min_{\pi'} C_{h+1}^{\pi,\pi',\rho}(s) \\
=& R_h(s,a) + [P_h V_{h+1}^\pi](s,a).
\end{aligned} \tag{19}
$$

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 equation 1 and the form of uncertainty set, we have

$$
\begin{aligned}
V_h^\pi(s) =& \min_{\widetilde{\pi} \in \Pi(\pi)} \mathbb{E}\left[\sum_{h'=h}^{H} R_{h'}(s_{h'}, a_{h'}) | s_h = s, a_{h'} \sim \widetilde{\pi}_{h'}(\cdot|s_{h'})\right] \\
=& \min_{\pi'} C_h^{\pi,\pi',\rho}(s) \\
=& \min_{{\pi'}_h} \min_{\{{\pi'}_{h'}\}_{h'=h+1}^{H}} C_h^{\pi,\pi',\rho}(s) \\
\geq& (1-\rho) \min_{\{{\pi'}_{h'}\}_{h'=h+1}^{H}} \mathbb{E}_{a \sim \pi_h(\cdot|s)} D_h^{\pi,\pi',\rho}(s,a) + \rho \min_{{\pi'}_h} \min_{\{{\pi'}_{h'}\}_{h'=h+1}^{H}} \mathbb{E}_{a \sim \pi'_h(\cdot|s)} D_h^{\pi,\pi',\rho}(s,a) \\
\geq& (1-\rho) \mathbb{E}_{a \sim \pi_h(\cdot|s)} \min_{\{{\pi'}_{h'}\}_{h'=h+1}^{H}} D_h^{\pi,\pi',\rho}(s,a) + \rho \min_{{\pi'}_h} \mathbb{E}_{a \sim \pi'_h(\cdot|s)} \min_{\{{\pi'}_{h'}\}_{h'=h+1}^{H}} D_h^{\pi,\pi',\rho}(s,a) \\
=& (1-\rho)[\mathbb{D}_{\pi_h} Q_h^\pi](s) + \rho \min_{a \in \mathcal{A}} Q_h^\pi(s,a).
\end{aligned} \tag{20}
$$

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

$$
\begin{aligned}
V_h^\pi(s) \leq& C_h^{\pi,\pi^-,\rho}(s) \\
=& (1-\rho)[\mathbb{D}_{\pi_h} D_h^{\pi,\pi^-,\rho}](s) + \rho \min_{a \in \mathcal{A}} D_h^{\pi,\pi^-,\rho}(s,a) \\
=& (1-\rho)[\mathbb{D}_{\pi_h} Q_h^\pi](s) + \rho \min_{a \in \mathcal{A}} Q_h^\pi(s,a),
\end{aligned} \tag{21}
$$

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 equation 4 is equivalent to

$$\max_{\pi} \min_{\widetilde{\pi}\in\Pi^{\rho}(\pi)} \mathbb{E}\left[\sum_{h'=h}^{H} R_{h'}(s_{h'}, a_{h'})|s_h = s, a_{h'} \sim \widetilde{\pi}_{h'}(\cdot|s_{h'})\right] = \max_{\pi}\min_{\pi'} C_h^{\pi,\pi',\rho}(s). \tag{22}$$

The control problem in the RHS of equation 4 is equivalent to

$$\min_{\widetilde{\pi}\in\Pi^{\rho}(\pi)} \max_{\pi} \mathbb{E}\left[\sum_{h'=h}^{H} R_{h'}(s_{h'}, a_{h'})|s_h = s, a_{h'} \sim \widetilde{\pi}_{h'}(\cdot|s_{h'})\right] = \min_{\pi'}\max_{\pi} C_h^{\pi,\pi',\rho}(s). \tag{23}$$

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

$$\begin{aligned}\max_{\pi}\min_{\pi'} C_H^{\pi,\pi',\rho}(s) =&(1-\rho)\max_{\pi}[\mathbb{D}_{\pi_H} R_H](s) + \rho\min_{\pi'}[\mathbb{D}_{\pi'_H} R_H](s)\\ =&(1-\rho)\max_{a\in\mathcal{A}} R_H(s,a) + \rho\min_{b\in\mathcal{A}} R_H(s,b),\end{aligned} \tag{24}$$

and

$$\begin{aligned}\min_{\pi'}\max_{\pi} C_H^{\pi,\pi',\rho}(s) =&(1-\rho)\max_{\pi}[\mathbb{D}_{\pi_H} R_H](s) + \rho\min_{\pi'}[\mathbb{D}_{\pi'_H} R_H](s)\\ =&(1-\rho)\max_{a\in\mathcal{A}} R_H(s,a) + \rho\min_{b\in\mathcal{A}} R_H(s,b).\end{aligned} \tag{25}$$

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

$$\begin{aligned}&\max_{\pi}\min_{\pi'}\sum_{s\in\mathcal{S}} w(s)C_H^{\pi,\pi',\rho}(s)\\ =&(1-\rho)\max_{\pi}\sum_{s\in\mathcal{S}} w(s)[\mathbb{D}_{\pi_H} R_H](s) + \rho\min_{\pi'}\sum_{s\in\mathcal{S}} w(s)[\mathbb{D}_{\pi'_H} R_H](s)\\ =&(1-\rho)\sum_{s\in\mathcal{S}} w(s)\max_{a\in\mathcal{A}} R_H(s,a) + \rho\sum_{s\in\mathcal{S}} w(s)\min_{b\in\mathcal{A}} R_H(s,b).\end{aligned} \tag{26}$$

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 $V_{h'}^*(s) = V_{h'}^{\pi^*}(s) = C_{h'}^{\pi^*,\pi^-,\rho}(s)$ and $Q_{h'}^*(s,a) = Q_{h'}^{\pi^*}(s,a) = D_{h'}^{\pi^*,\pi^-,\rho}(s,a)$ 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

$$\begin{aligned}Q_h^*(s,a) &= \max_{\pi}\min_{\pi'} D_h^{\pi,\pi',\rho}(s,a)\\ &= \max_{\pi}\min_{\pi'}(R_h(s,a) + [P_h C_{h+1}^{\pi,\pi',\rho}](s,a))\\ &= R_h(s,a) + [P_h(\max_{\pi}\min_{\pi'} C_{h+1}^{\pi,\pi',\rho})](s,a)\\ &= R_h(s,a) + [P_h V_{h+1}^*](s,a).\end{aligned} \tag{27}$$

and also $Q_h^*(s,a) = Q_h^{\pi^*}(s,a) = D_h^{\pi^*,\pi^-,\rho}(s,a)$.

From the robust Bellman equation, we have

$$\begin{aligned}\max_\pi V_h^\pi(s) =&\max_\pi\left((1-\rho)[\mathbb{D}_{\pi_h}Q_h^\pi](s)+\rho\min_{a\in\mathcal{A}}Q_h^\pi(s,a)\right)\\ \leq&(1-\rho)\max_{\pi_h}\max_{\{\pi_h\}_{h'=h+1}^H}[\mathbb{D}_{\pi_h}Q_h^\pi](s)+\rho\max_{\{\pi_h\}_{h'=h+1}^H}\min_{a\in\mathcal{A}}Q_h^\pi(s,a)\\ \leq&(1-\rho)\max_{\pi_h}\max_{\{\pi_h\}_{h'=h+1}^H}[\mathbb{D}_{\pi_h}Q_h^\pi](s)+\rho\min_{a\in\mathcal{A}}\max_{\{\pi_h\}_{h'=h+1}^H}Q_h^\pi(s,a)\\ \leq&(1-\rho)\max_{\pi_h}[\mathbb{D}_{\pi_h}Q_h^*](s)+\rho\min_{a\in\mathcal{A}}Q_h^*(s,a)\\ =&(1-\rho)\max_{a\in\mathcal{A}}Q_h^*(s,a)+\rho\min_{a\in\mathcal{A}}Q_h^*(s,a).\end{aligned}\tag{28}$$

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

$$\begin{aligned}\max_\pi V_h^\pi(s)\geq V_h^{\pi^*}(s) &= (1-\rho)[\mathbb{D}_{\pi_h^*}Q_h^{\pi^*}](s)+\rho\min_{a\in\mathcal{A}}Q_h^{\pi^*}(s,a)\\ &= (1-\rho)\max_{a\in\mathcal{A}}Q_h^{\pi^*}(s,a)+\rho\min_{a\in\mathcal{A}}Q_h^{\pi^*}(s,a)\\ &= (1-\rho)\max_{a\in\mathcal{A}}Q_h^*(s,a)+\rho\min_{a\in\mathcal{A}}Q_h^*(s,a).\end{aligned}\tag{29}$$

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 $C_h^{\pi^*,\pi^-,\rho}(s) = V_h^*(s)$.

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

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

for any policy π and any weighted function w. Hence

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

First, we have

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

Then, we can show

$$\begin{aligned}\max_\pi\sum_s w(s)\min_{\pi'}C_h^{\pi,\pi',\rho}(s) \quad&\geq\quad \sum_s w(s)\min_{\pi'}C_h^{\pi^*,\pi',\rho}(s)\\ &=\quad \sum_s w(s)C_h^{\pi^*,\pi^-,\rho}(s)\\ &=\quad \sum_s w(s)\max_\pi\min_{\pi'}C_h^{\pi,\pi',\rho}(s).\end{aligned}\tag{30}$$

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

$$\max_\pi\min_{\pi'}C_h^{\pi,\pi',\rho}(s) = C_h^{\pi^*,\pi^-,\rho}(s) = \max_\pi C_h^{\pi,\pi^-,\rho}(s)\geq\min_{\pi'}\max_\pi C_h^{\pi,\pi',\rho}(s).\tag{31}$$

By induction on $h = H,\cdots,1$, we prove Proposition 4.1.

C Proof for Action Robust Reinforcement Learning with Certificates

In this section, we prove Theorem 5.1. 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 6.1

When $N_h^k(s,a) \leq 1$, equation 9, equation 10 and equation 8 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))

$$\left|(\hat{P}_h^k - P_h)V_{h+1}^*(s,a)\right| \leq \sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k} V_{h+1}^*(s,a)\iota}{N_h^k(s,a)}} + \frac{7H\iota}{3(N_h^k(s,a))}, \tag{32}$$

and

$$\left|(\hat{P}_h^k - P_h)V_{h+1}^{\overline{\pi}^k}(s,a)\right| \leq \sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k} V_{h+1}^{\overline{\pi}^k}(s,a)\iota}{N_h^k(s,a)}} + \frac{7H\iota}{3(N_h^k(s,a))}. \tag{33}$$

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

$$\left|\hat{r}_h^k(s,a) - R_h(s,a)\right| \leq \sqrt{\frac{2Var(r_h^k(s,a))\iota}{N_h^k(s,a)}} + \frac{7\iota}{3(N_h^k(s,a))} \leq \sqrt{\frac{2\hat{r}_h^k(s,a)\iota}{N_h^k(s,a)}} + \frac{7\iota}{3(N_h^k(s,a))}, \tag{34}$$

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 6.2

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

$$\begin{aligned}
&\hat{r}_h^k(s,a) + \hat{P}_h^k \overline{V}_{h+1}(s,a) + \theta_h^k(s,a) - Q_h^*(s,a) \\
=&\hat{r}_h^k(s,a) + \hat{P}_h^k \overline{V}_{h+1}(s,a) + \theta_h^k(s,a) - R_h(s,a) - P_h V_{h+1}^*(s,a) \\
=&\hat{r}_h^k(s,a) - R_h(s,a) + \hat{P}_h^k\left(\overline{V}_{h+1} - V_{h+1}^*\right)(s,a) + (\hat{P}_h^k - P_h)V_{h+1}^*(s,a) + \theta_h^k(s,a) \\
\geq&(\hat{P}_h^k - P_h)V_{h+1}^*(s,a) + \sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s,a)\iota}{N_h^k(s,a)}} + \frac{\hat{P}_h^k\left(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k\right)(s,a)}{H} + \frac{8H^2\iota}{N_h^k(s,a)} \qquad (35) \\
\geq&\sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s,a)\iota}{N_h^k(s,a)}} + \frac{\hat{P}_h^k\left(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k\right)(s,a)}{H} + \frac{8H^2\iota}{N_h^k(s,a)} - \sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k} V_{h+1}^*(s,a)\iota}{N_h^k(s,a)}},
\end{aligned}$$

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

$$\begin{aligned}
&\left|\mathbb{V}_{\hat{P}^k_h}\left(\frac{\overline{V}^k_{h+1}+\underline{V}^k_{h+1}}{2}\right)(s,a)-\mathbb{V}_{\hat{P}^k_h}V^*_{h+1}(s,a)\right| \\
\leq&\left|[\hat{P}^k_h(\overline{V}^k_{h+1}+\underline{V}^k_{h+1})/2]^2-(\hat{P}^k_hV^*_{h+1})^2\right|(s,a)+\left|\hat{P}^k_h[(\overline{V}^k_{h+1}+\underline{V}^k_{h+1})/2]^2-\hat{P}^k_h(V^*_{h+1})^2\right|(s,a) \\
\leq&4H\hat{P}^k_h\left|(\overline{V}^k_{h+1}+\underline{V}^k_{h+1})/2-V^*_{h+1}\right|(s,a) \\
\leq&2H\hat{P}^k_h\left(\overline{V}^k_{h+1}-\underline{V}^k_{h+1}\right)(s,a)
\end{aligned} \tag{36}$$

and

$$\begin{aligned}
&\sqrt{\frac{2\mathbb{V}_{\hat{P}^k_h}V^*_{h+1}(s,a)\iota}{N^k_h(s,a)}} \\
\leq&\sqrt{\frac{2\mathbb{V}_{\hat{P}^k_h}[(\overline{V}^k_{h+1}+\underline{V}^k_{h+1})/2](s,a)\iota+4H\hat{P}^k_h\left(\overline{V}^k_{h+1}-\underline{V}^k_{h+1}\right)(s,a)\iota}{N^k_h(s,a)}} \\
\leq&\sqrt{\frac{2\mathbb{V}_{\hat{P}^k_h}[(\overline{V}^k_{h+1}+\underline{V}^k_{h+1})/2](s,a)\iota}{N^k_h(s,a)}}+\sqrt{\frac{4H\hat{P}^k_h\left(\overline{V}^k_{h+1}-\underline{V}^k_{h+1}\right)(s,a)\iota}{N^k_h(s,a)}} \\
\leq&\sqrt{\frac{2\mathbb{V}_{\hat{P}^k_h}[(\overline{V}^k_{h+1}+\underline{V}^k_{h+1})/2](s,a)\iota}{N^k_h(s,a)}}+\frac{\hat{P}^k_h\left(\overline{V}^k_{h+1}-\underline{V}^k_{h+1}\right)(s,a)}{H}+\frac{8H^2\iota}{N^k_h(s,a)}.
\end{aligned} \tag{37}$$

Plugging equation 37 back into equation 35, we have $\hat{r}^k_h(s,a)+\hat{P}^k_h\overline{V}_{h+1}(s,a)+\theta^k_h(s,a)\geq Q^*_h(s,a)$. Thus, $\overline{Q}^k_h(s,a)=\min\{H-h+1,\hat{r}^k_h(s,a)+\hat{P}^k_h\overline{V}^k_{h+1}(s,a)+\theta^k_h(s,a)\}\geq Q^*_h(s,a)$.

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

$$\begin{aligned}
\overline{V}^k_h(s)=&(1-\rho)\overline{Q}^k_h(s,\overline{\pi}^k_h(s))+\rho\overline{Q}^k_h(s,\underline{\pi}^k_h(s)) \\
\geq&(1-\rho)\overline{Q}^k_h(s,\pi^*_h(s))+\rho Q^*_h(s,\underline{\pi}^k_h(s)) \\
\geq&(1-\rho)Q^*_h(s,\pi^*_h(s))+\rho\min_{a\in\mathcal{A}}Q^*_h(s,a)=V^*_h(s).
\end{aligned} \tag{38}$$

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

$$\begin{aligned}
&\hat{r}^k_h(s,a)+\hat{P}^k_h\underline{V}_{h+1}(s,a)-\theta^k_h(s,a)-Q^{\overline{\pi}^k}_h(s,a) \\
=&\hat{r}^k_h(s,a)+\hat{P}^k_h\underline{V}_{h+1}(s,a)-\theta^k_h(s,a)-R_h(s,a)-P_hV^{\overline{\pi}^k}_{h+1}(s,a) \\
=&\hat{r}^k_h(s,a)-R_h(s,a)+\hat{P}^k_h\left(\underline{V}_{h+1}-V^{\overline{\pi}^k}_{h+1}\right)(s,a)+(\hat{P}^k_h-P_h)V^{\overline{\pi}^k}_{h+1}(s,a)-\theta^k_h(s,a) \\
\leq&(\hat{P}^k_h-P_h)V^{\overline{\pi}^k}_{h+1}(s,a)-\sqrt{\frac{2\mathbb{V}_{\hat{P}^k_h}[(\overline{V}^k_{h+1}+\underline{V}^k_{h+1})/2](s,a)\iota}{N^k_h(s,a)}}-\frac{\hat{P}^k_h\left(\overline{V}^k_{h+1}-\underline{V}^k_{h+1}\right)(s,a)}{H}-\frac{8H^2\iota}{N^k_h(s,a)} \\
\leq&\sqrt{\frac{2\mathbb{V}_{\hat{P}^k_h}V^{\overline{\pi}^k}_{h+1}(s,a)\iota}{N^k_h(s,a)}}-\sqrt{\frac{2\mathbb{V}_{\hat{P}^k_h}[(\overline{V}^k_{h+1}+\underline{V}^k_{h+1})/2](s,a)\iota}{N^k_h(s,a)}}-\frac{\hat{P}^k_h\left(\overline{V}^k_{h+1}-\underline{V}^k_{h+1}\right)(s,a)}{H}-\frac{8H^2\iota}{N^k_h(s,a)}\leq 0,
\end{aligned} \tag{39}$$

and

$$
\begin{aligned}
\underline{V}_h^k(s) =&(1-\rho)\underline{Q}_h^k(s,\overline{\pi}_h^k(s)) + \rho\underline{Q}_h^k(s,\underline{\pi}_h^k(s)) \\
\le&(1-\rho)Q_h^{\overline{\pi}^k}(s,\overline{\pi}_h^k(s)) + \rho\min_{a\in\mathcal{A}}\underline{Q}_h^k(s,a) \\
\le&(1-\rho)Q_h^{\overline{\pi}^k}(s,\overline{\pi}_h^k(s)) + \rho\underline{Q}_h^k(s,\arg\min_{a\in\mathcal{A}}Q_h^{\overline{\pi}^k}(s,a)) \\
\le&(1-\rho)Q_h^{\overline{\pi}^k}(s,\overline{\pi}_h^k(s)) + \rho\min_{a\in\mathcal{A}}Q_h^{\overline{\pi}^k}(s,a) = V_h^{\overline{\pi}^k}(s).
\end{aligned}
\tag{40}
$$

C.2 Regret Analysis

C.2.1 Proof of Lemma 6.3

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 6.2, the regret can be bounded by $Regret(K) := \sum_{k=1}^K (V_1^*(s_1^k) - V_1^{\overline{\pi}^k}(s_1^k)) \le \sum_{k=1}^K (\overline{V}_1^k(s_1^k) - \underline{V}_1^k(s_1^k))$.

By the update steps in Algorithm 1, we have

$$
\begin{aligned}
&\overline{V}_h^k(s_h^k) - \underline{V}_h^k(s_h^k) \\
=&(1-\rho)\overline{Q}_h^k(s_h^k,\overline{\pi}_h^k(s_h^k)) + \rho\overline{Q}_h^k(s_h^k,\underline{\pi}_h^k(s_h^k)) - (1-\rho)\underline{Q}_h^k(s_h^k,\overline{\pi}_h^k(s_h^k)) - \rho\underline{Q}_h^k(s_h^k,\underline{\pi}_h^k(s_h^k)) \\
\le&[\mathbb{D}_{\widetilde{\pi}_h^k}\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k) + 2\mathbb{D}_{\widetilde{\pi}_h^k}\theta_h(s_h^k) \\
=&[\mathbb{D}_{\widetilde{\pi}_h^k}\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k) - [\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k,a_h^k) + 2\mathbb{D}_{\widetilde{\pi}_h^k}\theta_h(s_h^k) + [\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k,a_h^k) \\
=&[\mathbb{D}_{\widetilde{\pi}_h^k}\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k) - [\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k,a_h^k) + 2\mathbb{D}_{\widetilde{\pi}_h^k}\theta_h(s_h^k) \\
&+ [\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k,a_h^k) - c_1P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k,a_h^k) \\
&+ c_1P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k,a_h^k) - c_2(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k) + c_2(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k).
\end{aligned}
\tag{41}
$$

According to the definition of $\theta_h(s,a)$,

$$
\begin{aligned}
&\overline{V}_h^k(s_h^k) - \underline{V}_h^k(s_h^k) \\
=&[\mathbb{D}_{\widetilde{\pi}_h^k}\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k) - [\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k,a_h^k) \\
&+ [\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k,a_h^k) - c_1P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k,a_h^k) \\
&+ c_1P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k,a_h^k) - c_2(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k) + c_2(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k) \\
&+ 2(1-\rho)\sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s_h^k,\overline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))}} + 2(1-\rho)\sqrt{\frac{2\hat{r}_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))}} \\
&+ (1-\rho)\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k,\overline{\pi}_h^k(s_h^k))/H + \frac{2(1-\rho)(24H^2+7H+7)\iota}{3N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k)))} \\
&+ 2\rho\sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s_h^k,\underline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k,\underline{\pi}_h^k(s_h^k))}} + 2\rho\sqrt{\frac{2\hat{r}_h^k(s_h^k,\underline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k,\underline{\pi}_h^k(s_h^k))}} \\
&+ \rho\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k,\underline{\pi}_h^k(s_h^k))/H + \frac{2\rho(24H^2+7H+7)\iota}{3N_h^k(s_h^k,\underline{\pi}_h^k(s_h^k)))}.
\end{aligned}
\tag{42}
$$

We noticed that

$$
\mathbb{D}_{\widetilde{\pi}_h^k}\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k) = (1-\rho)\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k,\overline{\pi}_h^k(s_h^k)) + \rho\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k,\underline{\pi}_h^k(s_h^k)). \tag{43}
$$

We can reorganize the equation 42 to

$$
\begin{aligned}
&\overline{V}_h^k(s_h^k) - \underline{V}_h^k(s_h^k) \\
=&(1+1/H)[\mathbb{D}_{\widetilde{\pi}_h^k}\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k) - (1+1/H)[\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k, a_h^k) \\
&+ \underbrace{(1+1/H)[\hat{P}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k, a_h^k) - c_1 P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k)}_{(a)} \\
&+ c_1 P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k) - c_2(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k) + c_2(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k) \\
&+ 2(1-\rho)\underbrace{\sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s_h^k, \overline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k, \overline{\pi}_h^k(s_h^k))}}}_{(b1)} + 2(1-\rho)\sqrt{\frac{2\hat{r}_h^k(s_h^k, \overline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k, \overline{\pi}_h^k(s_h^k))}} \\
&+ \frac{2(1-\rho)(24H^2+7H+7)\iota}{3N_h^k(s_h^k, \overline{\pi}_h^k(s_h^k)))} + 2\rho\underbrace{\sqrt{\frac{2\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s_h^k, \underline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k, \underline{\pi}_h^k(s_h^k))}}}_{(b2)} \\
&+ 2\rho\sqrt{\frac{2\hat{r}_h^k(s_h^k, \underline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k, \underline{\pi}_h^k(s_h^k))}} + \frac{2\rho(24H^2+7H+7)\iota}{3N_h^k(s_h^k, \underline{\pi}_h^k(s_h^k)))}.
\end{aligned} \tag{44}
$$

Bound of the error of the empirical probability estimator (a) Here, we set $c_1 = 1 + 1/H$. By Bennett's inequality, we have that w.p. $1 - S\delta$

$$
|\hat{P}_h^k(s'|s,a) - P_h(s'|s,a)| \leq \sqrt{\frac{2P_h(s'|s,a)\iota}{N_h^k(s,a)}} + \frac{\iota}{3N_h^k(s,a)} \tag{45}
$$

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

Thus, we have that

$$
\begin{aligned}
&(\hat{P}_h^k - P_h)(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s,a) \\
=&\sum_{s'}(\hat{P}_h^k(s'|s,a) - P_h(s'|s,a))(\overline{V}_{h+1}^k(s') - \underline{V}_{h+1}^k(s')) \\
\leq&\sum_{s'}\sqrt{\frac{2P_h(s'|s,a)\iota}{N_h^k(s,a)}}(\overline{V}_{h+1}^k(s') - \underline{V}_{h+1}^k(s')) + \frac{SH\iota}{3N_h^k(s,a)} \\
\leq&\sum_{s'}\left(\frac{P_h(s'|s,a)\iota}{H} + \frac{H}{2N_h^k(s,a)}\right)\left(\overline{V}_{h+1}^k(s') - \underline{V}_{h+1}^k(s')\right) + \frac{SH\iota}{3N_h^k(s,a)} \\
\leq&P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s,a)/H + \frac{SH^2}{2N_h^k(s,a)} + \frac{SH\iota}{3N_h^k(s,a)} \\
\leq&P_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s,a)/H + \frac{SH^2\iota}{N_h^k(s,a)},
\end{aligned} \tag{46}
$$

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 $\mathcal{E}^R \cap \mathcal{E}^{PV}$,

$$\begin{aligned}
&\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s,a) - \mathbb{V}_{P_h} C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s,a) \\
=&\hat{P}_h^k[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2]^2(s,a) - [\hat{P}_h^k(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2]^2(s,a) \\
&- P_h(C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho})^2(s,a) + (P_h C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho})^2(s,a) \\
\leq&[\hat{P}_h^k(\overline{V}_{h+1}^k)^2 - (\hat{P}_h^k\underline{V}_{h+1}^k)^2 - P_h(\underline{V}_{h+1}^k)^2 + (P_h\overline{V}_{h+1}^k)^2](s,a) \\
\leq&|(\hat{P}_h^k - P_h)(\overline{V}_{h+1}^k)^2|(s,a) + |(P_h\underline{V}_{h+1}^k)^2 - (\hat{P}_h^k\underline{V}_{h+1}^k)^2|(s,a) \\
&+ P_h|(\overline{V}_{h+1}^k)^2 - (\underline{V}_{h+1}^k)^2|(s,a) + |(P_h\overline{V}_{h+1}^k)^2 - (P_h\underline{V}_{h+1}^k)^2|(s,a),
\end{aligned} \tag{47}$$

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 $1-\delta$

$$\|\hat{P}_h^k(\cdot|s,a) - P_h(\cdot|s,a)\|_1 \leq \sqrt{\frac{2S\iota}{N_h^k(s,a)}} \tag{48}$$

holds for all s,a,h,k.

These terms can be bounded separately by

$$\begin{aligned}
&|(\hat{P}_h^k - P_h)(\overline{V}_{h+1}^k)^2|(s,a) \leq H^2\sqrt{\frac{2S\iota}{N_h^k(s,a)}}, \\
&|(P_h\underline{V}_{h+1}^k)^2 - (\hat{P}_h^k\underline{V}_{h+1}^k)^2|(s,a) \leq 2H|(P_h - \hat{P}_h^k)\underline{V}_{h+1}^k| \leq 2H^2\sqrt{\frac{2S\iota}{N_h^k(s,a)}}, \\
&P_h|(\overline{V}_{h+1}^k)^2 - (\underline{V}_{h+1}^k)^2|(s,a) \leq 2HP_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s,a), \\
&|(P_h\overline{V}_{h+1}^k)^2 - (P_h\underline{V}_{h+1}^k)^2|(s,a) \leq 2HP_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s,a),
\end{aligned} \tag{49}$$

where the first two inequality is due to equation 48. In addition, $3H^2\sqrt{\frac{2S\iota}{N_h^k(s,a)}} \leq 1 + \frac{9SH^4\iota}{2N_h^k(s,a)}$. Thus, we have

$$\begin{aligned}
&\mathbb{V}_{\hat{P}_h^k}[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s,a) - \mathbb{V}_{P_h} C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s,a) \\
\leq&4HP_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s,a) + 3H^2\sqrt{\frac{2S\iota}{N_h^k(s,a)}} \\
\leq&4HP_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s,a) + 1 + \frac{9SH^4\iota}{2N_h^k(s,a)}.
\end{aligned} \tag{50}$$

Then 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))}} \\
\leq&(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))} \\
\leq&(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}_{\widetilde{\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{51}
$$

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

Recursing on h Plugging equation 46 and equation 51 into equation 44and setting $c_1 = 1 + 1/H$ and $c_2 = (1 + 1/H)^3$, we have

$$\begin{aligned}
&\overline{V}^k_h(s^k_h) - \underline{V}^k_h(s^k_h) \\
\leq&(1+1/H)[\mathbb{D}_{\widetilde{\pi}^k_h}\hat{P}^k_h(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})](s^k_h) - (1+1/H)[\hat{P}^k_h(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})](s^k_h, a^k_h) \\
&+ (1/H + 1/H^2)P_h(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h, a^k_h) + \frac{(SH + SH^2)\iota}{N^k_h(s^k_h, a^k_h)} \\
&+ c_1 P_h(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h, a^k_h) - c_2(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_{h+1}) + c_2(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_{h+1}) \\
&+ 2(1-\rho)\sqrt{\frac{2\hat{r}^k_h(s^k_h, \overline{\pi}^k_h(s^k_h))\iota}{N^k_h(s^k_h, \overline{\pi}^k_h(s^k_h))}} + \frac{2(1-\rho)(24H^2 + 7H + 7)\iota}{3N^k_h(s^k_h, \overline{\pi}^k_h(s^k_h)))} \\
&+ 2\rho\sqrt{\frac{2\hat{r}^k_h(s^k_h, \underline{\pi}^k_h(s^k_h))\iota}{N^k_h(s^k_h, \underline{\pi}^k_h(s^k_h))}} + \frac{2\rho(24H^2 + 7H + 7)\iota}{3N^k_h(s^k_h, \underline{\pi}^k_h(s^k_h)))} \qquad (52) \\
&+ (1-\rho)\sqrt{\frac{8\mathbb{V}_{P_h} C^{\pi^{k*}, \underline{\pi}^k, \rho}_{h+1}(s^k_h, \overline{\pi}^k_h(s^k_h))\iota}{N^k_h(s^k_h, \overline{\pi}^k_h(s^k_h))}} + \rho\sqrt{\frac{8\mathbb{V}_{P_h} C^{\pi^{k*}, \underline{\pi}^k, \rho}_{h+1}(s^k_h, \underline{\pi}^k_h(s^k_h))\iota}{N^k_h(s^k_h, \underline{\pi}^k_h(s^k_h))}} \\
&+ \frac{\mathbb{D}_{\widetilde{\pi}^k_h} P_h(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h)}{H} + \frac{8(1-\rho)H^2\iota}{N^k_h(s^k_h, \overline{\pi}^k_h(s^k_h))} + \frac{8\rho H^2\iota}{N^k_h(s^k_h, \underline{\pi}^k_h(s^k_h))} \\
&+ (1-\rho)\sqrt{\frac{8}{N^k_h(s^k_h, \overline{\pi}^k_h(s^k_h))}} + \rho\sqrt{\frac{8}{N^k_h(s^k_h, \underline{\pi}^k_h(s^k_h))}} + \frac{6(1-\rho)\sqrt{SH^4\iota}}{N^k_h(s^k_h, \overline{\pi}^k_h(s^k_h))} + \frac{6\rho\sqrt{SH^4\iota}}{N^k_h(s^k_h, \underline{\pi}^k_h(s^k_h))}.
\end{aligned}$$

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

$$\begin{aligned}
&\overline{V}^k_h(s^k_h) - \underline{V}^k_h(s^k_h) \\
\leq&(1+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) - (1+1/H)[\hat{P}^k_h(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})](s^k_h, a^k_h) \\
&+ (1/H + 1/H^2)P_h(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h, a^k_h) + \frac{(SH + SH^2)\iota}{N^k_h(s^k_h, a^k_h)} \\
&+ c_1 P_h(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h, a^k_h) - c_2(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_{h+1}) + c_2(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_{h+1}) \\
&+ \frac{\mathbb{D}_{\widetilde{\pi}^k_h} P_h(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h, \overline{\pi}^k_h(s^k_h))}{H} + \mathbb{D}_{\widetilde{\pi}^k_h}\Theta^k_h(s^k_h) \\
\leq&(1+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) - (1+1/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)] \qquad (53) \\
&+ (1 + 3/H + 1/H^2)P_h(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h, a^k_h) - c_2(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_{h+1}) \\
&+ c_2(\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) \\
\leq&(1+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) - (1+1/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)] \\
&+ c_2 P_h(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_h, a^k_h) - c_2(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s^k_{h+1}) \\
&+ c_2(\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}$$

By induction of equation 44 on $h = 1, \cdots, H$ and $\overline{V}_{h+1}^k = \underline{V}_{h+1}^k = 0$, we have that

$$\begin{aligned}
Regret(K) \leq 21 \sum_{k=1}^{K} \sum_{h=1}^{H} (&\mathbb{D}_{\widetilde{\pi}_h^k} \hat{P}_h^k (\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k) - \hat{P}_h^k (\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k) \\
&+ \frac{1}{H} [\mathbb{D}_{\widetilde{\pi}_h^k} P_h (\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k) - P_h (\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k)] \\
&+ P_h (\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k) - (\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k) \\
&+ \frac{(SH + SH^2)\iota}{N_h^k(s_h^k, a_h^k)} + \mathbb{D}_{\widetilde{\pi}_h^k} \Theta_h^k(s_h^k)).
\end{aligned} \tag{54}$$

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

C.2.2 Proof of Lemma 6.4

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

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

$$\left| \sum_{k=1}^{K} \sum_{h=1}^{H} [\mathbb{D}_{\widetilde{\pi}_h^k} \hat{P}_h^k (\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k) - \hat{P}_h^k (\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k)] \right| \leq H\sqrt{2HK\iota}. \tag{55}$$

C.2.3 Proof of Lemma 6.5

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

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

$$\left| \sum_{k=1}^{K} \sum_{h=1}^{H} [\mathbb{D}_{\widetilde{\pi}_h^k} P_h (\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k) - P_h (\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k)] \right| \leq H\sqrt{2HK\iota}. \tag{56}$$

C.2.4 Proof of Lemma 6.6

Recall that $M_3 = \sum_{k=1}^K \sum_{h=1}^H (P_h^k (\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k) - (\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k))$.

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

$$\left| \sum_{k=1}^{K} \sum_{h=1}^{H} [(P_h^k - \hat{\mathbb{1}}_h^k)(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s_h^k, a_h^k) \right| \leq H\sqrt{2HK\iota}. \tag{57}$$

C.2.5 Proof of Lemma 6.7

We bounded $M_4 = \sum_{k=1}^K \sum_{h=1}^H [\frac{(SH+SH^2)\iota}{N_h^k(s_h^k, a_h^k)} + \mathbb{D}_{\widetilde{\pi}_h^k} \Theta_h^k(s_h^k)]$ 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.

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{(SH+SH^2)\iota}{N_h^k(s_h^k,a_h^k)}=(SH+SH^2)\iota\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{n=1}^{N_h^K(s,a)}\frac{1}{n}\le(SH+SH^2)SAH\iota^2. \tag{58}$$

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

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{32\iota}{N_h^k(s_h^k,\pi(s_h^k))}}=\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{n=1}^{N_h^K(s,a)}\sqrt{\frac{32\iota}{n}}\le 8H\sqrt{SAK\iota}. \tag{59}$$

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

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{46\sqrt{SH^4\iota}}{N_h^k(s_h^k,\pi(s_h^k))}=46\sqrt{SH^4\iota}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{n=1}^{N_h^K(s,a)}\frac{1}{n}\le 46S^{\frac{3}{2}}AH^3\iota^2. \tag{60}$$

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,

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\overline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))}}\\
\le&\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\overline{\pi}_h^k(s_h^k))\cdot\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{\iota}{N_h^k(s_h^k,\overline{\pi}_h^k(s_h^k))}}\\
\le&\sqrt{SAH\iota^2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\overline{\pi}_h^k(s_h^k))}.
\end{aligned} \tag{61}$$

Similarly,

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\underline{\pi}_h^k(s_h^k))\iota}{N_h^k(s_h^k,\underline{\pi}_h^k(s_h^k))}}\\
\le&\sqrt{SAH\iota^2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\underline{\pi}_h^k(s_h^k))}.
\end{aligned} \tag{62}$$

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

$$\begin{aligned}
&(1-\rho)\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\overline{\pi}_h^k(s_h^k))}+\rho\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\underline{\pi}_h^k(s_h^k))}\\
\le&\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}[(1-\rho)\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\overline{\pi}_h^k(s_h^k))+\rho\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\underline{\pi}_h^k(s_h^k))]}.
\end{aligned} \tag{63}$$

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

$$[\mathbb{D}_{\widetilde{\pi}^k_h}P_hV_{h+1}](s) = \sum_{s'}[(1-\rho)P_h(s'|s,\overline{\pi}^k_h(s)) + \rho P_h(s'|s,\underline{\pi}^k_h(s))]V_{h+1}(s'), \tag{64}$$

and

$$\begin{aligned}\mathbb{V}_{[\mathbb{D}_{\widetilde{\pi}^k_h}P_h]}V_{h+1}(s) = &\sum_{s'}[(1-\rho)P_h(s'|s,\overline{\pi}^k_h(s)) + \rho P_h(s'|s,\underline{\pi}^k_h(s))][V_{h+1}(s')]^2 \\ &- [\sum_{s'}\left((1-\rho)P_h(s'|s,\overline{\pi}^k_h(s)) + \rho P_h(s'|s,\underline{\pi}^k_h(s))\right)V_{h+1}(s')]^2.\end{aligned} \tag{65}$$

We have that

$$\begin{aligned}&\mathbb{V}_{[\mathbb{D}_{\widetilde{\pi}^k_h}P_h]}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k) \\ =&\sum_{s'}[(1-\rho)P_h(s'|s_h^k,\overline{\pi}^k_h(s_h^k)) + \rho P_h(s'|s_h^k,\underline{\pi}^k_h(s_h^k))][C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s')]^2 \\ &- [\sum_{s'}\left((1-\rho)P_h(s'|s_h^k,\overline{\pi}^k_h(s_h^k)) + \rho P_h(s'|s_h^k,\underline{\pi}^k_h(s_h^k))\right)C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s')]^2 \\ \geq&(1-\rho)\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\overline{\pi}^k_h(s_h^k)) + \rho\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\underline{\pi}^k_h(s_h^k)) \\ &+ (1-\rho)[P_hC_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\overline{\pi}^k_h(s_h^k))]^2 + \rho P_h[C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\underline{\pi}^k_h(s_h^k))]^2 \\ &- [\sum_{s'}(1-\rho)P_h(s'|s_h^k,\overline{\pi}^k_h(s_h^k))C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s') + \rho P_h(s'|s_h^k,\underline{\pi}^k_h(s_h^k))C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s')]^2 \\ \geq&(1-\rho)\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\overline{\pi}^k_h(s_h^k)) + \rho\mathbb{V}_{P_h}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,\underline{\pi}^k_h(s_h^k)),\end{aligned} \tag{66}$$

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

$$\begin{aligned}&\sum_{k=1}^K\sum_{h=1}^H\mathbb{V}_{[\mathbb{D}_{\widetilde{\pi}^k_h}P_h]}C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k) \\ =&\sum_{k=1}^K\sum_{h=1}^H\left([\mathbb{D}_{\widetilde{\pi}^k_h}P_h(C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho})^2](s_h^k) - \left([\mathbb{D}_{\widetilde{\pi}^k_h}P_hC_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}](s_h^k)\right)^2\right) \\ =&\sum_{k=1}^K\sum_{h=1}^H\left([\mathbb{D}_{\widetilde{\pi}^k_h}P_h(C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho})^2](s_h^k) - \left(C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_{h+1}^k)\right)^2\right) \\ &+\sum_{k=1}^K\sum_{h=1}^H\left(\left(C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_{h+1}^k)\right)^2 - \left([\mathbb{D}_{\widetilde{\pi}^k_h}P_hC_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}](s_h^k)\right)^2\right) \\ \leq&H^2\sqrt{2HK\iota} + \sum_{k=1}^K\sum_{h=1}^H\left((C_h^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k))^2 - \left([\mathbb{D}_{\widetilde{\pi}^k_h}P_hC_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}](s_h^k)\right)^2\right) - \sum_{k=1}^K(C_1^{\pi^{k*},\underline{\pi}^k,\rho}(s_1^k))^2 \\ \leq&H^2\sqrt{2HK\iota} + 2H\sum_{k=1}^K\sum_{h=1}^H|C_h^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k) - \mathbb{D}_{\widetilde{\pi}^k_h}P_hC_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k)| \\ \leq&H^2\sqrt{2HK\iota} + 2H\sum_{k=1}^K\left(C_1^{\pi^{k*},\underline{\pi}^k,\rho}(s_1^k) + \sum_{h=1}^H\left(C_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_{h+1}^k) - \mathbb{D}_{\widetilde{\pi}^k_h}P_hC_{h+1}^{\pi^{k*},\underline{\pi}^k,\rho}(s_h^k,a_h^k)\right)\right) \\ \leq&H^2\sqrt{2HK\iota} + 2H^2K + 2H^2\sqrt{2HK\iota} \\ \leq&3H^2K + 9H^3\iota/2,\end{aligned} \tag{67}$$

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 6.2 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 (ARQ-H) 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 uses a model-free method to update Q and V values.

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.

E Proof for model-free algorithm

In this section, we prove Theorem 7.1. 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 $t \geq 1$;
(2) $\sum_{i=1}^t \alpha_t^i = 0$ and $\alpha_t^0 = 1$ for $t = 0$;
(3) $\frac{1}{\sqrt{t}} \leq \sum_{i=1}^t \frac{\alpha_t^i}{\sqrt{t}} \leq \frac{2}{\sqrt{t}}$ for every $t \geq 1$;
(4) $\sum_{i=1}^t (\alpha_t^i)^2 \leq \frac{2H}{t}$ for every $t \geq 1$;
(5) $\sum_{t=i}^\infty \alpha_t^i \leq (1 + \frac{1}{H})$ for every $i \geq 1$.

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

$$\begin{aligned}\overline{Q}_h^k(s,a) &= \alpha_t^0(H-h+1) + \sum_{i=1}^t \alpha_t^i \left(r_h^{k_i} + \overline{V}_{h+1}^{k_i}(s_{h+1}^{k_i}) + b_i\right); \\ \underline{Q}_h^k(s,a) &= \sum_{i=1}^t \alpha_t^i \left(r_h^{k_i} + \underline{V}_{h+1}^{k_i}(s_{h+1}^{k_i}) - b_i\right).\end{aligned} \tag{68}$$

Thus,

$$\begin{aligned}(\overline{Q}_h^k - Q_h^*)(s,a) =& \alpha_t^0(H-h+1) + \sum_{i=1}^t \alpha_t^i \left(r_h^{k_i} + \overline{V}_{h+1}^{k_i}(s_{h+1}^{k_i}) + b_i\right) \\ & - \left(\alpha_t^0 Q_h^*(s,a) + \sum_{i=1}^t \alpha_t^i \left(R_h(s,a) + P_h V_{h+1}^*(s,a)\right)\right) \\ =& \alpha_t^0(H-h+1-Q_h^*(s,a)) + \sum_{i=1}^t \alpha_t^i \left((\overline{V}_{h+1}^{k_i} - V_{h+1}^*)(s_{h+1}^{k_i})\right) \\ & + \sum_{i=1}^t \alpha_t^i \left((r_h^{k_i} - R_h(s,a)) + V_{h+1}^*(s_{h+1}^{k_i}) - P_h V_{h+1}^*(s,a) + b_i\right),\end{aligned} \tag{69}$$

and similarly

$$\begin{aligned}(\underline{Q}_h^k - Q_h^{\overline{\pi}^k})(s,a) =& \sum_{i=1}^t \alpha_t^i \left(r_h^{k_i} + \underline{V}_{h+1}^{k_i}(s_{h+1}^{k_i}) - b_i\right) \\ & - \left(\alpha_t^0 Q_h^{\overline{\pi}^k}(s,a) + \sum_{i=1}^t \alpha_t^i \left(R_h(s,a) + P_h V_{h+1}^{\overline{\pi}^k}(s,a)\right)\right) \\ =& -\alpha_t^0 Q_h^{\overline{\pi}^k}(s,a) + \sum_{i=1}^t \alpha_t^i \left([P_h(\underline{V}_{h+1}^{k_i} - V_{h+1}^{\overline{\pi}^k})](s,a)\right) \\ & + \sum_{i=1}^t \alpha_t^i \left((r_h^{k_i} - R_h(s,a)) + \underline{V}_{h+1}^{k_i}(s_{h+1}^{k_i}) - P_h \underline{V}_{h+1}^{k_i}(s,a) - b_i\right).\end{aligned} \tag{70}$$

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

$$\begin{aligned}(\underline{Q}_h^{k'} - Q_h^{\overline{\pi}^k})(s,a) =& -\alpha_t^0 Q_h^{\overline{\pi}^k}(s,a) + \sum_{i=1}^{t'} \alpha_{t'}^i \left([P_h(\underline{V}_{h+1}^{k_i} - V_{h+1}^{\overline{\pi}^k})](s,a)\right) \\ & + \sum_{i=1}^{t'} \alpha_{t'}^i \left((r_h^{k_i} - R_h(s,a)) + \underline{V}_{h+1}^{k_i}(s_{h+1}^{k_i}) - P_h \underline{V}_{h+1}^{k_i}(s,a) - b_i\right).\end{aligned} \tag{71}$$

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

$$\left|\sum_{i=1}^t \alpha_t^i \left((r_h^{k_i} - R_h(s,a)) + \underline{V}_{h+1}^{k_i}(s_{h+1}^{k_i}) - P_h \underline{V}_{h+1}^{k_i}(s,a)\right)\right| \le H\sqrt{\sum_{i=1}^t (\alpha_t^i)^2 \iota/2} \le \sqrt{H^3\iota/t}. \tag{72}$$

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

$$\left|\sum_{i=1}^t \alpha_t^i \left((r_h^{k_i} - R_h(s,a)) + V_{h+1}^*(s_{h+1}^{k_i}) - P_h V_{h+1}^*(s,a)\right)\right| \le \sqrt{H^3\iota/t}. \tag{73}$$

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

$$\begin{aligned}
&\overline{V}_{h'}^{k'}(s) \ge V_{h'}^*(s) \ge V_{h'}^{\overline{\pi}^{k'}}(s) \ge \underline{V}_{h'}^{k'}(s) \ \forall (k',h',s) \in [k-1] \times [H+1] \times \mathcal{S},\\
&\overline{Q}_{h'}^{k'}(s,a) \ge Q_{h'}^*(s,a) \ge Q_{h'}^{\overline{\pi}^{k'}}(s,a) \ge \underline{Q}_{h'}^{k'}(s,a) \ \forall (k',h',s,a) \in [k-1] \times [H+1] \times \mathcal{S} \times \mathcal{A},\\
&\overline{V}_{h'}^{k}(s) \ge V_{h'}^*(s) \ge V_{h'}^{\overline{\pi}^{k}}(s) \ge \underline{V}_{h'}^{k}(s) \ \forall h' \ge h+1 \text{ and } s \in \mathcal{S},\\
&\overline{Q}_{h'}^{k}(s,a) \ge Q_{h'}^*(s,a) \ge Q_{h'}^{\overline{\pi}^{k}}(s,a) \ge \underline{Q}_{h'}^{k}(s,a) \ \forall h' \ge h+1 \text{ and } (s,a) \in \mathcal{S} \times \mathcal{A}.
\end{aligned} \tag{74}$$

We first show the monotonicity of Q values. We have

$$(\overline{Q}_h^k - Q_h^*)(s,a) \ge \alpha_t^0(H-h+1-Q_h^*(s,a)) + \sum_{i=1}^{t} \alpha_t^i \left((\overline{V}_{h+1}^{k_i} - V_{h+1}^*)(s_{h+1}^{k_i}) \right) \ge 0, \tag{75}$$

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

$$\begin{aligned}
(\underline{Q}_h^k - Q_h^{\overline{\pi}^k})(s,a) \le& -\alpha_t^0 Q_h^{\overline{\pi}^k}(s,a) + \sum_{i=1}^{t} \alpha_t^i \left([P_h(\underline{V}_{h+1}^{k_i} - V_{h+1}^{\overline{\pi}^k})](s,a) \right)\\
\le& -\alpha_t^0 Q_h^{\overline{\pi}^k}(s,a) + \sum_{i=1}^{t} \alpha_t^i \left([P_h(\underline{V}_{h+1}^{k} - V_{h+1}^{\overline{\pi}^k})](s,a) \right) \le 0.
\end{aligned} \tag{76}$$

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

$$\begin{aligned}
(\underline{Q}_h^{k'} - Q_h^{\overline{\pi}^k})(s,a) \le& -\alpha_t^0 Q_h^{\overline{\pi}^k}(s,a) + \sum_{i=1}^{t'} \alpha_{t'}^i \left([P_h(\underline{V}_{h+1}^{k_i} - V_{h+1}^{\overline{\pi}^k})](s,a) \right)\\
\le& -\alpha_t^0 Q_h^{\overline{\pi}^k}(s,a) + \sum_{i=1}^{t'} \alpha_{t'}^i \left([P_h(\underline{V}_{h+1}^{k} - V_{h+1}^{\overline{\pi}^k})](s,a) \right) \le 0.
\end{aligned} \tag{77}$$

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

$$\begin{aligned}
&(1-\rho)\max_a \overline{Q}_h^k(s,a) + \rho \overline{Q}_h^k(s, \arg\min_a \underline{Q}_h^k(s,a))\\
\ge&(1-\rho)\max_a \overline{Q}_h^k(s,a) + \rho Q_h^*(s, \arg\min_a \underline{Q}_h^k(s,a))\\
\ge&(1-\rho)\overline{Q}_h^k(s,\pi_h^*(s)) + \rho \min_{a\in\mathcal{A}} Q_h^*(s,a)\\
\ge&(1-\rho)Q_h^*(s,\pi_h^*(s)) + \rho \min_{a\in\mathcal{A}} Q_h^*(s,a) = V_h^*(s).
\end{aligned} \tag{78}$$

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

$$\overline{V}_h^k(s) = \min\{\overline{V}_h^{k-1}(s), (1-\rho)\max_a \overline{Q}_h^k(s,a) + \rho \overline{Q}_h^k(s, \arg\min_a \underline{Q}_h^k(s,a))\} \ge V_h^*(s). \tag{79}$$

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,

$$\begin{aligned}
\underline{V}_h^k(s) =&(1-\rho)\underline{Q}_h^{\tau+1}(s,\overline{\pi}_h^{\tau+1}(s)) + \rho\min_a \underline{Q}_h^{\tau+1}(s,a)\\
\leq&(1-\rho)Q_h^{\overline{\pi}^k}(s,\overline{\pi}_h^{\tau+1}(s)) + \rho\min_a \underline{Q}_h^{\tau+1}(s,a)\\
\leq&(1-\rho)Q_h^{\overline{\pi}^k}(s,\overline{\pi}_h^{k}(s)) + \rho\underline{Q}_h^{\tau+1}(s,\arg\min_{a\in\mathcal{A}} Q_h^{\overline{\pi}^k}(s,a))\\
\leq&(1-\rho)Q_h^{\overline{\pi}^k}(s,\overline{\pi}_h^{k}(s)) + \rho\min_{a\in\mathcal{A}} Q_h^{\overline{\pi}^k}(s,a) = V_h^{\overline{\pi}^k}(s).
\end{aligned} \tag{80}$$

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

$$Regret(K) := \sum_{k=1}^{K}(V_1^*(s_1^k) - V_1^{\overline{\pi}^k}(s_1^k)) \leq \sum_{k=1}^{K}(\overline{V}_1^k(s_1^k) - \underline{V}_1^k(s_1^k)). \tag{81}$$

By the update rules in Algorithm 2, we have

$$\begin{aligned}
&\overline{V}_h^k(s_h^k) - \underline{V}_h^k(s_h^k)\\
\leq&(1-\rho)\overline{Q}_h^k(s_h^k, \arg\max_a \overline{Q}_h^k(s_h^k,a)) + \rho\overline{Q}_h^k(s_h^k, \arg\min_a \underline{Q}_h^k(s_h^k,a))\\
&-(1-\rho)\underline{Q}_h^k(s_h^k, \arg\max_a \overline{Q}_h^k(s_h^k,a)) + \rho\underline{Q}_h^k(s_h^k, \arg\min_a \underline{Q}_h^k(s_h^k,a))\\
=&(1-\rho)[\overline{Q}_h^k - \underline{Q}_h^k](s_h^k, \overline{a}_h^k) + \rho[\overline{Q}_h^k - \underline{Q}_h^k](s_h^k, \underline{a}_h^k)\\
=&[\overline{Q}_h^k - \underline{Q}_h^k](s_h^k, a_h^k) + [\mathbb{D}_{\widetilde{\pi}_h^k}(\overline{Q}_h^k - \underline{Q}_h^k)](s_h^k) - [\overline{Q}_h^k - \underline{Q}_h^k](s_h^k, a_h^k).
\end{aligned} \tag{82}$$

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,

$$\begin{aligned}
\phi_h^k =&\overline{V}_h^k(s_h^k) - \underline{V}_h^k(s_h^k)\\
\leq&\alpha_{n_h^k}^0(H-h+1) + \sum_{i=1}^{n_h^k}\alpha_{n_h^k}^i\left(\overline{V}_{h+1}^{k_i(s_h^k,a_h^k)}(s_{h+1}^{k_i(s_h^k,a_h^k)}) - \underline{V}_{h+1}^{k_i(s_h^k,a_h^k)}(s_{h+1}^{k_i(s_h^k,a_h^k)}) + 2b_i\right)\\
&+[\mathbb{D}_{\widetilde{\pi}_h^k}(\overline{Q}_h^k - \underline{Q}_h^k)](s_h^k) - [\overline{Q}_h^k - \underline{Q}_h^k](s_h^k, a_h^k)\\
=&\alpha_{n_h^k}^0(H-h+1) + \sum_{i=1}^{n_h^k}\alpha_{n_h^k}^i(\phi_{h+1}^{k_i(s_h^k,a_h^k)} + 2b_i) + \xi_h^k\\
\leq&\alpha_{n_h^k}^0(H-h+1) + \sum_{i=1}^{n_h^k}\alpha_{n_h^k}^i\phi_{h+1}^{k_i(s_h^k,a_h^k)} + \xi_h^k + 4\sqrt{H^3\iota/n_h^k}.
\end{aligned} \tag{83}$$

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

$$\begin{aligned}
&\sum_{k=1}^{K}(\overline{V}_h^k(s_h^k) - \underline{V}_h^k(s_h^k)) \\
\leq&\sum_{k=1}^{K}\alpha_{n_h^k}^0(H-h+1) + \sum_{k=1}^{K}\sum_{i=1}^{n_h^k}\alpha_{n_h^k}^i\phi_{h+1}^{k_i(s_h^k,a_h^k)} + \sum_{k=1}^{K}\xi_h^k + \sum_{k=1}^{K}4\sqrt{H^3\iota/n_h^k} \\
=&\sum_{k=1}^{K}\alpha_{n_h^k}^0(H-h+1) + \sum_{k'=1}^{K}\phi_{h+1}^{k'}\sum_{t=n_h^{k'}+1}^{n_h^K}\alpha_t^{n_h^{k'}} + \sum_{k=1}^{K}\xi_h^k + \sum_{k=1}^{K}4\sqrt{H^3\iota/n_h^k} \\
\leq&\sum_{k=1}^{K}\alpha_{n_h^k}^0(H-h+1) + (1+1/H)\sum_{k=1}^{K}\phi_{h+1}^k + \sum_{k=1}^{K}\xi_h^k + \sum_{k=1}^{K}4\sqrt{H^3\iota/n_h^k}
\end{aligned} \tag{84}$$

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

$$\begin{aligned}
&\sum_{k=1}^{K}(\overline{V}_1^k(s_1^k) - \underline{V}_1^k(s_1^k)) \\
\leq&3\sum_{h=1}^{H}\sum_{k=1}^{K}\alpha_{n_h^k}^0(H-h+1) + 3\sum_{h=1}^{H}\sum_{k=1}^{K}\xi_h^k + \sum_{h=1}^{H}\sum_{k=1}^{K}12\sqrt{H^3\iota/n_h^k}
\end{aligned} \tag{85}$$

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 6.4, 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

$$\begin{aligned}
V_1^*(s_1) - V_1^{\pi^{out}}(s_1) \leq&\overline{V}_1^{K+1}(s_1) - \underline{V}_1^{K+1}(s_1) \\
=&\min_{k\in[K+1]}(\overline{V}_1^k(s_1^k) - \underline{V}_1^k(s_1^k)) \\
\leq&O\left(\frac{\sqrt{SAH^5\iota}}{K} + \frac{SAH^2}{K}\right).
\end{aligned} \tag{86}$$