MODEL-FREE OFFLINE REINFORCEMENT LEARNING WITH ENHANCED ROBUSTNESS

Chi Zhang[1], **Zain Ulabedeen Farhat**[1], **George K. Atia**[1,2], **Yue Wang**[1,2]
[1] Department of Electrical and Computer Engineering [2] Department of Computer Science
University of Central Florida
Orlando, FL 32816, USA
{chi.zhang,zainulabedeen.farhat,george.atia,yue.wang}@ucf.edu

ABSTRACT

Offline reinforcement learning (RL) has gained considerable attention for its ability to learn policies from pre-collected data without real-time interaction, which makes it particularly useful for high-risk applications. However, due to its reliance on offline datasets, existing works inevitably introduce assumptions to ensure effective learning, which, however, often lead to a trade-off between robustness to model mismatch and scalability to large environments. In this paper, we enhance both aspects with a novel double-pessimism principle, which conservatively estimates performance and accounts for both limited data and potential model mismatches, two major reasons for the previous trade-off. We then propose a universal, model-free algorithm to learn a policy that is robust to potential environment mismatches, which enhances robustness in a scalable manner. Furthermore, we provide a sample complexity analysis of our algorithm when the mismatch is modeled by the l_α-norm, which also theoretically demonstrates the efficiency of our method. Extensive experiments further demonstrate that our approach significantly improves robustness in a more scalable manner than existing methods.

1 INTRODUCTION

Traditional reinforcement learning (RL) (Sutton & Barto, 2018) optimizes an agent's performance through iterative trial-and-error interactions with the environment, and has shown significant success in many areas such as video games (Wei et al., 2022; Liu et al., 2022a). However, such an online learning scheme can be costly or unsafe in real-world applications. For instance, in domains including autonomous driving (Kiran et al., 2021), stock market trading (Kabbani & Duman, 2022), and healthcare (Yu et al., 2021), poor decisions can have significant consequences, making extensive explorations impractical. To address them, *offline RL* has been developed (Lange et al., 2012; Levine et al., 2020), enabling agents to learn from pre-collected datasets, offering a more reliable framework.

Since offline RL relies heavily on pre-collected datasets, the quality of these datasets largely determines performance. It is hence unclear whether satisfactory performance can be achieved for complex problems with a relatively limited dataset. In this context, two key challenges in improving offline RL performance have been studied. The first is **scalability**—the ability to handle large-scale problems. Without real-time interaction, learning an effective policy for large-scale problems from a limited dataset, which may not fully cover the entire state-action space, can be challenging. Recent research has focused on improving scalability by adapting model-free algorithms (Shi et al., 2022; Yan et al., 2022; Laroche et al., 2019; Fujimoto et al., 2019; Ghasemipour et al., 2021; Kumar et al., 2019; Wu et al., 2019; Siegel et al., 2020) and leveraging function approximation techniques (Ross & Bagnell, 2012; Liu et al., 2020; Xie et al., 2021a; Yin et al., 2021a; Xie & Jiang, 2021; Jiang & Huang, 2020). However, due to the complexity of large environments, many of these approaches assume that the dataset sufficiently represents the full deployment environment, typically presuming that the deployment environment is identical to the one from which the data was collected.

However, this assumption can be too restrictive. Static datasets only capture the environment at the time of data collection, but real-world applications frequently face environmental uncertainty due to perturbations or non-stationarity. This mismatch between the data collection and deployment environments, commonly known as the *sim-to-real gap* (Zhao et al., 2020), can cause significant performance degradation during deployment. Therefore, it is crucial to enhance the **robustness** of offline RL to ensure that the learned policies can perform reliably in the presence of such uncertainties. A promising solution is to adapt robust RL frameworks (Iyengar, 2005; Nilim & El Ghaoui, 2004) to the offline setting, as explored recently in (Shi & Chi, 2022; Blanchet et al., 2023). However, these methods often come at the cost of scalability. Due to their inherent structure, robust RL methods typically rely on dynamic planning, which requires knowledge of the full transition dynamics, and are predominantly model-based. This necessitates learning and storing a complete transition model, which is resource-intensive (Zhang et al., 2021a) and limits scalability for large-scale problems.

Recognizing the limitations of current methods and the challenges posed by large-scale problems and model uncertainty, a trade-off between robustness and scalability becomes apparent. Enhancing one typically comes at the expense of the other. This naturally leads to the following question:

Can we develop a unified framework that enhances both scalability and robustness in offline RL?

In this paper, we address this question by presenting a model-free algorithm to learn a policy that is both robust to model uncertainty and scalable to large-scale problems. Our method introduces a principle of double pessimism to simultaneously address two key sources of uncertainty: (1) the uncertainty arising from inaccurate estimations due to the underexplored datasets, and (2) model mismatch between the data collection and deployment environments. We then propose a streamlined conceptual framework, design a model-free algorithm, and provide the first theoretical guarantee of convergence and robustness of our approach. Our contributions can be summarized as follows.

- **A Double-Pessimism Principle for Offline RL with Model Mismatch.** We begin by framing the challenge of enhancing robustness in offline RL within an offline robust RL framework, where an uncertainty set captures potential environmental mismatches. To solve offline robust RL in a scalable manner, we propose the double-pessimism principle that does not require transition kernel estimations. This principle maintains a conservative estimate of robust performance, obtained directly from data collection without requiring model estimation. We then introduce the first model-free pessimistic robust Q-learning algorithm. Our algorithm optimizes performance under model mismatch using an offline dataset, while offering greater memory efficiency and more scalability than previous methods.
- **First and Near-Optimal Model-Free Algorithm for Offline Robust RL.** We provide a rigorous sample complexity analysis for our model-free double-pessimistic robust Q-learning algorithm under the widely used l_α-norm uncertainty set. Our analysis shows that, given a dataset satisfying the partial coverage condition (to be introduced later), our algorithm can identify an optimal robust policy with near-optimal sample complexity, comparable to that of model-based offline robust RL and model-free offline non-robust RL. This represents the first sample complexity analysis for model-free robust offline RL, demonstrating its applicability to large-scale problems that require high data efficiency.
- **Numerical Experimental Verification of Enhanced Robustness.** We conduct extensive numerical experiments to demonstrate the improvements in robustness achieved by our algorithms in both simulated environments (Archibald et al., 1995) and real physics-based Classic Control problems (Brockman et al., 2016). In each case, our algorithm consistently outperforms existing methods in handling model uncertainty, showcasing its enhanced ability to maintain stable performance across a wide range of environmental perturbations. Moreover, our approach demonstrates superior scalability stemming directly from our model-free algorithm design, as shown by its effectiveness in solving more complex Classic Control problems with robustness guarantees, which have proven difficult or unsolvable for previous model-based robust methods.

2 Preliminaries

2.1 Finite-Horizon Markov Decision Process (MDP)

A finite-horizon MDP is represented by $\mathcal{M} = \left(\mathcal{S}, \mathcal{A}, H, P \triangleq \{P_h\}_{h=1}^H, r \triangleq \{r_h\}_{h=1}^H\right)$, where $\mathcal{S}$ and $\mathcal{A}$ are the finite state and action spaces of size S and A, respectively, and H is the horizon length.

The probability transition kernel $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ and the reward function $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ are defined at each step h $(1 \leq h \leq H)$. At each step h, the agent starts in state s_h, takes action a_h, transitions to the next state s_{h+1} according to the transition kernel P_{h,s_h,a_h}, and receives a reward $r_h(s_h, a_h)$. This process terminates after H steps when the agent reaches state s_{H+1}.

A policy $\pi = \{\pi_h\}_{h=1}^H$ defines the strategy for selecting actions in different states, where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ specifies the probability distribution over actions at step h. The performance of an agent following a policy π is measured by the value function $V^{\pi,P} = \{V_h^{\pi,P}\}_{h=1}^H$, where

$$V_h^{\pi,P}(s) \triangleq \mathbb{E}_{\pi,P}\left[\sum_{t=h}^{H} r_t(s_t, a_t) \,\middle|\, s_h = s\right]. \tag{1}$$

The expectation is taken over the trajectory $\{s_h, a_h, r_h\}_{h=1}^H$ generated by executing the policy π and transitioning according to the transition kernel P: $a_h \sim \pi_h(s_h)$ and $s_{h+1} \sim P_{h,s_h,a_h}$.

2.2 Infinite-Horizon MDP

An infinite-horizon MDP is defined as $\mathcal{M} = \big(\mathcal{S}, \mathcal{A}, P, r, \gamma\big)$, where both the transition kernel P and the reward function r are stationary and do not change over time. The discount factor $\gamma < 1$ ensures the finiteness of the accumulated reward over an infinite horizon.

Due to its stationary nature, it suffices to consider only stationary policies $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, which specify the action-selection probabilities over the action space. The value function $V^{\pi,P}$ of a policy π with transition kernel P is defined as

$$V^{\pi,P}(s) \triangleq \mathbb{E}_{\pi,P}\left[\sum_{t=1}^{\infty} \gamma^t r_t(s_t, a_t) \,\middle|\, s_0 = s\right]. \tag{2}$$

2.3 Robust MDP

A finite-horizon robust MDP (RMDP) is specified by $\big(\mathcal{S}, \mathcal{A}, H, \mathcal{P} = \{\mathcal{P}_h\}, r\big)$, and an infinite-horizon RMDP is denoted by $\big(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma\big)$, where $\mathcal{P}$ is a set containing some transition kernels, named the uncertainty set. At each step, the environment transitions to the next state following an arbitrary kernel belonging to the uncertainty set, instead of a fixed one as in non-robust MDPs. In this paper, we consider the (s, a)-rectangular uncertainty set (Wiesemann et al., 2013), where $\mathcal{P}$ is independently defined for each state-action pair, with $\bigotimes$ denoting the Cartesian product:

$$\mathcal{P}_h = \bigotimes_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathcal{P}_{h,s,a} \text{ (finite-horizon)}, \quad \mathcal{P} = \bigotimes_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathcal{P}_{s,a} \text{ (infinite-horizon)}. \tag{3}$$

The performance of a policy in an RMDP is evaluated based on its worst-case value function over all the instances in the uncertainty set. Specifically, the finite-horizon robust value functions $V^\pi = \{V_h^\pi\}_{h=1}^H$ and the infinite-horizon robust value functions V^π are defined as

$$V_h^\pi(s) \triangleq \inf_{P\in\mathcal{P}} V_h^{\pi,P}(s) \text{ (finite-horizon)}, \quad V^\pi(s) \triangleq \inf_{P\in\mathcal{P}} V^{\pi,P}(s) \text{ (infinite-horizon)}$$

where the infimum is taken over the uncertainty set of transition kernels. For a given initial state distribution $\rho \in \Delta(\mathcal{S})$, we write the expected robust performance as

$$V_1^\pi(\rho) \triangleq \mathbb{E}_{s_1\sim\rho}[V_1^\pi(s_1)] \text{ (finite-horizon)}, \quad V^\pi(\rho) \triangleq \mathbb{E}_{s\sim\rho}[V^\pi(s)] \text{ (infinite-horizon)}. \tag{4}$$

The goal of an RMDP is to learn a policy that optimizes the worst-case performance, or equivalently, the robust value functions. Such a policy is referred to as an optimal robust policy:

$$\pi^* = \{\pi_h^*\} \triangleq \arg\max_\pi V_1^\pi(\rho), \text{(finite-horizon)}, \tag{5}$$

$$\pi^* \triangleq \arg\max_\pi V^\pi(\rho), \text{(infinite-horizon)}. \tag{6}$$

3 FORMULATION: ENHANCING ROBUSTNESS AND SCALABILITY

In this section, we develop our formulation, where we utilize RMDPs to formulate the offline RL problem against model mismatch.

In the offline setting, the dataset is collected under a fixed environment P (referred to as the nominal kernel) by executing some behavior policy μ. However, due to factors such as non-stationarity, unexpected perturbations, or adversarial attacks, the deployment environment may differ from P. To account for this model deviation and improve robustness, we construct an uncertainty set by perturbing the nominal kernel and aim to learn the optimal robust policy. Specifically, following (Xu & Mannor, 2010; Xu et al., 2010; Derman et al., 2021; Kumar et al., 2023), we define the uncertainty set (of (s, a)-pair) for modeling environmental perturbations as:

$$\mathcal{P}_{h,s,a} = \{P_{h,s,a} + q \in \Delta(\mathcal{S}) : q \in \mathcal{Q}_{h,s,a}\} \quad \text{(finite-horizon)}, \tag{7}$$

$$\mathcal{P}_{s,a} = \{P_{s,a} + q \in \Delta(\mathcal{S}) : q \in \mathcal{Q}_{s,a}\} \quad \text{(infinite-horizon)}, \tag{8}$$

for some set $\mathcal{Q}_{h,s,a}$, $\mathcal{Q}_{s,a}$ containing the possible model perturbations, and aim to learn the optimal robust policy for the corresponding RMDPs. This will not only provide an optimized lower bound on performance when the deployment environment lies within the uncertainty set, but also improves the robustness to model uncertainty (Pinto et al., 2017).

3.1 FINITE-HORIZON

In the finite-horizon setting, the dataset $\mathcal{D}$ consists of K episodes each of length H. These episodes are independently generated based on a certain behavior policy μ and the nominal kernel P:

$$\mathcal{D} = \{(s_1^k, a_1^k, r_1^k, ..., s_H^k, a_H^k, r_H^k, s_{H+1}^k)_{k=1,...,K}\}, \tag{9}$$

where $a_i^k \sim \mu(\cdot|s_i^k)$, $s_{i+1}^k \sim P_{i,s_i^k,a_i^k}$, and the initial state $s_1^k \sim \rho$.

Since the dataset is collected by a fixed policy under a single nominal environment, there exists a distribution shift between the data distribution, and the distribution induced by the optimal policy and the worst-case kernel. To guarantee that a provable efficient algorithm can be designed based on the dataset, we adopt a popular assumption on the distributional mismatch between the dataset distribution and the occupancy measure induced by the optimal policy π^*, as in (Shi & Chi, 2022).

Assumption 1 (Robust single-policy concentrability)**.** *The behavior policy μ satisfies that*

$$C^* \triangleq \max_{(s,a,P',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{P}\times[H]} \frac{d^{\pi^*}_{P',h}(s,a)}{d^{\mu}_{P,h}(s,a)} < +\infty, \tag{10}$$

where $d^{\pi}_{P,h}$ is the occupancy distribution induced by policy π and transition kernel P at step h.

In Assumption 1, we only require that the dataset covers the state-action pairs that are visited by the optimal policy, known as the partial coverage condition (Rashidinejad et al., 2021).

Our goal is then to learn an ϵ-optimal policy $\hat{\pi}$ for the RMDP with the uncertainty set defined as in equation 3 and equation 7, such that

$$V_1^{\pi^*}(\rho) - V_1^{\hat{\pi}}(\rho) \leq \epsilon. \tag{11}$$

3.2 INFINITE-HORIZON

In the infinite-horizon setting, the offline dataset contains a single trajectory of length T obtained by executing a behavior policy μ under the nominal kernel P:

$$\mathcal{D} = \{s_1, a_1, r_1, s_2, ..., s_T\}, \tag{12}$$

where $s_1 \sim \rho$, $a_i \sim \mu(\cdot|s_i)$ and $s_{i+1} \sim P_{s_i,a_i}$. For the infinite horizon setting, we adopt the following two assumptions on the behavior policy.

We first adopt the partial coverage assumption in (Blanchet et al., 2023; Wang et al., 2024c).

Assumption 2. *The behavior policy μ satisfies*

$$C^* \triangleq \max_{(s,a,P')\in\mathcal{S}\times\mathcal{A}\times\mathcal{P}} \frac{d_{P'}^{\pi^*}(s,a)}{d_P^{\mu}(s,a)} < +\infty, \tag{13}$$

where d_P^π denotes the occupancy distribution induced by policy π and transition kernel P.

We make an additional assumption on the behavior policy as follows.

Assumption 3. *The behavior policy μ is stationary, and the induced Markov chain under the nominal kernel is uniformly ergodic.*

Remark 1. *This assumption is commonly adopted in prior works (Wang et al., 2020; Yan et al., 2022; Li et al., 2020; Wang & Zou, 2020), as it ensures that the dataset includes all state-action pairs covered by the **behavior policy**, provided the dataset size exceeds a certain threshold. This assumption is required since the dataset consists of a single Markovian trajectory. When the dataset contains i.i.d. samples from the occupancy distribution d_P^μ, as in (Wang et al., 2024c; Li et al., 2022), such an assumption can be removed.*

Our goal is then to find an ϵ-optimal policy $\hat{\pi}$ through $\mathcal{D}$ for the RMDP with the uncertainty set defined in equation 3 and equation 8, such that

$$V^{\pi^*}(\rho) - V^{\hat{\pi}}(\rho) \leq \epsilon. \tag{14}$$

4 Double-Pessimism principle

In this section, we introduce our model-free algorithm for learning an optimal robust policy from an offline dataset. As we mentioned, two major challenges in offline RL are the two sources of uncertainty: one arising from the limited and under-explored dataset, and the other from the mismatch between the data collection and target environments. We aim to develop a unified double-pessimism principle to address both of them.

As suggested by previous studies on offline RL, e.g., (Rashidinejad et al., 2021; Li et al., 2022; Shi et al., 2022; Yan et al., 2022; Wang et al., 2024c), the uncertainty arising from the dataset can be addressed using a single-pessimism principle. This involves introducing a penalty term b_n, which depends on the visitation frequency of each state-action pair, to penalize less frequently visited pairs. By doing so, we obtain a conservative estimate of the value function under the nominal kernel.

However, addressing the uncertainty arising from model mismatch is particularly challenging, especially with a model-free approach. Most previous robust RL studies require that the estimation of the worst-case transition, $\sigma_{\mathcal{P}}(V) \triangleq \min_{p\in\mathcal{P}} pV$, be unbiased. This can be satisfied when the agent can freely generate data as needed (e.g., (Wang et al., 2023d; Liu et al., 2022b; Wang et al., 2023c;b)), yet is impractical in offline settings. To address this issue, we argue that another pessimism principle can be adopted, and that learning a policy robust to model mismatch does not require an unbiased estimator or an accurate solution to the worst-case. As long as the estimator provides a (not too pessimistic) lower bound on the worst-case, it is sufficient to account for the uncertainty due to model mismatch and still learn a robust policy. We therefore propose a model-free estimator that lower bounds $\sigma(V)$ to produce a conservative estimation as follows.

Definition 1. *For the uncertainty set $\mathcal{P}_{s,a}$, a function κ is referred to as a model-mismatch penalty function if for any non-negative vector V and a sample $s' \sim P_{s,a}$ from the nominal kernel,*

$$\mathbb{E}[V(s') - \kappa_{s,a}(V)] \leq \sigma_{\mathcal{P}_{s,a}}(V). \tag{15}$$

A universal design of the penalty function κ is provided in Appendix C. Such a penalty function ensures that at each step, the updated estimate represents a lower bound on the true worst-case scenario, resulting in a conservative estimation and enhancing robustness. We note that this additional pessimism may result in a more conservative policy, as the algorithm will estimate the robust value function more pessimistically. However, we argue that as long as the pessimism level is not too large, the learned policy will not be too conservative, maintaining a satisfactory performance and enhancing the robustness. More importantly, calculation of κ does not require any information of the model, but can be done in a data-driven and model-free fashion.

We then combine the two pessimism principles together, to develop our double-pessimism algorithm based on the Q-learning algorithm. For each sample (s, a, s'), we update the Q table by

$$Q(s,a) \leftarrow \quad (1-\eta)Q(s,a) + \eta\left(r(s,a) + \gamma V(s') - \underbrace{\gamma\kappa_{s,a}(V)}_{\text{model mismatch}} - \underbrace{b_n(V)}_{\text{limited dataset}}\right). \tag{16}$$

As we will show later, such an update rule incorporating the double-pessimism principle ensures that our estimation is conservative, and can effectively tackle the uncertainty in offline robust RL. More importantly, such an update rule does not require any information on the transition model, and hence can be adapted in a model-free manner and is more suitable for large-scale problems.

Based on this, we develop our model-free offline algorithms for both finite and infinite horizon cases. In the following sections, we present these algorithms and develop their sample complexity analysis.

5 Double-Pessimism Q-Learning for Finite-Horizon MDPs

Adopting the double-pessimism principle, we propose our algorithm for finite-horizon MDPs.

Algorithm 1 Double-Pessimism Q-Learning for finite-horizon RMDPs.

Input: $\mathcal{D}$, target success probability $1-\delta$, uncertainty set radius R, penalty function κ
Initialize: $Q_h(s,a) = 0$, $N_h(s,a) = 0$, $V_h(s) = 0, \forall s, a, h$
for $k = 1, \dots, K$ **do**
  Sample a trajectory $\{s_h, a_h, r_h\}_{h=1}^H$ from $\mathcal{D}_\mu$
  for $h = 1, \dots, H$ **do**
    $N_h(s_h, a_h) \leftarrow N_h(s_h, a_h) + 1$; $n \leftarrow N_h(s_h, a_h)$; $\eta_n \leftarrow \frac{H+1}{H+n}$
    $b_n \leftarrow c_b\sqrt{\frac{H^3 \log^2(SAKH/\delta)}{n}}$
    $Q_h(s_h,a_h) \leftarrow (1-\eta_n)Q_h(s_h,a_h) + \eta_n\Big\{r_h(s_h,a_h) + V_{h+1}(s_{h+1}) - \kappa_{h,s_h,a_h}(V_{h+1}) - b_n\Big\}$
    $V_h(s_h) \leftarrow \max\Big\{V_h(s_h),\, \max_a Q_h(s_h,a)\Big\}$
  end for
  $\widehat{\pi}_h^k(s) \leftarrow \arg\max_a Q_h(s,a), \forall s, h$
end for
$\widehat{\pi}_h(s) \leftarrow \widehat{\pi}_h^K(s), \forall s, h$
Output: $\widehat{\pi} = \{\hat{\pi}_h\}$

In our algorithm, the term κ is for conservative estimation of the worst-case performance within the uncertainty set, while the term b addresses the pessimism of the limited dataset. We track the visitation count of each state-action pair and construct the penalty term b based on these counts. As the dataset visits a pair more frequently, the associated uncertainty decreases and b decreases.

Remark 2. *Our algorithm design is universal and works for any uncertainty set models, as long as we have a penalty function κ satisfying equation 15, which is provided in Appendix C. However, since κ for different models requires individual studies, we mainly derive our theoretical analysis for the l_α-norm models (Kumar et al., 2023; Derman et al., 2021):*

$$\mathcal{P}_{h,s,a} = \left\{q \in \Delta(\mathcal{S}) : \|q - P_{h,s,a}\|_\alpha \le R_{h,s,a}\right\}. \tag{17}$$

We again emphasize that our double-pessimism principle and algorithm design can be extended further to other uncertainty set models. We provide a detailed discussion on κ in Appendix C.

Next, we develop our theoretical results for l_α-norm sets. We first show that equation 15 is satisfied by our design, and the algorithm results in a conservative estimation of the robust value function.

Lemma 1. *For the l_α-norm uncertainty set, set the penalty function κ as*

$$\kappa_{h,s,a}(V) \triangleq R_{h,s,a} \min_{w\in\mathbb{R}} \|we - V\|_\beta, \tag{18}$$

where $\beta = \frac{1}{1-\frac{1}{\alpha}}$ is the Hölder conjugate of α, and $e = (1, 1, ..., 1) \in \mathbb{R}^S$. Then, equation 15 is satisfied. Moreover, it holds that for all $(k, h, s) \in [K] \times [H] \times \mathcal{S}$,

$$V_h(s) \leq V_h^{\widehat{\pi}_h^k}(s) \leq V_h^\star(s). \tag{19}$$

The lemma provides a concrete construction of the penalty function for the l_α-norm model. More importantly, our model-free estimator and algorithm result in pessimistic estimations of robust value functions, tackling both uncertainty sources. In our next result, we show that our double-pessimism principle is effective in learning the optimal robust policy from the mismatched offline dataset.

Theorem 2. *For the l_α-norm uncertainty set, and any $\delta \in (0, 1)$, suppose that the behavior policy μ satisfies Assumption 1. When $T \triangleq HK > \tilde{\mathcal{O}}(SC^\star)$, the policy $\widehat{\pi}$ returned by Algorithm 1 satisfies*

$$V_1^{\pi^\star}(\rho) - V_1^{\widehat{\pi}}(\rho) \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{H^6 SC^\star}{T}}\right) \tag{20}$$

with probability at least $1 - \delta$. Here, π^* is the optimal robust policy w.r.t. a (possibly) relaxed l_α-norm uncertainty set (see Appendix C.2 for detailed discussion). $f(T) = \tilde{\mathcal{O}}(g(T))$ means that $|f(T)| \leq C \cdot g(T) \cdot (\log g(T))^k$ for some constants $C > 0$ and $k \geq 0$, when T is sufficiently large.

Our algorithm is the first model-free algorithm for offline RL under model mismatch with sub-optimality gap analysis. The sub-optimality gap we obtain in the previous result further implies that we can learn an ϵ-optimal policy as long as the size of the offline dataset T exceeds

$$\underbrace{\tilde{\mathcal{O}}\left(\frac{H^6 SC^\star}{\epsilon^2}\right)}_{\epsilon\text{-dependent}} + \underbrace{\tilde{\mathcal{O}}(SC^\star)}_{\text{burn-in cost}}. \tag{21}$$

Note that in the sample complexity, the second term, referred to as the burn-in cost, is a universal constant that does not depend on ϵ, while the first term asymptotically depends on ϵ. When ϵ becomes smaller, the first term dominates the overall complexity, resulting in an asymptotic complexity of $\tilde{\mathcal{O}}\left(\frac{H^6 SC^\star}{\epsilon^2}\right)$. A more detailed discussion of the complexity will be provided in Section 7.

Remark 3. *When the radius R is small, it holds that $\mathbb{E}[V(s') - \kappa_{s,a}(V)] = \sigma_{\mathcal{P}_{s,a}}(V)$ (see Theorem 1 in (Kumar et al., 2023)), hence Algorithm 1 converges to the optimal robust policy w.r.t. the original uncertainty set. For general uncertainty set and corresponding penalty function κ, Algorithm 1 may converge to the optimal robust policy w.r.t. a relaxed uncertainty set, as the estimation may be inaccurate. However, robustness can still be enhanced due to the additional pessimism. See Appendix C for further discussion.*

6 Double-Pessimism Q-Learning for Infinite-Horizon MDPs

In this section, we present our algorithm design and analysis for offline RL with infinite-horizon MDPs. Due to space limitation and similarities in algorithm design, the algorithm is deferred to Algorithm 3 in Appendix E.1. The algorithm follows a similar design as the finite-horizon one, where the two terms κ and b represent conservative penalties for the double-pessimism principle. Again, our algorithm design is universal, but we develop the sample complexity results only for l_α-norm models.

Theorem 3. *Consider the l_α-norm uncertainty set and any $\delta \in (0, 1)$. Suppose that the behavior policy μ satisfies Assumption 2 and Assumption 3. Then, the policy $\widehat{\pi}$ returned by Algorithm 3 satisfies*

$$V^\star(\rho) - V^{\widehat{\pi}}(\rho) \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{C^\star S}{T(1-\gamma)^5}} + \frac{C^\star S}{T(1-\gamma)^2} + \frac{C^\star}{T(1-\gamma)^3}\right) \tag{22}$$

with probability at least $1 - \delta$.

An ϵ-optimal robust policy can be learned as long as the size of the offline dataset exceeds

$$\tilde{\mathcal{O}}\left(\frac{SC^*}{(1-\gamma)^5\epsilon^2}\right). \tag{23}$$

This sample complexity matches the results of model-free offline non-robust RL (Yan et al., 2022) without variance reduction techniques, which implies the near-optimality of our method. Compared to model-based offline robust RL (Shi & Chi, 2022; Blanchet et al., 2023), our result matches theirs in terms of C^*, S, ϵ, but exhibits a higher order dependence on $(1-\gamma)$. We argue that, in general, model-free algorithms tend to have lower memory requirements but incur higher sample complexity compared to model-based approaches. A more detailed discussion will be provided in Section 7.

7 RELATED WORK

7.1 COMPARISON WITH PRIOR ARTS

In this section, we compare our work to the most closely related studies for tabular offline robust RL (Shi & Chi, 2022; Blanchet et al., 2023). The results are summarized in Table 1, where we only include the infinite horizon ones. Compared to previous studies, our method offers improved memory and computational complexity, while maintaining comparable sample complexity.

Reference	Memory complexity	Sample complexity	Computational complexity
Our Work	$\mathcal{O}(SA)$	$\tilde{\mathcal{O}}\left(\frac{SC^\star}{\epsilon^2(1-\gamma)^5}\right)$	Polynomial
(Blanchet et al., 2023)	$\mathcal{O}(S^2A)$	$\tilde{\mathcal{O}}\left(\frac{S^2C^\star}{\epsilon^2(1-\gamma)^4}\right)$	NP Hard
(Shi & Chi, 2022)	$\mathcal{O}(S^2A)$	$\tilde{\mathcal{O}}\left(\frac{SC^\star}{\epsilon^2 P_{\min}(1-\gamma)^4}\right)$	Polynomial

Table 1: Comparison with offline robust RL works. (Shi & Chi, 2022) is for the KL-divergence set.

First, both related works are model-based, which involves estimating and storing the transition model $\{\hat{P}_{s,a} : (s,a) \in \mathcal{S} \times \mathcal{A}\} \in \mathbb{R}^{S^2A}$. This approach thus requires an additional memory of size $\mathcal{O}(S^2A)$ to store the model, along with $\mathcal{O}(SA)$ space for the number of visited state-action pairs from the dataset. As a result, it becomes inefficient for large-scale problems or environments with complicated transition dynamics. In contrast, our model-free algorithm only requires $\mathcal{O}(SA)$-sized space for the number of visits. Such a reduced memory complexity enables our model-free algorithms to handle large-scale problems, scaling effectively to large-scale or even continuous problems.

In terms of computational complexity, the most related work (Blanchet et al., 2023) requires to solve a non-rectangular RMDP, which is generally NP-hard (Wiesemann et al., 2013). In contrast, our algorithm can be effectively implemented in polynomial time, which is much more practical. Compared to (Shi & Chi, 2022), our algorithm still enjoys lower computational complexity, since the update rule of the model-based approach requires computing the inner product $\hat{P}_{s,a}V$, whereas our model-free approach eliminates this computation and only requires a single vector entry $V(s')$. See Appendix C for a more detailed discussion.

In terms of sample complexity, both of our sample complexity results match the ones for offline non-robust Q-learning without variance reduction, illustrating our data efficiency and near-optimality. Our result improves the dependence on S compared to (Blanchet et al., 2023) under the l_∞-norm uncertainty set, showing the enhanced scalability to large-scale problems. On the other hand, it is the general observation that model-based methods tend to demonstrate better sample complexity in terms of $(1-\gamma)$ than model-free methods, especially when additional techniques like variance reduction are not employed. Such findings have been widely noted in various settings, for instance, when comparing robust RL with generative models ((Wang et al., 2024a) vs. (Shi et al., 2023)) and non-robust offline RL ((Yan et al., 2022) vs. (Li et al., 2022)).

On a side note, we note that our result for the finite-horizon setting exhibits a higher-order dependence on H (where we set $H = \frac{1}{1-\gamma}$ as the effective horizon in infinite setting). This is due to the non-stationary environment inherent in the finite-horizon setting, which is also consistent with findings from previous studies, such as in (Shi & Chi, 2022).

To summarize, our approach addresses existing gaps in offline RL by enhancing robustness to model mismatch, reducing memory requirements, and providing adaptability to large-scale problems, establishing a state-of-the-art method in the field.

7.2 OTHER RELATED WORKS

Offline RL without model mismatch. A significant body of offline RL works assumes identical collection and deployment environments. Based on that, many early works further rely on the global coverage assumption, where the behavior policy covers all state-action pairs (Scherrer, 2014; Chen & Jiang, 2019; Munos, 2005; Yin et al., 2021b; Yin & Wang, 2021a; Jiang, 2019; Wang et al., 2019; Liao et al., 2020; Liu et al., 2019; Zhang et al., 2020; Uehara et al., 2020; Duan et al., 2020; Xie & Jiang, 2020; Levine et al., 2020; Antos et al., 2007; Farahmand et al., 2010). This assumption is often too restrictive and unrealistic, as it requires complete coverage of state-action pairs in historical data (Gulcehre et al., 2020; Agarwal et al., 2020a; Fu et al., 2020). A more practical partial coverage setting is later proposed, allowing to learn from a less explored dataset. Under partial coverage, the optimal policy can still be learned by incorporating the pessimism principle to handle dataset uncertainty (Jin et al., 2021; Uehara & Sun, 2021; Xie et al., 2021a;b; Rashidinejad et al., 2021; Zanette et al., 2021; Yin & Wang, 2021b; Shi et al., 2022; Li et al., 2022; Zhan et al., 2022; Wang et al., 2023e; Kumar et al., 2020). Differently, we consider potential model mismatches.

Robust RL. Robust RL (Iyengar, 2005; Nilim & El Ghaoui, 2004; Xu & Mannor, 2010) aims to tackle the challenge of model mismatch in RL, by optimizing the worst-case performance over an uncertainty set. Existing work focuses mainly on the online setting (Wang & Zou, 2021; 2022; Wang et al., 2023a; Badrinath & Kalathil, 2021; Dong et al., 2022; Lu et al., 2024; Liu & Xu, 2024a) or with a generative model (Yang et al., 2021; Xu et al., 2023; Panaganti & Kalathil, 2022; Shi et al., 2023; Wang et al., 2024a;b; 2022). Offline robust RL, except for the two mentioned above, either relies on strong assumptions, such as global coverage or absorbing states (Panaganti et al., 2022; Yang et al., 2021), or employs fitted type algorithm designs (Yang et al., 2022; Panaganti et al., 2022; Liu et al., 2023). More importantly, most of them are model-based, while we develop the first model-free algorithm for offline robust RL. Another line of research aims to improve robustness and scalability through function approximation (Liu & Xu, 2024b; Wang et al., 2024a; Ma et al., 2022), yet we focus on the tabular setting to develop a more fundamental understanding of offline RL. Another line of robust RL aims to optimize the performance under the environment from a corrupted dataset collected under the same environment (Yang et al., 2023; Zhang et al., 2021b; 2022), which is different from our setting.

8 EXPERIMENTS

We use numerical experiments to demonstrate the advantages of our framework in terms of robustness. We consider two sets of environments: simulated MDPs with controllable transition dynamics and Classic Control environments. More experiments are further provided in Appendix B.

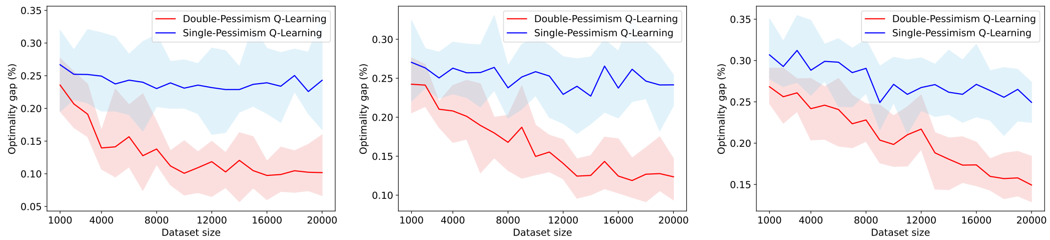
8.1 SIMULATION MDPS

We first evaluate the performance of our algorithm on the Garnet problem (Archibald et al., 1995), a randomly generated MDP $\mathcal{G}(a, b, c)$ with a states, b actions, and c branches (see Appendix A for a more detailed description). Both the nominal kernels and reward functions are generated randomly. The uncertainty set is constructed using the l_∞-norm, with the radius $R_{s,a} \in [0.1, 0.5]$.

We first generate a dataset of size N from the nominal kernel and apply our double-pessimism algorithm, with the single-pessimism baseline (Yan et al., 2022), to learn policies. We then compute the robust value functions of the learned policies and plot the difference between these values and the optimal robust value functions, referred to as the optimality gap, in Figure 1. The results are averaged over 10 times, with the maximum and minimum gaps as an envelope around the average value. The results show that our double-pessimism algorithm converges to the true optimal robust value as the dataset size increases, maintaining a lower optimality gap, while the single-pessimism approach results in a larger gap. These findings demonstrate that our double-pessimism principle significantly enhances robustness while remaining model-free and scalable.

8.2 CLASSIC CONTROL PROBLEMS

To further demonstrate the improvements in both scalability and robustness offered by our approach, we consider more complex Classic Control tasks from OpenAI Gym (Brockman et al., 2016),

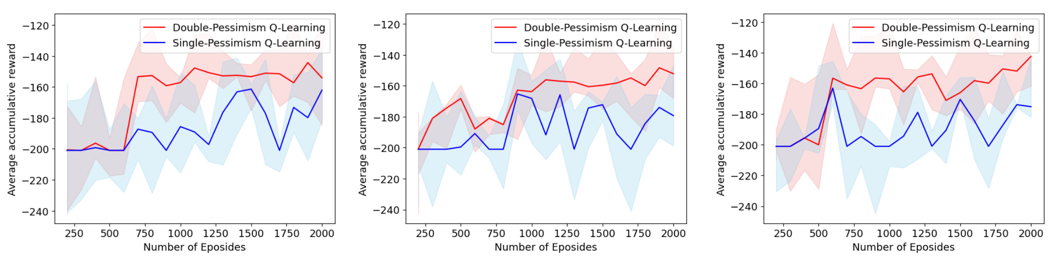


(a) $\mathcal{G}(20, 30, 20)$ (b) $\mathcal{G}(30, 50, 30)$ (c) $\mathcal{G}(50, 100, 50)$

Figure 1: Optimality gaps under different Garnet problems.

specifically MountainCar and CartPole (results are shown in Figure 4 in Appendix). The dynamics of these environments are indirectly controlled by their parameters, e.g., the length of the pole in CartPole, the gravity and the force in MountainCar, and it is of interest to improve the robustness against their uncertainty. Since these model mismatches are hard to model, model-based approaches become ineffective,yet our model-free method remains applicable and effective in such scenarios.

For each dataset generated under the nominal environment with the default parameters, we implemented our algorithm alongside the baseline (Yan et al., 2022) to learn policies. To evaluate the robustness of the learned policies, we test their performance in modified environments with parameter perturbations (Pinto et al., 2017; Wang & Zou, 2021), where we randomly perturbed these parameters within the range of $[-\tau, \tau]$ for 800 trials. As shown in Fig. 2, our double-pessimism algorithm maintains a higher average performance under environment perturbations, demonstrating superior robustness, which aligns with our theoretical findings. This illustrates the enhanced robustness achieved by our approach. Moreover, given the large-scale and complex dynamics of these environments which are difficult for model-based approaches, our model-free algorithm effectively addresses these challenges, further demonstrating the scalability of our method.



(a) $\tau = 0.02$, MountainCar-v0 (b) $\tau = 0.04$, MountainCar-v0 (c) $\tau = 0.06$, MountainCar-v0

Figure 2: Reward profiles with different parameter perturbations.

9 Conclusion

We explored offline RL with a focus on improving scalability and robustness simultaneously. We framed the problem as offline robust RL and developed a model-free algorithm to optimize the worst-case performance within an uncertainty set accounting for the possible model mismatch. To address two key challenges—uncertainty from the under explored dataset and model mismatch between data collection and deployment environments—we introduced a double-pessimism principle that conservatively estimates the agent's performance in a model-free manner. Building on this, we designed a universal model-free algorithm that eliminates the need for model estimation, adapts to various uncertainty sets, and scales to large problems. We further analyzed its performance for the widely studied l_α-norm uncertainty set, showing near-optimal data efficiency of our approach. Our approach significantly improves the robustness, scalability, and efficiency of offline RL compared to existing methods, pushing the boundaries of offline RL research.

ACKNOWLEDGMENT

This work was supported by DARPA under Agreement No. HR0011-24-9-0427 and NSF under Award CCF-2106339. The authors thank the anonymous reviewers whose constructive comments led to substantial improvement to the paper. The authors gratefully acknowledge Xinran Tang at the University of Central Florida for helpful discussions.

REFERENCES

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pp. 67–83. PMLR, 2020a.

Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pp. 104–114. PMLR, 2020b.

Andras Antos, Rémi Munos, and Csaba Szepesvari. Fitted Q-iteration in continuous action-space mdps. In *Neural Information Processing Systems*, 2007.

TW Archibald, KIM McKinnon, and LC Thomas. On the generation of Markov decision processes. *Journal of the Operational Research Society*, 46(3):354–361, 1995.

Kishan Panaganti Badrinath and Dileep Kalathil. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *Proc. International Conference on Machine Learning (ICML)*, pp. 511–520. PMLR, 2021.

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3): 334–334, 1997.

Jose Blanchet, Miao Lu, Tong Zhang, and Han Zhong. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *arXiv preprint arXiv:2305.09659*, 2023.

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

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019.

Esther Derman, Matthieu Geist, and Shie Mannor. Twice regularized MDPs and the equivalence between robustness and regularization. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Jing Dong, Jingwei Li, Baoxiang Wang, and Jingzhao Zhang. Online policy optimization for robust mdp. *arXiv preprint arXiv:2209.13841*, 2022.

Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *Proc. International Conference on Machine Learning (ICML)*, pp. 2701–2709. PMLR, 2020.

Amir Massoud Farahmand, Rémi Munos, and Csaba Szepesvári. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*, 2010.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062. PMLR, 2019.

Seyed Kamyar Seyed Ghasemipour, Dale Schuurmans, and Shixiang Shane Gu. Emaq: Expected-max q-learning operator for simple yet effective offline and online rl. In *International Conference on Machine Learning*, pp. 3682–3691. PMLR, 2021.

Caglar Gulcehre, Ziyu Wang, Alexander Novikov, Tom Le Paine, Sergio Gómez Colmenarejo, Konrad Zolna, Rishabh Agarwal, Josh Merel, Daniel Mankowitz, Cosmin Paduraru, et al. RL unplugged: Benchmarks for offline reinforcement learning. *arXiv preprint arXiv:2006.13888*, 2020.

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

Nan Jiang. On value functions and the agent-environment boundary. *arXiv preprint arXiv:1905.13341*, 2019.

Nan Jiang and Jiawei Huang. Minimax value interval for off-policy evaluation and policy optimization. *Advances in Neural Information Processing Systems*, 33:2747–2758, 2020.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 4868–4878, 2018.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pp. 5084–5096, 2021.

Taylan Kabbani and Ekrem Duman. Deep reinforcement learning approach for trading automation in the stock market. *IEEE Access*, 10:93564–93574, 2022.

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2021.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

Aviral Kumar, Justin Fu, George Tucker, and Sergey Levine. Stabilizing off-policy Q-learning via bootstrapping error reduction. *arXiv preprint arXiv:1906.00949*, 2019.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020.

Navdeep Kumar, Kfir Levy, Kaixin Wang, and Shie Mannor. An efficient solution to s-rectangular robust markov decision processes. *arXiv preprint arXiv:2301.13642*, 2023.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pp. 45–73. Springer, 2012.

Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, pp. 3652–3661. PMLR, 2019.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Gen Li, Laixi Shi, Yuxin Chen, Yuantao Gu, and Yuejie Chi. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *in preparation*, 2022.

Peng Liao, Zhengling Qi, and Susan Murphy. Batch policy learning in average reward Markov decision processes. *arXiv preprint arXiv:2007.11771*, 2020.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural trust region/proximal policy optimization attains globally optimal policy. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Ruo-Ze Liu, Zhen-Jia Pang, Zhou-Yu Meng, Wenhai Wang, Yang Yu, and Tong Lu. On efficient reinforcement learning for full-length game of starcraft ii. *Journal of Artificial Intelligence Research*, 75:213–260, 2022a.

Xiao-Yin Liu, Xiao-Hu Zhou, Guo-Tao Li, Hao Li, Mei-Jiang Gui, Tian-Yu Xiang, De-Xing Huang, and Zeng-Guang Hou. Micro: Model-based offline reinforcement learning with a conservative bellman operator. *arXiv preprint arXiv:2312.03991*, 2023.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*, 2020.

Zhishuai Liu and Pan Xu. Distributionally robust off-dynamics reinforcement learning: Provable efficiency with linear function approximation. *arXiv preprint arXiv:2402.15399*, 2024a.

Zhishuai Liu and Pan Xu. Minimax optimal and computationally efficient algorithms for distributionally robust offline reinforcement learning. *arXiv preprint arXiv:2403.09621*, 2024b.

Zijian Liu, Qinxun Bai, Jose Blanchet, Perry Dong, Wei Xu, Zhengqing Zhou, and Zhengyuan Zhou. Distributionally robust Q-learning. In *International Conference on Machine Learning*, pp. 13623–13643. PMLR, 2022b.

Miao Lu, Han Zhong, Tong Zhang, and Jose Blanchet. Distributionally robust reinforcement learning with interactive data collection: Fundamental hardness and near-optimal algorithm. *arXiv preprint arXiv:2404.03578*, 2024.

Xiaoteng Ma, Zhipeng Liang, Jose Blanchet, Mingwen Liu, Li Xia, Jiheng Zhang, Qianchuan Zhao, and Zhengyuan Zhou. Distributionally robust offline reinforcement learning with linear function approximation. *arXiv preprint arXiv:2209.06620*, 2022.

Rémi Munos. Error bounds for approximate value iteration. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, pp. 1006. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

Arnab Nilim and Laurent El Ghaoui. Robustness in Markov decision problems with uncertain transition matrices. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 839–846, 2004.

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.

Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. *arXiv preprint arXiv:2208.05129*, 2022.

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

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Stephane Ross and J Andrew Bagnell. Agnostic system identification for model-based reinforcement learning. *arXiv preprint arXiv:1203.1007*, 2012.

Bruno Scherrer. Approximate policy iteration schemes: A comparison. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1314–1322, 2014.

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, and Yuejie Chi. Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. *International Conference on Machine Learning*, pp. 19967–20025, 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. *arXiv preprint arXiv:2305.16589*, 2023.

Noah Y Siegel, Jost Tobias Springenberg, Felix Berkenkamp, Abbas Abdolmaleki, Michael Neunert, Thomas Lampe, Roland Hafner, and Martin Riedmiller. Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. *arXiv preprint arXiv:2002.08396*, 2020.

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

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and Q-function learning for off-policy evaluation. In *Proc. International Conference on Machine Learning (ICML)*, pp. 9659–9668. PMLR, 2020.

He Wang, Laixi Shi, and Yuejie Chi. Sample complexity of offline distributionally robust linear markov decision processes. *arXiv preprint arXiv:2403.12946*, 2024a.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2019.

Qiuhao Wang, Chin Pang Ho, and Marek Petrik. Policy gradient in robust mdps with global convergence guarantee, 2023a.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. A finite sample complexity bound for distributionally robust q-learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3370–3398. PMLR, 2023b.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. Sample complexity of variance-reduced distributionally robust q-learning. *arXiv preprint arXiv:2305.18420*, 2023c.

Yuanhao Wang, Kefan Dong, Xiaoyu Chen, and Liwei Wang. Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. In *International Conference on Learning Representations*, 2020.

Yudan Wang, Shaofeng Zou, and Yue Wang. Model-free robust reinforcement learning with sample complexity analysis. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024b. URL `https://openreview.net/forum?id=brZRvwK58H`.

Yue Wang and Shaofeng Zou. Finite-sample analysis of Greedy-GQ with linear function approximation under Markovian noise. In *Proc. International Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 11–20. PMLR, 2020.

Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 7193–7206, 2021.

Yue Wang and Shaofeng Zou. Policy gradient method for robust reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, volume 162, pp. 23484–23526. PMLR, 2022.

Yue Wang, Fei Miao, and Shaofeng Zou. Robust constrained reinforcement learning. *arXiv preprint arXiv:2209.06866*, 2022.

Yue Wang, Alvaro Velasquez, George K Atia, Ashley Prater-Bennette, and Shaofeng Zou. Model-free robust average-reward reinforcement learning. In *International Conference on Machine Learning*, pp. 36431–36469. PMLR, 2023d.

Yue Wang, Jinjun Xiong, and Shaofeng Zou. Achieving the asymptotically optimal sample complexity of offline reinforcement learning: A dro-based approach. *arXiv preprint arXiv:2305.13289*, 2023e.

Yue Wang, Jinjun Xiong, and Shaofeng Zou. Achieving the asymptotically minimax optimal sample complexity of offline reinforcement learning: A DRO-based approach. *Transactions on Machine Learning Research*, 2024c. ISSN 2835-8856. URL `https://openreview.net/forum?id=Y7FbGcjOuD`.

Hua Wei, Jingxiao Chen, Xiyang Ji, Hongyang Qin, Minwen Deng, Siqin Li, Liang Wang, Weinan Zhang, Yong Yu, Liu Linc, et al. Honor of kings arena: an environment for generalization in competitive reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 11881–11892, 2022.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. *arXiv preprint arXiv:2008.04990*, 2020.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pp. 11404–11413. PMLR, 2021.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *arXiv preprint arXiv:2106.06926*, 2021a.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021b.

Huan Xu and Shie Mannor. Distributionally robust Markov decision processes. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 2505–2513, 2010.

Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust pca via outlier pursuit. In *Advances in Neural Information Processing Systems*, pp. 2496–2504, 2010.

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.

Yuling Yan, Gen Li, Yuxin Chen, and Jianqing Fan. The efficacy of pessimism in asynchronous Q-learning. *arXiv preprint arXiv:2203.07368*, 2022.

Rui Yang, Chenjia Bai, Xiaoteng Ma, Zhaoran Wang, Chongjie Zhang, and Lei Han. Rorl: Robust offline reinforcement learning via conservative smoothing. *Advances in neural information processing systems*, 35:23851–23866, 2022.

Rui Yang, Han Zhong, Jiawei Xu, Amy Zhang, Chongjie Zhang, Lei Han, and Tong Zhang. Towards robust offline reinforcement learning under diverse data corruption. *arXiv preprint arXiv:2310.12955*, 2023.

Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Towards theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics. *arXiv preprint arXiv:2105.03863*, 2021.

Ming Yin and Yu-Xiang Wang. Optimal uniform ope and model-based offline reinforcement learning in time-homogeneous, reward-free and task-agnostic settings. *arXiv preprint arXiv:2105.06029*, 2021a.

Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34, 2021b.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1567–1575. PMLR, 2021a.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal offline reinforcement learning via double variance reduction. *arXiv preprint arXiv:2102.01748*, 2021b.

Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason D Lee. Offline reinforcement learning with realizability and single-policy concentrability. *arXiv preprint arXiv:2202.04634*, 2022.

Baohe Zhang, Raghu Rajan, Luis Pineda, Nathan Lambert, André Biedenkapp, Kurtland Chua, Frank Hutter, and Roberto Calandra. On the importance of hyperparameter optimization for model-based reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 4015–4023. PMLR, 2021a.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 4572–4583, 2020.

Xuezhou Zhang, Yiding Chen, Xiaojin Zhu, and Wen Sun. Robust policy gradient against strong data corruption. In *International Conference on Machine Learning*, pp. 12391–12401. PMLR, 2021b.

Xuezhou Zhang, Yiding Chen, Xiaojin Zhu, and Wen Sun. Corruption-robust offline reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 5757–5773. PMLR, 2022.

Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE symposium series on computational intelligence (SSCI)*, pp. 737–744. IEEE, 2020.

Ruida Zhou, Tao Liu, Min Cheng, Dileep Kalathil, PR Kumar, and Chao Tian. Natural actor-critic for robust reinforcement learning with function approximation. *Advances in neural information processing systems*, 36, 2024.

Zhengqing Zhou, Qinxun Bai, Zhengyuan Zhou, Linhai Qiu, Jose Blanchet, and Peter Glynn. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *Proc. International Conference on Artifical Intelligence and Statistics (AISTATS)*, pp. 3331–3339. PMLR, 2021.

A Experimental Setup of Section 8

A.1 Garnet Problems

For simulated MDP environments, we implement Algorithm 3 on Garnet problems $\mathcal{G}(20, 30, 20)$, $\mathcal{G}(30, 50, 30)$ and $\mathcal{G}(50, 100, 50)$. Here, the branch number denotes the number of states that can be achieved after taking an action. The uncertainty radius $R_{s,a}$ is randomly drawn from a uniform distribution ranging from 0.1 to 0.5 for all state-action pairs. The true robust expected values for the Garnet problems, over a certain state distribution, can be obtained via the model-based robust value iteration method. For each problem, we first generate a stochastic behavior policy with partial coverage over state-action pairs. To obtain a near-optimal stochastic behavior policy, we compute the Q-values for the nominal kernel, and adopt a softmax transformation to assign probabilities for all state-action pairs. The randomness (i.e., optimality) of the behavior policy is controlled via temperature parameter $t_{\rm b} = 1$. State-action pairs with probabilities $P_{s,a} \leq 0.03$ (for $\mathcal{G}(20, 30, 20)$), $P_{s,a} \leq 0.02$ (for $\mathcal{G}(30, 50, 30)$) and $P_{s,a} \leq 0.01$ (for $\mathcal{G}(50, 100, 50)$) are then excluded to achieve partial coverage. Finally, non-zero elements are re-normalized to maintain a valid probability distribution. By deploying the behavior policy on the nominal kernel, 10 datasets are generated at each dataset size from $T = 1000$ to $T = 20000$. We compared the double-pessimism method with the single-pessimism method in (Yan et al., 2022). We set $\gamma = 0.95$, $C_{\mathsf{b}} = 1 \times 10^{-4}$ and $\delta = 0.02$.

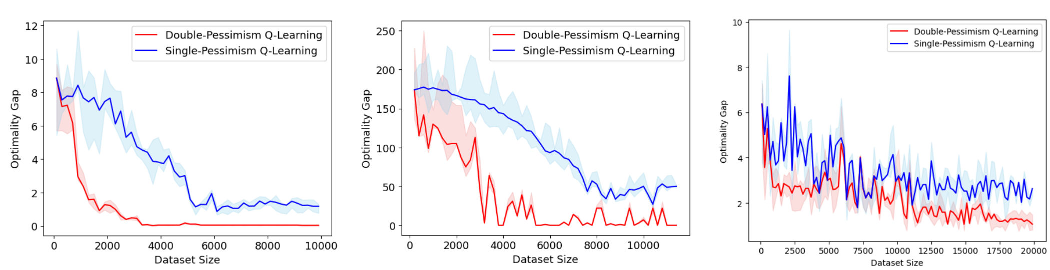
A.2 Classic Control Problems

Note in the Classic Control problems, the underlying uncertain environments may not be modeled using our perturbation-based uncertainty set in equation 8, but we can still implement our algorithms to enhance the robustness. We generate the dataset according to a random behavior policy, and implement Algorithm 3 with the radius $R = \tau$. In our experiments, we set $\gamma = 0.95$, $C_{\mathsf{b}} = 1 \times 10^{-4}$ and $\delta = 0.02$. After a policy is learned, we test its performance under a perturbed environment with the parameter randomly generated from $[-\tau, \tau]$ for 800 times, and plot the average performance among them.

B Additional Experiment Results

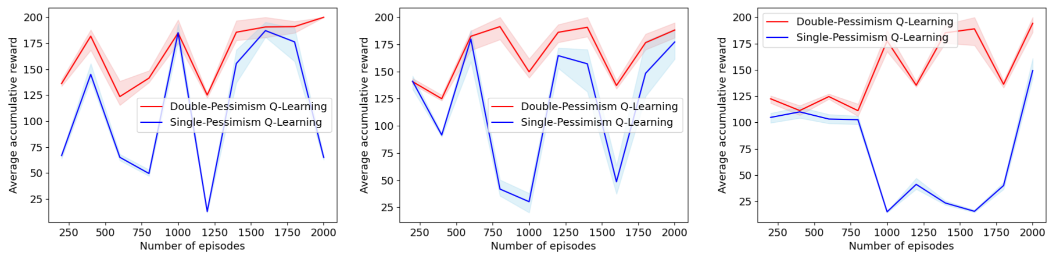
B.1 Comparisons in Tabular Environments

In this section, we include additional experiment results under three simulated environments. Specifically, we consider the Frozen-Lake and Taxi environments from OpenAI Gym (Brockman et al., 2016), and the American Option problem (Panaganti et al., 2022; Shi & Chi, 2022; Zhou et al., 2021). The transition dynamics of these environments can be directly controlled, and we construct l_∞-norm uncertainty sets centered at their nominal kernels. Similarly, we trained our double-pessimism Q-learning together with the single-pessimism baseline, and plotted the optimality gap between the learned and optimal robust value functions. As the results in Figure 3 show, our double-pessimism Q-learning effectively obtains the optimal robust policy, whereas the single-pessimism Q-learning only achieves sub-optimal performance. The results hence indicate that our additional pessimism effectively enhances robustness against model uncertainty, verifying our theoretical results.



(a) Frozen-Lake (b) Taxi (c) American Option

Figure 3: Optimality gaps under different Gym environments.



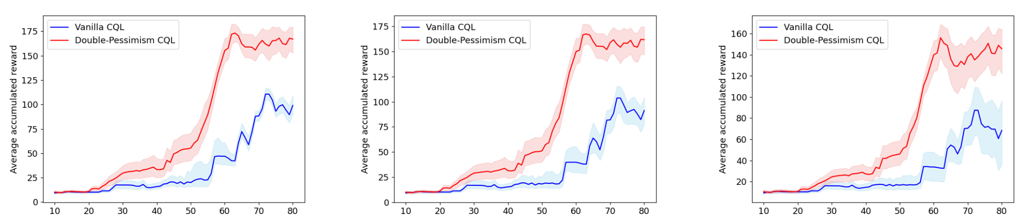
(a) $\tau = 0.6$, CartPole-v0 (b) $\tau = 0.7$, CartPole-v0 (c) $\tau = 0.8$, CartPole-v0

Figure 4: Reward profiles with different parameter perturbations.

B.2 Scalable Algorithm with Function Approximation: Double-Pessimism CQL

In this section, we extend the evaluation of our double-pessimism framework to large-scale problems using function approximation techniques. The algorithms presented earlier (Algorithm 1, Algorithm 3), while model-free, are designed for tabular settings and require memory space of $\mathcal{O}(SA)$ for the Q-table, making them less efficient for large-scale applications. To improve scalability, replacing the Q-table with low-dimensional function approximations (e.g., neural networks) to reduce memory costs is a widely adopted approach. On the other hand, existing offline RL algorithms like Conservative Q-learning (CQL, (Kumar et al., 2020)) and Implicit Q-learning (IQL, (Kostrikov et al., 2021)), along with others (Ross & Bagnell, 2012; Laroche et al., 2019; Fujimoto et al., 2019; Kumar et al., 2019; Agarwal et al., 2020b; Liu et al., 2020; Jin et al., 2021; Xie et al., 2021a; Yin et al., 2021a; Rashidinejad et al., 2021; Xie & Jiang, 2021; Jiang & Huang, 2020), have focused solely on offline RL **without model mismatch**, resulting in degraded performance when model mismatch is present.

Aiming to enhance both robustness and scalability, we design and evaluate a double-pessimism CQL algorithm, demonstrating that our framework is not limited to tabular settings but can also be integrated with function approximation or deep RL techniques, significantly improving robustness against model mismatch. Specifically, we employ the CQL method to impose pessimism on the limited dataset, and further incorporate an additional penalty term into the robust Bellman operator estimation to effectively mitigate model mismatch. Based on this construction, we can similarly design a double-pessimism CQL algorithm, from which enhanced robustness is expected.



(a) $\tau = 0.2$ (b) $\tau = 0.5$ (c) $\tau = 0.8$

Figure 5: Double-Pessimism CQL vs. Vanilla CQL under CartPole.

To validate the effectiveness of our double-pessimism principle, we compare our double-pessimism CQL with the vanilla single-pessimism CQL under CartPole from OpenAI Gym. The policy is trained in the nominal environment and evaluated in randomly perturbed environments (perturbation radius τ) over 800 trials. The results, shown in Figure 5, display the average performance as solid curves, with envelopes representing standard deviations.

As the results indicate, our double-pessimism CQL consistently outperforms the vanilla CQL in perturbed environments, demonstrating enhanced robustness. This experiment confirms the universal applicability of our double-pessimism framework in improving robustness, regardless of the specific algorithm used. It also highlights the scalability of our approach, which can be integrated with advanced deep offline RL algorithms for large-scale problems using function approximation.

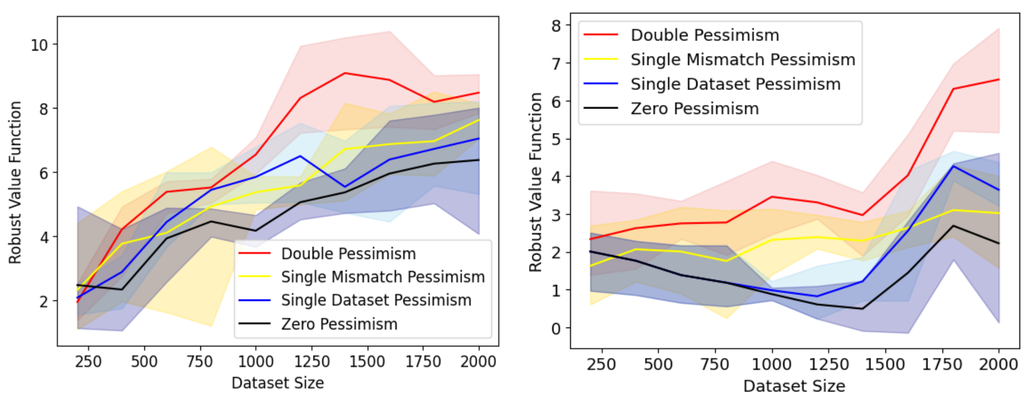
B.3 Ablation Experiments

Our double-pessimism principle addresses two key challenges: the first component tackles the limited dataset coverage in offline RL to handle out-of-distribution issues, while the second addresses model mismatch between the data generation and deployment environments.

In this section, we conduct ablation experiments to evaluate the effectiveness of this principle. Specifically, we compare four algorithms in an offline setting: vanilla Q-learning (with zero pessimism), robust Q-learning (with model-mismatch pessimism only), offline non-robust Q-learning (with dataset pessimism only), and our proposed offline robust Q-learning (with double pessimism). The experiments are conducted on two Garnet problems, where we evaluate the robust value functions of the learned policies with respect to an uncertainty set defined by the l_∞-norm.

The results are shown in Figure 6. The solid curve represents the average value across 10 independent runs, while the shaded area indicates the maximum and minimum values observed.

Our double-pessimism approach outperforms all four algorithms, including those with a single source of pessimism, demonstrating the effectiveness of our framework. Furthermore, the single-pessimism methods achieve better performance than the vanilla algorithm with no pessimism, highlighting the benefits of incorporating pessimism in offline robust RL. However, both are ultimately outperformed by our double-pessimism method, underscoring the importance of addressing both sources of uncertainty through the double-pessimism principle.



(a) $\mathcal{G}(30, 10, 30)$ (b) $\mathcal{G}(40, 20, 40)$

Figure 6: Robust value functions in Garnet problems.

C Further Discussion of κ

C.1 A Universal construction of κ

In this section, we discuss the design of the penalty function κ for universal uncertainty set models defined by some distribution divergence/distance functions $F(\cdot||\cdot)$:

$$\mathcal{P} = \{P + q \in \Delta(\mathcal{S}) : F(P + q||P) \leq R\}. \tag{24}$$

Note that this uncertainty set includes perturbed environments within a region centered around the nominal kernel, effectively modeling environmental uncertainty in practical applications. This is because, in practice, perturbed environments should not deviate significantly from the nominal kernel and should therefore fall within a defined region.

We first present the following theorem for a universal construction of the penalty function κ.

Theorem 4. *Let $\kappa(V)$ be the optimal value of the following constrained problem:*

$$\max_q -\sum q_i V_i, \quad s.t. \sum_i q_i = 0, \ \ F(P + q||P) \leq R. \tag{25}$$

Then, $\kappa(V)$ satisfies equation 15, i.e.,

$$PV - \kappa(V) \leq \sigma_{\mathcal{P}}(V). \tag{26}$$

Proof. Note that the problem in equation 25 is equivalent to the problem

$$\max_{q \in \mathcal{Q}} -qV, \text{ where } \mathcal{Q} = \{q \in \mathbb{R}^S, \sum_i q_i = 0, F(P+q||P) \leq R\}. \tag{27}$$

The proof is then straightforward by noting that $\mathcal{P} \subset \mathcal{Q}$, hence

$$PV - \kappa(V) \leq \min_{p \in \mathcal{P}} pV = \sigma_{\mathcal{P}}(V). \tag{28}$$

□

Such a result provides a universal construction of the penalty function κ, for the perturbed-based uncertainty set as in equation 24. Note that $\kappa(V)$ depends on P, which is unknown in practice, but any unbiased estimation of it is sufficient. To illustrate this and show the generality of our design, we develop a case study for the χ^2-divergence uncertainty set in the following section.

C.2 Case Study: l_α-norm Uncertainty set

In this section, we provide a more detailed discussion on the l_α-norm uncertainty set. As discussed, we consider the relaxed l_α-norm uncertainty set:

$$\tilde{\mathcal{P}}_{s,a} = \{P_{s,a} + q : \sum q_i = 0, \|q\|_\alpha \leq R_{s,a}\}, \tag{29}$$

where we relax the condition $P_{s,a} + q \geq 0$. Then the worst-case transition w.r.t. $\tilde{\mathcal{P}}$ can be derived as

$$\sigma_{\tilde{\mathcal{P}}_{s,a}}(V) = P_{s,a}V - \kappa(V), \tag{30}$$

where

$$\kappa(V) \triangleq R \min_{w \in \mathbb{R}} \|we - V\|_\beta, \tag{31}$$

with $\beta = \frac{1}{1-\frac{1}{\alpha}}$. For popular choices of α, the optimization problem in equation 31 has a closed-form solution, specified in Table 2 (Kumar et al., 2023). Note that for the three choices of $\alpha = 1, 2, \infty$,

α	$\kappa(v)$
∞	$\frac{\max_s v(s) - \min_s v(s)}{2}$
2	$\sqrt{\sum_s \left(v(s) - \frac{\sum_s v(s)}{S}\right)^2}$
1	$\sum_{i=1}^{\lfloor (S+1)/2 \rfloor} v(s_i) - \sum_{i=\lfloor (S+1)/2 \rfloor}^{S} v(s_i)$

Table 2: Penalty term for l_α-norm uncertainty set

the resulting penalty terms incur a computational complexity of $\mathcal{O}(S)$. When combined with our algorithm, this leads to an overall implementation complexity of $\mathcal{O}(SA)$ per step. In contrast, the model-based methods proposed in (Shi & Chi, 2022; Blanchet et al., 2023) have a computational cost of $\mathcal{O}(S^2A)$ per step (Kumar et al., 2023), highlighting the superior efficiency and scalability of our approach.

Our algorithm and theoretical result will then characterize the convergence to the optimal robust policy w.r.t. $\tilde{\mathcal{P}}$. More importantly, when the uncertainty radius R is small, the relaxation will not be effective, i.e., $\tilde{\mathcal{P}} = \mathcal{P}$ (Zhou et al., 2024).

C.3 CASE STUDY: χ^2 UNCERTAINTY SET

We adapt the construction we obtained to the widely used χ^2-divergence as a case study. The design of κ for other uncertainty sets can be obtained in a similar way.

Specifically, the uncertainty defined for the (s, a)-pair is

$$\mathcal{P}_{s,a} = \{P_{s,a} + q \in \Delta(\mathcal{S}) : D_{\chi^2}(P_{s,a} + q||P_{s,a}) \le R_{s,a}\}, \tag{32}$$

where $D_{\chi^2}(p||q) = \sum_i \frac{(p_i - q_i)^2}{q_i}$ is the χ^2-divergence. We aim to design a model-free function κ that serves as the penalty term to address the uncertainty from the model mismatch.

We first establish the following lemma.

Lemma 5. *The constrained problem*

$$\min_q \sum_i q_i V_i, \text{ s.t. } \sum_i q_i = 0, D_{\chi^2}(q + P_{s,a}||P_{s,a}) \le R_{s,a} \tag{33}$$

has the solution

$$-\sqrt{R_{s,a} \mathrm{Var}_{P_{s,a}}(V)}. \tag{34}$$

Proof. To simplify the notation, we omit the subscript s, a from $P_{s,a}$ and $R_{s,a}$. We note that if any entry $P_i = 0$, then any feasible $q_i = 0$, otherwise the χ^2-divergence will be infinite. Thus, we can simply ignore the i-th entry in this case and only consider the remaining ones. Hence, we assume $P_i > 0, \forall i$ without loss of generality.

Note that the condition $D_{\chi^2}(q + P||P) \le R$ is equivalent to

$$\sum_i \frac{q_i^2}{P_i} \le R, \tag{35}$$

hence the Lagrangian function L of the constrained problem is

$$L = \sum_i q_i V_i + \lambda \sum_i q_i + \mu \left(\sum_i \frac{q_i^2}{P_i} - R \right). \tag{36}$$

From the KKT conditions (Bertsekas, 1997), the solution q^* and the Lagrangian multipliers λ^* and μ^* must satisfy

$$V_i + \lambda^* + \mu^* \frac{2q_i^*}{P_i} = 0, \forall i. \tag{37}$$

We first show that if q^* is the optimal solution, then $D_{\chi^2}(q^* + P||P) < R$. To show that this statement always holds, our claim is that there exists an optimal solution such that $\mu^* = 0$, then we have

$$V_i + \lambda^* = 0, \forall i \tag{38}$$

and hence,

$$\sum_i q_i^* V_i = -\lambda^* \sum_i q_i^* = 0. \tag{39}$$

To prove that this claim is not possible, we provide a counterexample to demonstrate that $\mu^* = 0$ and $\sum_i q_i^* V_i \ne 0$, which is a contradiction:

For $V = [V_1, V_2]$ and $P_{s,a} = [p_1, p_2]$, where $p_1, p_2 \ne 0$. We have $q_2 = -q_1$, then

$$\frac{q_1^2}{p_1} + \frac{q_2^2}{p_2} < R. \tag{40}$$

Hence,

$$|q_1| < \sqrt{\frac{R}{\frac{1}{p_1} + \frac{1}{p_2}}} \tag{41}$$

The optimal value of the optimization problem is

$$\sum_i q_i^* V_i = q_1^*(V_1 - V_2). \tag{42}$$

Obviously, the optimization problem does not have an optimal solution, but instead an infimum. There always exists a feasible solution $\hat{q}$ such that

$$\sum_i \hat{q}_i V_i < 0, \tag{43}$$

which means that $\mu^* \neq 0$ always holds.

Thus,

$$q_i^* V_i = -\lambda^* q_i^* - 2\mu^* \frac{(q_i^*)^2}{P_i}, \forall i, \tag{44}$$

and hence,

$$\sum_i q_i^* V_i = -2\mu^* \sum_i \frac{(q_i^*)^2}{P_i} = -2\mu^* R, \tag{45}$$

where we use the constraint $\sum_i q_i^* = 0$ and $\sum_i \frac{(q_i^*)^2}{P_i} = R$.

Again, from equation 37, we have that

$$4(\mu^*)^2 \left(\frac{q_i^*}{P_i}\right)^2 = (V_i + \lambda^*)^2, \tag{46}$$

and hence,

$$\left(\frac{q_i^*}{P_i}\right)^2 = \frac{(V_i + \lambda^*)^2}{4(\mu^*)^2}. \tag{47}$$

Taking the sum over i implies that

$$\sum_i \frac{(q_i^*)^2}{P_i} = R = \sum P_i \frac{(V_i + \lambda^*)^2}{4(\mu^*)^2}, \tag{48}$$

and hence,

$$2\mu^* R = \sqrt{R \sum_i P_i (V_i + \lambda^*)^2}. \tag{49}$$

On the other hand, note that equation 37 further implies that

$$0 = \sum_i P_i V_i + \lambda^* \sum_i P_i, \tag{50}$$

and hence,

$$\lambda^* = -\sum_i P_i V_i. \tag{51}$$

Plugging in equation 49 implies that

$$2\mu^* R = \sqrt{R \mathrm{Var}_P(V)}. \tag{52}$$

Hence, from equation 45, the optimal solution of the constrained problem is then $-\sqrt{R\mathrm{Var}_P(V)}$, which completes the proof. $\square$

With the optimal solution to equation 33, we can then design the penalty function κ for the χ^2 uncertainty set defined as in equation 32. Firstly, we note that equation 33 is a relaxation of the support function over equation 32, therefore the optimal solution to equation 33 is not greater than

$\sigma_{\mathcal{P}}(V)$, and therefore is a pessimistic penalty of the model mismatch. We thus design the penalty function as

$$\kappa(V) = \sum_i P_i V_i - \sqrt{R\mathrm{Var}_P(V)}. \tag{53}$$

We note that in the model-free setting, it is straightforward to obtain an unbiased estimation of κ, which however requires more than 1 sample. Specifically, for n i.i.d. samples $(s, a, s'_i), i = 1, ..., n$, the model-free penalty function is defined as

$$\kappa(V) = \bar{V} - \sqrt{R}\sqrt{\frac{\sum_{i=1}^n (V(s'_i) - \bar{V})^2}{n-1}}, \tag{54}$$

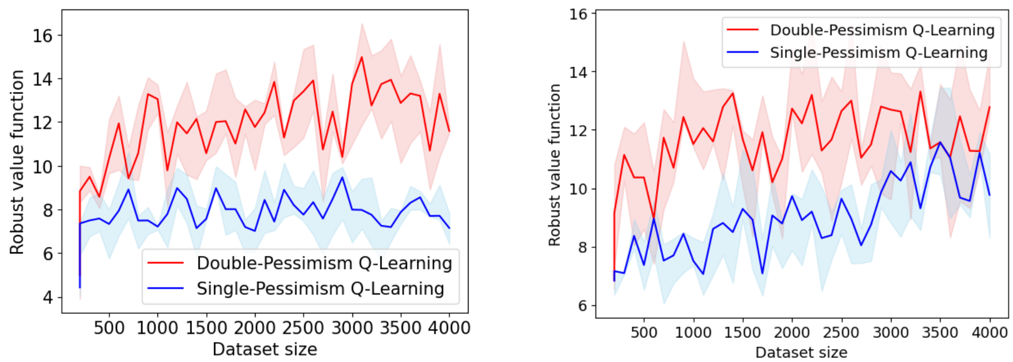
where $\bar{V} = \frac{\sum_i V(s'_i)}{n}$. Such a penalty function satisfies the condition equation 15 of the pessimism principle, and hence we can extend our model-free algorithms to the χ^2-divergence model. We present the algorithm for the infinite horizon in Algorithm 2. Different from Algorithm 3, for the

Algorithm 2 Double-Pessimism Q-Learning for infinite-horizon RMDPs with χ^2-divergence uncertainty set.

Input: $\mathcal{D}$, target success probability $1-\delta$, $\Gamma = \left\lceil \frac{4}{1-\gamma} \log \frac{ST}{\delta} \right\rceil$
Initialize: $Q_0(s,a) = 0$, $V_0(s) = 0$, $n_0(s,a) = 0, \forall s, a$
for $t = 1, ..., T$ **do**
 Sample 2 samples $(s_{t-1}, a_{t-1}, s_t^1)$,$(s_{t-1}, a_{t-1}, s_t^2)$ from $\mathcal{D}$
 $n_t(s_{t-1}, a_{t-1}) \leftarrow n_{t-1}(s_{t-1}, a_{t-1}) + 2$; $n_t(s,a) \leftarrow n_{t-1}(s,a), \forall (s,a) \neq (s_{t-1}, a_{t-1})$
 $n \leftarrow n_t(s,a)$; $\eta_n \leftarrow (\Gamma+1)/(\Gamma+n)$
 $b_n \leftarrow c_b \sqrt{\frac{\Gamma \log(ST/\delta)}{n(1-\gamma)^2}}$
 $M \leftarrow \frac{V_{t-1}(s_t^1) + V_{t-1}(s_t^2)}{2}$
 $\kappa \leftarrow -\sqrt{R_{s_t,a_t}\left((V_{t-1}(s_t^1) - M)^2 + (V_{t-1}(s_t^2) - M)^2\right)}$
 $Q_t(s_{t-1}, a_{t-1}) = (1-\eta_n) Q_{t-1}(s_{t-1}, a_{t-1}) + \eta_n \left\{ r(s_{t-1}, a_{t-1}) + \gamma M - \gamma\kappa - b_n \right\}$
 $Q_t(s,a) = Q_{t-1}(s,a)$ for all $(s,a) \neq (s_{t-1}, a_{t-1})$
 $V_t(s_{t-1}) = \max\left\{ \max_{a\in\mathcal{A}} Q_t(s_{t-1}, a),\ V_{t-1}(s_{t-1}) \right\}$,
 $V_t(s) = V_{t-1}(s)$ for all $s \neq s_{t-1}$.
end for
$\widehat{\pi}(s) = \arg\max_{a\in\mathcal{A}} Q_T(s,a), \forall s$
Output: $\widehat{\pi}$

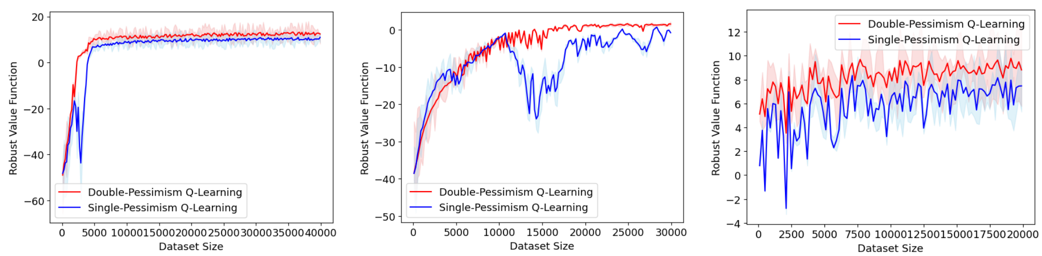
χ^2-divergence model, we require 2 samples at each step to estimate κ. However, the estimation does not required any information on $P_{s,a}$ and hence Algorithm 3 is still model-free.

Note that generally the penalty function κ is biased, thus the algorithm may not converge to the optimal robust policy. However, the robustness can still be enhanced due to the additional pessimism. We validate the effectiveness of our algorithm in optimizing performance under model mismatch in an offline setting through numerical experiments. Specifically, we implemented our algorithm alongside the baseline single-pessimism Q-learning algorithm (Yan et al., 2022) on Garnet problems with varying parameters, and three simulation environments: Frozen-Lake, Taxi, and American Option. Using datasets of different sizes, we computed the robust value function of the learned policy via dynamic programming (Iyengar, 2005), and plotted the results in Figure 7 and Figure 8. Each curve represents the average over 10 independent runs, with the shaded region indicating the maximum and minimum values. As demonstrated in the results, our double-pessimism Q-learning significantly outperforms the single-pessimism approach, showcasing the robustness of our algorithm to model mismatch and confirming the efficacy of our double-pessimism design.



(a) $\mathcal{G}(15, 8, 15)$ (b) $\mathcal{G}(20, 10, 20)$

Figure 7: Robust value functions of two Granet problems over χ^2-divergence uncertainty set. Solid lines represent the mean values over 10 independent runs. Shaded areas represent the maximum and minimum values.



(a) Frozen-Lake (b) Taxi (c) American Option

Figure 8: Robust value functions of three simulation environments over the χ^2-divergence uncertainty set. Solid lines represent the mean values over 10 independent runs. Shaded areas represent the maximum and minimum values.

D Analysis of the finite horizon setting

D.1 Notation

Recall the learning rate defined by

$$\eta_n = \frac{H+1}{H+n} \tag{55}$$

for the n-th visit of a given state-action pair at a given time step h. We further adopt two sequences of related quantities for any integers $N \geq 0$ and $n \geq 1$ from (Shi et al., 2022):

$$\eta_0^N \triangleq \begin{cases} \prod_{i=1}^{N}(1-\eta_i) = 0, & \text{if } N > 0, \\ 1, & \text{if } N = 0 \end{cases}, \tag{56}$$

$$\eta_n^N \triangleq \begin{cases} \eta_n \prod_{i=n+1}^{N}(1-\eta_i), & \text{if } N > n, \\ \eta_n, & \text{if } N = n, \\ 0, & \text{if } N < n \end{cases}. \tag{57}$$

It has been shown in (Shi et al., 2022; Yan et al., 2022) that

$$\sum_{n=0}^{N} \eta_n^N = 1. \tag{58}$$

We also introduce the following notation:

- $N_h^k(s,a)$, or simply N_h^k: The number of episodes that have visited the state-action pair (s,a) at step h before the start of the k-th episode.
- $k_h^n(s,a)$, or simply k^n: The index of the episode in which the state-action pair (s,a) is visited at step h for the n-th time. We adopt the convention that $k^0 = 0$.
- $P_h^k \in \{0,1\}^{1\times S}$: A row vector corresponding to the empirical transition at step h of the k-th episode, defined as

$$P_h^k(s) = \mathbf{1}\big(s = s_{h+1}^k\big) \qquad \text{for all } s \in \mathcal{S}. \tag{59}$$

- $\pi^k = \{\pi_h^k\}_{h=1}^H$ with $\pi_h^k(s) \triangleq \arg\max_a Q_h^k(s,a)$ for all $(h,s) \in [H]\times\mathcal{S}$: The deterministic greedy policy at the beginning of the k-th episode.
- $\widehat{\pi}$: The final output of the algorithm, corresponding to π^{K+1} as defined above. For simplicity in our analysis, we treat $\widehat{\pi}$ as π^K, which does not affect the result.

D.2 Lemmas for Theorem 2

In this section, we present the lemmas that are utilized in the proof of Theorem 2.

The first lemma demonstrates how our choice of the penalty term κ can address the uncertainty arising from model mismatch.

Lemma 6. *(Theorem 1 in (Kumar et al., 2023)) Let $\mathcal{P}_{s,a}$ be the uncertainty set defined using the l_α-norm. For any vector V, the following relationship holds:*

$$\sigma_{\mathcal{P}_{s,a}}(V) = P_{s,a}V - \kappa_{s,a}(V), \tag{60}$$

where κ is defined as in equation 18.

The following lemma provides properties concerning the learning rates and is adapted from (Jin et al., 2018; Li et al., 2021).

Lemma 7 (Lemma 1 in (Li et al., 2021))**.** *For any integer $N > 0$, the following properties hold:*

$$\frac{1}{N^a} \le \sum_{n=1}^{N} \frac{\eta_n^N}{n^a} \le \frac{2}{N^a} \qquad \textit{for all} \quad \frac{1}{2} \le a \le 1, \tag{61a}$$

$$\max_{1\le n\le N} \eta_n^N \le \frac{2H}{N}, \qquad \sum_{n=1}^{N} (\eta_n^N)^2 \le \frac{2H}{N}, \qquad \sum_{n=N}^{\infty} \eta_n^N \le 1 + \frac{1}{H}. \tag{61b}$$

The following lemmas concern the concentration properties of the sample generation.

The first lemma below is adapted from Xie et al. (2021b, Lemma A.1).

Lemma 8. *(Lemma 8 in (Shi et al., 2022)) Suppose $N \sim \mathsf{Binomial}(n,p)$, where $n \ge 1$ and $p \in [0,1]$. For any $\delta \in (0,1)$, we have*

$$\frac{p}{N \vee 1} \le \frac{8\log\left(\frac{1}{\delta}\right)}{n}, \tag{62}$$

and

$$N \ge \frac{np}{8\log\left(\frac{1}{\delta}\right)} \qquad \textit{if } np \ge 8\log\left(\frac{1}{\delta}\right), \tag{63a}$$

$$N \le \begin{cases} e^2 np & \textit{if } np \ge \log\left(\frac{1}{\delta}\right), \\ 2e^2\log\left(\frac{1}{\delta}\right) & \textit{if } np \le 2\log\left(\frac{1}{\delta}\right). \end{cases} \tag{63b}$$

with probability at least $1 - 4\delta$.

The following lemma is a standard concentration inequlity result.

Theorem 9 (Freedman's inequality (Freedman, 1975)). *Consider a filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$, and let $\mathbb{E}_k$ stand for the expectation conditioned on $\mathcal{F}_k$. Suppose that $Y_n = \sum_{k=1}^{n} X_k \in \mathbb{R}$, where $\{X_k\}$ is a real-valued scalar sequence obeying*

$$|X_k| \le R \qquad \text{and} \qquad \mathbb{E}_{k-1}\big[X_k\big] = 0 \qquad \text{for all } k \ge 1$$

for some quantity $R < \infty$. We also define

$$W_n \coloneqq \sum_{k=1}^{n} \mathbb{E}_{k-1}\left[X_k^2\right].$$

In addition, suppose that $W_n \le \sigma^2$ holds deterministically for some given quantity $\sigma^2 < \infty$. Then, for any positive integer $m \ge 1$, with probability at least $1-\delta$ one has

$$|Y_n| \le \sqrt{8\max\left\{W_n, \frac{\sigma^2}{2^m}\right\}\log\frac{2m}{\delta}} + \frac{4}{3}R\log\frac{2m}{\delta}. \tag{64}$$

The Freedman's inequality further implies several important results related to our problem.

Lemma 10. *Let $\{W_h^i \in \mathbb{R}^S \mid 1 \le i \le K, 1 \le h \le H+1\}$ be a collection of vectors satisfying the following properties:*

- *W_h^i is fully determined by the samples collected up to the end of the $(h-1)$-th step of the i-th episode;*
- *$\|W_h^i\|_\infty \le C_{\mathrm{w}}$.*

For any positive integer $N \ge H$, consider the following sequence:

$$X_i(s,a,h,N) \triangleq \eta^N_{N_h^i(s,a)}\big(P_h^i W_{h+1}^i - R_{s,a}\kappa(W_{h+1}^i) - \sigma_{h,s,a}(W_{h+1}^i)\big)\boldsymbol{1}\big\{(s_h^i, a_h^i) = (s,a)\big\}. \tag{65}$$

With probability at least $1-\delta$,

$$\left|\sum_{i=1}^{k} X_i(s,a,h,N)\right| \lesssim \sqrt{\frac{H}{N}C_{\mathrm{w}}^2}\log^2\frac{SAT}{\delta} \tag{66}$$

holds simultaneously for all $(k,h,s,a,N) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}\times[K]$.

Proof. Let $u_h^i(s,a,N) = \eta^N_{N_h^i(s,a)}$. From equation 61b in Lemma 7, we have

$$\big|u_h^i(s,a,N)\big| \le \frac{2H}{N} \triangleq C_{\mathrm{u}}.$$

Given that $\mathrm{Var}_{h,s,a}\big(W_{h+1}^{k_h^n(s,a)}\big) \le C_{\mathrm{w}}^2$, we can apply Lemma 7 from (Li et al., 2021) to obtain, with probability at least $1-\delta$,

$$\begin{aligned}
&\left|\sum_{i=1}^{k} X_i(s,a,h,N)\right| \\
&\lesssim \sqrt{C_{\mathrm{u}}\log^2\frac{SAT}{\delta}}\sqrt{\sum_{n=1}^{N_h^k(s,a)}\eta_n^N C_{\mathrm{w}}^2} + \left(C_{\mathrm{u}}C_{\mathrm{w}} + \sqrt{\frac{C_{\mathrm{u}}}{N}}C_{\mathrm{w}}\right)\log^2\frac{SAT}{\delta} \\
&\lesssim \sqrt{\frac{H}{N}\log^2\frac{SAT}{\delta}}\cdot C_{\mathrm{w}} + \frac{HC_{\mathrm{w}}}{N}\log^2\frac{SAT}{\delta} \\
&\lesssim \sqrt{\frac{HC_{\mathrm{w}}^2}{N}}\log^2\frac{SAT}{\delta},
\end{aligned}$$

where the final line uses equation 61b from Lemma 7 again. $\square$

Lemma 11. *Let* $\left\{W_h^k(s,a) \in \mathbb{R}^S \mid (s,a) \in \mathcal{S} \times \mathcal{A}, 1 \leq k \leq K, 1 \leq h \leq H+1\right\}$ *be a collection of vectors satisfying the following properties:*

- $W_h^k(s,a)$ *is fully determined by the given state-action pair* (s,a) *and the samples collected up to the end of the* $(k-1)$*-th episode;*
- $\|W_h^k(s,a)\|_\infty \leq C_{\mathrm{w}}$.

For any positive $C_{\mathrm{d}} \geq 0$*, consider the following sequences:*

$$X_{h,k} \triangleq C_{\mathrm{d}} \left[\frac{d_{P,h}^{\pi_\star}(s_h^k, a_h^k)}{d_{P,h}^{\mu}(s_h^k, a_h^k)} W_{h+1}^k(s_h^k, a_h^k) - \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{P,h}^{\pi_\star}(s,a) W_{h+1}^k(s,a)\right], \tag{67}$$

$$\overline{X}_{h,k} \triangleq C_{\mathrm{d}} \left[\frac{d_{P,h}^{\pi_\star}(s_h^k, a_h^k)}{d_{P,h}^{\mu}(s_h^k, a_h^k)} W_{h+1}^k(s_h^k, a_h^k) - \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{P,h}^{\pi_\star}(s,a) W_{h+1}^k(s,a)\right]. \tag{68}$$

Consider any $\delta \in (0,1)$*. Then with probability at least* $1-\delta$*,*

$$\left|\sum_{k=1}^K X_{h,k}\right| \leq \sqrt{\sum_{k=1}^K 8C_{\mathrm{d}}^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{P,h}^{\pi_\star}(s,a) \left[P_{h,s,a} W_{h+1}^k(s,a)\right]^2 \log\frac{2H}{\delta}} + 2C_{\mathrm{d}} C^\star C_{\mathrm{w}} \log\frac{2H}{\delta}, \tag{69}$$

$$\left|\sum_{k=1}^K \overline{X}_{h,k}\right| \leq \sqrt{\sum_{k=1}^K 8C_{\mathrm{d}}^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{P,h}^{\pi_\star}(s,a) P_{h,s,a} \left[W_{h+1}^k(s,a)\right]^2 \log\frac{2H}{\delta}} + 2C_{\mathrm{d}} C^\star C_{\mathrm{w}} \log\frac{2H}{\delta}, \tag{70}$$

hold simultaneously for all $h \in [H]$.

Proof. The proof similarly follows from (Shi et al., 2022). □

We then prove Lemma 1 showing the effectiveness of our double pessimism principle, i.e., that our estimation is a conservative estimation of the robust value function.

Lemma 12. *Consider any* $\delta \in (0,1)$*, and suppose that* $c_b > 0$ *is some sufficiently large constant. Then, with probability at least* $1-\delta$*,*

$$\left|\sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \Big(\sigma_{h,s,a}(V_{h+1}^{k^n(s,a)}) - P_h^{k^n(s,a)} V_{h+1}^{k^n(s,a)} + R_{s,a}\kappa(V_{h+1}^{k^n(s,a)})\Big)\right| \leq \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} b_n \tag{71}$$

holds simultaneously for all $(k,h,s,a) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}$*, and*

$$V_h^k(s) \leq V_h^{\pi^k}(s) \leq V_h^\star(s) \tag{72}$$

holds simultaneously for all $(k,h,s) \in [K]\times[H]\times\mathcal{S}$.

Proof. **Proof of inequality equation 71.** We show it by invoking Lemma 10. Let

$$W_{h+1}^i \coloneqq V_{h+1}^i,$$

which satisfies

$$\|W_{h+1}^i\|_\infty \leq H =: C_{\mathrm{w}}.$$

Note that it holds that

$$\begin{aligned}
&\sigma_{h,s,a}(V_{h+1}^{k^n(s,a)}) - P_h^{k^n(s,a)} V_{h+1}^{k^n(s,a)} + R_{s,a}\kappa(V_{h+1}^{k^n(s,a)}) \\
&= P_{h,s,a} V_{h+1}^{k^n(s,a)} - R_{s,a}\kappa(V_{h+1}^{k^n(s,a)}) - P_h^{k^n(s,a)} V_{h+1}^{k^n(s,a)} + R_{s,a}\kappa(V_{h+1}^{k^n(s,a)})
\end{aligned}$$

$$= P_{h,s,a}V_{h+1}^{k^n(s,a)} - P_h^{k^n(s,a)}V_{h+1}^{k^n(s,a)}, \tag{73}$$

where the first equation is from Lemma 6. Hence applying Lemma 10 implies that with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)}\Big(\sigma_{h,s,a}(V_{h+1}^{k^n(s,a)}) - P_h^{k^n(s,a)}V_{h+1}^{k^n(s,a)} + R_{s,a}\kappa(V_{h+1}^{k^n(s,a)})\Big)\\
&= \left|\sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)}\Big(P_{h,s,a} - P_h^{k^n(s,a)}\Big)V_{h+1}^{k^n(s,a)}\right|\\
&= \left|\sum_{i=1}^{k} X_i\big(s,a,h,N_h^k(s,a)\big)\right|\\
&\le c_b\sqrt{\frac{H^3\iota^2}{N_h^k(s,a)}}
\end{aligned} \tag{74}$$

holds simultaneously for all $(s,a,k,h)\in\mathcal{S}\times\mathcal{A}\times[K]\times[H]$, provided that the constant $c_b>0$ is large enough and that $N=N_h^k(s,a)>0$. When $N_h^k(s,a)=0$, we have the trivial bound

$$\left|\sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)}\Big(P_{h,s,a} - P_h^{k^n(s,a)}\Big)V_{h+1}^{k^n(s,a)}\right| = 0. \tag{75}$$

Additionally, from the definition $b_n = c_b\sqrt{\frac{H^3\iota^2}{n}}$, we observe that

$$\begin{cases}\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}b_n \in \left[c_b\sqrt{\frac{H^3\iota^2}{N_h^k(s,a)}}, 2c_b\sqrt{\frac{H^3\iota^2}{N_h^k(s,a)}}\right], & \text{if } N_h^k(s,a)>0\\ \sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}b_n = 0, & \text{if } N_h^k(s,a)=0\end{cases} \tag{76}$$

holds simultaneously for all $s,a,h,k\in\mathcal{S}\times\mathcal{A}\times[H]\times[K]$, which follows directly from the property equation 61a in Lemma 7.

Combining the above, equation 74 and equation 76 hence imply that

$$\begin{aligned}
&\left|\sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)}\Big(\sigma_{h,s,a}(V_{h+1}^{k^n(s,a)}) - P_h^{k^n(s,a)}V_{h+1}^{k^n(s,a)} + R_{s,a}\kappa(V_{h+1}^{k^n(s,a)})\Big)\right|\\
&\le \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} b_n.
\end{aligned}$$

Proof of inequality equation 72. Note that the second inequality of equation 72 is straightforward as

$$V_h^\pi(s)\le V^\star(s)$$

holds for any policy π. As a consequence, it suffices to establish the first inequality of equation 72:

$$V_h^k(s)\le V_h^{\pi^k}(s) \qquad \text{for all } (s,h,k)\in\mathcal{S}\times[H]\times[K]. \tag{77}$$

Define

$$k_o(h,k,s) := \max\Big\{l : l<k \text{ and } V_h^l(s)=\max_a Q_h^l(s,a)\Big\} \tag{78}$$

for any $(h,k,s)\in[H]\times[K]\times\mathcal{S}$, which denotes the index of the latest episode — before the end of the $(k-1)$-th episode — in which $V_h(s)$ has been updated. We abbreviate $k_o(h,k,s)$ as $k_o(h)$ whenever it is clear from the context.

We utilize an induction approach to show that. Assume that

$$V_\Gamma^{k'}(s)\le V_\Gamma^{\pi^{k'}}(s) \qquad \text{for all } (k',\Gamma,s)\in[k-1]\times[H+1]\times\mathcal{S}, \tag{79a}$$

$$V_\Gamma^k(s) \le V_\Gamma^{\pi^k}(s) \qquad \text{for all } \Gamma \ge h+1 \text{ and } s \in \mathcal{S}. \tag{79b}$$

We need to verify

$$V_h^k(s) \le V_h^{\pi^k}(s) \qquad \text{for all } s \in \mathcal{S}. \tag{80}$$

Step 1: base case.

Let us begin with the base case when $h+1 = H+1$ for all episodes $k \in [K]$. Recognizing the fact that $V_{H+1}^{\pi} = V_{H+1}^k = 0$ for any π and any $k \in [K]$, we directly arrive at

$$V_{H+1}^k(s) \le V_{H+1}^{\pi^k}(s) \qquad \text{for all } (k,s) \in [K] \times \mathcal{S}. \tag{81}$$

Step 2: induction. To justify equation 80 under the induction hypothesis equation 79, we decompose the difference term to obtain

$$\begin{aligned} V_h^{\pi^k}(s) - V_h^k(s) &= V_h^{\pi^k}(s) - \max\big\{ \max_a Q_h^k(s,a), V_h^{k-1}(s)\big\} \\ &= Q_h^{\pi^k}\big(s, \pi_h^k(s)\big) - \max\big\{ \max_a Q_h^k(s,a), V_h^{k_o(h)}(s)\big\}, \end{aligned} \tag{82}$$

where the last line holds since $V_h(s)$ has not been updated during episodes $k_o(h), k_o(h)+1, \cdots, k-1$ (in view of the definition of $k_o(h)$ in equation 78). We shall prove that the right-hand side of equation 82 is non-negative by discussing the following two cases separately.

Case 1. Consider the case where $V_h^k(s) = \max_a Q_h^k(s,a)$. Note that

$$\pi_h^k(s) = \arg\max_a Q_h^k(s,a), \qquad \text{when } V_h^k(s) = \max_a Q_h^k(s,a) \tag{83}$$

holds for all $(k,h) \in [K] \times [H]$, Thus

$$\begin{aligned} V_h^{\pi^k}(s) - V_h^k(s) &= Q_h^{\pi^k}\big(s, \pi_h^k(s)\big) - \max_a Q_h^k(s,a) \\ &= Q_h^{\pi^k}\big(s, \pi_h^k(s)\big) - Q_h^k\big(s, \pi_h^k(s)\big). \end{aligned} \tag{84}$$

To continue, we turn to controlling a more general term $Q_h^{\pi^k}(s,a) - Q_h^k(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Invoking the fact $\eta_0^{N_h^k} + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} = 1$ (see equation 56 and equation 58) leads to

$$Q_h^{\pi^k}(s,a) = \eta_0^{N_h^k} Q_h^{\pi^k}(s,a) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} Q_h^{\pi^k}(s,a).$$

This relation combined with equation 106 allows us to express the difference between $Q_h^{\pi^k}$ and Q_h^k as follows

$$\begin{aligned} & Q_h^{\pi^k}(s,a) - Q_h^k(s,a) \\ &= \eta_0^{N_h^k}\Big(Q_h^{\pi^k}(s,a) - Q_h^1(s,a)\Big) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\Big[Q_h^{\pi^k}(s,a) - r_h(s,a) - V_{h+1}^{k^n}(s_{h+1}^{k^n}) + R_{s,a}\kappa(V_{h+1}^{k^n}) + b_n\Big] \\ &\overset{\text{(a)}}{=} \eta_0^{N_h^k}\Big(Q_h^{\pi^k}(s,a) - Q_h^1(s,a)\Big) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\Big[P_{h,s,a}V_{h+1}^{\pi^k} - R_{s,a}\kappa(V_{h+1}^{\pi^k}) - V_{h+1}^{k^n}(s_{h+1}^{k^n}) + R_{s,a}\kappa(V_{h+1}^{k^n}) + b_n\Big] \\ &\overset{\text{(b)}}{\ge} \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\Big[P_{h,s,a}V_{h+1}^{\pi^k} - R_{s,a}\kappa(V_{h+1}^{\pi^k}) - V_{h+1}^{k^n}(s_{h+1}^{k^n}) + R_{s,a}\kappa(V_{h+1}^{k^n}) + b_n\Big] \\ &\overset{\text{(c)}}{=} \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\Big[\sigma_{h,s,a}(V_{h+1}^{\pi^k}) - \sigma_{h,s,a}(V_{h+1}^{k^n}) + \sigma_{h,s,a}(V_{h+1}^{k^n}) - V_{h+1}^{k^n}(s_{h+1}^{k^n}) + R_{s,a}\kappa(V_{h+1}^{k^n}) + b_n\Big] \\ &\overset{\text{(d)}}{\ge} \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\Big[\Big(P_{h,s,a} - P_h^{k^n}\Big)V_{h+1}^{k^n} + b_n\Big]. \end{aligned} \tag{85}$$

Here, (a) invokes the robust Bellman equation $Q_h^{\pi^k}(s,a) = r_h(s,a) + \sigma_{h,s,a}(V_{h+1}^{\pi^k})$; (b) holds since $Q_h^{\pi^k}(s,a) \geq 0 = Q_h^1(s,a)$; (c) is from Lemma 6; and (d) comes from the fact

$$V_{h+1}^{\pi^k} \geq V_{h+1}^k \geq V_{h+1}^{k^n},$$

owing to the induction hypothesis in equation 79 as well as the monotonicity of V_{h+1} in Lemma 12. Consequently, it follows from equation 85 that

$$\begin{aligned}
&Q_h^{\pi^k}(s,a) - Q_h^k(s,a)\\
&\geq \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \left(P_{h,s,a} - P_h^{k^n(s,a)}\right) V_{h+1}^{k^n(s,a)} + \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} b_n\\
&\geq \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} b_n - \left| \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \left(P_{h,s,a} - P_h^{k^n(s,a)}\right) V_{h+1}^{k^n(s,a)} \right|\\
&\geq 0
\end{aligned} \tag{86}$$

for all state-action pair (s,a), where the last inequality holds due to the bound in equation 71 in Lemma 12. Plugging the above result into equation 84 directly establishes that

$$V_h^{\pi^k}(s) - V_h^k(s) = Q_h^{\pi^k}\left(s, \pi^k(s)\right) - Q_h^k\left(s, \pi^k(s)\right) \geq 0. \tag{87}$$

Case 2. When $V_h^k(s) = V_h^{k_o(h)}(s)$, it indicates that

$$V_h^{k_o(h)}(s) = \max_a Q_h^{k_o(h)}(s,a), \qquad \pi_h^{k_o(h)}(s) = \arg\max_a Q_h^{k_o(h)}(s,a), \tag{88}$$

which follows from the definition of $k_o(h)$ in equation 78 and the corresponding fact in equation 83. We also make note of the fact that

$$\pi_h^k(s) = \pi_h^{k_o(h)}(s), \tag{89}$$

which holds since $V_h(s)$ (and hence $\pi_h(s)$) has not been updated during episodes $k_o(h), k_o(h)+1, \cdots, k-1$ (in view of the definition equation 78). Combining the above two results, we can show that

$$\begin{aligned}
V_h^{\pi^k}(s) - V_h^k(s) &= Q_h^{\pi^k}\left(s, \pi_h^k(s)\right) - V_h^{k_o(h)}(s) = Q_h^{\pi^k}\left(s, \pi_h^k(s)\right) - \max_a Q_h^{k_o(h)}(s,a)\\
&= Q_h^{\pi^k}\left(s, \pi_h^{k_o(h)}(s)\right) - Q_h^{k_o(h)}\left(s, \pi_h^{k_o(h)}(s)\right)\\
&\geq 0,
\end{aligned} \tag{90}$$

where the final line can be verified using exactly the same argument as in the previous case to show equation 85 and then equation 87. Here, we omit the proof of this step for brevity.

To conclude, substituting the relations equation 87 and equation 90 in the above two cases back into equation 82, we arrive at

$$V_h^{\pi^k}(s) - V_h^k(s) \geq 0$$

as desired in equation 80. This immediately completes the induction argument. □

Lemma 13. *With probability at least* $1-\delta$*, it holds that*

$$\begin{aligned}
&\sum_{k=1}^K \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{P,h}^{\pi^\star}(s,a) \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \left(\sigma_{h,s,a}(V_{h+1}^\star) - \sigma_{h,s,a}(V_{h+1}^{k_h^n(s,a)})\right) \qquad (91)\\
&\leq \left(1+\frac{1}{H}\right) \sum_{k=1}^K \sum_{s\in\mathcal{S}} d_{P,h+1}^{\pi^\star}(s) \left(V_{h+1}^\star(s) - V_{h+1}^k(s)\right) + 24\sqrt{H^2 C^\star K \log\frac{2H}{\delta}} + 12HC^\star \log\frac{2H}{\delta}.
\end{aligned}$$

Proof. It is sufficient to show that

$$A_h \triangleq \sum_{k=1}^K \underbrace{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{P,h}^{\pi^\star}(s,a) \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \left(\sigma_{h,s,a}(V_{h+1}^\star) - \sigma_{h,s,a}(V_{h+1}^{k_h^n(s,a)})\right)}_{=:A_{h,k}} \tag{92}$$

$$\leq \sum_{k=1}^{K} \underbrace{\left(1+\frac{1}{H}\right) \sum_{s\in\mathcal{S}} d_{P,h+1}^{\pi^\star}(s)\left(V_{h+1}^\star(s)-V_{h+1}^k(s)\right)}_{=:B_{h,k}} + 24\sqrt{H^2C^\star K\log\frac{2H}{\delta}} + 12HC^\star\log\frac{2H}{\delta}.$$

Define two auxiliary sequences $\{Y_{h,k}\}_{k=1}^K$ and $\{Z_{h,k}\}_{k=1}^K$ which are the empirical estimates of $A_{h,k}$ and $B_{h,k}$, respectively. For any time step h in episode k, $Y_{h,k}$ and $Z_{h,k}$ are defined as follows

$$Y_{h,k} := \frac{d_{P,h}^{\pi^\star}(s_h^k,a_h^k)}{d_{P,h}^{\mu}(s_h^k,a_h^k)} \sum_{n=1}^{N_h^k(s_h^k,a_h^k)} \eta_n^{N_h^k(s_h^k,a_h^k)} \left(\sigma_{h,s_h^k,a_h^k}(V_{h+1}^\star) - \sigma_{h,s_h^k,a_h^k}(V_{h+1}^{k^n(s_h^k,a_h^k)})\right),$$

$$Z_{h,k} := \left(1+\frac{1}{H}\right)\frac{d_{P,h}^{\pi^\star}(s_h^k,a_h^k)}{d_{P,h}^{\mu}(s_h^k,a_h^k)} \left(\sigma_{h,s_h^k,a_h^k}(V_{h+1}^\star) - \sigma_{h,s_h^k,a_h^k}(V_{h+1}^k)\right).$$

Note that

$$\begin{aligned}\sum_{k=1}^K Y_{h,k} &= \sum_{k=1}^K \frac{d_{P,h}^{\pi^\star}(s_h^k,a_h^k)}{d_{P,h}^{\mu}(s_h^k,a_h^k)} \sum_{n=1}^{N_h^k(s_h^k,a_h^k)} \eta_n^{N_h^k(s_h^k,a_h^k)} \left(\sigma_{h,s_h^k,a_h^k}(V_{h+1}^\star) - \sigma_{h,s_h^k,a_h^k}(V_{h+1}^{k^n(s_h^k,a_h^k)})\right) \\ &\overset{\text{(i)}}{=} \sum_{l=1}^K \frac{d_{P,h}^{\pi^\star}(s_h^l,a_h^l)}{d_{P,h}^{\mu}(s_h^l,a_h^l)} \left\{\sum_{N=N_h^l(s_h^l,a_h^l)}^{N_h^K(s_h^l,a_h^l)} \eta^N_{N_h^l(s_h^l,a_h^l)}\right\} \left(\sigma_{h,s_h^l,a_h^l}(V_{h+1}^\star) - \sigma_{h,s_h^l,a_h^l}(V_{h+1}^l)\right) \end{aligned} \tag{93}$$

$$\leq \left(1+\frac{1}{H}\right)\sum_{k=1}^K \frac{d_{P,h}^{\pi^\star}(s_h^k,a_h^k)}{d_{P,h}^{\mu}(s_h^k,a_h^k)} \left(\sigma_{h,s_h^k,a_h^k}(V_{h+1}^\star) - \sigma_{h,s_h^k,a_h^k}(V_{h+1}^k)\right) = \sum_{k=1}^K Z_{h,k}. \tag{94}$$

Here, (a) holds by replacing $k^n(s_h^k,a_h^k)$ with l and gathering all terms that involve $V_{h+1}^\star - V_{h+1}^l$; in the last line, we have invoked the property $\sum_{N=n}^{N_h^K(s,a)}\eta_n^N \leq \sum_{N=n}^{\infty}\eta_n^N = 1+1/H$ (see equation 61b) together with the fact $V_{h+1}^\star - V_{h+1}^l \geq 0$ (see Lemma 12), and have further replaced l with k.

With the above relation in hand, in order to verify equation 93, we further decompose A_h into several terms

$$\begin{aligned} A_h = \sum_{k=1}^K A_{h,k} &= \sum_{k=1}^K Y_{h,k} + \sum_{k=1}^K (A_{h,k}-Y_{h,k}) \overset{\text{(a)}}{\leq} \sum_{k=1}^K Z_{h,k} + \sum_{k=1}^K (A_{h,k}-Y_{h,k}) \\ &= \sum_{k=1}^K B_{h,k} + \sum_{k=1}^K (Z_{h,k}-B_{h,k}) + \sum_{k=1}^K (A_{h,k}-Y_{h,k}) \end{aligned} \tag{95}$$

where (a) follows from equation 94.

As a result, it remains to control $\sum_{k=1}^K (Z_{h,k}-B_{h,k})$ and $\sum_{k=1}^K (A_{h,k}-Y_{h,k})$ separately in the following.

Step 1: controlling $\sum_{k=1}^K (A_{h,k}-Y_{h,k})$. We shall first control this term by means of Lemma 11. Specifically, consider

$$W_{h+1}^k(s,a) := \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \left(\sigma_{h,s,a}(V_{h+1}^\star) - \sigma_{h,s,a}(V_{h+1}^{k^n(s,a)})\right), \qquad C_{\mathrm{d}} := 1 \tag{96}$$

which satisfies

$$\left\|W_{h+1}^k(s,a)\right\|_\infty \leq \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \left(\left\|V_{h+1}^\star\right\|_\infty + \left\|V_{h+1}^{k^n(s,a)}\right\|_\infty\right) \leq 2H =: C_{\mathrm{w}}. \tag{97}$$

Here, we use the fact that $\eta_0^{N_h^k}+\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}=1$ (see equation 56 and equation 58). Then, applying Lemma 11 with equation 96, we have with probability at least $1-\delta$, the following inequality holds true

$$\begin{aligned}\left|\sum_{k=1}^{K}(A_{h,k}-Y_{h,k})\right| &=: \left|\sum_{k=1}^{K}X_{h,k}\right| \\ &\le \sqrt{\sum_{k=1}^{K}8C_{\mathrm{d}}^2C^\star\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_{P,h}^{\pi^\star}(s,a)\left[W_{h+1}^k(s,a)\right]^2\log\frac{2H}{\delta}}+2C_{\mathrm{d}}C^\star C_{\mathrm{w}}\log\frac{2H}{\delta} \\ &\le 16\sqrt{H^2C^\star K\log\frac{2H}{\delta}}+4HC^\star\log\frac{2H}{\delta},\end{aligned} \tag{98}$$

where the last inequality is from $\left|W_{h+1}^k(s,a)\right|\le\left\|V_{h+1}^*-V_{h+1}^{k^n(s,a)}\right\|_\infty\le H$.

Step 2: controlling $\sum_{k=1}^K(Z_{h,k}-B_{h,k})$. Similarly, we shall control $\sum_{k=1}^K(Z_{h,k}-B_{h,k})$ by invoking Lemma 11.

Recall that

$$\begin{aligned}Z_{h,k}-B_{h,k} &= \left(1+\frac{1}{H}\right)\frac{d_{P,h}^{\pi^\star}(s_h^k,a_h^k)}{d_{P,h}^{\mu}(s_h^k,a_h^k)}\left(\sigma_{h,s_h^k,a_h^k}(V_{h+1}^\star)-\sigma_{h,s_h^k,a_h^k}(V_{h+1}^k)\right) \\ &\quad-\left(1+\frac{1}{H}\right)\sum_{s\in\mathcal{S}}d_{P,h+1}^{\pi^\star}(s)\left(V_{h+1}^\star(s)-V_{h+1}^k(s)\right),\end{aligned} \tag{99}$$

and let us consider

$$W_{h+1}^k(s,a):=\sigma_{h,s_h^k,a_h^k}(V_{h+1}^\star)-\sigma_{h,s_h^k,a_h^k}(V_{h+1}^k),\qquad C_{\mathrm{d}}:=\left(1+\frac{1}{H}\right)\le 2 \tag{100}$$

which satisfies

$$\left\|W_{h+1}^k(s,a)\right\|_\infty\le\left\|V_{h+1}^\star\right\|_\infty+\left\|V_{h+1}^k\right\|_\infty\le 2H=:C_{\mathrm{w}}. \tag{101}$$

Similarly, in view of Lemma 11, we can show that with probability at least $1-\delta$,

$$\begin{aligned}\left|\sum_{k=1}^{K}(B_{h,k}-Z_{h,k})\right| &= \left|\sum_{k=1}^{K}X_{h,k}\right| \\ &\le 16\sqrt{H^2C^\star K\log\frac{2H}{\delta}}+8HC^\star\log\frac{2H}{\delta}.\end{aligned} \tag{102}$$

Step 3: putting all this together. Substitution results in equation 98 and equation 102 back into equation 95 completes the proof of equation 93 as follows

$$\begin{aligned}A_h &\le \sum_{k=1}^{K}B_{h,k}+\left|\sum_{k=1}^{K}(Z_{h,k}-B_{h,k})\right|+\left|\sum_{k=1}^{K}(A_{h,k}-Y_{h,k})\right| \\ &\le \sum_{k=1}^{K}B_{h,k}+24\sqrt{H^2C^\star K\log\frac{2H}{\delta}}+12HC^\star\log\frac{2H}{\delta}.\end{aligned}$$

This hence completes the proof. □

Lemma 14. *Denote the term* $\sum_{k=1}^K\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_{P,h}^{\pi^\star}(s,a)\eta_0^{N_h^k(s,a)}H+2\sum_{k=1}^K\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_{P,h}^{\pi^\star}(s,a)\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}b_n$ *by* I_h. *Consider any* $\delta\in(0,1)$. *With probability at least* $1-\delta$, *we have*

$$\begin{aligned}&\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\left(I_h+24\sqrt{H^2C^\star K\log\frac{2H}{\delta}}+12HC^\star\log\frac{2H}{\delta}\right) \\ &\lesssim H^2SC^\star\iota+\sqrt{H^5SC^\star K\iota^3},\end{aligned} \tag{103}$$

where we recall that $\iota:=\log\left(\frac{SAT}{\delta}\right)$.

Proof. The proof can be obtained by directly following the proof in (Shi et al., 2022), and is hence omitted here.

□

D.3 Proof of Theorem 2

We then proceed to the proof.

Theorem 15. *(Restatement of Theorem 2) Consider any $\delta \in (0,1)$. Suppose that the behavior policy μ satisfies Assumption 1. There exists some universal constant c_a, such that if we set $\iota := \log\left(\frac{SAT}{\delta}\right)$ and set $T > SC^\star \iota$, then the policy $\widehat{\pi}$ returned by Algorithm 1 satisfies*

$$V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) \le c_{\mathrm{a}} \sqrt{\frac{H^6 SC^\star \iota^3}{T}} \tag{104}$$

with probability at least $1-\delta$.

Proof. For any state-action pair (s,a), according to the update rule specified in Algorithm 1, we have

$$\begin{aligned} &Q_h^k(s,a) = Q_h^{k^{N_h^k}+1}(s,a) \\ &= \left(1-\eta_{N_h^k}\right) Q_h^{k^{N_h^k}}(s,a) + \eta_{N_h^k}\left\{ r_h(s,a) + V_{h+1}^{k^{N_h^k}}\left(s_{h+1}^{k^{N_h^k}}\right) - R_{s,a}\kappa(V_{h+1}^{k^{N_h^k}}) - b_{N_h^k} \right\}, \end{aligned} \tag{105}$$

where the first identity holds because $k^{N_h^k}$ denotes the most recent episode before k that visits (s,a) at step h, and the learning rate is defined as in equation 55. Note that $k > k^{N_h^k}$ always holds. Applying the above relation recursively and using the notation defined in equation 56, we obtain

$$Q_h^k(s,a) = \eta_0^{N_h^k} Q_h^1(s,a) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\left(r_h(s,a) + V_{h+1}^{k^n}\left(s_{h+1}^{k^n}\right) - R_{s,a}\kappa(V_{h+1}^{k^n}) - b_n \right). \tag{106}$$

Applying Lemma 12, the optimality gap term equation 104 can be decomposed as follows

$$\begin{aligned} &V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) \\ &= \mathop{\mathbb{E}}_{s_1\sim\rho}\left[V_1^\star(s_1)\right] - \mathop{\mathbb{E}}_{s_1\sim\rho}\left[V_1^{\pi^K}(s_1)\right] \\ &\overset{\text{(a)}}{\le} \mathop{\mathbb{E}}_{s_1\sim\rho}\left[V_1^\star(s_1)\right] - \mathop{\mathbb{E}}_{s_1\sim\rho}\left[V_1^{K}(s_1)\right] \\ &\overset{\text{(b)}}{\le} \frac{1}{K}\sum_{k=1}^K \left(\mathop{\mathbb{E}}_{s_1\sim\rho}\left[V_1^\star(s_1)\right] - \mathop{\mathbb{E}}_{s_1\sim\rho}\left[V_1^{k}(s_1)\right] \right) \\ &= \frac{1}{K}\sum_{k=1}^K \sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s)\left(V_1^\star(s) - V_1^k(s)\right), \end{aligned} \tag{107}$$

where (a) follows from Lemma 12 (i.e., $V_1^{\pi^K}(s) \ge V_1^K(s)$ for all $s \in \mathcal{S}$), (b) results from the monotonicity property in Lemma 12, and the final equality holds because $d_1^{\pi^\star}(s) = \rho(s)$.

We then bound the right-hand side of equation 107. Since $\pi^\star$ is a deterministic policy, $d_{P,h}^{\pi^\star}(s) = d_{P,h}^{\pi^\star}(s, \pi^\star(s))$. And from the fact that $V_h^k(s) \ge \max_a Q_h^k(s,a) \ge Q_h^k(s, \pi_h^\star(s))$ and $V_h^\star(s) = Q_h^\star(s, \pi_h^\star(s))$, we have that

$$\begin{aligned} &\sum_{k=1}^K \sum_{s\in\mathcal{S}} d_{P,h}^{\pi^\star}(s)\left(V_h^\star(s) - V_h^k(s)\right) \\ &= \sum_{k=1}^K \sum_{s\in\mathcal{S}} d_{P,h}^{\pi^\star}(s, \pi_h^\star(s))\left(V_h^\star(s) - V_h^k(s)\right) \end{aligned}$$

$$\leq \sum_{k=1}^{K} \sum_{s\in\mathcal{S}} d_{P,h}^{\pi^\star}(s,\pi_h^\star(s)) \left(Q_h^\star\big(s,\pi_h^\star(s)\big) - Q_h^k\big(s,\pi_h^\star(s)\big)\right)$$
$$= \sum_{k=1}^{K} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{P,h}^{\pi^\star}(s,a) \left(Q_h^\star(s,a) - Q_h^k(s,a)\right), \tag{108}$$

for any $h \in [H]$, where the last identity holds because

$$d_{P,h}^{\pi^\star}(s,a) = 0 \qquad \text{for any } a \neq \pi_h^\star(s). \tag{109}$$

To further bound the term $Q_h^\star(s,a) - Q_h^k(s,a)$ in equation 108, we first adapt equation 58 and have that

$$\begin{aligned} Q_h^\star(s,a) &= \sum_{n=0}^{N_h^k} \eta_n^{N_h^k} Q_h^\star(s,a) \\ &= \eta_0^{N_h^k} Q_h^\star(s,a) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} Q_h^\star(s,a) \\ &= \eta_0^{N_h^k} Q_h^\star(s,a) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \left(r_h(s,a) + \sigma_{h,s,a}(V_{h+1}^\star)\right), \end{aligned} \tag{110}$$

where the second line follows from the robust Bellman's optimality equation. Combining equation 106 and equation 110 implies that

$$\begin{aligned} &Q_h^\star(s,a) - Q_h^k(s,a) \\ &= \eta_0^{N_h^k} \left(Q_h^\star(s,a) - Q_h^1(s,a)\right) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \left(\sigma_{h,s,a}(V_{h+1}^\star) - V_{h+1}^{k^n}(s_{h+1}^{k^n}) + R_{s,a}\kappa(V_{h+1}^{k^n}) + b_n\right) \\ &\overset{(a)}{=} \eta_0^{N_h^k} \left(Q_h^\star(s,a) - Q_h^1(s,a)\right) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} b_n + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \big(\sigma_{h,s,a}(V_{h+1}^\star) - \sigma_{h,s,a}(V_{h+1}^{k^n})\big) \\ &\quad + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \big(P_{h,s,a} - P_h^{k^n}\big) V_{h+1}^{k^n} \end{aligned} \tag{111}$$
$$\leq \eta_0^{N_h^k} H + 2\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} b_n + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \big(\sigma_{h,s,a}(V_{h+1}^\star) - \sigma_{h,s,a}(V_{h+1}^{k^n})\big), \tag{112}$$

where (a) is from Lemma 6 and the definition of $P_h^{k^n} V_{h+1}^{k^n} = V_{h+1}^{k^n}(s_{h+1}^{k^n})$, and the last inequality follows from the fact $Q_h^\star(s,a) - Q_h^1(s,a) = Q_h^\star(s,a) - 0 \leq H$ and equation 71 in Lemma 12. Plug equation 112 in equation 108, we have that

$$\begin{aligned} &\sum_{k=1}^{K} \sum_{s\in\mathcal{S}} d_{P,h}^{\pi^\star}(s) \left(V_h^\star(s) - V_h^k(s)\right) \\ &\leq \underbrace{\sum_{k=1}^{K} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{P,h}^{\pi^\star}(s,a) \eta_0^{N_h^k(s,a)} H + 2\sum_{k=1}^{K} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{P,h}^{\pi^\star}(s,a) \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} b_n}_{=: I_h} \\ &\qquad + \sum_{k=1}^{K} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{P,h}^{\pi^\star}(s,a) \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \big(\sigma_{h,s,a}(V_{h+1}^\star) - \sigma_{h,s,a}(V_{h+1}^{k^n})\big). \end{aligned} \tag{113}$$

We then bound the last term on the right-hand side of equation 113. By applying Lemma 13, it implies that

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_{P,h}^{\pi^\star}(s)\left(V_h^\star(s)-V_h^k(s)\right)\\
&\leq \left(1+\frac{1}{\Gamma}\right)\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_{P,h+1}^{\pi^\star}(s)\left(V_{h+1}^\star(s)-V_{h+1}^k(s)\right)\\
&\quad + I_h + 24\sqrt{H^2C^\star K\log\frac{2H}{\delta}} + 12HC^\star\log\frac{2H}{\delta}.
\end{aligned}
\tag{114}
$$

Recursively applying equation 114 over the time steps $h = H, H-1, \cdots, 1$ with the terminal condition $V_{H+1}^k = V_{H+1}^\star = 0$ further implies that

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s)\left(V_1^\star(s)-V_1^k(s)\right)\\
&\leq \max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_{P,h}^{\pi^\star}(s)\left(V_h^\star(s)-V_h^k(s)\right)\\
&\leq \sum_{h=1}^{H}\left(1+\frac{1}{\Gamma}\right)^{h-1}\left(I_h + 24\sqrt{H^2C^\star K\log\frac{2H}{\delta}} + 12HC^\star\log\frac{2H}{\delta}\right).
\end{aligned}
\tag{115}
$$

Finally, to bound the right-hand side of equation 115, we combine Lemma 14 and equation 107, which yields

$$
\begin{aligned}
&V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho)\\
&\leq \frac{1}{K}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s)\left(V_1^\star(s)-V_1^k(s)\right)\\
&\leq \frac{1}{K}\max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_{P,h}^{\pi^\star}(s)\left(V_h^\star(s)-V_h^k(s)\right)\\
&\leq \frac{c_{\mathrm{a}}}{2}\sqrt{\frac{H^5SC^\star\iota^3}{K}} + \frac{c_{\mathrm{a}}}{2}\frac{H^2SC^\star\iota}{K} = \frac{c_{\mathrm{a}}}{2}\sqrt{\frac{H^6SC^\star\iota^3}{T}} + \frac{c_{\mathrm{a}}}{2}\frac{H^3SC^\star\iota}{T}\\
&\leq c_{\mathrm{a}}\sqrt{\frac{H^6SC^\star\iota^3}{T}}
\end{aligned}
\tag{116}
$$

for some sufficiently large constant $c_{\mathrm{a}} > 0$, where the last inequality is valid as long as $T > SC^\star\iota$.

This hence completes the proof of Theorem 2. □

E ANALYSIS OF THE INFINITE HORIZON SETTING

E.1 ALGORITHM FOR INFINITE HORIZON

In this section, we present the analysis of the infinite horizon robust MDPs.

E.2 NOTATION

The notation used in the proof for the infinite horizon setting is largely similar to that used in the finite horizon case. For any state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, we define:

$$
P_{s,a} = P(\cdot \mid s,a) \in \mathbb{R}^{1\times S}
$$

to be the (s,a)-th row of a probability transition matrix $P \in \mathbb{R}^{SA\times S}$.

Algorithm 3 Double-Pessimism Q-Learning for infinite-horizon RMDPs.

Input: $\mathcal{D}$, target success probability $1-\delta$, uncertainty set radius R, $\Gamma = \left\lceil \frac{4}{1-\gamma} \log \frac{ST}{\delta} \right\rceil$, penalty function κ
Initialize: $Q_0(s,a) = 0$, $V_0(s) = 0$, $n_0(s,a) = 0, \forall s, a$
for $t = 1, ..., T$ **do**
Sample a sample (s_{t-1}, a_{t-1}, s_t) from $\mathcal{D}$
$n_t(s_{t-1}, a_{t-1}) \leftarrow n_{t-1}(s_{t-1}, a_{t-1}) + 1$; $n_t(s,a) \leftarrow n_{t-1}(s,a), \forall (s,a) \neq (s_{t-1}, a_{t-1})$
$n \leftarrow n_t(s,a)$; $\eta_n \leftarrow (\Gamma+1)/(\Gamma+n)$
$b_n \leftarrow c_b \sqrt{\frac{\Gamma \log(SAT/\delta)}{n(1-\gamma)^2}}$
$Q_t(s_{t-1}, a_{t-1}) = (1-\eta_n) Q_{t-1}(s_{t-1}, a_{t-1}) + \eta_n \Big\{ r(s_{t-1}, a_{t-1}) + \gamma V_{t-1}(s_t) - \gamma \kappa_{s_{t-1}, a_{t-1}}(V_{t-1}) - b_n \Big\}$
$Q_t(s,a) = Q_{t-1}(s,a)$ for all $(s,a) \neq (s_{t-1}, a_{t-1})$
$V_t(s_{t-1}) = \max \Big\{ \max_{a \in \mathcal{A}} Q_t(s_{t-1}, a),\ V_{t-1}(s_{t-1}) \Big\}$,
$V_t(s) = V_{t-1}(s)$ for all $s \neq s_{t-1}$.
end for
$\widehat{\pi}(s) = \arg\max_{a \in \mathcal{A}} Q_T(s,a), \forall s$
Output: $\widehat{\pi}$

For any $t \geq 0$, we define $P_t \in \mathbb{R}^{SA \times S}$ to be an empirical probability transition matrix, given by:

$$P_t(s' \mid s,a) = \begin{cases} 1, & \text{if } (s,a,s') = (s_{t-1}, a_{t-1}, s_t) \\ 0, & \text{otherwise} \end{cases} \tag{117}$$

for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$.

For any deterministic policy π, we introduce two probability transition kernels: $P_\pi : \mathcal{S} \to \Delta(\mathcal{S})$ and $P^\pi : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S} \times \mathcal{A})$, defined as follows:

$$P_\pi(s' \mid s) = P(s' \mid s, \pi(s)), \tag{118a}$$

$$P^\pi(s', a' \mid s, a) = \begin{cases} P(s' \mid s,a), & \text{if } a' = \pi(s') \\ 0, & \text{otherwise} \end{cases} \tag{118b}$$

for any $(s,a), (s',a') \in \mathcal{S} \times \mathcal{A}$.

Additionally, we define $\rho^{\pi^\star}$ to be a distribution over $\mathcal{S} \times \mathcal{A}$ such that:

$$\rho^{\pi^\star}(s,a) = \begin{cases} \rho(s), & \text{if } a = \pi^\star(s) \\ 0, & \text{otherwise} \end{cases} \tag{119}$$

For any sequence $\{a_i\}_{i=n_1}^{n_2}$ and two integers m_1 and m_2, we define:

$$\sum_{i=m_1}^{m_2} a_i = \begin{cases} \sum_{i=\max\{n_1,m_1\}}^{\min\{n_2,m_2\}} a_i, & \text{if } \max\{n_1,m_1\} \leq \min\{n_2,m_2\} \\ 0, & \text{otherwise} \end{cases}$$

E.3 Lemmas for Theorem 3

Lemma 16. *(Lemma 4.1 in (Jin et al., 2018), Lemma 1 in (Li et al., 2021)) Recall the learning rates are*

$$\eta_0^t \coloneqq \prod_{j=1}^{t} (1-\eta_j) \qquad \textit{and} \qquad \eta_i^t \coloneqq \begin{cases} \eta_i \prod_{j=i+1}^{t} (1-\eta_j), & \textit{if } t > i, \\ \eta_i, & \textit{if } t = i, \\ 0, & \textit{if } t < i, \end{cases} \tag{120}$$

where $\eta_j = (\Gamma+1)/(\Gamma+j)$. Then

1. *For any integer $t \geq 1$, $\sum_{i=1}^{t} \eta_i^t = 1$ and $\eta_0^t = 0$.*
2. *For any integer $t \geq 1$ and any $1/2 \leq a \leq 1$,*
$$\frac{1}{t^a} \leq \sum_{i=1}^{t} \frac{1}{i^a} \eta_i^t \leq \frac{2}{t^a}.$$
3. *For any integer $t \geq 1$,*
$$\max_{i \in [t]} \eta_i^t \leq \frac{2\Gamma}{t} \qquad \textit{and} \qquad \sum_{i=1}^{t} \left(\eta_i^t\right)^2 \leq \frac{2\Gamma}{t}.$$
4. *For any integer $i \geq 1$,*
$$\sum_{t=i}^{\infty} \eta_i^t = 1 + \frac{1}{\Gamma}.$$

We then present the following lemma to establish an upper bound on $Q^\star - Q_t$, and simultaneously justify that the value function estimate V_t is always a pessimistic view of V^{π_t} (and hence $V^\star$).

Lemma 17. *With probability exceeding $1-\delta$, for all $s \in \mathcal{S}$ and $t \in [T]$, it holds that*

$$Q^\star\big(s, \pi^\star(s)\big) - Q_t\big(s, \pi^\star(s)\big) \leq \gamma \sum_{i=1}^{n} \eta_i^n \big(\sigma_{s,\pi^\star(s)}(V^\star) - \sigma_{s,\pi^\star(s)}(V_{k_i})\big) + \beta_n\big(s, \pi^\star(s)\big), \tag{121}$$

where $n = n_t(s, \pi^\star(s))$ and we define

$$\beta_n\big(s, \pi^\star(s)\big) \equiv \beta_n \coloneqq 3c_b \sqrt{\frac{\Gamma\iota}{n\,(1-\gamma)^2}};$$

in addition, we also have

$$V_t(s) \leq V^{\pi_t}(s) \leq V^\star(s), \qquad \forall s \in \mathcal{S}. \tag{122}$$

Proof. **Proof of equation 121.** Consider any given pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ and denote $n = n_t(s,a)$, the total number of times that (s,a) has been visited prior to time t. Set $k_0 = -1$, and let

$$k_i \coloneqq \min\Big\{\big\{0 \leq k < T : k > k_{i-1}, (s_k, a_k) = (s,a)\big\}, T\Big\} \tag{123}$$

for each $1 \leq i \leq T$. Clearly, each k_i is a stopping time. In view of the update rule, we have

$$Q_t\,(s,a) = \sum_{i=1}^{n} \eta_i^n \Big\{ r\,(s,a) + \gamma V_{k_i}\,(s_{k_i+1}) - \gamma\kappa(V_{k_i}) - b_i\,(s,a) \Big\},$$

which together with the robust Bellman optimality equation gives

$$\begin{aligned} &(Q^\star - Q_t)\,(s,a) \\ &= r\,(s,a) + \gamma\sigma_{s,a}(V^\star) - \sum_{i=1}^{n} \eta_i^n \Big\{ r\,(s,a) + \gamma V_{k_i}\,(s_{k_i+1}) - \gamma\kappa(V_{k_i}) - b_i\,(s,a) \Big\} \end{aligned}$$

$$= \gamma\sigma_{s,a}(V^\star) - \sum_{i=1}^{n}\eta_i^n\Big\{\gamma V_{k_i}\left(s_{k_i+1}\right) - \gamma\kappa(V_{k_i}) - b_i\left(s,a\right)\Big\}$$
$$= \sum_{i=1}^{n}\eta_i^n\gamma\left(\sigma_{s,a}(V^\star) - \sigma_{s,a}(V_{k_i})\right) + \sum_{i=1}^{n}\eta_i^n\gamma\Big(\big(P - P_{k_i}\big)V_{k_i}\Big)\left(s,a\right) + \sum_{i=1}^{n}\eta_i^n b_i\left(s,a\right), \quad (124)$$

where the last two lines are valid since $\sum_{i=1}^{n}\eta_i^n = 1$ (cf. Lemma 16) and Lemma 6.

Henceforth, we only focus on the case where $a = \pi^\star(s)$. Define $\mathcal{F}_i$ to be the σ-field generated by $\{(s_i, a_i)\}_{i=0}^{k_i}$. It is straightforward to check that for any $1 \le \tau \le T$,

$$\Big\{\mathbf{1}_{k_i<T}\Big(\big(P - P_{k_i}\big)V_{k_i}\Big)\big(s,\pi^\star(s)\big)\Big\}_{i=1}^{\tau}$$

is a martingale difference sequence with respect to $\{\mathcal{F}_i\}_{i\ge0}$. Then, we can invoke the Azuma-Hoeffding inequality together with the basic bound $\|V_{k_i}\|_\infty \le \frac{1}{1-\gamma}$ to show that for any fixed $s \in \mathcal{S}$ and $\tau \in [T]$,

$$\left|\sum_{i=1}^{\tau}\mathbf{1}_{k_i<T}\eta_i^\tau\Big(\big(P - P_{k_i}\big)V_{k_i}\Big)\big(s,\pi^\star(s)\big)\right| \lesssim \frac{1}{1-\gamma}\sqrt{\sum_{i=1}^{\tau}\left(\eta_i^\tau\right)^2\log\frac{ST}{\delta}}$$
$$\lesssim \sqrt{\frac{\Gamma}{\tau\left(1-\gamma\right)^2}\log\frac{ST}{\delta}}$$

holds with probability exceeding $1 - \delta/(ST)$. Here, the last line utilizes Lemma 16. Taking the union bound over $\tau \le T$ allows us to replace τ with $n = n_t(s,a)$ in the above inequality, namely, for any fixed $s \in \mathcal{S}$ and $a \in \mathcal{A}$, with probability exceeding $1 - \delta/S$ we have

$$\left|\sum_{i=1}^{n}\eta_i^n\gamma\Big(\big(P - P_{k_i}\big)V_{k_i}\Big)\big(s,\pi^\star(s)\big)\right| \lesssim \sqrt{\frac{\Gamma\iota}{n\left(1-\gamma\right)^2}} \quad (125)$$

holds for all $n = n_t(s,\pi^\star(s))$ with $1 \le t \le T$. In view of Lemma 16, for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$ we know that

$$c_b\sqrt{\frac{\Gamma\iota}{n_t(s,a)\left(1-\gamma\right)^2}} \le \sum_{i=1}^{n_t(s,a)}\eta_i^{n_t(s,a)}b_i\left(s,a\right) \le 2c_b\sqrt{\frac{\Gamma\iota}{n_t(s,a)\left(1-\gamma\right)^2}}. \quad (126)$$

Therefore, when c_b is sufficiently large, it follows that

$$\left(Q^\star - Q_t\right)\big(s,\pi^\star(s)\big) \le \gamma\sum_{i=1}^{n}\eta_i^n\left(\sigma_{s,\pi^\star(s)}(V^\star) - \sigma_{s,\pi^\star(s)}(V_{k_i})\right) + 3c_b\sqrt{\frac{\Gamma\iota}{n\left(1-\gamma\right)^2}}.$$

Taking the union bound over $s \in \mathcal{S}$ and defining

$$\beta_n\big(s,\pi^\star(s)\big) \coloneqq 3c_b\sqrt{\frac{\Gamma\iota}{n\left(1-\gamma\right)^2}},$$

we can conclude that with probability exceeding $1 - \delta$,

$$\left(Q^\star - Q_t\right)\big(s,\pi^\star(s)\big) \le \gamma\sum_{i=1}^{n}\eta_i^n\left(\sigma_{s,\pi^\star(s)}(V^\star) - \sigma_{s,\pi^\star(s)}(V_{k_i})\right) + \beta_n\big(s,\pi^\star(s)\big)$$

for all $s \in \mathcal{S}$ and $t \in [T]$.

Proof of equation 122. Note that $V^\star \ge V^{\pi_t}$ holds trivially due to the optimality of $V^\star$. We are therefore left with showing $V^{\pi_t} \ge V_t$. Suppose for the moment that with probability exceeding $1 - \delta$, for all $s \in \mathcal{S}$, $t \in [T]$ and $j \in [t]$, it holds that

$$\left(Q^{\pi_t} - Q_j\right)\big(s,\pi_t(s)\big) \ge \gamma\left(\sigma_{s,\pi_t(s)}(V^{\pi_t}) - \sigma_{s,\pi_t(s)}(V_j)\right)\mathbf{1}\left\{n_t\big(s,\pi_t(s)\big) \ge 1\right\}; \quad (127)$$

the proof of this claim (127) is deferred to later. As a consequence, for every $s \in \mathcal{S}$ and $t \in [T]$, there exists $j(t) \in [t]$ such that

$$
\begin{aligned}
\left(V^{\pi_t} - V_t\right)(s) &\overset{\text{(a)}}{=} Q^{\pi_t}\big(s, \pi_t(s)\big) - Q_{j(t)}\big(s, \pi_t(s)\big) \\
&\overset{\text{(b)}}{=} Q^{\pi_t}\big(s, \pi_t(s)\big) - Q_{j(t)}\big(s, \pi_{j(t)}(s)\big) \\
&\overset{\text{(c)}}{\geq} \min\left\{\gamma\left(\sigma_{s,\pi_t(s)}(V^{\pi_t}) - \sigma_{s,\pi_t(s)}(V_{j(t)})\right), 0\right\} \\
&\overset{\text{(d)}}{\geq} \min\left\{\gamma\left(\sigma_{s,\pi_{j(t)}(s)}(V^{\pi_t}) - \sigma_{s,\pi_{j(t)}(s)}(V_t)\right), 0\right\}.
\end{aligned}
$$

Here, (a) and (b) hold since the update rule asserts that there must exist some $j(t) \leq t$ such that $V_t(s) = V_{j(t)}(s) = Q_{j(t)}(s, \pi_{j(t)}(s))$ and $\pi_t(s) = \pi_{j(t)}(s)$; (c) utilizes (127); and (d) follows from the monotonicity of V_t in t (by construction). By setting

$$
s_{\min} \coloneqq \arg\min_{s \in \mathcal{S}} \left(V^{\pi_t} - V_t\right)(s),
$$

we can deduce that

$$
\begin{aligned}
\left(V^{\pi_t} - V_t\right)(s_{\min}) &\geq \min\left\{\gamma\left(\sigma_{s_{\min},\pi_{j(t)}(s_{\min})}(V^{\pi_t}) - \sigma_{s_{\min},\pi_{j(t)}(s_{\min})}(V_t)\right), 0\right\} \\
&\geq \min\left\{\gamma \min_{s \in \mathcal{S}} \left(V^{\pi_t} - V_t\right)(s), 0\right\} \\
&= \min\left\{\gamma\left(V^{\pi_t} - V_t\right)(s_{\min}), 0\right\},
\end{aligned}
$$

which together with the assumption $0 < \gamma < 1$ immediately gives

$$
\left(V^{\pi_t} - V_t\right)(s_{\min}) \geq 0.
$$

Given that $\left(V^{\pi_t} - V_t\right)(s) \geq \left(V^{\pi_t} - V_t\right)(s_{\min})$ for every $s \in \mathcal{S}$, we conclude the proof.

Now we show equation 127. First of all, if $n_t\big(s, \pi_t(s)\big) = 0$, then for all $j \in [t]$, $Q_j\big(s, \pi_t(s)\big) = 0$ since it is never updated; therefore, (127) holds true. From now on, we shall only focus on the case when $n_t\big(s, \pi_t(s)\big) \geq 1$.

Consider any $s \in \mathcal{S}$, $t \in [T]$ and $j \in [t]$. For the moment, let us define $\{k_i\}_{i=1}^T$ w.r.t. the state-action pair $\big(s, \pi_t(s)\big)$ in the same way as (123). We can then repeat the argument in (124) to decompose

$$
\begin{aligned}
&\big(Q^{\pi_t} - Q_j\big)\big(s, \pi_t(s)\big) \\
&= \big(r + \gamma\sigma(V^{\pi_t})\big)\big(s, \pi_t(s)\big) - \sum_{i=1}^{n_j(s,\pi_t(s))} \eta_i^{n_j(s,\pi_t(s))}\Big\{r\big(s, \pi_t(s)\big) + \gamma V_{k_i}(s_{k_i+1}) - R_s^{\pi_t(s)}\kappa(V^{\pi_t}) - b_i\big(s, \pi_t(s)\big)\Big\} \\
&= \sum_{i=1}^{n_j(s,\pi_t(s))} \eta_i^{n_j(s,\pi_t(s))}\gamma\Big\{\left(\sigma_{s,\pi_t(s)}(V^{\pi_t}) - \sigma_{s,\pi_t(s)}(V_{k_i})\right) + \Big(\big(P - P_{k_i}\big)V_{k_i}\Big)\big(s, \pi_t(s)\big)\Big\} \\
&\quad + \sum_{i=1}^{n_j(s,\pi_t(s))} \eta_i^{n_j(s,\pi_t(s))} b_i\big(s, \pi_t(s)\big) \\
&\geq \left(\sum_{i=1}^{n_j(s,\pi_t(s))} \eta_i^{n_j(s,\pi_t(s))}\right)\gamma \min_{1 \leq i \leq n}\left(\sigma_{s,\pi_t(s)}(V^{\pi_t}) - \sigma_{s,\pi_t(s)}(V_{k_i})\right) \\
&\quad + \sum_{i=1}^{n_j(s,\pi_t(s))} \eta_i^{n_j(s,\pi_t(s))}\gamma\Big(\big(P - P_{k_i}\big)V_{k_i}\Big)\big(s, \pi_t(s)\big) \\
&\qquad + \sum_{i=1}^{n_j(s,\pi_t(s))} \eta_i^{n_j(s,\pi_t(s))} b_i\big(s, \pi_t(s)\big) \\
&\geq \gamma\left(\sigma_{s,\pi_t(s)}(V^{\pi_t}) - \sigma_{s,\pi_t(s)}(V_t)\right)
\end{aligned}
$$

$$+ \sum_{i=1}^{n_j(s,\pi_t(s))} \eta_i^{n_j(s,\pi_t(s))}\gamma\Big(\big(P-P_{k_i}\big)V_{k_i}\Big)\big(s,\pi_t(s)\big) + c_b\sqrt{\frac{\Gamma\iota}{n_j\big(s,\pi_t(s)\big)\,(1-\gamma)^2}}.$$

Here, the last inequality follows from (126), as well as the facts that $\sum_{i=1}^{n_j(s,\pi_t(s))}\eta_i^{n_j(s,\pi_t(s))} = 1$ (cf. Lemma 16) and that V_t is non-decreasing in t. It thus boils down to showing that for every $s \in \mathcal{S}$, $t \in [T]$ and $j \in [t]$,

$$\sum_{i=1}^{n_j(s,\pi_t(s))} \eta_i^{n_j(s,\pi_t(s))}\gamma\Big(\big(P-P_{k_i}\big)V_{k_i}\Big)\big(s,\pi_t(s)\big) \lesssim \sqrt{\frac{\Gamma\iota}{n_j\big(s,\pi_t(s)\big)\,(1-\gamma)^2}}. \tag{128}$$

If this were true and if c_b is sufficiently large, then we could combine the above two inequalities to conclude the proof of (127).

We then prove the inequality equation 128. Notice that for all $(s,\pi_t(s))$ such that $n_t(s,\pi_t(s)) \geq 1$, it must appear at least once in the sample trajectory. Therefore it suffices to show that for all $0 \leq l < T$ and $t \in [T]$, it holds that

$$\sum_{i=1}^{n_t(s_l,a_l)} \eta_i^{n_t(s_l,a_l)}\gamma\Big(\big(P-P_{k_i}\big)V_{k_i}\Big)(s_l,a_l) \lesssim \sqrt{\frac{\Gamma\iota}{n_t(s_l,a_l)\,(1-\gamma)^2}},$$

where we abuse the notation by defining $\{k_i\}_{i=1}^T$ for the state-action pair (s_l,a_l) in the same way as (123). Furthermore, it suffices to only check those (s_l,a_l) in the sample trajectory that were visited for the first time, i.e., $n_l(s_l,a_l) = 0$ and $n_{l+1}(s_l,a_l) = 1$. It is straightforward to check that, for any $1 \leq \tau \leq T$,

$$\Big\{\mathbf{1}_{k_i<T}\Big(\big(P-P_{k_i}\big)V_{k_i}\Big)(s_l,a_l)\Big\}_{i=1}^{\tau}$$

is a martingale difference sequence with respect to $\{\mathcal{F}_i\}_{i\geq 0}$, where $\mathcal{F}_i$ is the σ-field generated by $\{(s_i,a_i)\}_{i=0}^{k_i}$. Then, we can invoke the Azuma-Hoeffding inequality to show that: for any such (s_l,a_l) and any $\tau \in [T]$, with probability exceeding $1-\delta/T^2$,

$$\left|\sum_{i=1}^{\tau}\mathbf{1}_{k_i<T}\eta_i^{\tau}\Big(\big(P-P_{k_i}\big)V_{k_i}\Big)(s_l,a_l)\right| \lesssim \frac{1}{1-\gamma}\sqrt{\sum_{i=1}^{\tau}\left(\eta_i^{\tau}\right)^2\log\frac{T}{\delta}} \lesssim \sqrt{\frac{\Gamma\iota}{\tau\,(1-\gamma)^2}}.$$

Taking the union bound over $\tau \in [T]$ allows us to replace τ with $n_t(s_l,a_l)$ in the above inequality, namely, this shows that for any such (s_l,a_l), with probability exceeding $1-\delta/T$ we have

$$\left|\sum_{i=1}^{n_t(s_l,a_l)}\eta_i^{n_t(s_l,a_l)}\Big(\big(P-P_{k_i}\big)V_{k_i}\Big)(s_l,a_l)\right| \lesssim \sqrt{\frac{\Gamma\iota}{n_t(s_l,a_l)\,(1-\gamma)^2}}$$

for all $t \in [T]$. Taking the union bound over all such (s_l,a_l) (which are concerned with at most T pairs), we see that with probability exceeding $1-\delta$,

$$\left|\sum_{i=1}^{n_t(s_l,a_l)}\eta_i^{n_t(s_l,a_l)}\Big(\big(P-P_{k_i}\big)V_{k_i}\Big)(s_l,a_l)\right| \lesssim \sqrt{\frac{\Gamma\iota}{n_t(s_l,a_l)\,(1-\gamma)^2}}$$

is valid for any $0 \leq j < T$ and any $t \in [T]$. This establishes the inequality equation 128, thus concluding the proof. $\square$

Next, we define two disjoint sets of state-action pairs, divided based on the associated occupancy probability induced by the behavior policy:

$$\mathcal{I} \coloneqq \left\{\big(s,\pi^\star(s)\big) \mid s \in \mathcal{S}, \mu_{\mathsf{b}}\big(s,\pi^\star(s)\big) \geq \frac{\delta}{ST}\right\}, \tag{129a}$$

$$\mathcal{I}^c \coloneqq \left\{\big(s,\pi^\star(s)\big) \mid s \in \mathcal{S}, \mu_{\mathsf{b}}\big(s,\pi^\star(s)\big) < \frac{\delta}{ST}\right\}. \tag{129b}$$

It turns out that the state-action pairs in $\mathcal{I}^c$ are rarely visited, as formalized by the following lemma.

Lemma 18. *(Lemma 3 in (Yan et al., 2022))With probability exceeding* $1-\delta$*, we have*

$$\mathcal{I}^c \cap \left\{(s_t, a_t)\right\}_{t=t_{\mathsf{mix}}(\delta)}^{T} = \varnothing.$$

Lemma 19. *(Lemma 5 in (Yan et al., 2022)) We can construct an auxiliary set of random variables* $\left\{\left(s_k^i, a_k^i\right) : 1 \le k \le K-1\right\}$ *satisfying*

$$\left\{ \left(s_k^i, a_k^i\right) : 1 \le k \le K-1\right\} \overset{\text{i.i.d.}}{\sim} \mu_{\mathsf{b}}, \tag{130a}$$

$$\mathbb{P}\Big\{ \left(s_k^i, a_k^i\right) = \left(s_{k\tau+i}, a_{k\tau+i}\right) \quad \textit{for all } 1 \le k \le K-1\Big\} \ge 1 - \frac{\delta}{T}, \tag{130b}$$

and

$$\left(s_k^i, a_k^i\right) \ \textit{is independent of} \left\{ \left(s_t, a_t\right) : 0 \le t \le (k-1)\,\tau + i\right\}. \tag{130c}$$

Lemma 20. *(Lemma 4 in (Yan et al., 2022)) Let* $\Gamma = \left\lceil \frac{4}{1-\gamma} \log \frac{ST}{\delta} \right\rceil$ *for some* $0 < \delta < 1$*. For any vector with non-negative entries* $V \in \mathbb{R}^d$ *, we have*

$$\sum_{j=0}^{\infty} \left[\gamma\left(1+\frac{1}{\Gamma}\right)^3\right]^j \left\langle \rho(P_{\pi^\star})^j, V\right\rangle \lesssim \frac{1}{1-\gamma}\left\langle d_\rho^\star, V\right\rangle + \frac{\delta}{ST^4\,(1-\gamma)}\left\|V\right\|_\infty. \tag{131}$$

E.4 Proof of Theorem 3

Following (Yan et al., 2022), we similarly define the following terms first:

$$\alpha_j \coloneqq \left[\gamma\left(1+\frac{1}{\Gamma}\right)^3\right]^j \sum_{t=1}^{T} \left\langle \rho(P_{\pi^\star})^j, V^\star - V_t\right\rangle,$$

$$\theta_j \coloneqq \left[\gamma\left(1+\frac{1}{\Gamma}\right)^3\right]^j \sum_{t=1}^{T}\sum_{s\in\mathcal{S}} \left[\rho(P_{\pi^\star})^j\right]\left(s, \pi^\star(s)\right) \min\left\{\beta_{n_t(s,\pi^\star(s))}\big(s,\pi^\star(s)\big), \frac{1}{1-\gamma}\right\},$$

$$\xi_j \coloneqq \left[\gamma\left(1+\frac{1}{\Gamma}\right)^3\right]^j \sum_{t=1}^{t_{\mathsf{mix}}(\delta)} \left\langle \rho(P_{\pi^\star})^j, V^\star - V_t\right\rangle + \left[\gamma\left(1+\frac{1}{\Gamma}\right)^3\right]^{j+1} \left\langle \rho(P_{\pi^\star})^{j+1}, V^\star - V_0\right\rangle,$$

$$\begin{aligned}\psi_j \coloneqq \left[\gamma\left(1+\frac{1}{\Gamma}\right)^3\right]^j \sum_{t=t_{\mathsf{mix}}(\delta)}^{T} \Bigg[&\sum_{s\in\mathcal{S}, a\in\mathcal{A}} \left[\rho^{\pi^\star}(P^{\pi^\star})^j\right](s,a) \sum_{i=1}^{n_t(s,a)} \eta_i^{n_t(s,a)} P_{s,a}\left(V^\star - V_{k_i(s,a)}\right) \\ &- \left(1+\frac{1}{\Gamma}\right)\frac{\left[\rho^{\pi^\star}(P^{\pi^\star})^j\right](s_t,a_t)}{\mu_{\mathsf{b}}\left(s_t,a_t\right)} \sum_{i=1}^{n_t(s_t,a_t)} \eta_i^{n_t(s_t,a_t)} P_{s_t,a_t}\left(V^\star - V_{k_i(s_t,a_t)}\right)\Bigg],\end{aligned}$$

$$\begin{aligned}\phi_j \coloneqq \gamma^{j+1}\left(1+\frac{1}{\Gamma}\right)^{3j+2} \sum_{t=0}^{T} \mathbf{1}_{(s_t,a_t)\in\mathcal{I}} \Bigg[&\frac{\left[\rho^{\pi^\star}(P^{\pi^\star})^j\right](s_t,a_t)}{\mu_{\mathsf{b}}\left(s_t,a_t\right)} P_{s_t,a_t}\left(V^\star - V_t\right) \\ &- \left(1+\frac{1}{\Gamma}\right)\sum_{s\in\mathcal{S}, a\in\mathcal{A}} \left[\rho^{\pi^\star}(P^{\pi^\star})^j\right](s,a)\, P_{s,a}\left(V^\star - V_t\right)\Bigg],\end{aligned}$$

where we recall the definition of $\mathcal{I}$ in equation 129.

We then proceed to the proof.

Theorem 21. *(Restatement of Theorem 3) Consider any* $\delta \in (0,1)$*. Suppose that the behavior policy* μ *satisfies Assumption 2. The policy* $\widehat{\pi}$ *returned by Algorithm 3 satisfies*

$$V^\star(\rho) - V^{\widehat{\pi}}(\rho) \le \tilde{\mathcal{O}}\left(\sqrt{\frac{C^\star S}{T\,(1-\gamma)^5}} + \frac{C^\star S}{T\,(1-\gamma)^2} + \frac{C^\star}{T\,(1-\gamma)^3}\right). \tag{132}$$

with probability at least $1-\delta$*.*

Proof. Note that

$$V^\star(\rho) - V^{\widehat{\pi}}(\rho) = \left\langle \rho, V^\star - V^{\widehat{\pi}} \right\rangle \overset{\text{(a)}}{\leq} \langle \rho, V^\star - V_T \rangle \overset{\text{(b)}}{\leq} \frac{1}{T}\sum_{t=1}^{T} \langle \rho, V^\star - V_t \rangle \overset{\text{(c)}}{=} \frac{1}{T}\alpha_0. \tag{133}$$

Here, (a) holds true according to Lemma 17; (b) follows from the monotonicity of V_t in t (by construction); and (c) follows simply from the definition of α_0. We then turn attention to bounding α_0, towards which we observe that

$$\begin{aligned}
\alpha_0 &= \sum_{t=1}^{t_{\mathsf{mix}}(\delta)-1} \langle \rho, V^\star - V_t \rangle + \sum_{t=t_{\mathsf{mix}}(\delta)}^{T} \sum_{s\in\mathcal{S}} \rho(s) \min\left\{ Q^\star\big(s,\pi^\star(s)\big) - V_t(s), \frac{1}{1-\gamma} \right\} \\
&\leq \sum_{t=1}^{t_{\mathsf{mix}}(\delta)-1} \langle \rho, V^\star - V_t \rangle + \sum_{t=t_{\mathsf{mix}}(\delta)}^{T} \sum_{s\in\mathcal{S}} \rho(s) \min\left\{ Q^\star\big(s,\pi^\star(s)\big) - Q_t\big(s,\pi^\star(s)\big), \frac{1}{1-\gamma} \right\} \\
&\leq \sum_{t=1}^{t_{\mathsf{mix}}(\delta)} \langle \rho, V^\star - V_t \rangle + \underbrace{\gamma \sum_{t=t_{\mathsf{mix}}(\delta)}^{T} \sum_{s\in\mathcal{S}} \rho(s) \sum_{i=1}^{n_t(s,\pi^\star(s))} \eta_i^{n_t(s,\pi^\star(s))} \big(\sigma_{s,\pi^\star(s)}(V^\star) - \sigma_{s,\pi^\star(s)}(V_{k_i}) \big)}_{=:\,\zeta} \\
&\quad + \underbrace{\sum_{t=1}^{T} \sum_{s\in\mathcal{S}} \rho(s) \min\left\{ \beta_{n_t(s,\pi^\star(s))}\big(s,\pi^\star(s)\big), \frac{1}{1-\gamma} \right\}}_{=:\,\frac{\theta_0}{1+R}}.
\end{aligned}$$

Here, the first identity holds since $V^\star(s) = Q^\star\big(s,\pi^\star(s)\big)$ and $0 \leq V^\star(s) - V_t(s) \leq 1/(1-\gamma)$ for all $s \in \mathcal{S}$, the second line relies on the fact that $V_t(s) \geq \max_a Q_t(s,a) \geq Q_t(s,\pi^\star(s))$, while the last line invokes Lemma 17. With probability exceeding $1-\delta$, the first term ζ can be upper bounded by

$$\begin{aligned}
\zeta &\leq \gamma \sum_{t=t_{\mathsf{mix}}(\delta)}^{T} \sum_{s\in\mathcal{S}} \rho(s) \sum_{i=1}^{n_t(s,\pi^\star(s))} \eta_i^{n_t(s,\pi^\star(s))} \big(\sigma_{s,\pi^\star(s)}(V^\star) - \sigma_{s,\pi^\star(s)}(V_{k_i}) \big) \\
&= \gamma \sum_{t=t_{\mathsf{mix}}(\delta)}^{T} \sum_{s\in\mathcal{S},a\in\mathcal{A}} \mu_{\mathsf{b}}(s,a) \frac{\rho^{\pi^\star}(s,a)}{\mu_{\mathsf{b}}(s,a)} \sum_{i=1}^{n_t(s,a)} \eta_i^{n_t(s,a)} P_{s,\pi^\star(s)} \left(V^\star - V_{k_i} \right) \\
&\quad - \gamma \sum_{t=t_{\mathsf{mix}}(\delta)}^{T} \sum_{s\in\mathcal{S},a\in\mathcal{A}} \mu_{\mathsf{b}}(s,a) \frac{\rho^{\pi^\star}(s,a)}{\mu_{\mathsf{b}}(s,a)} R_{s,a} \sum_{i=1}^{n_t(s,a)} \eta_i^{n_t(s,a)} \left(\kappa(V^\star) - \kappa(V_{k_i}) \right) \\
&\leq \gamma \sum_{t=t_{\mathsf{mix}}(\delta)}^{T} \sum_{s\in\mathcal{S},a\in\mathcal{A}} \mu_{\mathsf{b}}(s,a) \frac{\rho^{\pi^\star}(s,a)}{\mu_{\mathsf{b}}(s,a)} \sum_{i=1}^{n_t(s,a)} \eta_i^{n_t(s,a)} P_{s,\pi^\star(s)} \left(V^\star - V_{k_i} \right) \\
&\quad + 2\gamma \sum_{t=t_{\mathsf{mix}}(\delta)}^{T} \sum_{s\in\mathcal{S},a\in\mathcal{A}} \mu_{\mathsf{b}}(s,a) \frac{\rho^{\pi^\star}(s,a)}{\mu_{\mathsf{b}}(s,a)} R_{s,a} \sum_{i=1}^{n_t(s,a)} \eta_i^{n_t(s,a)} \left(V^\star - V_{k_i} \right),
\end{aligned}$$

where we utilize the fact that $V^* \geq V_{k_i}$ and κ is 1-Lipschitz. Hence we further have that

$$\begin{aligned}
&\alpha_0 \\
&\leq (1+R)\gamma \sum_{t=t_{\mathsf{mix}}(\delta)}^{T} \sum_{s\in\mathcal{S},a\in\mathcal{A}} \mu_{\mathsf{b}}(s,a) \frac{\rho^{\pi^\star}(s,a)}{\mu_{\mathsf{b}}(s,a)} \sum_{i=1}^{n_t(s,a)} \eta_i^{n_t(s,a)} P_{s,\pi^\star(s)} \left(V^\star - V_{k_i} \right) \\
&\quad + (1+R) \sum_{t=1}^{t_{\mathsf{mix}}(\delta)} \langle \rho, V^\star - V_t \rangle + \theta_0 \\
&\overset{\text{(a)}}{\leq} \gamma(1+R)\left(1+\frac{1}{\Gamma}\right) \sum_{t=t_{\mathsf{mix}}(\delta)}^{T} \mathbf{1}\{(s_t,a_t)\in\mathcal{I}\} \frac{\rho^{\pi^\star}(s_t,a_t)}{\mu_{\mathsf{b}}(s_t,a_t)} \sum_{i=1}^{n_t(s_t,a_t)} \eta_i^{n_t(s_t,a_t)} P_{s_t,a_t}\big(V^\star - V_{k_i(s_t,a_t)} \big) + \psi_0
\end{aligned}$$

$$
\begin{aligned}
&+ (1+R) \sum_{t=1}^{t_{\mathsf{mix}}(\delta)} \langle \rho, V^\star - V_t \rangle + \theta_0 \\
&\overset{\text{(b)}}{\asymp} \gamma \left(1 + \frac{1}{\Gamma}\right) \sum_{t=t_{\mathsf{mix}}(\delta)}^{T} \mathbf{1}\{(s_t,a_t) \in \mathcal{I}\} \frac{\rho^{\pi^\star}(s_t,a_t)}{\mu_{\mathsf{b}}(s_t,a_t)} \left(\sum_{j=n_t(s_t,a_t)}^{n_T(s_t,a_t)} \eta^j_{n_t(s_t,a_t)} \right) P_{s_t,a_t}(V^\star - V_t) + \psi_0 \\
&+ (1+R) \sum_{t=1}^{t_{\mathsf{mix}}(\delta)} \langle \rho, V^\star - V_t \rangle + \theta_0 \\
&\overset{\text{(c)}}{\leq} \gamma \left(1 + \frac{1}{\Gamma}\right)^2 \sum_{t=0}^{T} \mathbf{1}\{(s_t,a_t) \in \mathcal{I}\} \frac{\rho^{\pi^\star}(s_t,a_t)}{\mu_{\mathsf{b}}(s_t,a_t)} P_{s_t,a_t}(V^\star - V_t) + \psi_0 \\
&+ (1+R) \sum_{t=1}^{t_{\mathsf{mix}}(\delta)} \langle \rho, V^\star - V_t \rangle + \theta_0 \\
&= \gamma \left(1 + \frac{1}{\Gamma}\right)^3 \sum_{t=0}^{T} \sum_{s\in\mathcal{S},a\in\mathcal{A}} \rho^{\pi^\star}(s,a) P_{s,a}(V^\star - V_t) + \psi_0 + \phi_0 + (1+R) \sum_{t=1}^{t_{\mathsf{mix}}(\delta)} \langle \rho, V^\star - V_t \rangle + \theta_0 \\
&= \gamma \left(1 + \frac{1}{\Gamma}\right)^3 \sum_{t=0}^{T} \langle \rho P_{\pi^\star}, V^\star - V_t \rangle + \psi_0 + \phi_0 + (1+R) \sum_{t=1}^{t_{\mathsf{mix}}(\delta)} \langle \rho, V^\star - V_t \rangle + \theta_0 \\
&\leq \alpha_1 + \psi_0 + \phi_0 + \gamma \left(1 + \frac{1}{\Gamma}\right)^3 \langle \rho P_{\pi^\star}, V^\star - V_0 \rangle + (1+R) \sum_{t=1}^{t_{\mathsf{mix}}(\delta)} \langle \rho, V^\star - V_t \rangle + \theta_0,
\end{aligned}
$$

where we remind the reader of our notation $\rho^{\pi^\star}$ in equation 119. Here, (a) is valid (i.e., $\rho(s_t,a_t)/\mu_{\mathsf{b}}(s,a)$ is well defined for $t \geq t_{\mathsf{mix}}(\delta)$) due to Lemma 18; (b) holds by grouping the terms in the previous line; and (c) utilizes Lemma 16 and the property that $V^\star \geq V_t$ (cf. Lemma 17). Therefore, we arrive at

$$
\begin{aligned}
\alpha_0 &\leq \sum_{t=1}^{t_{\mathsf{mix}}(\delta)} \langle \rho, V^\star - V_t \rangle + \zeta + \theta_0 \\
&\leq (1+R) \sum_{t=1}^{t_{\mathsf{mix}}(\delta)} \langle \rho, V^\star - V_t \rangle + \alpha_1 + \psi_0 + \phi_0 + \gamma \left(1 + \frac{1}{\Gamma}\right)^3 \langle \rho P_{\pi^\star}, V^\star - V_0 \rangle + \theta_0 \\
&= \alpha_1 + \xi_0 + \theta_0 + \psi_0 + \phi_0,
\end{aligned}
$$

where we have used the definition of ξ_0. Repeat the same argument to reach

$$\alpha_j \leq \alpha_{j+1} + \xi_j + \theta_j + \psi_j + \phi_j$$

for all $j \geq 1$. This in turn allows us to conclude that

$$\alpha_0 \leq \underbrace{\limsup_{j\to\infty} \alpha_j}_{=:\,\alpha} + \underbrace{\sum_{j=0}^{\infty} \xi_j}_{=:\,\xi} + \underbrace{\sum_{j=0}^{\infty} \theta_j}_{=:\,\theta} + \underbrace{\sum_{j=0}^{\infty} \psi_j}_{=:\,\psi} + \underbrace{\sum_{j=0}^{\infty} \phi_j}_{=:\,\phi}. \tag{134}$$

We will then bound the terms α, ξ, θ, ψ and ϕ separately in the subsequent steps. Our proofs are similar to the ones in (Yan et al., 2022), hence we omit the repeated part.

Bounding α. The bound is similar to (Yan et al., 2022). It is first observed that

$$\alpha = \limsup_{j\to\infty} \left[\gamma \left(1 + \frac{1}{\Gamma}\right)^3\right]^j \sum_{t=1}^{T} \left\langle \rho (P_{\pi^\star})^j, V^\star - V_t \right\rangle \leq \frac{T}{1-\gamma} \limsup_{k\to\infty} \left[\gamma \left(1 + \frac{1}{\Gamma}\right)^3\right]^k = 0.$$

Bounding ξ.

By utilizing (131), it holds that

$$\xi = \sum_{t=1}^{t_{\mathsf{mix}}(\delta)} \left\{ \sum_{j=0}^{\infty} \left[\gamma \left(1+\frac{1}{\Gamma}\right)^3 \right]^j \left\langle \rho P_{\pi^\star}^j, V^\star - V_t \right\rangle \right\} + \sum_{j=0}^{\infty} \left[\gamma \left(1+\frac{1}{\Gamma}\right)^3 \right]^{j+1} \left\langle \rho (P_{\pi^\star})^{j+1}, V^\star - V_0 \right\rangle$$
$$\lesssim \frac{1}{1-\gamma} \sum_{t=0}^{t_{\mathsf{mix}}(\delta)} \left\langle d_\rho^\star, V^\star - V_t \right\rangle + \frac{1}{ST^4 (1-\gamma)} \frac{t_{\mathsf{mix}}(\delta)+1}{1-\gamma}$$
$$\lesssim \frac{t_{\mathsf{mix}}}{(1-\gamma)^2} \log\frac{1}{\delta} + \frac{t_{\mathsf{mix}}}{T^4 (1-\gamma)^2} \log\frac{1}{\delta}.$$

Bounding θ. Following (Yan et al., 2022), we have that Note that

$$\theta = \sum_{j=0}^{\infty} \left[\gamma \left(1+\frac{1}{\Gamma}\right)^3 \right]^j \sum_{t=1}^{T} \sum_{s\in\mathcal{S}} \left[\rho (P_{\pi^\star})^j \right](s) \min\left\{ \beta_{n_t(s,\pi^\star(s))}, \frac{1}{1-\gamma} \right\}$$
$$\lesssim \frac{C^\star S t_{\mathsf{mix}} \iota}{(1-\gamma)^2} + \sqrt{\frac{C^\star S T \iota^2}{(1-\gamma)^5}}.$$

Bounding ψ. Note that

$$\psi = \sum_{j=0}^{\infty} \gamma \left[\gamma \left(1+\frac{1}{\Gamma}\right)^3 \right]^j \sum_{t=t_{\mathsf{mix}}(\delta)}^{T} \Bigg[\sum_{s\in\mathcal{S},a\in\mathcal{A}} \left[\rho^{\pi^\star} (P^{\pi^\star})^j \right](s,a) \sum_{i=1}^{n_t(s,a)} \eta_i^{n_t(s,a)} P_{s,a} \left(V^\star - V_{k_i(s,a)}\right)$$
$$- \left(1+\frac{1}{\Gamma}\right) \frac{\left[\rho^{\pi^\star} (P^{\pi^\star})^j \right](s_t,a_t)}{\mu_{\mathsf{b}}(s_t,a_t)} \sum_{i=1}^{n_t(s_t,a_t)} \eta_i^{n_t(s_t,a_t)} P_{s_t,a_t} \left(V^\star - V_{k_i(s_t,a_t)}\right) \Bigg]$$
$$= \sum_{t=t_{\mathsf{mix}}(\delta)}^{T} \Bigg[\sum_{s\in\mathcal{S},a\in\mathcal{A}} \widetilde{d}(s,a) \sum_{i=1}^{n_t(s,a)} \eta_i^{n_t(s,a)} P_{s,a} \left(V^\star - V_{k_i(s,a)}\right)$$
$$- \left(1+\frac{1}{\Gamma}\right) \frac{\widetilde{d}(s_t,a_t)}{\mu_{\mathsf{b}}(s_t,a_t)} \sum_{i=1}^{n_t(s_t,a_t)} \eta_i^{n_t(s_t,a_t)} P_{s_t,a_t} \left(V^\star - V_{k_i(s_t,a_t)}\right) \Bigg].$$

Here,

$$\widetilde{d}(s,a) := \sum_{j=0}^{\infty} \gamma \left[\gamma \left(1+\frac{1}{\Gamma}\right)^3 \right]^j \left[\rho^{\pi^\star} (P^{\pi^\star})^j \right](s,a)$$

for any $(s,a) \in \mathcal{S}\times\mathcal{A}$. Note that this equation exactly matches with Step 2.4 in (Yan et al., 2022), hence the remaining proof similarly follows, and is omitted here. Specifically, we have that

$$\psi \lesssim \frac{C^\star t_{\mathsf{mix}} \iota}{(1-\gamma)^3} \log^2\left(\frac{T}{\delta}\right) + \frac{C^\star S t_{\mathsf{mix}}}{(1-\gamma)^2} \log\left(\frac{T}{\delta}\right).$$

Bounding ϕ. Similar to (Yan et al., 2022), we can employ an analogous argument to show that ϕ can be bounded as

$$\phi \lesssim \frac{C^\star t_{\mathsf{mix}} \iota}{(1-\gamma)^3} \log^2\left(\frac{T}{\delta}\right) + \frac{C^\star S t_{\mathsf{mix}}}{(1-\gamma)^2} \log\left(\frac{T}{\delta}\right).$$

Now, plugging the bounds on α, θ, ψ and ϕ further implies that

$$\alpha_0 \le \alpha + \xi + \theta + \psi + \phi$$
$$\lesssim \sqrt{\frac{C^\star S T \iota^2}{(1-\gamma)^5}} + \frac{C^\star S t_{\mathsf{mix}} \iota}{(1-\gamma)^2} + \frac{C^\star t_{\mathsf{mix}} \iota}{(1-\gamma)^3} \log^2\left(\frac{T}{\delta}\right).$$

We then invoke equation 133 to conclude that

$$V^{\star}\left(\rho\right)-V^{\widehat{\pi}}\left(\rho\right)\leq\frac{\alpha_{0}}{T}\lesssim\sqrt{\frac{C^{\star}S\iota^{2}}{T\left(1-\gamma\right)^{5}}}+\frac{C^{\star}St_{\mathsf{mix}}\iota}{T\left(1-\gamma\right)^{2}}+\frac{C^{\star}t_{\mathsf{mix}}\iota^{2}}{T\left(1-\gamma\right)^{3}}.$$

This hence completes the proof. $\square$