PROVABLY SAMPLE-EFFICIENT ROBUST REINFORCEMENT LEARNING WITH AVERAGE REWARD

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust reinforcement learning (RL) under the average-reward criterion is essential for long-term decision-making, particularly when the environment may differ from its specification. However, a significant gap exists in understanding the finite-sample complexity of these methods, as most existing work provides only asymptotic guarantees. This limitation hinders their principled understanding and practical deployment, especially in data-limited scenarios. We close this gap by proposing **Robust Halpern Iteration (RHI)**, a new algorithm designed for robust Markov Decision Processes (MDPs) with transition uncertainty characterized by ℓ_p-norm and contamination models. Our approach offers three key advantages over previous methods: *(1). Weaker Structural Assumptions:* RHI only requires the underlying robust MDP to be communicating, a less restrictive condition than the commonly assumed ergodicity or irreducibility; *(2). No Prior Knowledge:* Our algorithm operates without requiring any prior knowledge of the robust MDP; *(3). State-of-the-Art Sample Complexity:* To learn an ϵ-optimal robust policy, RHI achieves a sample complexity of $\tilde{\mathcal{O}}\left(\frac{SA\mathcal{H}^2}{\epsilon^2}\right)$, where S and A denote the numbers of states and actions, and $\mathcal{H}$ is the robust optimal bias span. This result represents the tightest known bound. Our work hence provides essential theoretical understanding of sample efficiency of robust average reward RL.

1 INTRODUCTION

Reinforcement Learning (RL) seeks to find an optimal policy for an agent interacting with an environment to maximize a cumulative reward. While RL has achieved remarkable success in controlled settings like board games (Silver et al., 2016; Zha et al., 2021) and video games (Wei et al., 2022; Liu et al., 2022a), its deployment in real-world applications is often hindered by a significant performance drop. This issue, known as the "Sim-to-Real" gap (Zhao et al., 2020; Peng et al., 2018; Tobin et al., 2017), stems from mismatches between the training (simulation) and deployment (real-world) environments. In contrast to games where these environments are identical, practical scenarios are fraught with model discrepancies arising from modeling errors, environmental perturbations, or even adversarial attacks (Henderson et al., 2018; Rajeswaran et al., 2016; Zhang et al., 2018). Such mismatches can render a learned policy highly suboptimal, severely undermining the reliability of RL in practice. To address this critical reliability challenge, the framework of (distributionally) robust RL was developed (Bagnell et al., 2001; Nilim & El Ghaoui, 2004; Iyengar, 2005). Instead of assuming a single, perfectly known environment model, robust RL considers an uncertainty set of plausible transition dynamics. The objective is to find a policy that optimizes performance for the worst-case model within this set. This "worst-case" approach yields a policy with formal performance guarantees across all considered environmental variations, making it inherently more resilient and robust to model mismatch and enhancing its generalizability (Pinto et al., 2017; Zhang et al., 2025).

Beyond robustness, the choice of the reward criterion fundamentally shapes the RL problem. The discounted-reward criterion, while mathematically elegant and widely studied, can be myopic due to its exponential down-weighting of future rewards, potentially leading to poor long-term outcomes (Schwartz, 1993; Seijen & Sutton, 2014; Tsitsiklis & Roy, 1997; Abounadi et al., 2001). In contrast, numerous real-world applications–such as queuing control, portfolio optimization, and communication networks (Kober et al., 2013; Lu et al., 2018; Chen et al., 2022; Wu et al., 2023; Moody & Saffell, 2001; Charpentier et al., 2021; Masoudi, 2021; Li & Hai, 2024)–demand policies that are evaluated

based on their long-term, steady-state performance when executed over an extended period of time. This practical necessity underscores the importance of the average-reward criterion, which does not discount the future reward and thus captures the long-term performance (Sigaud & Buffet, 2013). In this paper, we focus on the intersection of these two needs: developing robust RL algorithms under the average-reward criterion, to ensure performance of RL systems under model mismatch.

Robust RL under the average-reward criterion, however, is more challenging than its discounted-reward counterpart and remains relatively understudied. The primary difficulties stem from its reliance on the limiting behavior of stochastic processes, leading to analytical and algorithmic complications. Recent work has highlighted these issues, including the non-contractive nature of the associated Bellman operator, the high dimensionality of the solution space, and the instability of standard iterative algorithms (Wang et al., 2023g; Grand-Clement et al., 2023). Therefore, a critical gap in the literature persists: existing studies are predominantly asymptotic or planning based, leaving the crucial finite-sample properties of data-driven robust average-reward RL largely unexplored.

A natural strategy to obtain finite-sample results is to reduce the average-reward problem to its discounted counterpart, thereby leveraging the rich literature on robust discounted-reward RL (Wang et al., 2022; Zurek & Chen, 2023). This approach is theoretically supported by the convergence of the robust discounted value function to the average-reward value function as the discount factor approaches one (Wang et al., 2023f). However, these reduction-based methods are often suboptimal (Grand-Clément & Petrik, 2023) or require additional prior knowledge (Roch et al., 2025). While other recent works have proposed direct methods, they typically rely on strong structural assumptions, such as irreducibility, which induce a contraction property (Xu et al., 2025a;b). To circumvent these limitations, in this paper, we propose a direct approach, **Robust Halpern Iteration (RHI)**, which enables a practical, model-free implementation and achieves a near-optimal sample complexity. Our contributions are summarized as follows.

Theoretical Foundation for Communicating Robust AMDPs. We relax the restrictive structural assumptions common in prior work, such as irreducibility (Xu et al., 2025a) and ergodicity (Chen et al., 2025), by analyzing robust AMDPs under the weaker *communicating* condition (Bertsekas, 2011). Within this more general framework, we first establish that the optimal robust average reward is constant across all states. We then provide fundamental guarantees for the corresponding robust Bellman equation, proving its solvability and the optimality of its solution. Crucially, we formally derive the equivalence between solving this equation and finding an optimal robust policy, which provides the theoretical foundation for our algorithm's design and analysis.

A Near-Optimal, Model-Free Algorithm for Robust Average-Reward RL. We propose the Robust Halpern Iteration (RHI), a direct algorithm that bypasses the complexities of reduction-based approaches. Inspired by Halpern Iteration from the optimization literature (Halpern, 1967; Lieder, 2021; Lee et al., 2025), our method integrates two key technical innovations: (1) leveraging a quotient space to manage the high dimensionality of the robust Bellman equation's solution space and tackle the double unknown variables in the equation, and (2) designing a novel estimator for the robust average-reward Bellman operator. We provide a rigorous finite-sample analysis for RHI under both contamination (Wang & Zou, 2021; 2022; Jiao & Li, 2024) and ℓ_p-norm (Kumar et al., 2023; Zhang et al., 2025) uncertainty models. Under our communicating assumption, we prove that RHI finds an ϵ-optimal policy with a sample complexity of $\tilde{\mathcal{O}}\left(\frac{SA\mathcal{H}^2}{\epsilon^2}\right)$, where S and A are the sizes of the state and action spaces, and $\mathcal{H}$ is the span of the robust optimal bias. This result establishes the tightest near-optimal sample complexity bound for robust average-reward RL.

Empirical Validation. We validate the practical performance of RHI by conducting experiments across three common uncertainty models: contamination, total variation (ℓ_∞-norm), and ℓ_2-norm. Our results demonstrate that RHI consistently and efficiently converges to the optimal robust average reward, computed based on the RRVI method (Wang et al., 2023g). These empirical findings corroborate our theoretical analysis and validate the convergence of RHI in practice.

2 Preliminaries and problem formulation

Discounted reward MDPs. A discounted reward Markovian decision process (DMDP) $(\mathcal{S}, \mathcal{A}, \mathsf{P}, r, \gamma)$ is specified by: a state space $\mathcal{S}$, an action space $\mathcal{A}$, a nominal (stationary) transition kernel $\mathsf{P} =$

$\{\mathsf{P}_s^a \in \Delta(\mathcal{S}), a \in \mathcal{A}, s \in \mathcal{S}\}$[1], where P_s^a is the distribution of the next state over $\mathcal{S}$ upon taking action a in state s (with $\mathsf{P}_{s,s'}^a$ denoting the probability of transitioning to s'), a reward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$, and a discount factor $\gamma \in [0, 1)$. At each time step t, the agent at state s_t takes an action a_t, the environment then transitions to the next state s_{t+1} according to $\mathsf{P}_{s_t}^{a_t}$, and produces a reward signal $r_t = r(s_t, a_t)$ to the agent.

A stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is a distribution over $\mathcal{A}$ for any given state s. The agent follows the policy by taking an action following the distribution $\pi(s)$. The accumulative reward of a stationary policy π starting from $s \in \mathcal{S}$ for DMDPs is measured by the discounted value function: $V_{\gamma,\mathsf{P}}^{\pi}(s) \triangleq \mathbb{E}_{\pi,\mathsf{P}}\left[\sum_{t=0}^{\infty} \gamma^t r_t | S_0 = s\right]$.

Average reward MDPs. Unlike DMDPs, average reward MDPs (AMDPs) do not discount the rewards over time and instead measure the accumulative reward by considering the behavior of the underlying Markov process under the steady-state distribution. Specifically, the average reward (or the gain) of a policy π starting from $s \in \mathcal{S}$ is

$$g_{\mathsf{P}}^{\pi}(s) \triangleq \liminf_{n\to\infty} \mathbb{E}_{\pi,\mathsf{P}}\left[\frac{1}{n}\sum_{t=0}^{n-1} r_t | S_0 = s\right]. \tag{1}$$

The bias or the relative value function for an AMDP is defined as the cumulative difference over time between the immediate reward and the average reward:

$$h_{\mathsf{P}}^{\pi}(s) \triangleq \mathbb{E}_{\pi,\mathsf{P}}\left[\sum_{t=0}^{\infty} (r_t - g_{\mathsf{P}}^{\pi}) | S_0 = s\right]. \tag{2}$$

Distributionally robust MDPs. In distributionally robust MDPs, the transition kernel is not fixed but, instead, belongs to a designated uncertainty set denoted as $\mathcal{P}$. Following an action, the environment undergoes a transition to the next state based on an arbitrary transition kernel $\mathsf{P} \in \mathcal{P}$. In this paper, we mainly focus on the (s, a)-rectangular uncertainty set (Nilim & El Ghaoui, 2004; Iyengar, 2005; Wiesemann et al., 2013), where $\mathcal{P} = \bigotimes_{s,a} \mathcal{P}_s^a$, with $\mathcal{P}_s^a \subseteq \Delta(\mathcal{S})$ defined independently over all state-action pairs. In most studies, the uncertainty set is defined through some distribution divergence:

$$\mathcal{P}_s^a = \{q \in \Delta(\mathcal{S}) : D(q||\mathsf{P}_s^a) \le R\}, \tag{3}$$

where D is some distribution divergence like total variation, P_s^a is the centroid of the uncertainty set, referred to as the nominal kernel, and R is the radius of the uncertainty set for the given state and action, measuring the level of uncertainties. In most studies, the nominal kernel can be viewed as the simulation, and all training data are generated under it. In this paper, we mainly consider two widely studied models:

$$\text{Contamination model: } \mathcal{P}_s^a = \{(1 - R)\mathsf{P}_s^a + Rq : q \in \Delta(\mathcal{S})\}, \tag{4}$$

$$\ell_p\text{-norm model: } \mathcal{P}_s^a = \{q \in \Delta(\mathcal{S}) : \|q - \mathsf{P}_s^a\|_p \le R\}. \tag{5}$$

Robust MDPs aim to optimize the worst-case performance over the uncertainty set. With the discounted reward criterion, the robust DMDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$ consider the robust discounted value function of a policy π, which is the worst-case discounted value function over all possible transition kernels:

$$V_{\gamma,\mathcal{P}}^{\pi}(s) \triangleq \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\pi,\mathsf{P}}\left[\sum_{t=0}^{\infty} \gamma^t r_t | S_0 = s\right]. \tag{6}$$

The discounted robust value functions are shown to be the unique solution to the robust discounted Bellman equation (Iyengar, 2005), where $\sigma_{\mathcal{P}_s^a}(V) \triangleq \min_{\mathsf{P}\in\mathcal{P}_s^a} \mathsf{P}V$:

$$V(s) = \sum_a \pi(a|s)(r(s, a) + \gamma\sigma_{\mathcal{P}_s^a}(V)). \tag{7}$$

When the long-term performance under uncertainty is concerned, we focus on the robust AMDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$. The worst-case performance is then measured by the following robust average reward:

$$g_{\mathcal{P}}^{\pi}(s) \triangleq \min_{\mathsf{P}\in\mathcal{P}} \liminf_{n\to\infty} \mathbb{E}_{\pi,\mathsf{P}}\left[\frac{1}{n}\sum_{t=0}^{n-1} r_t | S_0 = s\right] = \min_{\mathsf{P}\in\mathcal{P}} g_{\mathsf{P}}^{\pi}(s). \tag{8}$$

[1] $\Delta(\mathcal{S})$: the $(|\mathcal{S}| - 1)$-dimensional probability simplex on $\mathcal{S}$.

The robust AMDP aims to find an optimal policy w.r.t. it: $\pi^* \triangleq \arg\max_{\pi\in\Pi} g^\pi_{\mathcal{P}}(s),$ for any $s \in \mathcal{S}$, and we denote the optimal robust average reward by $g^*_{\mathcal{P}} \triangleq \max_\pi g^\pi_{\mathcal{P}}$. Moreover, we define the optimal robust bias span for the robust AMDP as

$$\mathcal{H} \triangleq \max_{\mathsf{P}\in\mathcal{P}} \mathbf{Sp}(h^{\pi^*}_{\mathsf{P}}) \tag{9}$$

where $h^{\pi^*}_{\mathsf{P}}$ is the bias defined in equation 2 and $\mathbf{Sp}(h) \triangleq \max_s h(s) - \min_s h(s)$ is the Span semi-norm.

Problem formulation. We consider the standard *generative model setting* (Panaganti & Kalathil, 2022; Shi et al., 2023; Xu et al., 2023), where the learner assumes access to a simulator to generate i.i.d. samples under any state-action pair, following the nominal kernel P. We study the sample complexity from the nominal kernel for identifying an ϵ-optimal policy π for the robust AMDP:

$$g^{\pi^*}_{\mathcal{P}}(s) - g^\pi_{\mathcal{P}}(s) \le \epsilon, \forall s \in \mathcal{S}. \tag{10}$$

3 Communicating RAMDPs

In this work, we consider robust AMDPs with compact uncertainty sets and satisfying the robust communicating assumption, which can be viewed as an extension of the standard weakly communicating condition in standard MDPs, e.g., (Bertsekas, 2011; Wan et al., 2021; Wan & Sutton, 2022; Zurek & Chen, 2024; 2023; Wang et al., 2022; Zhang & Xie, 2023).[2]

Assumption 3.1. The uncertainty set $\mathcal{P}$ is compact. Moreover, for any transition kernel $\mathsf{P} \in \mathcal{P}$, and any two states $s \neq s' \in \mathcal{S}$, there exists a stationary policy π and some positive integer N, such that $\mathsf{P}^\pi(S_N = s' | S_0 = s) > 0$.

The robust communicating assumption assumes that for any kernel $\mathsf{P} \in \mathcal{P}$, any state s' can be reached from any other state s under some policy. Note that this policy may vary depending on the specific state pair and transition kernel. This condition is substantially weaker than the ergodic or irreducible assumptions made in previous robust AMDP literature Chen et al. (2025); Xu et al. (2025b;a), which require that all states inter-communicate under **any** stationary policy. It also differs from the unichain assumption (Wang et al., 2023f;g; Roch et al., 2025), which permits transient states but requires all recurrent states to form a single communicating class under **any** stationary deterministic policy. While neither our communicating assumption nor the unichain assumption strictly contains the other, however, our theoretical results can be directly applied to the unichain setting.

We then characterize structures of robust AMDPs under Assumption 3.1. Specifically, we mainly focus on the following robust Bellman equation of $(Q, g) \in \mathbb{R}^{SA} \times \mathbb{R}$:

$$Q(s,a) = r(s,a) - g + \sigma_{\mathcal{P}^a_s}(Q_{\max}), \tag{11}$$

where $\cdot_{\max} : \mathbb{R}^{SA} \to \mathbb{R}^S$ is a mapping that maps any SA-dimensional vector Q to a S-dimensional vector $Q_{\max} \in \mathbb{R}^S$ with entry $Q_{\max}(s) = \max_{a\in\mathcal{A}} Q(s,a)$. This equation plays a central part in unichain robust AMDP studies, and we extend the results to our communicating setting.

Theorem 3.2. *Consider a robust AMDP satisfying Assumption 3.1. Then it holds that:*

(1). The optimal robust average reward $g^*_{\mathcal{P}}$ is a constant, i.e., $g^*_{\mathcal{P}}(s_1) = g^*_{\mathcal{P}}(s_2), \forall s_1 \neq s_2$;

(2). The robust Bellman equation in 11 has a solution (Q^*, g^*), and the solution g^* is the optimal robust average reward, i.e., $g^* = g^*_{\mathcal{P}}(s)$;

(3). The greedy policy π^* w.r.t. Q, i.e., $\pi^*(s) \in \arg\max_a Q(s,a)$, is an optimal robust policy.

Our results extend the results for unichain robust AMDPs in (Wang et al., 2023f;g). Specifically, denote $\mathcal{T}_{\mathcal{P},g}(Q)(s,a) \triangleq r(s,a) - g + \sigma_{\mathcal{P}^a_s}(Q_{\max})$, then the robust Bellman equation equation 11 can be rewritten as $Q = \mathcal{T}_{\mathcal{P},g(Q)}$. As proved, the optimal policy π^* can be obtained from the solution Q^* to equation 11: $\pi^*(s) \in \arg\max_a Q^*(s,a)$, thus obtaining the optimal policy for our communicating robust AMDP is equivalent to solving the equation $Q = \mathcal{T}_{\mathcal{P},g^*_{\mathcal{P}}}(Q)$.

Based on this fundamental result, we develop a sample efficient algorithm to effectively solve equation 11, thus finding the optimal robust policy.

[2]Our communicating assumption is slightly stronger than the standard weakly communicating condition, which allows transient states to exist.

4 ROBUST HALPERN ITERATION (RHI) FOR ROBUST AMDPS

In this section, we design our data-driven robust Halpern Iteration (RHI) algorithm to solve equation 11. We will show later that, our RHI algorithm does not require any prior information of the robust AMDP, and achieves a near-optimal sample complexity.

As discussed in Section 2, finding the optimal policy for a robust AMDP is equivalent to solving the corresponding robust Bellman equation (11): $Q = \mathcal{T}_{\mathcal{P},g^*}(Q) = \mathcal{T}_{\mathcal{P}}(Q) - g^*_{\mathcal{P}}$, where $\mathcal{T}_{\mathcal{P}}(Q) \triangleq r + \sigma_{\mathcal{P}}(Q_{\max})$. However, solving this equation is highly challenging. Firstly, the equation has two unknown variables: Q and $g^*_{\mathcal{P}}$; Since $g^*_{\mathcal{P}}$ is unknown, the operator $\mathcal{T}_{\mathcal{P},g^*_{\mathcal{P}}}$ is not readily feasible. Moreover, different from the irreducible or ergodic cases where the operator $\mathcal{T}_{\mathcal{P},g}$ is a contraction, it is a *non-expansion* under our setting, invalidating the previous methods. Finally, the non-linear structure of $\mathcal{T}_{\mathcal{P},g}$ (compared to the linear structure of the non-robust operator) further results in a complicated solution space to the Bellman equation (Wang et al., 2023g). In the following, we address these challenges sequentially, and propose our RHI algorithm.

Curse of dual variables. To address the issue of solving an equation with two unknown variables, we first claim that, even if we do not know the value of $g^*_{\mathcal{P}}$, we can still obtain the optimal policy through a proximal equation. Our claim is based on the following result, where we show that a near-optimal policy can be identified by approximating the solution to the robust Bellman equation (11) w.r.t. the Span semi-norm.

Lemma 4.1. *Under Assumption 3.1, let $Q \in \mathbb{R}^{SA}$ and π be the greedy policy w.r.t. Q, i.e., $\pi(s) \in \arg\max_{a\in\mathcal{A}} Q(s,a)$. Then, for every state $s \in \mathcal{S}$, it holds that:*

$$0 \le g^*_{\mathcal{P}} - g^{\pi}_{\mathcal{P}}(s) \le \boldsymbol{Sp}(\mathcal{T}_{\mathcal{P},g^*_{\mathcal{P}}}(Q) - Q) = \boldsymbol{Sp}(\mathcal{T}_{\mathcal{P}}(Q) - Q). \tag{12}$$

The result thus implies that, to obtain the optimal policy π^*, exactly solving equation 11 is not necessary; instead, it suffices to find a weaker solution Q such that $\mathcal{T}_{\mathcal{P},g^*_{\mathcal{P}}}(Q) - Q = ce$, for some constant $c \in \mathbb{R}$ and the all-one vector $e = (1, ..., 1) \in \mathbb{R}^{SA}$ (note that the solution Q to equation 11 also satisfies the equation with $c = 0$). Moreover, we show that this equation, and hence finding the optimal policy, are further equivalent to solving the proximal equation that only contains one variable:

$$\mathcal{T}_{\mathcal{P}}(Q) - Q = ce, \text{ for some } c \in \mathbb{R}, \tag{13}$$

since it is sufficient to find an arbitrary solution to equation 13 for some c. Noting that the span semi-norm is invariant to constant shifts, and inspired by previous studies of non-robust AMDPs (Zhang et al., 2021; Lee et al., 2025), we instead consider the embedded equation in the quotient space w.r.t. identical vectors. Namely, we define a relation between two vectors $v, w \in \mathbb{R}^{SA}$: $v \sim w$ if $v - w = ce$ for some c, which can be directly verified to be an equivalence relation. We thus construct the quotient space $E \triangleq \mathbb{R}^{SA}/\sim$, and the embedded equation of equation 13 on E becomes:

$$[\mathcal{T}_{\mathcal{P}}(Q)] = [Q], \text{ where } [\cdot] \text{ denotes the equivalence class of } \cdot. \tag{14}$$

Thus, solving a robust AMDP is equivalent to solving equation 14 in the quotient space E. Notably, *this equation only contains one variable and has a much easier structure*.

Non-contraction. The second challenge is that the robust Bellman operator $\mathcal{T}_{\mathcal{P}}$ is not a contraction, but rather only a non-expansion, even in the quotient space E. This invalidates the previous approaches for the discounted setting or average reward setting with stronger assumptions (Chen et al., 2025; Xu et al., 2025a;b), which utilize the Banach-Picard iteration to find the unique fixed point of the contracted operator. To address this issue and find a solution to the non-expansion equation 13, we adopt the Halpern iteration (Halpern, 1967) from the stochastic approximation area. Specifically, to solve an equation $x = T(x)$ for a non-expansion operator x, the Halpern iteration recursively updates the algorithms through $x^{k+1} = (1 - \beta_{k+1})x^0 + \beta_{k+1}T(x^k)$, which is a convex combination between $T(x^k)$ and the initialization x^0. Halpern iteration has been studied in optimization areas (Halpern, 1967; Sabach & Shtern, 2017; Lieder, 2021; Park & Ryu, 2022; Contreras & Cominetti, 2023) and more recently in non-robust RL (Lee et al., 2025; Lee & Ryu, 2025).

Based on the Halpern iteration, we can similarly develop our RHI algorithm in the quotient space as $[Q^{k+1}] = [(1 - \beta_{k+1})Q^0 + \beta_{k+1}\mathcal{T}_{\mathcal{P}}(Q)]$. We show in the following result that it will converge to some solution to equation 14, and hence find an optimal policy, when the robust AMDP is known.

Theorem 4.2. *Consider the exact robust Halpern iteration* $[Q^{k+1}] = [(1-\beta_{k+1})Q^0 + \beta_{k+1}\mathcal{T}_{\mathcal{P}}(Q)]$, *with* $\beta_k = \frac{k}{k+2}$. *Set* π^k *to be the greedy policy w.r.t.* Q^k. *Then,*

$$\boldsymbol{Sp}(\mathcal{T}_{\mathcal{P}}(Q^k) - Q^k) \to 0, \text{ and } g^*_{\mathcal{P}} - g^{\pi^k}_{\mathcal{P}} \to 0, \text{ as } k \to \infty. \tag{15}$$

This result hence implies the asymptotic convergence of our RHI algorithm, even if the operator may not be a contraction. Notably, the convergence result utilizes the solvability of the robust Bellman equation, which we derived under our weaker communicating setting.

Efficient data-driven algorithm. The above convergence of RHI can be obtained when we exactly know the uncertainty set $\mathcal{P}$. However, in the learning setting where we do not know the worst-case kernel, we only have access to samples from the nominal kernel. This stands as the most challenging problem in the robust RL setting, since estimating the robust Bellman operator from nominal samples can be challenging, known as off-dynamic learning (Eysenbach et al., 2020; Liu & Xu, 2024; Holla, 2021). Note that the robust Bellman operator captures the dynamics under the worst-case transition kernel, which is generally different from the nominal kernel. To address this issue, a multi-level Monte-Carlo (MLMC) approach was introduced in previous works (Liu et al., 2022b; Wang et al., 2023g). However, MLMC generally results in an infinitely large sample complexity, and only guarantees asymptotic convergence, hence it cannot be applied.

To effectively estimate the robust Bellman operator while maintaining a tractable sample complexity, we propose a recursive sampling technique, inspired by (Lee et al., 2025; Jin et al., 2024b). In particular, we utilize the nominal samples to estimate the difference between two steps: $\mathcal{T}_{\mathcal{P}}(Q^k) - \mathcal{T}_{\mathcal{P}}(Q^{k-1})$. Notably, although $\mathcal{T}_{\mathcal{P}}$ is an off-dynamic term, the difference term $\mathcal{T}_{\mathcal{P}}(Q^k) - \mathcal{T}_{\mathcal{P}}(Q^{k-1})$ can be efficiently estimated under the uncertainty sets we considered, thus enabling our algorithm design. Moreover, this sampling scheme allows us to re-use the samples from previous steps, and hence improves sample efficiency. Based on this technique, we design a concrete sampling subroutine, **R-SAMPLE**, for two types of uncertainty sets: contamination model in equation 4 and ℓ_p-norm model in equation 5. We further incorporate our R-SAMPLE sampling algorithm to propose our RHI algorithm in Algorithm 1. In our algorithm, we utilize the sampling scheme to estimate the difference between two steps, and then re-use the estimation T^{k-1} of the Bellman operator for the previous step to construct the estimation T^k for the current step.

Algorithm 1 Robust Halpern Iteration (RHI)

Input: $Q^0 = 0 \in \mathbb{R}^{SA},\ \delta \in (0,1),\ c_0 = 10 \cdot \ln^2(2),\ \beta_0 = 0$
$\alpha = \ln(2|\mathcal{S}||\mathcal{A}|(n+1)/\delta)$
$T^{-1} = r; h^{-1} = 0$
for $k = 0, \ldots, n$ **do**
  $c_k = 5(k+2)\ln^2(k+2),\ \beta_k = k/(k+2)$
  $Q^k = (1-\beta_k)\, Q^0 \ +\ \beta_k\, T^{k-1}$
  $h^k = Q^k_{\max}$
  $m_k = \max\{\lceil \alpha c_k \mathbf{Sp}(h^k - h^{k-1})^2/\epsilon^2 \rceil, 1\}$
  $D^k = \text{R-SAMPLE}(h^k, h^{k-1}, m_k)$ `See Appendix C for the algorithm`
  $T^k = T^{k-1} + D^k$
end for
$\pi^n(s) \in \arg\max_{a \in A}\ Q^n(s,a) \quad \forall\, s \in S$
Output: π^n

We then derive the sample complexity analysis for our RHI algorithm.

Theorem 4.3 (Performance of RHI)**.** *Consider a robust AMDP defined by contamination or* ℓ_p*-norm, satisfying Assumption 3.1 (or the unichain assumption (Wang et al., 2023g)). Set the step sizes* $c_k = 5(k+2)\ln^2(k+2)$ *and* $\beta_k = k/(k+2)$. *Then, with probability at least* $1-\delta$, *the output policy* π^n *is* ϵ*-optimal:*

$$g^*_{\mathcal{P}} - g^{\pi^n}_{\mathcal{P}}(s) \le \epsilon, \tag{16}$$

as long as the total iteration number n *exceeds* $\frac{\mathcal{H}}{\epsilon}$, *resulting in a total sample complexity of*

$$\tilde{\mathcal{O}}\left(\frac{SA\mathcal{H}^2}{\epsilon^2}\right). \tag{17}$$

Our result is the first finite sample complexity guarantee for robust AMDPs under communicating assumptions, without any prior knowledge requirement. Hence, it underscores the sample efficiency and applicability of our algorithm. Our complexity result represents the state-of-the-art in robust average reward RL (see Section 5.1 for a detailed comparison with prior works).

We note that the minimax optimal sample complexity for *non-robust* AMDPs is $\tilde{\Omega}\left(\frac{SAH}{\epsilon^2}\right)$ (Wang et al., 2022), where H is the non-robust optimal span. Noting that non-robust AMDPs are special cases of robust ones, our sample complexity result matches this minimax optimal complexity in all terms except for $\mathcal{H}$, and is thus near-optimal. We also highlight that, the minimax optimal complexity for non-robust AMDPs is achievable only with prior knowledge of H or other MDP parameters (Zurek & Chen, 2023; Sapronov & Yudin, 2024; Wang et al., 2023b;c); and when there is no such knowledge, non-robust algorithms also are sub-optimal (Jin et al., 2024a; Lee et al., 2025). We leave it as future research to investigate the minimax lower bound for robust AMDPs, if it is achievable without any prior knowledge, and if it can be extended to other uncertainty sets.

Remark 4.4. *Implementing our RHI algorithm does not require any prior knowledge, except that the total iteration number, n, depends on $\mathcal{H}$. Although it is common in sample complexity analysis to have an iteration number that depends on unknown underlying parameters, e.g., (Li et al., 2021a;b; Wang et al., 2024d), its concrete and practical implementations can still be challenging. To address this issue, we further modify Algorithm 1 to employ a doubling trick (Auer et al., 1995; Besson & Kaufmann, 2018; Lee et al., 2025), and propose our Parameter-Free RHI (PF-RHI) algorithm. PF-RHI is completely independent of $\mathcal{H}$, while maintaining the same sample complexity. We defer the discussion to Appendix E,*

5 Related Work

5.1 Comparisons with Prior Results

In this section, we first compare with the most related works on finite sample complexity analysis of robust average-reward RL, including (Grand-Clément & Petrik, 2024; Roch et al., 2025; Xu et al., 2025b;a; Chen et al., 2025). The comparison is summarized in Table 1.

In (Grand-Clément & Petrik, 2024; Roch et al., 2025; Chen et al., 2025), reduction-based methods are developed. In these works, a robust discounted reward RL with some specific discount factor (referred to as a reduction factor) is constructed, and its optimal robust policy is shown to be near-optimal under average reward. Thus, the sample complexity of robust average reward RL is then equivalent to that of the corresponding discounted RL with the reduction factor. In (Grand-Clément & Petrik, 2024), an upper bound on the reduction factor is derived as $\gamma \leq 1 - \frac{C}{S^S m^{S^2}}$, when the nominal kernels are rational, i.e., $\mathsf{P}_s^a = n_{s,a}/m_{s,a}$ with $n_{s,a}, m_{s,a} \in \mathbb{N}$, and m is the smallest denominator among all kernel entries. However, coupling this bound with existing sample-complexity results for robust DMDPs yields exponential sample complexity for robust AMDPs. In (Chen et al., 2025), the reduction factor is set to a sample-number dependent value, and the corresponding sample complexity is derived. However, their results require stronger assumptions on the AMDP structure (uniformly ergodic) and the radius of the uncertainty set (the radius has to be small), limiting the applicability. More recently, a reduction factor $\gamma = 1 - \frac{\epsilon}{\mathcal{H}}$ is developed in (Roch et al., 2025) and sample complexity that matches ours is derived under the unichain setting. However, this reduction factor depends on the robust optimal span $\mathcal{H}$, requiring its knowledge even before learning. In practice, access to such knowledge is infeasible, and even its estimation can be challenging and inefficient (Zurek & Chen, 2023; Tuynman et al., 2024).

Another line of work (Xu et al., 2025b;a) utilizes the truncated multi-level Monte-Carlo method developed in (Wang et al., 2024b) to directly find the optimal policy. However, both works assume the underlying robust AMDP is irreducible, under which the robust Bellman operator becomes a γ-contraction w.r.t. the Span, and the sample complexity can be derived. Their method relies heavily on the contraction (which does not hold in our setting), and so cannot be applied.

Hence, compared to these prior works, our method enjoys three major advantages: (1). We require the *weakest* AMDP structure, communicating–all prior work imposes stronger structures; (2). We do not require *any* prior knowledge of the robust AMDP (like $\mathcal{H}$ in (Roch et al., 2025)); (3). We enjoy the tightest sample complexity (noting that $\mathcal{H} \leq t_m$, i.e., the mixing time (Wang et al., 2022; Roch et al., 2025)). Thus, our RHI method represents the state-of-the-art in robust average reward RL.

Algorithm	AMDP Structure	Uncertainty Set	Sample Complexity
(Grand-Clément & Petrik, 2024)	$\mathsf{P} \in \mathbb{Q}$	N/A	Exponential
Chen et al. (2025)	Uniformly ergodic	KL	$\tilde{\mathcal{O}}\left(\frac{SAt_m^2}{p_\wedge \epsilon^2}\right)$
Xu et al. (2025b)	Irreducible & aperiodic	TV	$\tilde{\mathcal{O}}\left(\frac{SAt_m^2}{(1-\gamma)^2\epsilon^2}\right)$
Xu et al. (2025a)	Irreducible & aperiodic	TV	$\tilde{\mathcal{O}}\left(\frac{SAt_m^2}{(1-\gamma)^2\epsilon^2}\right)$
Roch et al. (2025)	Unichain	TV	$\tilde{\mathcal{O}}\left(\frac{SA\mathcal{H}^2}{\epsilon^2}\right)$
Ours	Communicating/unichain	l_p	$\tilde{\mathcal{O}}\left(\frac{SA\mathcal{H}^2}{\epsilon^2}\right)$

Table 1: Comparison with prior results. t_m denotes the robust mixing time; γ in (Xu et al., 2025a;b) is the contraction coefficient under the irreducibility assumption.

5.2 Other Related work

Robust RL with average reward. Studies on robust RL with average reward are relatively limited. Early research focused on dynamic programming (DP) methods in robust AMDPs. These investigations, initiated by (Tewari & Bartlett, 2007) for specific finite-interval uncertainty sets, were subsequently extended to more general uncertainty models in works such as (Wang et al., 2023f; Grand-Clement et al., 2023; Wang & Si, 2025). These foundational studies were instrumental in revealing the fundamental structure of robust AMDPs and illustrating their connections to robust DMDPs. As an alternative method, (Chatterjee et al., 2023) recently proposed a game-theoretic approach for finding the optimal policy. Building on the understanding of robust AMDP structures, the focus also extends to learning algorithms, where (Wang et al., 2023g) introduced a model-free algorithm with asymptotic convergence guarantees. However, all of these aforementioned approaches focus on asymptotic convergence only, leaving finite-sample complexity analyses largely unaddressed.

Robust RL with discounted rewards. Robust DMDPs were first studied in foundational works such as (Iyengar, 2005; Nilim & El Ghaoui, 2004; Bagnell et al., 2001; Wiesemann et al., 2013; Lim et al., 2013). These initial investigations typically assumed a fully known uncertainty set and developed solutions based on robust DP. Since then, extensive theoretical research has significantly adapted and extended these concepts to various learning paradigms where the uncertainty set or the nominal model might be unknown or learned from data. Prominent research directions include analyses in settings with generative models (Yang et al., 2022; Panaganti & Kalathil, 2022; Xu et al., 2023; Shi et al., 2023; Zhou et al., 2021; Wang et al., 2023d; Liang et al., 2023; Liu et al., 2022b; Wang et al., 2023e; 2024b; 2023a; Kumar et al., 2023; Derman et al., 2021), investigations into offline learning from fixed datasets (Shi & Chi, 2022; Liu & Xu, 2024; Wang et al., 2024a;c), and developments within online learning frameworks involving exploration (Wang & Zou, 2021; Lu et al., 2024; Ghosh et al., 2025; He et al., 2025). A key focus across these diverse settings is often to provide rigorous finite-sample complexity guarantees or convergence rates, characterized under different assumptions regarding the structure of the uncertainty set and the nature of data access.

Non-robust RL with average reward. The study of non-robust AMDPs originated with foundational model-based DP techniques, such as Policy Iteration and Value Iteration, which assume a known model (Puterman, 2014; Bertsekas, 2011). Subsequently, research shifted towards model-free RL algorithms. These include adaptations of Q-learning and SARSA, like RVI Q-learning (Abounadi et al., 2001; Wan et al., 2021; Wan & Sutton, 2022), designed to learn optimal policies directly from interaction data without requiring explicit model knowledge (Dewanto et al., 2020).

Beyond asymptotic convergence, sample complexity for achieving near-optimal policies in (non-robust) AMDPs is extensively studied. A significant body of work is based on the reduction framework, which transforms the AMDP into a DMDP using a carefully chosen discount factor. However, selecting an appropriate discount factor typically requires prior knowledge of crucial MDP parameters, such as the span of the bias function (Zurek & Chen, 2023; Wang et al., 2022; Zurek & Chen, 2024; Sapronov & Yudin, 2024; Jin & Sidford, 2021) or various mixing time constants (Wang et al., 2023b;c). Notable progress has been made under such assumptions, for instance, (Zurek & Chen, 2023; Sapronov & Yudin, 2024) demonstrate that if the bias span is known and used to set the

reduction factor, the resulting sample complexity matches the minimax optimal rate for weakly communicating MDPs (Wang et al., 2022). Alongside reduction-based methods, direct approaches that do not involve conversion to DMDPs, but still require prior knowledge, have also been recently developed (Zhang & Xie, 2023; Li et al., 2024). Recognizing that the prerequisite of prior knowledge can be restrictive and impractical, and that estimating these parameters accurately is challenging (Tuynman et al., 2024), another line of research investigates AMDPs without prior knowledge, achieving sub-optimal sample complexity (Lee et al., 2025; Jin et al., 2024a; Lee & Ryu, 2025; Tuynman et al., 2024).

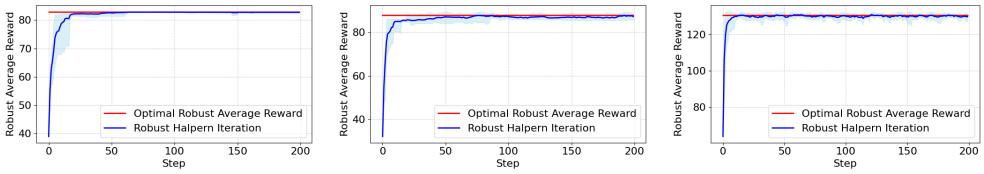
Extending these diverse frameworks and insights to robust AMDPs is, however, particularly challenging. This difficulty stems from the greater complexity inherent in the robust average-reward paradigm, including issues such as the non-linearity of the robust Bellman operator and a more intricate, high-dimensional solution space for the robust Bellman equation (Wang et al., 2023g).

6 Experiment Results

We conduct experiments to validate our theoretical results and evaluate the empirical performance of RHI. We consider the Garnet problem (Archibald et al., 1995) G(20,15) with 20 states and 15 actions, where nominal transition kernels are randomly generated. We consider three uncertainty sets: the contamination model, the ℓ_∞-norm model (total variation), and the ℓ_2-norm model.

After each iteration of our RHI algorithm, we derive the greedy policy based on the current Q-value estimates from RHI. The robust average reward of this derived policy is then calculated using the RRVI algorithm from (Wang et al., 2023g) and recorded. For comparison, we establish a baseline consisting of the optimal robust average reward, also computed via the RRVI algorithm. Each experimental configuration is repeated for 10 independent runs. All of our experiments require minimal compute resources and are implemented using Google Colab. We present the mean robust average reward across these runs where the shaded region in Figure 1 is the standard deviation.

As depicted in Figure 1, the experimental results demonstrate that our RHI algorithm effectively converges to the optimal robust average reward, thereby corroborating our theoretical findings.



(a) Contamination model (b) ℓ_∞-norm model (c) ℓ_2-norm model

Figure 1: Performance of RHI.

7 Conclusion

Robust reinforcement learning under the average-reward criterion suffers from the significant challenge of developing efficient algorithms with finite-sample guarantees, thus hindering its application in data-limited environments. This generally resulted from complexity of the problem setting and the limitations of prior approaches, which often relied on stronger structural assumptions, or required impractical prior knowledge. Therefore, we introduced Robust Halpern Iteration (RHI), a novel model-free algorithm for finding near-optimal policies in robust AMDPs. Key advantages of RHI are its ability to bypass the need for prior knowledge of specific MDP parameters or strong AMDP structures, which are common prerequisites for prior methods. We theoretically established that RHI achieves a sample complexity of $\tilde{\mathcal{O}}\left(\frac{SA\mathcal{H}^2}{\epsilon^2}\right)$ to find an ϵ-optimal policy, under the contamination/unichain conditions and ℓ_p-norm/contamination uncertainty sets. Our result is near-optimal, enhancing the applicability of average-reward robust RL in those data-intensive and real-world applications.

REFERENCES

Jinane Abounadi, Dimitrib Bertsekas, and Vivek S Borkar. Learning algorithms for Markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 40(3):681–698, 2001.

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.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th annual foundations of computer science*, pp. 322–331. IEEE, 1995.

J Andrew Bagnell, Andrew Y Ng, and Jeff G Schneider. Solving uncertain Markov decision processes. *Carnegie Mellon University, Technical Report*, 2001.

Dimitri P Bertsekas. Dynamic Programming and Optimal Control 3rd edition, volume II. *Belmont, MA: Athena Scientific*, 2011.

Lilian Besson and Emilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018.

Arthur Charpentier, Romuald Elie, and Carl Remlinger. Reinforcement learning in economics and finance. *Computational Economics*, pp. 1–38, 2021.

Krishnendu Chatterjee, Ehsan Kafshdar Goharshady, Mehrdad Karrabi, Petr Novotný, and undefinedorđe Žikelić. Solving long-run average reward robust mdps via stochastic games. *arXiv preprint arXiv:2312.13912*, 2023.

Pengcheng Chen, Shichao Liu, Xiaozhe Wang, and Innocent Kamwa. Physics-shielded multi-agent deep reinforcement learning for safe active voltage control with photovoltaic/battery energy storage systems. *IEEE Transactions on Smart Grid*, 14(4):2656–2667, 2022.

Zijun Chen, Shengbo Wang, and Nian Si. Sample complexity of distributionally robust average-reward reinforcement learning. *arXiv preprint arXiv:2505.10007*, 2025.

Juan Pablo Contreras and Roberto Cominetti. Optimal error bounds for non-expansive fixed-point iterations in normed spaces. *Mathematical Programming*, 199(1):343–374, 2023.

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.

Vektor Dewanto, George Dunn, Ali Eshragh, Marcus Gallagher, and Fred Roosta. Average-reward model-free reinforcement learning: a systematic review and literature mapping. *arXiv preprint arXiv:2010.08920*, 2020.

Benjamin Eysenbach, Swapnil Asawa, Shreyas Chaudhari, Sergey Levine, and Ruslan Salakhutdinov. Off-Dynamics Reinforcement Learning: Training for Transfer with Domain Classifiers. *arXiv preprint arXiv:2006.13916*, 2020.

Debamita Ghosh, George K Atia, and Yue Wang. Provably near-optimal distributionally robust reinforcement learning in online settings. *arXiv preprint arXiv:2508.03768*, 2025.

Julien Grand-Clément and Marek Petrik. Reducing blackwell and average optimality to discounted MDPs via the blackwell discount factor. *arXiv preprint arXiv:2302.00036*, 2023.

Julien Grand-Clément and Marek Petrik. Reducing blackwell and average optimality to discounted MDPs via the blackwell discount factor. *Advances in Neural Information Processing Systems*, 36, 2024.

Julien Grand-Clement, Marek Petrik, and Nicolas Vieille. Beyond discounted returns: Robust markov decision processes with average and blackwell optimality. *arXiv preprint arXiv:2312.03618*, 2023.

Benjamin Halpern. Fixed points of nonexpanding maps. *Bulletin of the American Mathematical Society*, 73(6):957–961, 1967. Publisher: American Mathematical Society.

Yiting He, Zhishuai Liu, Weixin Wang, and Pan Xu. Sample complexity of distributionally robust off-dynamics reinforcement learning with online interaction. In *Proc. International Conference on Machine Learning (ICML)*, 2025.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.

Joshua Arvind Holla. *On the Off-Dynamics Approach to Reinforcement Learning*. McGill University (Canada), 2021.

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

Yuchen Jiao and Gen Li. Minimax-Optimal Multi-Agent Robust Reinforcement Learning. *arXiv preprint arXiv:2412.19873*, 2024. doi: 10.48550/arXiv.2412.19873. URL `http://arxiv.org/abs/2412.19873`.

Ying Jin, Ramki Gummadi, Zhengyuan Zhou, and Jose Blanchet. Feasible q-learning for average reward reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1630–1638. PMLR, 2024a.

Yujia Jin and Aaron Sidford. Towards tight bounds on the sample complexity of average-reward mdps. In *International Conference on Machine Learning*, pp. 5055–5064. PMLR, 2021.

Yujia Jin, Ishani Karmarkar, Aaron Sidford, and Jiayi Wang. Truncated variance reduced value iteration. *arXiv preprint arXiv:2405.12952*, 2024b.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

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.

Jongmin Lee, Mario Bravo, and Roberto Cominetti. Near-optimal sample complexity for mdps via anchoring. *arXiv preprint arXiv:2502.04477*, 2025.

Jonmin Lee and Ernest K Ryu. Optimal non-asymptotic rates of value iteration for average-reward markov decision processes. *arXiv preprint arXiv:2504.09913*, 2025.

Gen Li, Changxiao Cai, Yuxin Chen, Yuantao Gu, Yuting Wei, and Yuejie Chi. Is Q-learning minimax optimal? a tight sample complexity analysis. *arXiv preprint arXiv:2102.06548*, 2021a.

Gen Li, Changxiao Cai, Yuxin Chen, Yuantao Gu, Yuting Wei, and Yuejie Chi. Tightening the dependence on horizon in the sample complexity of q-learning. In *International Conference on Machine Learning*, pp. 6296–6306. PMLR, 2021b.

Haifeng Li and Mo Hai. Deep reinforcement learning model for stock portfolio management based on data fusion. *Neural Processing Letters*, 56(2):108, 2024.

Tianjiao Li, Feiyang Wu, and Guanghui Lan. Stochastic first-order methods for average-reward markov decision processes. *Mathematics of Operations Research*, 2024.

Zhipeng Liang, Xiaoteng Ma, Jose Blanchet, Jiheng Zhang, and Zhengyuan Zhou. Single-trajectory distributionally robust reinforcement learning. *arXiv preprint arXiv:2301.11721*, 2023.

Felix Lieder. On the convergence rate of the Halpern-iteration. *Optimization Letters*, 15(2):405–418, 2021.

Shiau Hong Lim, Huan Xu, and Shie Mannor. Reinforcement learning in robust Markov decision processes. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 701–709, 2013.

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.

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

Zijian Liu, Qinxun Bai, Jose Blanchet, Perry Dong, Wei Xu, Zhengqing Zhou, and Zhengyuan Zhou. Distributionally robust Q-learning. In *Proc. International Conference on Machine Learning (ICML)*, 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.

Renzhi Lu, Seung Ho Hong, and Xiongfeng Zhang. A dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach. *Applied energy*, 220:220–230, 2018.

Mohammad Amin Masoudi. *Robust Deep Reinforcement Learning for Portfolio Management*. PhD thesis, Université d'Ottawa/University of Ottawa, 2021.

John Moody and Matthew Saffell. Learning to trade via direct reinforcement. *IEEE transactions on neural Networks*, 12(4):875–889, 2001.

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.

Jisun Park and Ernest K Ryu. Exact optimal accelerated complexity for fixed-point iterations. In *International Conference on Machine Learning*, pp. 17420–17457. PMLR, 2022.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 3803–3810. IEEE, 2018.

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.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Aravind Rajeswaran, Sarvjeet Ghotra, Balaraman Ravindran, and Sergey Levine. Epopt: Learning robust neural network policies using model ensembles. *arXiv preprint arXiv:1610.01283*, 2016.

Zachary Roch, George Atia, and Yue Wang. A reduction framework for distributionally robust reinforcement learning under average reward. In *Proc. International Conference on Machine Learning (ICML)*. PMLR, 2025.

Shoham Sabach and Shimrit Shtern. A first order method for solving convex bilevel optimization problems. *SIAM Journal on Optimization*, 27(2):640–660, 2017.

Yuri Sapronov and Nikita Yudin. Optimal approximation of average reward markov decision processes. In *International Conference on Computational Optimization*, 2024.

Anton Schwartz. A reinforcement learning method for maximizing undiscounted rewards. In *ICML*, volume 93, pp. 298–305, 1993.

Harm Seijen and Rich Sutton. True online td (lambda). In *International Conference on Machine Learning*, pp. 692–700. PMLR, 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, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. *arXiv preprint arXiv:2305.16589*, 2023.

Olivier Sigaud and Olivier Buffet. *Markov decision processes in artificial intelligence*. John Wiley & Sons, 2013.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Ambuj Tewari and Peter L Bartlett. Bounded parameter Markov decision processes with average reward criterion. In *International Conference on Computational Learning Theory*, pp. 263–277. Springer, 2007.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

J. N. Tsitsiklis and B. Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, May 1997.

Adrienne Tuynman, Rémy Degenne, and Emilie Kaufmann. Finding good policies in average-reward markov decision processes without prior knowledge. *arXiv preprint arXiv:2405.17108*, 2024.

Yi Wan and Richard S Sutton. On convergence of average-reward off-policy control algorithms in weakly-communicating MDPs. *arXiv preprint arXiv:2209.15141*, 2022.

Yi Wan, Abhishek Naik, and Richard S Sutton. Learning and planning in average-reward Markov decision processes. In *Proc. International Conference on Machine Learning (ICML)*, pp. 10653–10662. PMLR, 2021.

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

Jinghan Wang, Mengdi Wang, and Lin F Yang. Near sample-optimal reduction-based policy learning for average reward mdp. *arXiv preprint arXiv:2212.00603*, 2022.

Kaixin Wang, Uri Gadot, Navdeep Kumar, Kfir Levy, and Shie Mannor. Bring your own (non-robust) algorithm to solve robust mdps by estimating the worst kernel. *arXiv e-prints*, pp. arXiv–2306, 2023a.

Shengbo Wang and Nian Si. Bellman optimality of average-reward robust markov decision processes with a constant gain. *arXiv preprint arXiv:2509.14203*, 2025.

Shengbo Wang, Jose Blanchet, and Peter Glynn. Optimal sample complexity for average reward Markov decision processes. *arXiv preprint arXiv:2310.08833*, 2023b.

Shengbo Wang, Jose Blanchet, and Peter Glynn. Optimal sample complexity of reinforcement learning for mixing discounted Markov decision processes. *arXiv preprint arXiv:2302.07477*, 2023c.

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, 2023d.

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

Yudan Wang, Shaofeng Zou, and Yue Wang. Model-free robust reinforcement learning with sample complexity analysis. In *Proc. International Conference on Uncertainty in Artificial Intelligence (UAI)*, 2024b.

Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 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, Alvaro Velasquez, George Atia, Ashley Prater-Bennette, and Shaofeng Zou. Robust average-reward Markov decision processes. In *Proc. Conference on Artificial Intelligence (AAAI)*, 2023f.

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, 2023g.

Yue Wang, Zhongchang Sun, and Shaofeng Zou. A unified principle of pessimism for offline reinforcement learning under model mismatch. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024c.

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*, 2024d. 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.

Huayi Wu, Zhao Xu, Minghao Wang, Jian Zhao, and Xu Xu. Two-stage voltage regulation in power distribution system using graph convolutional network-based deep reinforcement learning in real time. *International Journal of Electrical Power & Energy Systems*, 151:109158, 2023.

Yang Xu, Swetha Ganesh, and Vaneet Aggarwal. Efficient Q-learning and actor-critic methods for robust average reward reinforcement learning. *arXiv preprint arXiv:2506.07040*, 2025a.

Yang Xu, Washim Uddin Mondal, and Vaneet Aggarwal. Finite-sample analysis of policy evaluation for robust average reward reinforcement learning. *arXiv preprint arXiv:2502.16816*, 2025b.

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

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

Daochen Zha, Jingru Xie, Wenye Ma, Sheng Zhang, Xiangru Lian, Xia Hu, and Ji Liu. Douzero: Mastering doudizhu with self-play deep reinforcement learning. In *international conference on machine learning*, pp. 12333–12344. PMLR, 2021.

Chi Zhang, Zain Ulabedeen Farhat, George K. Atia, and Yue Wang. Model-free offline reinforcement learning with enhanced robustness. In *Proc. International Conference on Learning Representations (ICLR)*, 2025.

Chiyuan Zhang, Oriol Vinyals, Rémi Munos, and Samy Bengio. A study on overfitting in deep reinforcement learning. *arXiv preprint arXiv:1804.06893*, 2018.

Sheng Zhang, Zhe Zhang, and Siva Theja Maguluri. Finite sample analysis of average-reward TD learning and Q-learning. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 1230–1242, 2021.

Zihan Zhang and Qiaomin Xie. Sharper model-free reinforcement learning for average-reward Markov decision processes. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 5476–5477. PMLR, 2023.

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.

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.

Matthew Zurek and Yudong Chen. Span-based optimal sample complexity for average reward mdps. *arXiv preprint arXiv:2311.13469*, 2023.

Matthew Zurek and Yudong Chen. The plug-in approach for average-reward and discounted mdps: Optimal sample complexity analysis. *arXiv preprint arXiv:2410.07616*, 2024.

A Preliminaries and Proof Organization

To facilitate the analysis and understanding of our work, we specify the notations as follows.

System Characteristics. We consider a robust AMDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$ centered around the nominal kernel P with the following properties:

- The uncertainty set $\mathcal{P} = \bigotimes_{s\in\mathcal{S}, a\in\mathcal{A}} \mathcal{P}_s^a$ is SA-rectangular (Nilim & El Ghaoui, 2004; Iyengar, 2005; Wiesemann et al., 2013), where $\mathcal{P}_s^a \subseteq \Delta(\mathcal{S})$ is defined independently $\forall (s,a) \in \mathcal{S}\times\mathcal{A}$.
- Each $\mathcal{P}_s^a$ is simultaneously compact and convex.
- The robust system is communicating, meaning that for any arbitrary transition kernel $\mathsf{P} \in \mathcal{P}$ and $s_1, s_2 \in \mathcal{S}$ s.t. $s_1 \neq s_2$, there exists some stationary policy π and integer N s.t. $\mathsf{P}^\pi(S_N = s_2 | S_0 = s_1) > 0$.
- The learner in the robust system has access to a generative model or simulator (Panaganti & Kalathil, 2022; Shi et al., 2023; Xu et al., 2023) to generate i.i.d. samples for any state-action pair under the nominal kernel P.

Additional notation.

- We define a stationary policy as $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, and subsequently define the finite set of all stationary policies as Π such that $\pi \in \Pi$.
- Since the worst-case robust average reward under the time varying model is equivalent to the one under the stationary model (Wang et al., 2023f), we therefore focus on this time invariant model. For a given stationary policy, $\pi \in \Pi$, satisfying equation 8, we define the set of minimizing (worst-case) transition kernels as $\Omega_g^\pi \triangleq \{\mathsf{P} \in \mathcal{P} : g_{\mathsf{P}}^\pi = g_{\mathcal{P}}^\pi\}$, where $g_{\mathsf{P}}^\pi(s) \triangleq \liminf_{T\to\infty} \mathbb{E}_{\pi,\mathsf{P}}\big[\frac{1}{T}\sum_{n=0}^{T-1} r_t | S_0 = s\big]$.
- We use r to denote the SA-dimensional vector, whose (s,a)-th entry is $r(s,a)$. We use $\mathsf{P}_{s,s'}^a$ to denote the transition probability from s to s' under the action a of some transition kernel P.
- Given a policy π, a reward r and a transition kernel P, we denote the induced reward and state-transition kernel by $r_\pi \in \mathbb{R}^S$ and $\mathsf{P}_\pi \in \mathbb{R}^{S\times S}$:

$$r^\pi(s) = \sum_a \pi(a|s) r(s,a), (\mathsf{P}^\pi)_{s,s'} = \sum_a \pi(a|s) \mathsf{P}_{s,s'}^a. \tag{18}$$

- For a vector $V \in \mathbb{R}^S$, we use $\mathsf{P}V$ to denote an SA-dimensional vector as

$$(\mathsf{P}V)_{s,a} = \mathsf{P}_s^a V. \tag{19}$$

  Specifically, for $Q \in \mathbb{R}^{SA}$, $Q_{\max} \in \mathbb{R}^S$, and

$$(\mathsf{P}(Q_{\max}))_{s,a} = \mathsf{P}_s^a(Q_{\max}) = \sum_{s'} \mathsf{P}_{s,s'}^a \max_b \{Q(s',b)\}. \tag{20}$$

- For an uncertainty set $\mathcal{P}$, we denote the robust Bellman operator $\mathcal{T}_{\mathcal{P}}(Q) : \mathbb{R}^{SA} \to \mathbb{R}^{SA}$ as

$$\mathcal{T}_{\mathcal{P}}(Q)(s,a) = r(s,a) + \sigma_{\mathcal{P}_s^a}(Q_{\max}). \tag{21}$$

B Proof of Theorem 3.2

Theorem B.1. *(Restatement of Theorem 3.2) Consider a robust AMDP satisfying Assumption 3.1. Then it holds that:*

(1). The optimal robust average reward $g_{\mathcal{P}}^*$ is a constant, i.e., $g_{\mathcal{P}}^*(s_1) = g_{\mathcal{P}}^*(s_2), \forall s_1 \neq s_2$;

(2). The robust Bellman equation in 11 has a solution (Q^*, g^*), and the solution g^* is the optimal robust average reward, i.e., $g^* = g_{\mathcal{P}}^*(s)$;

(3). The greedy policy π w.r.t. Q, i.e., $\pi(s) \in \arg\max_a Q(s,a)$, is an optimal robust policy.

Proof. **Proof of (1).** Note that Assumption 3.1 implies that for any $\mathsf{P} \in \mathcal{P}$, the non-robust MDP $(\mathcal{S}, \mathcal{A}, \mathsf{P}, r)$ is weakly accessibility (Zurek & Chen, 2023), thus the optimal average reward g^*_{P} is a constant (Bertsekas, 2011).

We then apply Theorem 3.5 from (Grand-Clement et al., 2023), which shows that the optimal robust average gain, $g^*_{\mathcal{P}}$, is the value of the zero-sum stochastic game between the agent and the environment, and the following saddle-point equilibrium exists:

$$g^*_{\mathcal{P}} = \sup_{\pi} \inf_{\mathsf{P} \in \mathcal{P}} g^{\pi}_{\mathsf{P}} = \inf_{\mathsf{P} \in \mathcal{P}} \sup_{\pi} g^{\pi}_{\mathsf{P}}. \tag{22}$$

Since for a fixed P, $\sup_\pi g^\pi_{\mathsf{P}} = g^*_{\mathsf{P}}$ is a constant, thus the RHS of equation 22 is also a constant, as the infimum over a set of scalar constants is itself a scalar constant. This hence proves that the optimal robust gain $g^*_{\mathcal{P}}$ is a constant, independent of the initial state.

Proof of (2). As $g^*_{\mathcal{P}}$ is a constant under our setting, it satisfies the initial-state-independent condition in (Wang & Si, 2025), thus part (2) can be directly obtained by applying the results in (Wang & Si, 2025).

Proof of (3). Since (Q^*, g^*) is a solution to the robust Bellman equation equation 11, the pair (h^*, g^*), where $h^*(s) \triangleq \max_a Q^*(s,a)$ satisfies the following equation:

$$h^*(s) + g^* = \sum_a \pi(a|s)(r(s,a) + \sigma_{\mathcal{P}^a_s}(h^*)).$$

Let $\mathsf{P}_\pi \in \mathcal{P}$ be the worst-case transition kernel for policy π. Then it holds that

$$h^*(s) + g^* = \sum_a \pi(a|s)(r(s,a) + \sigma_{\mathcal{P}^a_s}(h^*)) = \sum_a \pi(a|s)(r(s,a) + (\mathsf{P}_\pi)^a_s(h^*)),$$

i.e.,

$$h^* = r_\pi - g^* + (\mathsf{P}_\pi)^\pi h^*. \tag{23}$$

Multiplying this inequality by $((\mathsf{P}_\pi)^\pi)^k$ and taking a sum further implies that

$$g^* = \frac{\sum_{k=0}^{n-1}((\mathsf{P}_\pi)^\pi)^k r_\pi}{n} + \frac{(((\mathsf{P}_\pi)^\pi)^n - I)h^*}{n}. \tag{24}$$

We then take $\liminf$ on both sides, and it implies that

$$g^* = \liminf_{n\to\infty} \frac{\sum_{k=0}^{n-1}((\mathsf{P}_\pi)^\pi)^k r_\pi}{n} = g^\pi_{\mathcal{P}}, \tag{25}$$

since $(\mathsf{P}^k - I)h^*$ is bounded and finite. Since (2) implies that $g^* = g^*_{\mathcal{P}}$, thus $g^\pi_{\mathcal{P}} = g^*_{\mathcal{P}}$, and the greedy policy π is optimal.

We hence complete the proof. □

C SAMPLING ALGORITHM

In this section, we present a method to approximate the robust Bellman operator $T^k \approx \mathcal{T}_{\mathcal{P}}(Q^k)$ by sampling from the nominal kernel P. Our method is based on the concrete closed-form of the support function $\sigma_{\mathcal{P}}(\cdot)$ over the two considered uncertainty sets.

ℓ_p-norm sets. When the uncertainty set is defined through the ℓ_p-norm as in equation 5, it is shown that the robust Bellman operator has the following closed-form solution in (Kumar et al., 2023):

$$\mathcal{T}_{\mathcal{P}}(Q^k) = r + \mathsf{P}(Q^k_{\max}) - R\kappa(Q^k_{\max}), \tag{26}$$

with some penalty function κ that is independent from P. We defer the constructions of κ to Remark D.2.

Contamination set. With contamination set in equation 4, it holds that (Wang & Zou, 2021):

$$\mathcal{T}_{\mathcal{P}}(Q^k) = r + (1-R)\mathsf{P}(Q^k_{\max}) + R\min_s(Q^k_{\max}). \tag{27}$$

Note that for both uncertainty sets, the difference $\mathcal{T}_{\mathcal{P}}(Q_1) - \mathcal{T}_{\mathcal{P}}(Q_2)$ further can be derived, which facilitates our estimation. To re-use the pre-collected samples to enhance sample efficiency, we further develop our difference-based algorithm.

Specifically, for the ℓ_p-norm case, let $h^k = Q^k_{\max}$ and $h^{k-1} = Q^{k-1}_{\max}$, and we set the difference terms $d^k = h^k - h^{k-1}$, and $K^k = \kappa(h^k) - \kappa(h^{k-1})$. Then it holds that

$$\mathcal{T}_{\mathcal{P}}(Q^k) - \mathcal{T}_{\mathcal{P}}(Q^{k-1}) = \mathsf{P}d^k + K^k. \tag{28}$$

Hence it suffices to estimate $\mathsf{P}d^k$ in our algorithm. We present our robust sampling algorithm (R-SAMPLE) as follows.

Algorithm 2 R-SAMPLE(h^k, h^{k-1}, m)

Input: $h^k,\ h^{k-1},\ m$
for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
  if ℓ_p-norm uncertainty set **then**
    Compute $d^k = h^k - h^{k-1}$ and $K^k = \kappa(h^k) - \kappa(h^{k-1})$
    $D^k(s,a) = \frac{1}{m}\sum_{j=1}^m d^k(s_j) - RK^k(s,a)$ with $s_j \overset{\text{iid}}{\sim} \mathsf{P}^a_s$
  end if
  if Contamination uncertainty set **then**
    Compute $d^k = h^k - h^{k-1}$ and $K^k = \min_s(h^k) - \min_s(h^{k-1})$
    $D^k(s,a) = \frac{1-R}{m}\sum_{j=1}^m d^k(s_j) + RK^k(s,a)$ with $s_j \overset{\text{iid}}{\sim} \mathsf{P}^a_s$
  end if
end for
Output: D^k

D Proofs for RHI

D.1 Analysis of RHI

Lemma D.1 (Restatement of Lemma 4.1)**.** *Under Assumption 3.1, let $Q \in \mathbb{R}^{SA}$ and π be the greedy policy w.r.t. Q, i.e., $\pi(s) \in \arg\max_{a\in\mathcal{A}} Q(s,a)$. Then for every state $s \in \mathcal{S}$, it holds that:*

$$0 \le g^*_{\mathcal{P}} - g^{\pi}_{\mathcal{P}}(s) \le \boldsymbol{Sp}(\mathcal{T}_{\mathcal{P},g^*_{\mathcal{P}}}(Q) - Q) = \boldsymbol{Sp}(\mathcal{T}_{\mathcal{P}}(Q) - Q).$$

Proof. Denote $h(s) \triangleq Q_{\max}(s) = \max_a Q(s,a)$. Since π is greedy w.r.t Q, it follows that $h(s) = Q(s,\pi(s))$ for all $s \in \mathcal{S}$. We first denote the worst-case transition kernel of h over $\mathcal{P}$ by P, and its induced kernel by P_π, i.e.,

$$(\mathsf{P}_\pi h)(s) = \min_{\mathsf{P}\in\mathcal{P}^{\pi(s)}_s} \mathbb{E}_{s'\sim\mathsf{P}}[h(s')] = \sigma_{\mathcal{P}^{\pi(s)}_s}(h), \quad \forall s \in \mathcal{S}. \tag{29}$$

The robust average reward under Assumption 3.1, $g^{\pi}_{\mathcal{P}}$, exists and is the average reward under the worst-case kernel P_π, thus it holds that

$$g^{\pi}_{\mathcal{P}} = g^{\pi}_{\mathsf{P}_\pi} = \mathsf{P}^{\infty}_{\pi} r_\pi, \tag{30}$$

where $r_\pi = r(s,\pi(s))$ and $\mathsf{P}^{\infty}_{\pi}$ is the Cesaro limit of P_π (Puterman, 2014). Note that it holds that $\mathsf{P}^{\infty}_{\pi} = \mathsf{P}^{\infty}_{\pi}\mathsf{P}_\pi$ (Puterman, 2014), thus applying this fact to equation 30 yields that

$$g^{\pi}_{\mathcal{P}} = \mathsf{P}^{\infty}_{\pi}(r_\pi + \mathsf{P}_\pi h - h). \tag{31}$$

We further note that the $(s',\pi(s'))$-th entry of $(\mathcal{T}_{\mathcal{P}}(Q) - Q)$ is in fact $(r_\pi(s') + (\mathsf{P}_\pi h)(s') - h(s'))$, thus it holds that

$$\min_{s'\in\mathcal{S},a'\in\mathcal{A}} (\mathcal{T}_{\mathcal{P}}(Q) - Q)(s',a') \le (\mathcal{T}_{\mathcal{P}}(Q) - Q)(s',\pi(s)) = (r_\pi(s) + (\mathsf{P}_\pi h)(s') - h(s')),$$

and by multiplying by P_π^∞ recursively and equation 31 we have that

$$\min_{s'\in\mathcal{S},a'\in\mathcal{A}}(\mathcal{T}_{\mathcal{P}_\pi}(Q)-Q)(s',a') \le \mathsf{P}_\pi^\infty\,(r_\pi+(\mathsf{P}_\pi h)-h) = g_{\mathcal{P}}^\pi. \tag{32}$$

On the other hand, denote the optimal robust policy as π^* and its associated optimal average reward as $g_{\mathcal{P}}^*$. Let $\mathsf{P}_{\pi^*}\in\mathcal{P}$ be the corresponding worst-case transition kernel. Similar to equations 30-31, we have that,

$$g_{\mathcal{P}}^* = g_{\mathsf{P}_{\pi^*}}^{\pi^*} = \mathsf{P}_{\pi^*}^\infty r_{\pi^*} = \mathsf{P}_{\pi^*}^\infty(r_{\pi^*}+\mathsf{P}_{\pi^*}h-h), \tag{33}$$

We introduce an auxiliary function $h'\in\mathbb{R}^S$ as $h'(s')\triangleq Q(s',\pi^*(s'))$ for all $s'\in\mathcal{S}$. By definition of h and h', we have $h'(s')\le h(s')$ which implies that $-h(s')\le -h'(s')$. Substituting this in equation 33 implies that for all $s\in\mathcal{S}$,

$$\begin{aligned} g_{\mathcal{P}}^*(s) &= \mathsf{P}_{\pi^*}^\infty(r_{\pi^*}(s')+(\mathsf{P}_{\pi^*}h)(s')-h(s')) \\ &\le \mathsf{P}_{\pi^*}^\infty(r_{\pi^*}(s')+(\mathsf{P}_{\pi^*}h)(s')-h'(s')). \end{aligned} \tag{34}$$

Now we note that $(r_{\pi^*}(s')+(\mathsf{P}_{\pi^*}h)(s')-h'(s'))$ is exactly the $(s',\pi^*(s'))$-th entry of $(\mathcal{T}_{\mathcal{P}}(Q)-Q)$, then it holds that

$$\begin{aligned} g_{\mathcal{P}}^*(s) &= \mathsf{P}_{\pi^*}^\infty(r_{\pi^*}(s')+(\mathsf{P}_{\pi^*}h)(s')-h(s')) \\ &\le \mathsf{P}_{\pi^*}^\infty(r_{\pi^*}(s')+(\mathsf{P}_{\pi^*}h)(s')-h'(s')) \\ &\le \mathsf{P}_{\pi^*}^\infty\cdot\max_{s'\in\mathcal{S},a'\in\mathcal{A}}(r(s',a')+(\mathsf{P}_{\pi^*}h)(s')-h'(s')) \\ &= \max_{s',a'}(\mathcal{T}_{\mathcal{P}}(Q)-Q)(s',a') \end{aligned} \tag{35}$$

Thus together with equation 32, it implies that

$$g_{\mathcal{P}}^*-g_{\mathcal{P}}^\pi(s)\le\max_{s',a'}(\mathcal{T}_{\mathcal{P}}(Q)-Q)(s',a')-\min_{s',a'}(\mathcal{T}_{\mathcal{P}}(Q)-Q)(s',a') = \mathbf{Sp}(\mathcal{T}_{\mathcal{P}}(Q)-Q). \tag{36}$$

It hence completes the proof by noting that $\mathbf{Sp}(\mathcal{T}_{\mathcal{P},g_{\mathcal{P}}^*}(Q)-Q)=\mathbf{Sp}(\mathcal{T}_{\mathcal{P}}(Q)-Q)$ since $g_{\mathcal{P}}^*$ is a constant per Theorem 3.2. □

Remark D.2. *Let $F_{s,a}\subset\mathcal{S}$ be a subset of forbidden states, namely when the system is at state $s\in\mathcal{S}$ and taking action $a\in\mathcal{A}$, it is unfeasible for the system to transition to certain other states. Formally, by denoting the nominal kernel as $\tilde{\mathsf{P}}$ we have*

$$\tilde{\mathsf{P}}(s'|s,a)=\mathsf{P}(s'|s,a)=0,\quad\forall\mathsf{P}\in\mathcal{P},\forall s'\in F_{s,a}.$$

We can then rewrite our kernel noise in equation 3 as

$$\mathcal{P}_s^a=\left\{\mathsf{P}\mid||\mathsf{P}||_p=R,\sum_{s'}\mathsf{P}(s')=0,\mathsf{P}(s'')=0,\forall s''\in F_{s,a}\right\}$$

Under consideration of the ℓ_p-norm model in equation 5 it can be shown

$$\begin{aligned} \kappa(h,s,a) &= \min_{||\mathsf{P}||_p=R,\sum_{s'}\mathsf{P}(s')=0,\mathsf{P}(s'')=0,\forall s''\in F_{s,a}}\langle\mathsf{P},h\rangle \\ &= \min_{\omega\in\mathbb{R}}||u-\omega\mathbf{1}||_p,\quad \textit{where } u(s)=h(s)\mathbf{1}(s\notin F_{s,a}), \\ &= \kappa_p(u). \end{aligned}$$

For a concrete example within the context of our empirical results for the ℓ_∞ (total variation) model in Figure 1b, we have

$$\kappa_\infty(h,s,a)=\frac{\max_{s\notin F_{s,a}}h(s)-\min_{s\notin F_{s,a}}h(s)}{2}.$$

This construction of κ is what allows us to directly apply Theorem 8 from (Kumar et al., 2023) by considering the ℓ_p-ball of transition kernels $||\tilde{\mathsf{P}}-\mathsf{P}||_p\le R$ for all $\mathsf{P}\in\mathcal{P}$ and the nominal kernel $\tilde{\mathsf{P}}$ as turning into a penalty on the next state's value function. Adding this penalty during sampling in Algorithm 2 allows us to effectively sample from the worst-case kernel with only access to the nominal environment.

We then present the proofs for our Robust Halpern Iteration for ℓ_p-normed Robust AMDPs. The proof for contamination models can be derived similarly and is hence omitted.

Proposition D.3. *Let $c_k > 0$ with $2\sum_{k=0}^{\infty} c_k^{-1} \leq 1$ and T^k, Q^k the iterates generated by RHI$(Q^0, n, \epsilon, \delta)$. Then, with probability at least $1-\delta$ we have that $\|T^k - \mathcal{T}_{\mathcal{P}}(Q^k)\|_\infty \leq \epsilon$ simultaneously for all $k = 0, 1, \ldots, n$.*

Proof. We fix an (s,a)-pair in our analysis, denote $Y^i \triangleq D^i - \mathsf{P}d^i - K^i$ and $X^k \triangleq \sum_{i=0}^{k} Y^i$. Recall that $d^i = h^i - h^{i-1}$, then it holds that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and any i,

$$\sigma_{\mathcal{P}_s^a}(h^i) - \sigma_{\mathcal{P}_s^a}(h^{i-1}) = \mathsf{P}d^i - R\kappa(h^i) + R\kappa(h^{i-1}), \tag{37}$$

where the $R\kappa(\cdot)$ is the penalty term from (Kumar et al., 2023), which we discuss further in Remark D.2.

Since $h^{-1} = 0$ by the initialization of RHI, we have the robust Bellman operator as

$$\mathcal{T}_{\mathcal{P}}(Q^k)(s,a) = r(s,a) + \sigma_{\mathcal{P}_s^a}(h^k) = r(s,a) - R\kappa(h^k, s, a) + \sum_{s' \in \mathcal{S}} \mathsf{P}(s'|s,a)h(s'). \tag{38}$$

We further denote that $K^i \triangleq R\kappa(h^i) - R\kappa(h^{i-1})$, and from equation 28 we have that

$$\mathcal{T}_{\mathcal{P}}(Q^k) = r + \sum_{i=0}^{k} K^i. \tag{39}$$

We then consider the estimation error. Recall that $T^k(s,a) = T^{k-1}(s,a) + D^k(s,a)$, thus

$$\begin{aligned}
&T^k(s,a) - \mathcal{T}_{\mathcal{P}_s^a}(Q^k)(s,a) \\
&= T^{k-1}(s,a) - \mathcal{T}_{\mathcal{P}_s^a}(Q^{k-1})(s,a) + D^k(s,a) - \mathsf{P}d^k - K^k \\
&= T^{k-1}(s,a) - \mathcal{T}_{\mathcal{P}_s^a}(Q^{k-1})(s,a) + Y^k(s,a) \\
&= X^k(s,a),
\end{aligned} \tag{40}$$

due to our initialization.

We then estimate $\mathbb{P}(\|X^k(s,a)\|_\infty \geq \epsilon)$, $\forall (s,a)$ by adapting the arguments of the Azuma-Hoeffding inequality as in (Lee et al., 2025). We consider the filtration $\mathcal{F}^k = \sigma(\{D^i\}_{i=0}^k)$. Since h^k, d^k, and m_k are $\mathcal{F}_{k-1}$-measurable and the relation between the robust and non-robust Bellman operators (Kumar et al., 2023), it follows that $\mathbb{E}[Y^k(s,a)|\mathcal{F}_{k-1}] = 0$ for all (s,a) during sampling. Thus the sequence $\{X^k(s,a)\}_{k\geq 0}$ is a $\mathcal{F}^k$-martingale. Using Markov's inequality and the tower property of conditional expectation yields that for every $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $\lambda > 0$,

$$\begin{aligned}
\mathbb{P}(X^k(s,a) \geq \epsilon) &\leq e^{-\lambda\epsilon}\mathbb{E}[\exp(\lambda X^k(s,a))] \\
&= e^{-\lambda\epsilon}\mathbb{E}\big[\exp(\lambda X^{k-1}(s,a))\mathbb{E}[\exp(\lambda Y^k(s,a))|\mathcal{F}^{k-1}]\big].
\end{aligned} \tag{41}$$

Moreover, since K^i is deterministic and independent from P, it holds that $Y^k = \frac{1}{m_k}\left(\sum_j^{m_k} d^k(s_{k,j}^{s,a})\right) - \mathsf{P}d^k(s,a)$. Now since $d^k(s_{k,j}^{s,a}) \in [\min_{s'} d^k(s'),\ \max_{s'} d^k(s')]$ and $\mathbb{E}[Y^k(s,a)|\mathcal{F}_{k-1}] = 0$, Hoeffding's inequality yields that

$$\begin{aligned}
\mathbb{E}[\exp(\lambda Y^k(s,a))|\mathcal{F}_{k-1}] &= \prod_{j=1}^{m_k} \mathbb{E}\left[\exp(\lambda Y_j^k)|\mathcal{F}^{k-1}\right] \\
&\leq \exp\left(\frac{1}{2}\lambda^2 \mathbf{Sp}(d^k)^2/m_k\right),
\end{aligned} \tag{42}$$

where $Y_j^k = \frac{1}{m_k}\left(d^k(s_{k,j}^{s,a})\right) - \frac{1}{m_k}\mathsf{P}d^k(s,a)$, and the last inequality is due to the fact that $|Y_j^k| \leq \frac{\mathbf{Sp}(d^k)}{m_k}$.

Combining equation 41 and equation 42 along with $m_k \geq \alpha c_k \mathbf{Sp}(d^k)^2/\epsilon^2$, it can be derived that

$$\mathbb{E}[\exp(\lambda X^k(s,a))] \leq \exp\left(\frac{1}{2}\lambda^2\epsilon^2 \sum_{i=0}^{k} c_i^{-1}/\alpha\right). \tag{43}$$

Combining this with $\sum_{i=0}^{\infty} c_i^{-1} \leq \frac{1}{2}$, we have $\mathbb{P}(X^k(s,a) \geq \epsilon) \leq \exp(-\lambda\epsilon + \frac{1}{4}\lambda^2\epsilon^2/\alpha)$. Taking $\lambda = 2\alpha/\epsilon$ we can obtain

$$\mathbb{P}(X^k(s,a) \geq \epsilon) \leq \exp(-\alpha) = \frac{\delta}{2|\mathcal{S}||\mathcal{A}|(n+1)}.$$

Synonymously, we can find the same bound for $\mathbb{P}(X^k(s,a) \leq -\epsilon)$ s.t. $\mathbb{P}(|X^k(s,a)| \geq \epsilon) \leq \delta/(|\mathcal{S}||\mathcal{A}|(n+1))$. The proof is hence completed by taking the union bound over all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and over all iterations k. □

We then derive our analysis under the event specified, i.e.,

$$\|T^k(s,a) - \mathcal{T}_{\mathcal{P}_s^a}(Q^k)\|_\infty \leq \epsilon \quad \forall k = 0, 1, \ldots, n, \text{ and } (s,a) \in \mathcal{S} \times \mathcal{A}, \tag{44}$$

which holds with probability $(1-\delta)$ by Proposition D.3.

Moreover, we note that from our R-SAMPLE algorithm, it holds that $\mathbf{Sp}(D^k) \leq \mathbf{Sp}(d^k)$. Combining this fact with the nonexpansivity of the $\max$ operator implies that

$$\mathbf{Sp}(T^k - T^{k-1}) = \mathbf{Sp}(D^k) \leq \mathbf{Sp}(d^k) = \mathbf{Sp}(h^k - h^{k-1}) \leq \mathbf{Sp}(Q^k - Q^{k-1}). \tag{45}$$

We first provide two lemmas.

Lemma D.4. *Let Q^* be a solution to the robust Bellman equation $Q^* = \mathcal{T}_{\mathcal{P}}(Q^*)$. Under the event in equation 44, it holds that*

$$\boldsymbol{Sp}(Q^k - Q^*) \leq \boldsymbol{Sp}(Q^0 - Q^*) + \frac{2}{3}\epsilon k, \quad \forall\, k = 0, 1, \ldots, n. \tag{46}$$

Proof. By the update rule of RHI, at iteration k, it holds that $Q^k = (1-\beta_k)Q^0 + \beta_k T^{k-1}$ with $\beta_k = \frac{k}{k+2}$. We thus have that

$$\begin{aligned}\mathbf{Sp}(Q^k - Q^*) &\leq (1-\beta_k)\mathbf{Sp}(Q^0 - Q^*) + \beta_k \mathbf{Sp}(T^{k-1} - Q^*) \\ &= \frac{2}{k+2}\mathbf{Sp}(Q^0 - Q^*) + \frac{k}{k+2}\mathbf{Sp}(T^{k-1} - Q^*).\end{aligned} \tag{47}$$

Using the invariance of $\mathbf{Sp}(\cdot)$ by the addition of constants and the nonexpansivity of $\mathcal{T}_{\mathcal{P}}$, we can then apply the triangle inequality along with the fact that $\mathbf{Sp}(\cdot) \leq 2\|\cdot\|_\infty$ and the bound in equation 44 to obtain,

$$\mathbf{Sp}(T^{k-1} - Q^*) = \mathbf{Sp}\big(T^{k-1} - \mathcal{T}_{\mathcal{P}}(Q^*)\big) \leq 2\epsilon + \mathbf{Sp}(Q^{k-1} - Q^*). \tag{48}$$

We can then plug this back into equation 47 to get

$$\mathbf{Sp}(Q^k - Q^*) \leq \frac{2}{k+2}\mathbf{Sp}(Q^0 - Q^*) + \frac{k}{k+2}\big(2\epsilon + \mathbf{Sp}(Q^{k-1} - Q^*)\big).$$

Set $\theta_k = (k+1)(k+2)\mathbf{Sp}(Q^k - Q^*)$, then we have $\theta_k \leq \theta_0(k+1) + 2\epsilon k(k+1) + \theta_{k-1}$. Through induction we can get that,

$$\begin{aligned}\theta_k &\leq \theta_0 \sum_{i=1}^{k}(i+1) + 2\epsilon \sum_{i=1}^{k} i(i+1) + \theta_0 \\ &= \theta_0 \frac{1}{2}(k+1)(k+2) + \frac{2}{3}\epsilon k(k+1)(k+2).\end{aligned}$$

Dividing both sides by $(k+1)(k+2)$ hence completes the proof. □

Lemma D.5. *Under the event in equation 44. We denote $\rho_k \triangleq 2\boldsymbol{Sp}(Q^0 - Q^*) + \frac{2}{3}\epsilon k$, then for all $k = 1, 2, \ldots, n$, we have*

$$\boldsymbol{Sp}(Q^k - Q^{k-1}) \leq \frac{2}{k(k+1)} \sum_{i=1}^{k} \rho_{i+2}.$$

Proof. We have shown two equations:

$$Q^k = \frac{2}{k+2}Q^0 + \frac{k}{k+2}T^{k-1}, \quad Q^{k-1} = \frac{2}{k+1}Q^0 + \frac{k-1}{k+1}T^{k-2}. \tag{49}$$

We then subtract them and have that

$$Q^k - Q^{k-1} = \frac{2}{(k+1)(k+2)}\big(T^{k-1} - Q^0\big) + \frac{k-1}{k+1}\big(T^{k-1} - T^{k-2}\big). \tag{50}$$

By $\mathbf{Sp}(Q^k - Q^*) \leq \mathbf{Sp}(Q^0 - Q^*) + \frac{2}{3}\epsilon k$ (from Lemma D.4) and equation 48, we then have that

$$\begin{aligned}\mathbf{Sp}(T^{k-1} - Q^0) &\leq \mathbf{Sp}(T^{k-1} - Q^*) + \mathbf{Sp}(Q^* - Q^0) \\ &\leq \rho_{k+2}.\end{aligned}$$

Substituting this into equation 50 and using equation 45 yields

$$\mathbf{Sp}(Q^k - Q^{k-1}) \leq \frac{2}{(k+1)(k+2)}\rho_{k+2} + \frac{k-1}{k+1}\mathbf{Sp}(Q^{k-1} - Q^{k-2}).$$

We further set $\tilde{\theta}_k = k(k+1)\mathbf{Sp}(Q^k - Q^{k-1})$, and it holds that

$$\begin{aligned}\tilde{\theta}_k &\leq \frac{2k}{k+2}\rho_{k+2} + k(k-1)\mathbf{Sp}(Q^{k-1} - Q^{k-2}) \\ &\leq \frac{2k}{k+2}\rho_{k+2} + \tilde{\theta}_{k-1} \\ &\leq 2\rho_{k+2} + \tilde{\theta}_{k-1} \\ &\leq 2\sum_{i=1}^{k}\rho_{i+2}.\end{aligned}$$

Dividing both sides by $k(k+1)$ implies that

$$\mathbf{Sp}(Q^k - Q^{k-1}) \leq \frac{2}{k(k+1)}\sum_{i=1}^{k}\rho_{i+2}, \tag{51}$$

which completes the proof. □

Theorem D.6 (Restatement of Theorem 4.2)**.** *Consider the exact robust Halpern iteration* $[Q^{k+1}] = [(1-\beta_{k+1})Q^0 + \beta_{k+1}\mathcal{T}_{\mathcal{P}}(Q)]$*, with* $\beta_k = \frac{k}{k+2}$*. Set* π^k *to be the greedy policy w.r.t.* Q^k*. Then,*

$$\boldsymbol{Sp}(\mathcal{T}_{\mathcal{P}}(Q^k) - Q^k) \to 0, \text{ and } g^*_{\mathcal{P}} - g^{\pi^k} \to 0, \text{ as } k \to \infty. \tag{52}$$

Proof. By Lemma D.1, we have that $g^*_{\mathcal{P}} - g^{\pi^k}_{\mathcal{P}} \leq \mathbf{Sp}(\mathcal{T}_{\mathcal{P}}(Q^k) - Q^k)$, thus it suffices to show that $\mathbf{Sp}(\mathcal{T}_{\mathcal{P}}(Q^k) - Q^k) \to 0$.

We derive our analysis under the event in Proposition D.3, that with probability at least $(1-\delta)$, we have that $\|T^k - \mathcal{T}_{\mathcal{P}}(Q^k)\|_\infty \leq \epsilon$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and for all $k = 0, 1, \ldots, n$.

For ease of reading, we drop the brackets from the equivalence class notations. Our RHI updates as $Q^k = (1-\beta_k)Q^0 + \beta_k T^{k-1} = \frac{2}{k+2}Q^0 + \frac{k}{k+2}T^{k-1}$ in the quotient space, which implies that for each $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have the following decomposition

$$\begin{aligned}&\mathcal{T}_{\mathcal{P}}(Q^k) - Q^k \\ &= \frac{2}{k+2}\underbrace{\Big(\mathcal{T}_{\mathcal{P}}(Q^k) - Q^0\Big)}_{\textbf{Term 1}} + \frac{k}{k+2}\underbrace{\Big(\mathcal{T}_{\mathcal{P}}(Q^k) - \mathcal{T}_{\mathcal{P}}(Q^{k-1})\Big)}_{\textbf{Term 2}} + \frac{k}{k+2}\underbrace{\Big(\mathcal{T}_{\mathcal{P}}(Q^{k-1}) - T^{k-1}\Big)}_{\textbf{Term 3}}.\end{aligned} \tag{53}$$

We then bound the three terms.

Term 1:
Recall that $\rho_k = 2\mathbf{Sp}(Q^0 - Q^*) + \frac{2}{3}\epsilon k$. From the invariance of $\mathbf{Sp}(\cdot)$ by additive constants, the triangle inequality, the nonexpansivity of $\mathcal{T}_\mathcal{P}(\cdot)$ under the span seminorm, and Lemma D.4, it yields

$$\begin{aligned}\mathbf{Sp}(\mathcal{T}_\mathcal{P}(Q^k) - Q^0) &\leq \mathbf{Sp}(Q^k - Q^*) + \mathbf{Sp}(Q^* - Q^0) \\ &\leq \rho_k, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}.\end{aligned}$$

Term 2:
This term can be bounded through a similar approach to Lemma D.5 as

$$\begin{aligned}\mathbf{Sp}\big(\mathcal{T}_\mathcal{P}(Q^k) - \mathcal{T}_\mathcal{P}(Q^{k-1})\big) &= \mathbf{Sp}(Q^k - Q^{k-1}) \\ &\leq \frac{2}{k(k+1)} \sum_{i=1}^{k} \rho_{i+2}, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}.\end{aligned}$$

Term 3:
From Proposition D.3 we have that

$$\mathbf{Sp}(\mathcal{T}_\mathcal{P}(Q^{k-1}) - T^{k-1}) \leq \epsilon, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

We then combine all three terms in equation equation 53, and we have that

$$\begin{aligned}
&g^*_\mathcal{P} - g^{\pi^k}_\mathcal{P}(s) \\
&\overset{Lemma\ 4.1}{\leq} \mathbf{Sp}(\mathcal{T}_\mathcal{P}(Q^k) - Q^k) && (54)\\
&\overset{equation\ 53}{\leq} \frac{2}{k+2}\rho_k + \frac{k}{k+2}\left[\frac{2}{k(k+1)}\sum_{i=1}^{k}\rho_{i+2}\right] + \frac{k}{k+2}(2\epsilon) && (55)\\
&\overset{(a)}{\leq} \frac{4}{k+2}\mathbf{Sp}(Q^0 - Q^*) + \frac{4\epsilon k}{3(k+2)} + \frac{2}{(k+1)(k+2)}\left[\sum_{i=1}^{k}\left(2\|Q^0 - Q^*\|_\infty + \frac{2}{3}\epsilon(i+2)\right)\right] + 2\epsilon && (56)\\
&\overset{(b)}{\leq} \frac{4}{k+2}\mathbf{Sp}(Q^0 - Q^*) + \frac{4\epsilon k}{3(k+2)} + \frac{4k}{(k+1)(k+2)}\mathbf{Sp}(Q^0 - Q^*) + \frac{2\epsilon(k^2+5k)}{3(k+2)(k+1)} + 2\epsilon && (57)\\
&\leq \frac{4(1+2k)}{(k+2)(k+1)}\mathbf{Sp}(Q^0 - Q^*) + \frac{4(3k^2+8k+3)}{3(k+2)(k+1)}\epsilon && (58)\\
&\leq \frac{8\mathbf{Sp}(Q^0 - Q^*)}{k+2} + 4\epsilon, && (59)
\end{aligned}$$

where inequality (a) is from the definition of ρ_k, inequality (b) is from $\mathbf{Sp}(\cdot) \leq 2\|\cdot\|_\infty$.

The proof is thus completed by letting $k \to \infty$ and $\epsilon \to 0$. □

Theorem D.7 (Restatement of Theorem 4.3 - Performance of RHI)**.** *Consider a robust AMDP defined by contamination or ℓ_p-norm, satisfying Assumption 3.1. Set the step sizes $c_k = 5(k+2)\ln^2(k+2)$ and $\beta_k = k/(k+2)$. Then, with probability at least $1-\delta$, the output policy π^n is ϵ-optimal:*

$$g^*_\mathcal{P} - g^{\pi^n}_\mathcal{P}(s) \leq \epsilon, \tag{60}$$

as long as the total iteration number n exceeds $\frac{\mathcal{H}}{\epsilon}$, resulting in the total sample complexity of

$$\tilde{\mathcal{O}}\left(\frac{SA\mathcal{H}^2}{\epsilon^2}\right). \tag{61}$$

Proof. Using the fact that $\mathbf{Sp}(d^k) \leq \mathbf{Sp}(Q^k - Q^{k-1})$ in equation 45 and Lemma D.5, we can derive

$$\begin{aligned}\mathbf{Sp}(d^k) &\leq \frac{2}{k(k+1)} \sum_{i=1}^{k} \rho_{i+2} \\ &= \frac{4}{k+1}\mathbf{Sp}(Q^0 - Q^*) + \frac{2(k+5)}{3(k+1)}\epsilon \\ &\leq \frac{4}{k+1}\mathbf{Sp}(Q^0 - Q^*) + 2\epsilon.\end{aligned}$$

Since in RHI, in each step k, we sample m_k samples for each (s,a)-pair, thus the total sample complexity is $SA|\sum_{k=0}^{n} m_k$. Note that

$$m_k = \max\{\lceil \alpha c_k \mathbf{Sp}(d^k)^2/\epsilon^2 \rceil, 1\} \leq 1 + \alpha c_k \mathbf{Sp}(d^k)^2/\epsilon^2, \tag{62}$$

thus we have that

$$\begin{aligned}&\sum_{k=0}^{n} m_k \\ &\leq (n+1) + \frac{\alpha}{\epsilon^2}\sum_{k=0}^{n} c_k \mathbf{Sp}(d^k)^2 \\ &\leq (n+1) + \frac{10\alpha}{\epsilon^2}\ln^2(2)\mathbf{Sp}(Q^0)^2 + \frac{5\alpha}{\epsilon^2}\sum_{k=1}^{n}(k+2)\ln^2(k+2)\Big(\frac{4\mathbf{Sp}(Q^0 - Q^*)}{k+1} + 2\epsilon\Big)^2 \\ &\leq (n+1) + \frac{10\alpha}{\epsilon^2}\ln^2(2)\mathbf{Sp}(Q^0)^2 + \sum_{k=1}^{n}\frac{240\alpha}{\epsilon^2(k+1)}\ln^2(k+2)\mathbf{Sp}(Q^0 - Q^*)^2 + 40\alpha\sum_{k=1}^{n}(k+2)\ln^2(k+2) \\ &\leq \mathcal{O}\Big(\alpha\mathbf{Sp}(Q^0)^2/\epsilon^2 + \alpha\ln^3(n+2)\mathbf{Sp}(Q^0 - Q^*)^2/\epsilon^2 + \alpha n^2\ln^2(n+2)\Big),\end{aligned} \tag{63}$$

where the penultimate line uses $(a+b)^2 \leq 2a^2 + 2b^2$ and $\frac{k+2}{k+1} \leq \frac{3}{2}$, and the final equality by integral estimation of the sums. Recalling that $\alpha = \ln(2|\mathcal{S}||\mathcal{A}|(n+1)/\delta)$, $L = \ln\left(\frac{2|\mathcal{S}||\mathcal{A}|(n+1)}{\delta}\right)\log^3(n+2)$, $Q^0 = 0$, and since $n \geq \mathcal{H}/\epsilon$, $\mathbf{Sp}(Q^*) \leq \mathcal{H}$, it holds that

$$SA|\sum_{k=0}^{n} m_k \leq \tilde{O}(\frac{SA\mathcal{H}^2}{\epsilon^2}), \tag{64}$$

which completes the proof. □

E PF-RHI: A PARAMETER-FREE VARIANT OF RHI

In this section, we present a fully implementable framework for our Robust Halpern Iteration (RHI) algorithm for diverse and unknown problem settings.

As we mention in Remark 4.4, our RHI algorithm does not require any prior knowledge of the underlying robust AMDP, yet the total number of iterations necessary to generate an ϵ-optimal policy is dependent on $\mathcal{H}$. In practice, such an iteration number may need to be pre-set, and it may be infeasible to set for RHI.

In order to bridge this theoretical finite sample complexity result with the nuances of practical application for varying size problem settings, we now extend our RHI algorithm to a more general and implementable framework: PF-RHI, presented in Algorithm 3. Notably, our PF-RHI do not require any knowledge of $\mathcal{H}$ (even the iteration number); and we will show that it finds an ϵ-optimal policy with identical total sample complexity results as RHI, $\tilde{O}\Big(\frac{SA\mathcal{H}^2}{\epsilon^2}\Big)$.

Note that in our PF-RHI, in each episode i, we run the RHI for n_i steps, and output Q^{n_i} and T^{n_i}. PF-RHI will terminate if the span $\mathbf{Sp}(T^{n_i} - Q^{n_i})$ is small enough. Hence we do not specify iteration number, and thus no knowledge of $\mathcal{H}$ is needed.

Algorithm 3 Implementable Robust Halpern Iteration (PF-RHI)

Input $Q^0 \in \mathbb{R}^{S\times A},\ \epsilon > 0,\ \delta \in (0,1),\ i = 0$
repeat
  Set $n_i = 2^i,\ \delta_i = \delta/c_i$
  Set $\alpha_i = \ln(2|\mathcal{S}||\mathcal{A}|(n_i+1)/\delta_i),\ Q^0 = 0,\ T^{-1} = r,\ h^{-1} = 0,\ c_0 = 10\cdot\ln^2(2),\ \beta_0 = 0$
  for $k = 0,\dots,n_i$ **do**
    $c_k = 5(k+2)\ln^2(k+2),\ \beta_k = k/(k+2)$
    $Q^k = (1-\beta_k)\,Q^0 \ +\ \beta_k\,T^{k-1}$
    $h^k = \max_A(Q^k)$
    $d^k = h^k - h^{k-1}$
    $m_k = \max\{\lceil \alpha_i c_k \mathbf{Sp}(h^k - h^{k-1})^2/\epsilon^2\rceil, 1\}$
    $D^k = \text{R-SAMPLE}(h^k, h^{k-1}, m_k)$
    $T^k = T^{k-1} + D^k$
  end for
  $\pi^{n_i}(s) \in \arg\max_{a\in A}\ Q^{n_i}(s,a) \quad \forall\, s \in S$
  $i = i + 1$
until $\mathbf{Sp}(T^{n_i} - Q^{n_i}) \le 14\epsilon$
Output: $Q^{n_i}, T^{n_i}, \pi^{n_i}$

E.1 Analysis of PF-RHI

To facilitate our analysis of PF-RHI, we first present some useful notations as follows.

$$\mu \triangleq \mathbf{Sp}(Q^0 - Q^*),$$
$$\nu \triangleq \mathbf{Sp}(Q^0 - Q^*) + \mathbf{Sp}(Q^0),$$
$$\zeta \triangleq \max\{\mathbf{Sp}(r), \mathbf{Sp}(Q^0)\}.$$

We then define the following random variables:

$$N = \inf\{n_i \in \mathbb{N} : \mathbf{Sp}(T^{n_i} - Q^{n_i}) \le 14\epsilon\}, \quad \text{and}$$
$$I = \inf\{i \in \mathbb{N} : \mathbf{Sp}(T^{n_i} - Q^{n_i}) \le 14\epsilon\},$$

and it holds that $N = 2^I$.

We set $i_0 \in \mathbb{N}$ be the smallest integer s.t. $n_{i_0} \ge \mathbf{Sp}(Q^0 - Q^*)/\epsilon = \mu/\epsilon$. Then either $i_0 = 0$ and $n_{i_0} = 1$, or $n_{i_0-1} = n_{i_0}/2 < \mu/\epsilon$, which, when combined, imply that $n_{i_0} \le 2(1+\mu/\epsilon)$.

With these, we further define the additional random events:

$$S_i = \{\mathbf{Sp}(T^{n_i} - Q^{n_i}) \le 14\epsilon,\ \forall (s,a) \in \mathcal{S}\times\mathcal{A}\}, \quad \text{and}$$
$$G_i = \{\|T^k - \mathcal{T}_{\mathcal{P}}(Q^k)\|_\infty \le \epsilon,\ \forall k = 0,1,\dots,n_i,\ \forall (s,a) \in \mathcal{S}\times\mathcal{A}\}$$

where T^k and Q^k are generated by the inner-loop $k = 0,1,\dots,n_i$ during the i-th iteration of PF-RHI. During this specific iteration i, let M_i be the number of samples generated so that $M \triangleq \sum_{i=0}^{I} M_i$ where M and M_i are random variables.

Lemma E.1. *It holds that*

$$\mathbb{P}(S_i) \ge \mathbb{P}(G_i) \ge 1 - \delta_i, \quad \forall i \ge i_0, \forall (s,a) \in \mathcal{S}\times\mathcal{A}. \tag{65}$$

Proof. Note that Proposition D.3 directly implies $\mathbb{P}(G_i) \ge 1 - \delta_i$. Moreover, for $i \ge i_0$ and for all $\xi \in G_i$, from Theorem D.6 we have that

$$\mathbf{Sp}\big(T^{n_i}(\xi) - Q^{n_i}(\xi)\big) \le \mathbf{Sp}\big(T^{n_i}(\xi) - \mathcal{T}_{\mathcal{P}}(Q^{n_i})(\xi)\big) + \mathbf{Sp}\big(\mathcal{T}_{\mathcal{P}}(Q^{n_i})(\xi) - Q^{n_i}(\xi)\big) \tag{66}$$
$$\le 2\epsilon + \frac{8\mathbf{Sp}(Q^0 - Q^*)}{n_i + 2} + 4\epsilon \tag{67}$$
$$\le 14\epsilon, \tag{68}$$

thus $G_i \subseteq S_i$, which completes the proof. □

Proposition E.2. *It holds that*

$$\mathbb{E}[N] \leq 2(1+\mu/\epsilon)/(1-\delta).$$

Namely, N is finite almost surely and PF-RHI$(Q^0, \epsilon, \delta, i=0)$ stops with probability 1 after a finite number of iterations.

Proof. In each iteration i, in PF-RHI$(Q^0, \epsilon, \delta, i=0)$, we reinitialize $Q^0 = 0$ prior to the inner for loop where $k = 0, 1, \ldots, n_i$. This implies that the events $\{S_i : i \in \mathbb{N}\}$ are mutually independent. Thus,

$$\mathbb{P}(I=i) = \mathbb{P}\Big(\bigcap_{j=0}^{i-1} S_j^c \cap S_i\Big) = \prod_{j=0}^{i-1} \mathbb{P}(S_j^c) \cdot \mathbb{P}(S_i).$$

Now from Lemma E.1, it holds that $\mathbb{P}(S_i^c) \leq \mathbb{P}(G_i^c) \leq \delta_i$ for all $i \geq i_0$, which implies that $\mathbb{P}(I=i) \leq \prod_{j=i_0}^{i-1} \delta_j$.

Moreover, by definition of c, we have that $2\sum_{i=0}^{\infty} c_i^{-1} \leq 1$, thus $\delta_j = \delta/c_j \leq \delta/2$, implying that $\mathbb{P}(I=i) \leq (\delta/2)^{i-i_0}$. Using this and the fact that $n_i = n_{i_0} 2^{i-i_0}$, it holds that

$$\begin{aligned}
\mathbb{E}[N] &= \sum_{i=0}^{\infty} n_i \mathbb{P}(N = n_i) \\
&\leq n_{i_0} + \sum_{i=i_0+1}^{\infty} n_{i_0} 2^{i-i_0} \mathbb{P}(I=i) \\
&\leq n_{i_0}\big(1 + \sum_{i=i_0+1}^{\infty} \delta^{i-i_0}\big).
\end{aligned}$$

The proof is then completed by the bound of $n_{i_0} \leq 2(1+\mu/\epsilon)$, which implies that $\mathbb{E}[N] \leq 2(1+\mu/\epsilon)/(1-\delta)$. □

Theorem E.3. *Let $c_k = 5(k+2)\ln^2(k+2)$ and $\beta_k = k/(k+2)$ hold. Let $n_i = N$ so that (Q^N, T^N, π^N) is returned by PF-RHI$(Q^0, \epsilon, \delta, i=0)$. Then with probability of at least $(1-\delta)$, we have for all $s \in \mathcal{S}$,*

$$g_{\mathcal{P}}^* - g_{\mathcal{P}}^{\pi^N}(s) \leq \textbf{\textit{Sp}}(\mathcal{T}_{\mathcal{P}}(Q^N) - Q^N) \leq 16\epsilon,$$

Which obtains a sample and time complexity of $\mathcal{O}\big(\hat{L}|\mathcal{S}||\mathcal{A}|(\nu^2/\epsilon^2+1)\big)$, with $\hat{L} = \ln\big(4|\mathcal{S}||\mathcal{A}|(1+\mu/\epsilon)/\delta\big)\log^4(2(1+\mu/\epsilon))$.

Proof. As Lemma D.1 implies that $0 \leq g_{\mathcal{P}}^* - g_{\mathcal{P}}^{\pi^N} \leq \mathbf{Sp}(\mathcal{T}_{\mathcal{P}}(Q^N) - Q^N)$.

We first define $A = \{I \leq i_0\}$ and $B = \bigcap_{i=0}^{\infty} G_i$, where $G_i = \{\|T^k - \mathcal{T}_{\mathcal{P}}(Q^k)\|_\infty \leq \epsilon,\ \forall k = 0, 1, \ldots, n_i, \forall (s,a) \in \mathcal{S} \times \mathcal{A}$. We claim that, $\mathbb{P}(A \cap B) \geq (1-\delta)$. To prove this, note that from Proposition D.3, $\mathbb{P}(G_i^c) \leq \delta_i$. Thus

$$\mathbb{P}(B^c) = \mathbb{P}\Big(\bigcup_{i=1}^{\infty} G_i^c\Big) \leq \sum_{i=1}^{\infty} \delta/c_i \leq \delta/2. \tag{69}$$

Then Lemma E.1 implies that $\mathbb{P}(A) \geq \mathbb{P}(S_{i_0}) \geq \mathbb{P}(G_{i_0}) \geq 1 - \delta_{i_0} \geq 1 - \delta/2$, thus combining this with equation 69 implies $\mathbb{P}(A^c \cup B^c) \leq \delta$, which proves our claim.

We then use the definitions of $N,\ I$, and G_i, which imply that

$$\begin{aligned}
\mathbf{Sp}(\mathcal{T}_{\mathcal{P}}(Q^N) - Q^N) &\leq \mathbf{Sp}(\mathcal{T}_{\mathcal{P}}(Q^N) - T^N) + \mathbf{Sp}(T^N - Q^N) \\
&\leq 2\epsilon + 14\epsilon \\
&= 16\epsilon.
\end{aligned}$$

Applying equation 63 further implies the total sample complexity, $M \triangleq \sum_{i=0}^{I} M_i$ can be bounded as

$$M \leq \sum_{i=0}^{i_0} M_i \tag{70}$$
$$= |\mathcal{S}||\mathcal{A}| \sum_{i=0}^{i_0} \mathcal{O}\Big(\alpha_i \mathbf{Sp}(Q^0)^2/\epsilon^2 + \alpha_i \ln^3(n_i+2)\mathbf{Sp}(Q^0 - Q^*)^2/\epsilon^2 + \alpha_i n_i^2 \ln^2(n_i+2)\Big),$$

where $\alpha_i = \ln(2|\mathcal{S}||\mathcal{A}|(n_i+1)/\delta_i)$ is the parameter defined at iteration i (prior to the inner for loop) of PF-RHI. Moreover, since $n_{i_0}^2 \leq 4(1+\mu/\epsilon)^2 = \mathcal{O}\big(\mathbf{Sp}(Q^0 - Q^*)^2/\epsilon^2 + 1\big)$, we have that

$$\begin{aligned} M &\leq |\mathcal{S}||\mathcal{A}|(i_0+1)\mathcal{O}\big(\alpha_{i_0}\mathbf{Sp}(Q^0)^2/\epsilon^2 + \alpha_{i_0}\log^3(n_{i_0}+2)\mathbf{Sp}(Q^0-Q^*)^2/\epsilon^2 + \alpha_{i_0} n_{i_0}^2 \log^2(n_{i_0}+2)\big) \\ &\leq |\mathcal{S}||\mathcal{A}|\alpha_{i_0}\log^4(n_{i_0}+2)\ \mathcal{O}\big(\mathbf{Sp}(Q^0)^2/\epsilon^2 + 2\mathbf{Sp}(Q^0-Q^*)^2/\epsilon^2 + 1\big) \\ &\leq \hat{L}|\mathcal{S}||\mathcal{A}|\mathcal{O}(\nu^2/\epsilon^2+1), \end{aligned}$$

which completes the proof. □

Corollary E.4. *Let $n_i = N$. Then with probability of at least $(1-\delta)$, for all $s \in \mathcal{S}$, it holds that*

$$g_{\mathcal{P}}^* - g_{\mathcal{P}}^{\pi^N}(s) \leq \mathbf{Sp}(\mathcal{T}_{\mathcal{P}}(Q^N) - Q^N) \leq \epsilon.$$

This results in a sample and time complexity of $\mathcal{O}(\tilde{L}|\mathcal{S}||\mathcal{A}|\mathcal{H}^2/\epsilon^2) = \tilde{\mathcal{O}}(SA\mathcal{H}^2/\epsilon^2)$, where we define $\tilde{L} = \ln(2|\mathcal{S}||\mathcal{A}|\mathcal{H}/(\epsilon\delta))\ln^4(\mathcal{H}/\epsilon)$ and $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic terms.

Proof. Note that

$$\mathbf{Sp}(Q^0 - Q^*) = \mathbf{Sp}(Q^*) = \mathbf{Sp}(r + \mathsf{P}h^*) \leq \mathbf{Sp}(r) + \mathbf{Sp}(h^*),$$

which is due to $Q^0 = 0$ at each iteration i and the nonexpansivity of the map $Q \mapsto \max_{\mathcal{A}}(Q) = h$. Moreover, since $2(1+\mu/\epsilon) = \mathcal{O}\big(\mathbf{Sp}(h)^2/\epsilon\big)$, combining with Theorem E.3, the result follows by verifying the definition of $\tilde{L}$. □

We then derive the results under expecations.

Lemma E.5. *For an arbitrary fixed iteration $i \in \mathbb{N}$ of PF-RHI$(Q^0, \epsilon, \delta, i = 0)$, let $M_i = |\mathcal{S}||\mathcal{A}| \sum_{j=0}^{n_i} m_j$ be the number of samples obtained during iteration i. We have*

$$M_i \leq |\mathcal{S}||\mathcal{A}|\mathcal{O}\big(n_i + (\zeta/\epsilon)^2 \alpha_i n_i^2 \log^2(n_i+2)\big),$$

where $\alpha_i = \ln(2|\mathcal{S}||\mathcal{A}|(n_i+1)/\delta_i)$.

Proof. By using induction, for $k=0$ we have by initialization $d^0 = \max_{\mathcal{A}}(Q^0)$ and $T^{-1} = r$. By using both equation 45 and equation 50 along with the induction hypothesis for $k \geq 0$,

$$\begin{aligned} \mathbf{Sp}(d^k) &\leq \mathbf{Sp}(Q^k - Q^{k-1}) \\ &\leq \frac{2}{(k+1)(k+2)}\mathbf{Sp}(T^{k-1} - Q^0) + \frac{k-1}{k+1}\mathbf{Sp}(d^{k-1}) \\ &\leq \frac{2}{(k+1)(k+2)}\big((k+1)\zeta + \zeta\big) + \frac{k-1}{k+1}\zeta \\ &= \zeta. \end{aligned}$$

This implies that

$$\begin{aligned} \mathbf{Sp}(T^k) &\leq \mathbf{Sp}(T^{k-1}) + \mathbf{Sp}(D^k) \\ &\leq (k+1)\zeta + \mathbf{Sp}(d^k) \\ &\leq (k+2)\zeta. \end{aligned}$$

Thus for a fixed $i \in \mathbb{N}$ in PF-RHI, we can bound M_i as

$$\begin{aligned} M_i &\leq |\mathcal{S}||\mathcal{A}|\big((n_i+1) + (\alpha_i/\epsilon^2)\sum_{j=0}^{n_i} c_j \mathbf{Sp}(d^j)^2\big) \\ &\leq |\mathcal{S}||\mathcal{A}|\big((n_i+1) + 5(\zeta/\epsilon)^2\alpha_i \sum_{j=0}^{n_i}(j+2)\ln^2(j+2)\big) \\ &= |\mathcal{S}||\mathcal{A}|\mathcal{O}\big(n_i + (\zeta/\epsilon)^2\alpha_i n_i^2 \log^2(n_i+2)\big), \end{aligned}$$

which completes the proof. □

Theorem E.6. *Assume that the robust-AMDP satisfies Assumption 3.1, and that the sequences $c_k = 5(k+2)\ln^2(k+2)$ and $\beta_k = k/(k+2)$ hold. Let $n_i = N$ so that (Q^N, T^N, π^N) is the output of PF-RHI$(Q^0, \epsilon, \delta, i=0)$. Then for every $s \in \mathcal{S}$ we have,*

$$\mathbb{E}\big[g_{\mathcal{P}}^* - g_{\mathcal{P}}^{\pi^N}(s)\big] \leq 16\epsilon + \delta \boldsymbol{Sp}(r),$$

which yields an expected sample and time complexity of

$$\tilde{\mathcal{O}}\big(|\mathcal{S}||\mathcal{A}|(\nu^2/\epsilon^2 + 1 + \delta(1+\mu/\epsilon)^2(1+(\zeta/\epsilon)^2)\big).$$

Proof. We start our proof similar to Theorem E.3 by considering the events $A = \{I \leq i_0\}$ and $B = \bigcap_{i=1}^{\infty} G_i$. From Theorem E.3, under $A \cap B$, for every $s \in \mathcal{S}$ it holds that $g_{\mathcal{P}}^* - g_{\mathcal{P}}^{\pi^N}(s) \leq 16\epsilon$ with probability $\mathbb{P}(A \cap B) \geq 1 - \delta$.

On the other hand, under $(A \cap B)^c$, we have the trivial bound of $g_{\mathcal{P}}^* - g_{\mathcal{P}}^{\pi^N}(s) \leq \mathbf{Sp}(r),\ \forall s \in \mathcal{S}$.

Hence the two cases together imply that

$$\mathbb{E}[g_{\mathcal{P}}^* - g_{\mathcal{P}}^{\pi^N}(s)] \leq 16\epsilon + \delta \mathbf{Sp}(r), \quad \forall s \in \mathcal{S}.$$

Similar to Theorem E.3, we wish to estimate the sample complexity like $M = \sum_{i=0}^{I} M_i$ for each iteration i of PF-RHI. We accomplish this by considering the infinite disjoint union of all indexes $i > i_0$, or more formally $A^c = \bigsqcup_{i=i_0+1}^{\infty}\{I = i\}$ which yields

$$\mathbb{E}[M] = \underbrace{\mathbb{E}[M|A\cap B]\mathbb{P}(A\cap B)}_{\textbf{Term 1}} + \underbrace{\mathbb{E}[M|A\cap B^c]\mathbb{P}(A\cap B^c)}_{\textbf{Term 2}} + \underbrace{\sum_{i=i_0+1}^{\infty}\mathbb{E}[M|I=i]\mathbb{P}(I=i)}_{\textbf{Term 3}}.$$

Term 1:
We use the result derived from the proof of Theorem E.3 on the event $(A \cap B)$ and the fact that $\mathbb{P}(A \cap B) \leq 1$. By defining $\hat{L} = \ln\big(4|\mathcal{S}||\mathcal{A}|(1+\mu/\epsilon)/\delta\big)$, we have that

$$\mathbb{E}[M|A\cap B]\mathbb{P}(A\cap B) = \mathcal{O}\big(\hat{L}|\mathcal{S}||\mathcal{A}|(\nu^2/\epsilon^2+1)\big). \tag{71}$$

Term 2:
We can combine the result in Lemma E.5 with $\mathbb{P}(A\cap B^c) \leq \mathbb{P}(B^c) \leq \delta$, and $n_{i_0} \leq 2(1+\mu/\epsilon)$ to obtain the following result:

$$\begin{aligned} \mathbb{E}[M|A\cap B^c]\mathbb{P}(A\cap B^c) &\leq \delta|\mathcal{S}||\mathcal{A}|\sum_{i=0}^{i_0}\mathcal{O}\big(n_i + (\zeta/\epsilon)^2\alpha_i n_i^2\log^2(n_i+2)\big) \\ &\leq \delta|\mathcal{S}||\mathcal{A}|\mathcal{O}\big(n_{i_0} + (\zeta/\epsilon)^2\alpha_{i_0}n_{i_0}^2\log^3(n_{i_0}+2)\big) \\ &\leq \delta|\mathcal{S}||\mathcal{A}|\mathcal{O}\big(n_{i_0} + \hat{L}(\zeta/\epsilon)^2 n_{i_0}^2\big). \end{aligned} \tag{72}$$

The final inequality holds by using the definition of $\hat{L}$ and that $\alpha_{i_0}\log^3(n_{i_0}+2) \leq \mathcal{O}(\hat{L})$.

Term 3:
To bound this term, we can again employ the result of Lemma E.5 along with defining $Z \triangleq$

$\sum_{i=i_0+1}^{\infty} \mathbb{E}[M|I=i]\mathbb{P}(I=i)$ to have that

$$\begin{aligned} Z &\leq |\mathcal{S}||\mathcal{A}| \sum_{i=i_0+1}^{\infty} \mathcal{O}\big(n_i + (\zeta/\epsilon)^2 \alpha_i n_i^2 \log^2(n_i+2)\big)\mathbb{P}(I=i) \\ &\leq |\mathcal{S}||\mathcal{A}| \sum_{i=i_0+1}^{\infty} \mathcal{O}\big(n_i + \hat{L}(\zeta/\epsilon)^2 i^3 n_i^2\big)\mathbb{P}(I=i), \end{aligned}$$

where the final inequality follows from using the re-initializations of $n_i,\ \delta_i,$ and α_i in PF-RHI to obtain $\alpha_i = \mathcal{O}(\hat{L}+i) \leq \hat{L}\mathcal{O}(i)$, where $\log\big((n_i+1)c_i\big) = \mathcal{O}(i)$, and likewise $\log^2(n_i+2) = \mathcal{O}(i^2)$. With this in place, recall that $n_i = n_{i_0} 2^{i-i_0}$. From Proposition E.2, for $i \geq i_0+1$ we have that $\mathbb{P}(I=i) \leq \prod_{j=i_0}^{i-1} \delta_j \leq \mathcal{O}\big(\delta \prod_{j=i_0}^{i-1} \frac{1}{j+2}\big)$. Therefore, we can denote the following

$$S_1 \triangleq \sum_{i=i_0+1}^{\infty} 2^{i-i_0} \prod_{j=i_0}^{i-1} \frac{1}{j+2}, \tag{73}$$

$$S_2 \triangleq \sum_{i=i_0+1}^{\infty} 2^{2(i-i_0)} i^3 \prod_{j=i_0}^{i-1} \frac{1}{j+2}, \tag{74}$$

which allows us to show that

$$Z \leq \delta|\mathcal{S}||\mathcal{A}|\mathcal{O}\big(S_1 n_{i_0} + S_2 \hat{L}(\zeta/\epsilon)^2 n_{i_0}^2\big).$$

However, we can calculate equation 73 and equation 74 using their incomplete Gamma functions like,

$$\begin{aligned} S_1 &= e^2 2^{-(i_0+1)}[\Gamma(i_0+2) - \Gamma(i_0+2,2)] \\ &\leq \frac{(e^2-3)}{2} \end{aligned} \tag{75}$$

$$\begin{aligned} S_2 &= 84 + 4i_0(i_0+5) + 67e^4 2^{-2(i_0+1)}[\Gamma(i_0+2) - \Gamma(i_0+2,4)] \\ &= \mathcal{O}\big((i_0+1)^2\big). \end{aligned} \tag{76}$$

With equation 75 and equation 76, we can finally bound Z as

$$Z \leq \delta|\mathcal{S}||\mathcal{A}|\mathcal{O}\big(n_{i_0} + \hat{L}(\zeta/\epsilon)^2 n_{i_0}^2 (i_0+1)^2\big). \tag{77}$$

We can then find the total expected value of the sample complexity by combining equation 71, equation 72, and equation 77 by rearranging similar order terms and disregarding the logarithmic terms to obtain:

$$\begin{aligned} \mathbb{E}[M] &\leq |\mathcal{S}||\mathcal{A}|\mathcal{O}\big(\hat{L}(\nu^2/\epsilon^2+1) + \delta n_{i_0} + \delta \hat{L}(\zeta/\epsilon)^2 n_{i_0}^2 (i_0+1)^2\big) \\ &= |\mathcal{S}||\mathcal{A}|\tilde{\mathcal{O}}\big((\nu^2/\epsilon^2+1) + \delta(1+\mu/\epsilon)^2(1+(\zeta/\epsilon)^2)\big), \end{aligned}$$

which completes the proof. □

Corollary E.7. *Assume that the robust-AMDP satisfies Assumption 3.1, that the sequences* $c_k = 5(k+2)\ln^2(k+2)$ *and* $\beta_k = k/(k+2)$ *hold,* $r(s,a) \in [0,1]\ \forall (s,a) \in \mathcal{S}\times\mathcal{A}$*, and* $\mathcal{H} \geq 1$*. Let* $n_i = N$ *such that* $N \geq \mathcal{H}/\epsilon$ *so that* (Q^N, T^N, π^N) *is returned by PF-RHI*$(Q^0, \epsilon/17, \delta, i=0)$ *with* $Q^0 = 0,\ \epsilon \leq 1$*, and* $\delta = \epsilon^2/17$*. We have for every* $s \in \mathcal{S}$*,*

$$\mathbb{E}[g_{\mathcal{P}}^* - g_{\mathcal{P}}^{\pi^N}(s)] \leq \epsilon,$$

where we obtain an expected sample complexity of $\tilde{\mathcal{O}}\big(|\mathcal{S}||\mathcal{A}|\mathcal{H}^2/\epsilon^2\big)$*.*

Proof. The proof is directly derived by applying the value of δ in Theorem E.6. □