A Reduction Framework for Distributionally Robust Reinforcement Learning under Average Reward

Zachary Roch [1] George Atia [1 2] Yue Wang [1 2]

Abstract

Robust reinforcement learning (RL) under the average reward criterion, which seeks to optimize long-term system performance in uncertain environments, remains a largely unexplored area. To address this challenge, we propose a reduction-based framework that transforms robust average reward optimization into the more extensively studied robust discounted reward optimization by employing a specific discount factor. Our framework provides two key advantages. **Data Efficiency:** We design a model-based reduction algorithm that achieves near-optimal sample complexity, enabling efficient identification of optimal robust policies; **Scalability:** By bypassing the inherent challenges of scaling up average reward optimization, our framework facilitates the design of scalable, convergent algorithms for robust average reward optimization leveraging function approximation. Our algorithmic design, supported by theoretical and empirical analyses, provides a concrete solution to robust average reward RL with the first data efficiency and scalability guarantees, highlighting the framework's potential to optimize long-term performance under model uncertainty in practical problems.

1. Introduction

Reinforcement Learning (RL) aims to optimize an agent's performance by identifying a policy that maximizes cumulative rewards based on a specified criterion while interacting with an environment. Despite its remarkable success in applications such as synthetic control problems, board games (Silver et al., 2016; Zha et al., 2021), and video games (Wei et al., 2022; Liu et al., 2022a), RL often experiences significant performance degradation in practical settings. This phenomenon, known as the Sim-to-Real gap, arises from discrepancies between the training environment and the deployment environment. In simulation-based applications like games, the training and deployment environments are typically identical and vanilla RL performs well. However, in real-world scenarios, differences such as modeling errors, perturbations, partial observability, and potential adversarial attacks introduce model mismatches between the training and deployment environments, resulting in suboptimal policies and poor performance outcomes, undermining the reliability of RL in practical applications.

To address this issue, a framework of robust RL was introduced (Bagnell et al., 2001; Nilim & El Ghaoui, 2004; Iyengar, 2005). It deviates from vanilla RL by considering a set of environment transition dynamics instead of a fixed one, and its goal is to optimize performance under the worst-case scenario across these models, which provides performance guarantees across all uncertain environments within the defined uncertainty set, making the policy more robust to model mismatches and more generalizable.

On the other hand, different reward criteria in (robust) RL can result in substantially distinct problem settings. Among these, the discounted reward criterion is the most extensively studied. While it offers elegant mathematical properties, its focus on short-term rewards can lead to suboptimal long-term performance due to the exponential decay of rewards. In practical applications such as queuing control, supply chain inventory management, and communication networks (Kober et al., 2013), however, evaluating policies based on their long-term average performance becomes crucial. This highlights the importance of optimizing the long-term average reward in environments with uncertainty, motivating our focus on robust RL under average reward in this paper.

Despite its practical importance, robust RL under average reward is generally more complex than its discounted counterpart due to its dependence on the limiting behavior of the underlying stochastic processes, and is hence understudied. Recent work (Wang et al., 2023e; Grand-Clement et al., 2023) further emphasizes its inherent challenges, including

[1] Department of Electrical and Computer Engineering [2] Department of Computer Science, University of Central Florida, Orlando, Florida. Correspondence to: Zachary Roch <zachary.roch@ucf.edu>, George Atia <george.atia@ucf.edu>, Yue Wang <yue.wang@ucf.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

the non-contracted nature of the Bellman operator, the high dimensionality of the solution space, and the difficulties in relaxing underlying assumptions.

To address these challenges, a natural approach is to draw on insights from the extensive studies of robust RL for discounted rewards as an intermediate step. This idea has been validated in (Wang et al., 2023d; Grand-Clement et al., 2023), which show that, under certain assumptions, the performance of discounted robust RL asymptotically converges to that of average reward as the discount factor approaches 1. While this convergence highlights the potential of using discounted robust RL to study the average reward setting, there remains a lack of results demonstrating its practical applicability. Key questions regarding its efficiency, effectiveness, and scalability remain unanswered, leaving gaps in understanding its implementation and real-world impact.

In this paper, we explore this approach in greater depth and propose a reduction-based framework for concrete algorithm implementation. This framework facilitates the use of various robust discounted RL algorithms to address robust average reward RL problems. To assess the practicality of our framework, we evaluate its performance across two key dimensions: data efficiency and scalability. Our contributions are summarized as follows.

Reduction of robust average reward RL to discounted one: Under a standard assumption (Assumption 3.1), we propose a reduction-based framework that shows how robust average reward optimization under model uncertainty can be equivalently addressed through robust discounted RL with a specific discount factor. While prior work has explored asymptotic convergence, no practical guidance has been offered on selecting a reduction discount factor to guarantee the optimality of the resulting policy. Our framework provides a concrete choice of the reduction discount factor, ensuring that the robust policy learned for the discounted reward is also optimal under the average reward criterion. This universal framework deepens the understanding of the fundamental connections between average and discounted rewards while enabling robust average reward optimization.

Design of data-efficient reduction algorithms: Building on our reduction framework, we present the first model-based algorithm for robust RL under average reward, applicable to various uncertainty set models, with a thorough sample complexity analysis. We study the total number of samples required to learn an ϵ-optimal robust policy for the average reward criterion under different uncertainty set structures. Specifically, we provide detailed analyses for total variation, χ^2 divergence, and Kullback–Leibler divergence uncertainty sets, demonstrating that our reduction algorithms achieve near-optimal sample complexity. These results highlight the practical potential of our framework in data-intensive settings, offering the first finite-sample complexity characterization of robust RL with average reward.

Design of scalable reduction algorithms: To further validate the practical applicability of our framework, we adapt it to design scalable algorithms for robust average reward RL. After identifying key challenges in scaling up robust average reward RL, we show that our reduction framework circumvents these difficulties, enabling the design of efficient algorithms for large-scale problems. We evaluate our algorithms in large-scale MuJoCo environments, showcasing the capability of our framework to optimize long-term rewards under model uncertainty in complex systems. These results underscore the potential of our framework to efficiently solve large-scale, real-world problems.

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 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 action subject to 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[2]

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

It is also useful to define the following relative value function or bias for an AMDP:

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

which is the cumulative difference over time between the immediate reward and the average reward.

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

[2] The limit may not exist, but under the assumption we made, such a limit exists (Puterman, 1994).

Robust MDPs. In 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 $\mathsf{P} \in \mathcal{P}$. We specifically concentrate on the (s, a)-rectangular uncertainty set (Nilim & El Ghaoui, 2004; Iyengar, 2005), 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.

Robust MDPs aim to optimize the worst-case performance over the uncertainty set. The robust discounted value function of a policy π is defined as the worst-case discounted value function over all possible transition kernels:

$$V^\pi_{\gamma,\mathcal{P}}(s) \triangleq \min_{\kappa\in\bigotimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\pi,\kappa}\left[\sum_{t=0}^{\infty} \gamma^t r_t | S_0 = s\right], \quad (3)$$

where $\kappa = (\mathsf{P}_0, \mathsf{P}_1 ...) \in \bigotimes_{t\geq 0} \mathcal{P}$. The discounted robust value functions are shown to be the unique solution to the robust discounted Bellman equation (Iyengar, 2005):

$$V(s) = \sum_a \pi(a|s)(r(s,a) + \gamma\sigma_{\mathcal{P}_s^a}(V)), \quad (4)$$

where $\sigma_{\mathcal{P}_s^a}(V) \triangleq \min_{\mathsf{P}\in\mathcal{P}_s^a} \mathsf{P}V$ is the support function of V on the uncertainty set $\mathcal{P}_s^a$.

In scenarios where the long-term performance under model uncertainty is concerned, we focus on the following worst-case average reward:

$$g^\pi_{\mathcal{P}}(s) \triangleq \min_{\kappa\in\bigotimes_{t\geq 0}\mathcal{P}} \lim_{n\to\infty} \mathbb{E}_{\pi,\kappa}\left[\frac{1}{n}\sum_{t=0}^{n-1} r_t | S_0 = s\right], \quad (5)$$

to which we refer as the robust average reward. The robust AMDP aims to find an optimal policy w.r.t. it, that is, $\pi^* \triangleq \arg\max_{\pi\in\Pi} g^\pi_{\mathcal{P}}(s)$, for any $s \in \mathcal{S}$.

In (Wang et al., 2023d), it is shown that the robust discounted value functions converge to the robust average reward w.r.t. the same MDP as the discount factor approaches 1:

$$\lim_{\gamma\to 1}(1-\gamma)V^\pi_{\gamma,\mathcal{P}} = g^\pi_{\mathcal{P}} \,. \quad (6)$$

Hence, the robust AMDP can be approximately solved through the corresponding robust DMDP with a sufficiently large discount factor, known as the reduction method. However, selecting a discount factor to ensure near-optimality under the average reward remains unclear, leaving the adaptation of the reduction method uncertain.

In this paper, our goal is to develop a concrete reduction framework and design algorithms for optimizing the robust average reward, and to demonstrate the practical applicability of our framework.

3. Reduction Framework for Robust AMDPs

Our framework aims to reduce the robust average reward problem to a robust discounted reward one, leveraging well-developed algorithms in this space. The convergence (6) and the existence of a robust Blackwell optimal policy (Wang et al., 2023d; Grand-Clément & Petrik, 2024) (a policy that optimizes the robust discounted reward for any $\gamma > \gamma_{bw}$) further inspires us to reduce a robust AMDP to a robust DMDP with a sufficiently large discount factor. However, existing results focus on asymptotic convergence, leaving the choice of discount factor for a desired level of accuracy unresolved. In this section, we study the relationship between the two robust MDPs and determine a specific reduction discount factor.

We first adopt the following compactness and unichain assumption, which is commonly used in robust AMDPs.

Assumption 3.1. For any $s \in \mathcal{S}, a \in \mathcal{A}$, the uncertainty set $\mathcal{P}_s^a$ is a compact subset of $\Delta(\mathcal{S})$. Moreover, any deterministic policy π and any kernel $\mathsf{P} \in \mathcal{P}$ induce a unichain Markovian process[3].

Due to the Heine–Borel theorem (Dugac, 1989), the first part of Assumption 3.1 is satisfied if the uncertainty set is closed as it is always bounded. We remark that many standard uncertainty sets satisfy this assumption, e.g., those defined by ϵ-contamination (Huber, 1965), finite interval (Tewari & Bartlett, 2007), total-variation (Rahimian et al., 2022) and KL-divergence (Hu & Hong, 2013).

The second part of Assumption 3.1 imposes additional structure on the underlying MDP, an assumption commonly used in non-robust AMDP studies due to their inherent complexity (e.g., (Puterman, 1994; Wan et al., 2021; Zhang & Ross, 2021; Lan, 2020; Zhang et al., 2021b)). For robust AMDPs, the unichain assumption ensures the solvability of the average reward robust Bellman equation (Wang et al., 2023d;e), which plays an essential role in our analysis. Under this assumption, the stationary distribution η^π_{P} always exists and does not depend on the initial state, and the average reward is identical for all starting states (Bertsekas, 2011), i.e., $g^\pi_{\mathsf{P}}(s_1) = g^\pi_{\mathsf{P}}(s_2), \forall s_1, s_2 \in \mathcal{S}$.

Inspired by non-robust AMDP studies (Wang et al., 2022; Zurek & Chen, 2023), we further extend the concept of the optimal bias span (Bartlett & Tewari, 2012) therein to the robust setting.

Definition 3.2. For a robust AMDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$, its robust

[3] A Markovian process is a unichain if it contains exactly one recurrent class and possibly some transient states (Puterman, 1994; Bertsekas, 2011).

optimal bias span is defined as[4]

$$\mathcal{H} \triangleq \max_{\mathsf{P}\in\mathcal{P}} \mathbf{Sp}(h^{\pi^*}_{\mathsf{P}}) \tag{7}$$

where $h^{\pi^*}_{\mathsf{P}}$ is the relative value function as in (2), and $\mathbf{Sp}(V) \triangleq \max_i V(i) - \min_i V(i)$ is the Span semi-norm.

Remark 3.3. *We assume $\mathcal{H}$ is known, which can be viewed as a robust extension from the common assumption of knowledge of non-robust span in non-robust AMDPs studies (Wang et al., 2022; Zurek & Chen, 2023; Zhang & Xie, 2023; Wang et al., 2023a). Our framework and results remain valid if $\mathcal{H}$ is replaced with any upper bound on $\mathcal{H}$, such as the corresponding robust extensions of the mixing time or the diameter of a non-robust MDP (Wang et al., 2022). Additional discussion on estimating $\mathcal{H}$ is provided in Section 10. Specifically, we can derive an upper bound on $\mathcal{H}$ for robust MDPs with some additional structures.*

Next, we present our reduction framework.

Theorem 3.4. *(Reduction Framework) For any ϵ, set $\gamma := 1-\frac{\epsilon}{\mathcal{H}}$, then any ϵ_γ-optimal policy[5] $\hat{\pi}_\gamma$ for the robust DMDP $(\mathcal{S},\mathcal{A},\mathcal{P},r,\gamma)$ is also an $\mathcal{O}(\epsilon)$-optimal policy for the corresponding robust AMDP $(\mathcal{S},\mathcal{A},\mathcal{P},r)$:*

$$g^{\pi^*}_{\mathcal{P}} - g^{\hat{\pi}_\gamma}_{\mathcal{P}} \le \left(8+\frac{5\epsilon_\gamma}{\mathcal{H}}\right)\epsilon.$$

While we defer the full proof to the appendix, the intuition behind the proof of our reduction framework relies on showing a bound on the convergence error of the robust discounted value function to the average reward. As shown in Lemma 11.1, under Assumption 3.1 with $\gamma\in(0,1)$ and for any stationary $\pi\in\Pi$, $\|g^\pi_{\mathcal{P}} - (1-\gamma)V^\pi_{\gamma,\mathcal{P}}\|_\infty \le \mathbf{Sp}\big((1-\gamma)V^\pi_{\gamma,\mathcal{P}}\big)$. We then show that we can bound the ϵ_γ-optimal robust discounted value function induced by π_γ as $\mathbf{Sp}\big((1-\gamma)V^{\pi_\gamma}_{\gamma,\mathcal{P}}\big)\le\epsilon$. Similarly, we show we can bound $\mathbf{Sp}\big((1-\gamma)V^{\pi^*}_{\gamma,\mathcal{P}}\big)\le\epsilon$, before we combine these two results with that of Lemma 11.1 to derive our bound in the Theorem 3.4.

The result in Theorem 3.4 shows that we provide a concrete choice for the reduction discount factor, ensuring that the robust DMDP and robust AMDP share the same near-optimal policy. Our framework allows any algorithm designed for robust DMDPs to be directly applied to solve robust AMDPs, with theoretical performance guarantees, effectively bypassing the challenges of robust AMDPs by transforming them into the well-studied DMDP domain.

In the following sections, we investigate the applicability of our reduction framework from two key perspectives: data efficiency and scalability. These demonstrate the framework's ability to optimize long-term performance in data-intensive and large-scale practical scenarios.

4. Sample Complexity for Robust RL under Average Reward

In this section, we study the data efficiency of our reduction framework for robust RL under average reward, to characterize the total number of samples required to identify an ϵ-optimal policy π through our reduction framework.

We first present a general result on the sample complexity of robust AMDP reduction. Specifically, leveraging Theorem 3.4, the sample complexity of robust AMDP algorithms aligns with that of robust DMDP algorithms with a specific discount factor. This result is formally stated as follows:

Theorem 4.1. *Consider any algorithm $\mathcal{Y}$ optimizing the robust discounted reward. Denote the sample complexity of $\mathcal{Y}$ to identify an ϵ_γ-optimal policy (w.r.t. the discounted reward) by $\mathcal{N}(\mathcal{S},\mathcal{A},\mathcal{P},\gamma,\epsilon_\gamma)$. Then, we can identify an ϵ-optimal policy for the robust average reward through reduction and algorithm $\mathcal{Y}$, with a sample complexity of*

$$\mathcal{N}\left(\mathcal{S},\mathcal{A},\mathcal{P},1-\frac{\epsilon}{\mathcal{H}},\mathcal{H}\right). \tag{8}$$

This result holds universally, regardless of the uncertainty set models or algorithms used for robust DMDPs. More importantly, it provides a basis for studying sample complexity and data requirements for optimizing the average reward under model uncertainty. In the following, we analyze the sample complexity of robust AMDPs under different uncertainty sets, offering a concrete understanding of their data efficiency and our framework.

Remark 4.2. *Although Theorem 4.1 holds for general uncertainty sets, existing sample complexity studies of robust RL focus on the 'ball-structured' uncertainty sets:*

$$\mathcal{P}^a_s = \{q\in\Delta(\mathcal{S}) : D(q||\mathsf{P}^a_s)\le R\}, \tag{9}$$

where P^a_s is the nominal kernel, R is the uncertainty radius indicating the uncertainty level, and D is some distribution distance measure or divergence function. Hence, we similarly focus on uncertainty sets with this structure.

In our subsequent analysis, we focus on the generative model setting, where the agent can arbitrarily generate samples following the nominal kernel P. This setting has been widely adopted for sample complexity analysis under both non-robust RL (Agarwal et al., 2020; Li et al., 2020; Zurek & Chen, 2023) and robust RL (Shi et al., 2023; Panaganti & Kalathil, 2022; Wang et al., 2023e). While robust RL has also been explored in other settings, such as offline (Shi & Chi, 2022; Blanchet et al., 2024; Wang et al., 2024b; Panaganti et al., 2022; Liu & Xu, 2024; Wang et al., 2024a) and

[4] We prove in the Appendix that the value $\mathcal{H}$ exists and is finite.

[5] For a robust DMDP, an ϵ_γ-optimal policy is some policy π_γ such that $\max_\pi V^\pi_{\gamma,\mathcal{P}}(s) - V^{\pi_\gamma}_{\gamma,\mathcal{P}}(s)\le\epsilon_\gamma, \forall s\in\mathcal{S}$.

online (Lu et al., 2024) scenarios, and sample complexity results for robust average reward under these settings can be directly derived from Theorem 4.1, we concentrate on the generative setting, allowing us to focus on the challenges of the robust average reward framework itself rather than the complexities of data collection under restricted settings.

We develop a model-based reduction meta-algorithm for robust AMDPs. Specifically, after generating the data, we construct an estimate of the nominal kernel, and build an empirical uncertainty set centered around it with the same D and R in (9). We then solve a DMDP with a specific discount factor to identify a near-optimal policy. Our algorithm is presented in Algorithm 1.

Algorithm 1 Model-based algorithm for robust AMDPs

```
1: Input: $N$ nominal samples $\{(s,a,s_i')_{i=1}^N, s_i' \sim \mathsf{P}_s^a\}$ under each $(s,a)$ pair, uncertainty level $R$, robust bias span $\mathcal{H}$, and accuracy level $\epsilon$
2: Initialization: $Q \leftarrow 0$
3: Estimate transition model $\hat{\mathsf{P}}_{s,s'}^a = \frac{\sum_i \mathbf{1}_{(s,a,s_i')=(s,a,s')}}{N}$
4: Construct empirical uncertainty set $\hat{\mathcal{P}}$ centered at $\hat{\mathsf{P}}$: $\hat{\mathcal{P}}_s^a = \{q \in \Delta(\mathcal{S}) : D(q||\hat{\mathsf{P}}_s^a) \leq R\}$
5: Set $\gamma \leftarrow 1 - \frac{\epsilon}{\mathcal{H}}$, $\epsilon_\gamma \leftarrow \mathcal{H}$
6: Obtain an $\epsilon_\gamma$-optimal policy $\hat{\pi}_\gamma$ for the robust DMDP $(\mathcal{S}, \mathcal{A}, r, \hat{\mathcal{P}}, \gamma)$ with value iteration
7: Output: $\hat{\pi}_\gamma$
```

Remark 4.3. *In Line 6 of Algorithm 1, we need to identify an ϵ_γ-optimal policy for robust DMDPs, which can be done through robust value/policy iteration (Panaganti & Kalathil, 2022; Yang et al., 2021; Shi et al., 2023). For commonly used uncertainty sets, e.g., when D is total variation or χ^2 divergence, the algorithms can be implemented with polynomial computational complexity (Iyengar, 2005) and exponentially fast convergence rate.*

Next, we present the sample complexity of Algorithm 1 for the robust AMDP in the following theorem.

Theorem 4.4. *Consider an uncertainty set defined by total variation (TV) or χ^2 divergence (CS). Let C be some universal constant. If the total number of samples satisfies*

$$NSA \geq \frac{CSA\min\{\frac{1}{R}, \mathcal{H}^{1+\boldsymbol{1}_{R<\frac{1}{\mathcal{H}}}}\}\log\left(\frac{SAN}{\delta}\right)}{\epsilon^2}, (TV)$$

$$NSA \geq \frac{CSA\mathcal{H}^2(1+R)\log\left(\frac{SAN}{\delta}\right)}{\epsilon^2}, (CS)$$

then, with probability at least $1-4\delta$, $\hat{\pi}_\gamma$ is ϵ-optimal under the robust average reward.

The result shows that Algorithm 1 requires at most $\tilde{\mathcal{O}}\left(\frac{SA\mathcal{H}^2}{\epsilon^2}\right)$ samples to identify an ϵ-optimal policy for both robust AMDPs. We note that the minimax lower bound on the sample complexity for non-robust AMDPs is $\tilde{\Omega}\left(\frac{SAH}{\epsilon^2}\right)$, with H being the non-robust optimal span (Zurek & Chen, 2023; Wang et al., 2022; Jin & Sidford, 2021). Thus, our results are near-optimal under these uncertainty sets, aligning with the lower bound in terms of S, A, ϵ, with an additional dependence on $\mathcal{H}$. Notably, this is the first sample complexity analysis for robust RL under average reward, offering insights into data requirements for long-term reward optimization under model uncertainty. Furthermore, our framework demonstrates strong data efficiency, requiring nearly minimal samples to optimize the long-term reward, underscoring its potential in data-intensive scenarios.

Remark 4.5. *Theorem 4.4* ***is not*** *a direct combination of Theorem 4.1 with existing sample complexity of robust DMDPs (Panaganti & Kalathil, 2022; Shi et al., 2023). Specifically, the existing sample complexity of a robust DMDP with TV set is $\tilde{\mathcal{O}}\left(\frac{SA}{(1-\gamma)^2R\epsilon_\gamma^2}\right)$, and $\tilde{\mathcal{O}}\left(\frac{SA}{(1-\gamma)^4\epsilon_\gamma^2}\right)$ for CS set. By setting $\epsilon_\gamma = \mathcal{H}$ and $\gamma = 1-\frac{\epsilon}{\mathcal{H}}$, the resulting average reward complexity is $\tilde{\mathcal{O}}\left(\frac{SA}{\epsilon^2R}\right)$ and $\tilde{\mathcal{O}}\left(\frac{SA\mathcal{H}^2}{\epsilon^4}\right)$, respectively. Such higher complexity results are due to the higher dependence on $(1-\gamma)$ in the DMDP complexity, which becomes ϵ-order in the average reward setting through our framework. To achieve tighter results, we need to further tighten the complexity result for robust DMDPs. Specifically, we showed that with a reward perturbation technique (Li et al., 2020; Wang et al., 2022; Zurek & Chen, 2023) and more careful analysis involving the connection between robust DMDPs and AMDPs, we can reduce both sample complexities to $\tilde{\mathcal{O}}\left(\frac{SA\mathcal{H}^2}{(1-\gamma)^2\epsilon_\gamma^2}\right)$, which further result in the near-optimal complexity in Theorem 4.4.*

We can further obtain sample complexity for optimizing robust AMDPs for other types of uncertainty sets, by combining Theorem 4.1 with existing results for DMDPs. For instance, combining with (Panaganti & Kalathil, 2022), Algorithm 1 requires $\mathcal{O}\left(\frac{S^2A\mathcal{H}^2\log\left(\frac{SAN}{\delta}\right)}{\epsilon^4}\exp\left(\frac{\mathcal{H}}{\epsilon}\right)\right)$ samples to identify an ϵ-optimal policy under the Kullback-Leibler (KL) divergence. The results in (Clavier et al., 2023) imply a sample complexity of $\mathcal{O}\left(\frac{SA\mathcal{H}^2\log\left(\frac{SAN}{\delta}\right)}{\epsilon^4}\right)$ for the l_p-normed uncertainty set. Our framework thus provides the first concrete sample complexity characterization for robust AMDPs, establishing a foundation for studying robust long-term performance optimization.

Remark 4.6. *As demonstrated in prior studies on non-robust AMDPs (Zurek & Chen, 2023; Wang et al., 2022; 2023b), tightening the complexity bounds for non-robust DMDPs leads to optimal AMDP complexity in the corresponding reduction framework. We therefore attribute our sub-optimal sample complexity to the loose bounds of robust DMDPs rather than limitations in our reduction framework.*

Refining the complexity analysis of robust DMDPs to align with the optimal bounds for robust AMDPs is an important direction for future research.

5. Scalable Robust RL for Average Reward

In this section, we explore scalable approaches for optimizing the average reward under model uncertainty, aiming to facilitate long-term performance optimization in practical, large-scale problems. Specifically, we show that our framework overcomes the major challenges in solving large scale robust AMDPs, and further enables us to design scalable algorithms for the robust average reward, greatly enhancing the scalability and applicability of our methods.

When the problem scales are large, function approximation (FA) techniques have been extensively studied. FA methods aim to approximate the value functions by some low-dimensional function class, $\mathcal{F} = \{f_\theta(s) : \theta \in \Theta \subseteq \mathbb{R}^d, d \ll S\}$, to find some $\theta^* \in \Theta$ such that $V(s) \approx f_{\theta^*}(s), \forall s$. Two commonly studied function classes for FA are the linear function class and neural networks (Cai et al., 2019; Bhatnagar et al., 2009; Wai et al., 2019). We focus on linear FA to illustrate the scalability of our framework, but our method can be directly extended to neural networks or other function classes.

Linear FA is based on a set of feature functions $\big\{\phi : \mathcal{S} \to \mathbb{R}^d\big\}$. In robust DMDPs, the robust value function is approximated using a linear function: $V_\theta(s) \triangleq \phi(s)^\top \theta \approx V^*_{\gamma,\mathcal{P}}(s)$, where $\theta \in \mathbb{R}^d$ is some weight vector. Despite extensive studies on linear FA in robust DMDPs (Tamar et al., 2014; Xu & Mannor, 2010; Wang & Zou, 2021; Zhou et al., 2024; Roy et al., 2017; Badrinath & Kalathil, 2021), it remains largely understudied for (robust) AMDPs. In the non-robust average reward policy evaluation problem, we aim to approximate the bias h^π_{P} to estimate g^π_{P}, by approximating the solution to the non-robust Bellman equation

$$h(s) = \sum_a \pi(a|s)(r(s,a) - g + \mathsf{P}^a_s h). \tag{10}$$

However, two major challenges hinder the study of linear FA for average reward MDPs. The first is that (10) admits **non-unique solutions** (Puterman, 1994; Wan et al., 2021; Wan & Sutton, 2022). Specifically, besides the average reward and relative value function pair $(g^\pi_{\mathsf{P}}, h^\pi_{\mathsf{P}})$, any pair[6] $(g^\pi_{\mathsf{P}}, h^\pi_{\mathsf{P}} + ce)$ with any $c \in \mathbb{R}$ is also a solution to (10). This implies that the weight vector may not be unique, leading to a divergent algorithm. A common approach to address this issue is to impose additional assumptions on the feature functions (Zhang et al., 2021a; Tsitsiklis & Van Roy, 1999; Konda & Tsitsiklis, 1999; Yu & Bertsekas, 2009), ensuring that the all-one vector e does not lie within the span of $\{\phi(s)\}$, i.e., $e \neq \Phi\theta, \forall \theta$. Under this assumption, there exists a unique θ that minimizes the approximation loss. We note that the robust average reward also encounters this issue, but the higher dimensionality of its solution space (compared to the one-dimensional solution space in the non-robust setting) (Wang et al., 2024c) may result in more restrictive assumptions, making it even more challenging to directly apply linear FA to the robust average reward setting.

Another challenge arises from **unstable convergence** under the average reward. Even in the tabular setting, algorithm design and convergence analysis remain limited due to the instability caused by the Span semi-norm multi-step contraction of the Bellman operator (Puterman, 1994), compared to the norm contraction in the discounted setting. Existing studies address this issue either by introducing additional offset functions to stabilize convergence (Wan et al., 2021; Wan & Sutton, 2022; Puterman, 1994; Bertsekas, 2011) or by only ensuring convergence to a solution set (Zhang et al., 2021b). As demonstrated in (Wang et al., 2023d;e), similar challenges persist in the robust average reward setting, and addressing them is expected to involve even greater complexity.

Noting these two issues, it can be much more challenging to directly apply FA techniques to the robust average reward. However, our reduction framework simplifies this by transforming the complex robust average reward problem into the more manageable discounted reward setting. This bypasses the aforementioned difficulties and facilitates the design of scalable algorithms. As an immediate application, we extend our reduction framework to the robust natural actor-critic (NAC) algorithm with linear FA (Zhou et al., 2024), resulting in a robust NAC algorithm for robust AMDPs, detailed in Algorithm 2. Algorithm 2 consists of two key steps: (i) a robust TD step to update the weight vector θ for value approximation, and (ii) an actor step to update the policy[7], neither of which suffers from the issues mentioned above.

Note that the feature vectors in Algorithm 2 are predefined, e.g., by tile coding (Sutton, 1995), Fourier Basis (Konidaris et al., 2011), or randomly generated (Ghavamzadeh et al., 2010), by the learner prior to training.

We further characterize the convergence of Algorithm 2 (see Theorem 14.2 for the formal theorem statement).

Theorem 5.1. *(Informal) Under some additional standard assumptions, set $T = \tilde{\mathcal{O}}(\frac{\mathcal{H}}{\epsilon})$ in Algorithm 2, then π_{w_T} is an $\mathcal{O}(\epsilon + \epsilon_c)$-optimal policy, where ϵ_c is the critic error due to the representation power of the linear function class. The total sample complexity is $\tilde{\mathcal{O}}(\mathcal{H}^2\epsilon^{-3})$.*

Our results demonstrate that, without requiring any addi-

[6] $e = (1, 1, ..., 1)$ is the all-one vector.

[7] The detailed algorithms are provided in Section 14.

Algorithm 2 Reduction Robust Natural Actor-Critic

```
Input: $T, \mathcal{H}$, base functions $\{\phi\}$
Initialization: $\theta_0$ for value function approximation and $w_0$ for policy parametrization
$\gamma \leftarrow 1 - \frac{\epsilon}{\mathcal{H}}$
for $t = 0, 1, \dots, T-1$ do
    Robust critic updates $\theta_t$ with Algorithm 3
    Robust natural actor updates $w_{t+1}$ with Algorithm 4
end for
Output: $w_T$
```

tional assumptions on the base functions, our framework enables the optimization of robust average reward with function approximation with a stable convergent algorithm, underscoring its scalability. This represents the first solution for large-scale robust long-term reward optimization, offering both convergence and performance guarantees.

Remark 5.2. *In addition to function approximation, designing model-free algorithms for robust AMDPs can also improve scalability. Our framework also facilitates model-free algorithm design, which we discuss in detail in Section 15.*

6. Numerical Experiments

We first assess the effectiveness and efficiency of our framework in tabular settings, aiming to verify that with the reduction discount factor in Theorem 3.4, optimizing the robust DMDP approximately optimizes the robust AMDP.

Our experiments are conducted on the Garnet problem (Archibald et al., 1995), where the nominal transition kernel and reward functions are randomly generated from Gaussian distributions. We consider a Garnet problem with 20 states and 8 actions, constructing uncertainty sets using Total Variation (TV) and Chi-Square (CS) divergences, both with radius $R = 0.2$. Additional results for other problems are provided in Section 9.

To verify the effectiveness of our framework, we optimize robust DMDPs with different discount factors and plot the robust average reward $g_{\mathcal{P}}^{\pi_\gamma}$ of the learned policy π_γ using Algorithm 1 from (Wang et al., 2023d). As a baseline, we plot the optimal robust average reward $g_{\mathcal{P}}^*$, obtained via Algorithm 2 from (Wang et al., 2023d). Additionally, we estimate $\mathcal{H}$ using Algorithm 2 from (Wang et al., 2023e) and compute the corresponding reduction factor for different accuracy levels ϵ, as described in Theorem 3.4. The results, shown in Figure 1, indicate that the robust average reward can be approximately optimized through a robust DMDP with a sufficiently large discount factor. Furthermore, the robust average reward corresponding to each ϵ falls within the prescribed accuracy level, confirming the effectiveness of our framework and the validity of the reduction factor.

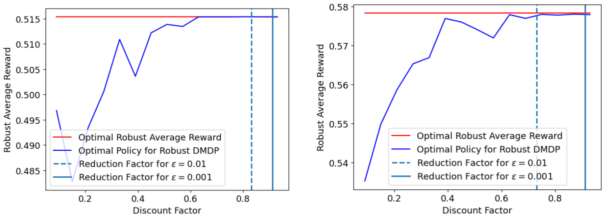

Figure 1. Effectiveness under the Garnet problem.

Next, we evaluate the data efficiency of our model-based framework under the same settings by optimizing empirical robust DMDPs with a reduction factor corresponding to $\epsilon = 0.01$ across different dataset sizes. The results, presented in Figure 2, show that our model-based algorithm converges to the optimal policy as dataset size N increases and achieves near-optimal performance with a limited amount of data. These findings highlight the data efficiency of our framework, demonstrating its ability to optimize long-term rewards with fewer samples. Since the computational cost for finding the worst performance (i.e., robust policy evaluation) in TV and CS uncertainty sets are $O\big(S\log(S)\big)$ (Iyengar, 2005) under the tabular setting, their estimation can be tractable.

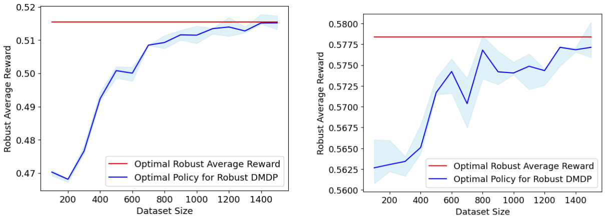

Figure 2. Efficiency under the Garnet problem.

We then demonstrate the scalability of our framework in large-scale problems by showing that robust average reward optimization can be equivalently formulated as a robust DMDP with a sufficiently large reduction factor. Specifically, we implement Algorithm 2 in the MuJoCo simulation environments (Todorov et al., 2012) for robust DMDPs with varying discount factors and evaluate the average reward performance under model uncertainty. Implementation details are provided in Section 14.

Our experiments are conducted in two continuous large-scale environments: Walker2d-v3 and Hopper-v3. We first train our algorithms in the nominal environments to learn the optimal policy for the given robust discounted reward. To estimate the robust average reward, we deploy the learned policies in perturbed environments, where at each evaluation epoch, we randomly sample parameters within the perturbation interval and apply them to each joint. The average reward is then computed over 30 independently perturbed environments.

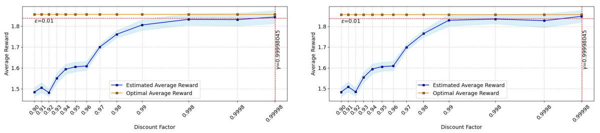

Figure 3. Scalability under Walker2d-v3.

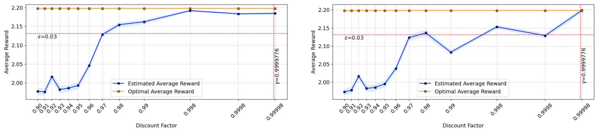

Figure 4. Scalability under Hopper-v3.

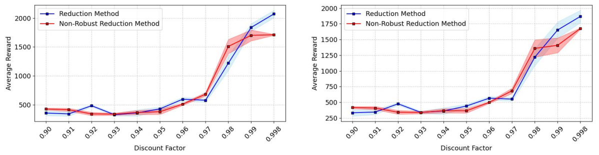

Figure 5. Neural network approximation under Walker2d-v3.

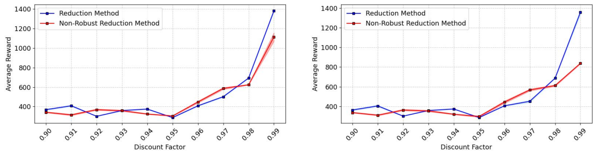

Figure 6. Neural network approximation under Hopper-v3.

As noted in Remark 3.3, we assume we have the knowledge of $\mathcal{H}$, however, we would like to note that in practice this is rarely the case. In lieu of a pre-obtained $\mathcal{H}$, we can equivalently use any upper-bound on $\mathcal{H}$ and see that our results still hold. In our experimentation on Walker2d-v3 and Hopper-v3, we opted to estimate the diameter as it could be easily obtained. For each of the independently perturbed environments, we sampled 1,000 randomly obtained trajectories and recorded how many steps it took the given trajectory to terminate. Based on this, we recorded the highest value for the number of steps across each of the 30 environments before averaging these values together to obtain an estimate for the diameter. The results, presented in Figures 3 and 4, show that as the discount factor increases, robust DMDP optimization aligns more closely with robust average reward optimization. In both environments, performance stabilizes when γ is large, further validating the effectiveness of our framework. Moreover, our results demonstrate that, when combined with function approximation techniques, our framework effectively scales to large-scale problems.

We emphasize that our reduction framework is independent of any specific discounted algorithm used, thus it is **not** limited to only linear approximation. To verify this claim, we present additional experimentation using neural network approximation in Figures 5 and 6. As the results show, our reduction framework remains valid even with neural network approximation, and it is more robust than the non-robust reduction method. We present additional results for this in Section 14.

7. Related Work

7.1. Comparison with Prior Art

In this section, we compare our results with the most related work (Grand-Clément & Petrik, 2024), which also explores reduction of robust AMDPs to DMDPs. However, it focuses on bounding the Blackwell discount factor. The Blackwell discount factor, denoted as γ_{bw}, is defined such that the optimal robust policy for the DMDP with discount factor γ_{bw} is also optimal to any robust DMDP with discount factor $\gamma \geq \gamma_{bw}$. This implies that any optimal policy for the robust γ_{bw}-DMDP also optimizes the robust AMDP.

With this notation, a robust AMDP can similarly be reduced to a robust DMDP with γ_{bw}. To enable this reduction, (Grand-Clément & Petrik, 2024) derives an upper bound on γ_{bw}. While this reduction-based framework shares some similarity with our approach, our method offers several advantages compared to their approach.

First, the results in (Grand-Clément & Petrik, 2024) are only applied to robust MDPs that meet two conditions: (1) the uncertainty sets are defined using l_1- or l_∞-norms; and (2) the nominal kernels are rational numbers, i.e., $\mathsf{P}_s^a = \frac{n_{s,a}}{m_{s,a}}$, for some $n_{s,a}, m_{s,a} \in \mathbb{N}$. In contrast, we only require the uncertainty set to be compact and the Markov chain induced by each kernel in it to be a unichain. The restrictions in their work are due to the reliance on the separation bounds of algebraic numbers for a rational polynomial in their proofs. In contrast, we developed a more detailed structural characterization of robust AMDPs, allowing us to obtain more general results. More importantly, the resulting sample complexity from (Grand-Clément & Petrik, 2024) is less favorable than ours. The robust Blackwell discount factor is bounded by $\gamma_{bw} \leq 1 - \frac{C}{S^S m^{S^2}}$, where m is the minimal denominator of the nominal kernel. Using this result in our reduction framework leads to an exponentially large sample complexity, making the results impractical. We note that the Blackwell optimality can be overly stringent for merely solving a robust AMDP, and hence it results in significantly worse sample complexity compared to ours.

7.2. Other Related Work

Robust AMDPs. Robust AMDPs studies are quite limited. Model-based robust AMDPs were first studied in (Tewari & Bartlett, 2007) for a specific finite-interval uncertainty set, which is further extended to more general models in recent works including (Wang et al., 2023d; Grand-Clement et al., 2023). A game-based method is also proposed in (Chatterjee et al., 2023). These works reveal the fundamental structure of robust AMDPs, illustrating their connections to robust DMDPs. However, all of them are model-based with asymptotic convergence, whereas we developed sample complexity analysis.

Robust DMDPs. Robust DMDPs were studied in (Iyengar, 2005; Nilim & El Ghaoui, 2004; Bagnell et al., 2001; Wiesemann et al., 2013; Lim et al., 2013), where the uncertainty set is assumed to be fully known. This inspired model-based methods for robust MDPs, where the learner first estimates a model, then solves the estimated model using robust dynamic programming (Zhang et al., 2021c; Panaganti & Kalathil, 2022; Yang et al., 2021; Shi et al., 2023). The studies were also extended to the model-free setting for more practical settings (Roy et al., 2017; Badrinath & Kalathil, 2021; Wang & Zou, 2021; 2022; Liu et al., 2022b; Zhou et al., 2021; Goyal & Grand-Clement, 2018; Kaufman & Schaefer, 2013; Ho et al., 2018; 2021; Wang et al., 2024c). Our work shows that the sample complexity of solving robust AMDPs can be transformed to that of solving DMDPs, enabling us to leverage the extensive prior work on the discounted setting.

Non-robust AMDPs. Early contributions to non-robust AMDPs involve fundamental characterizations of the problem and the development of model-based methods (Puterman, 1994; Bertsekas, 2011). Recently, model-free methods in the tabular setting, e.g., (Abounadi et al., 2001; Wan et al., 2021; Wan & Sutton, 2022), have been developed and demonstrated convergence to the optimal average reward. The sample complexity of non-robust AMDPs has been a recent focus (Wang et al., 2022; Zhang & Xie, 2023). Among them, similar reduction-based methods are considered in, e.g. (Wang et al., 2022; 2023b; Zurek & Chen, 2023), achieving the optimal complexity. Extending such frameworks to robust settings is notably challenging, due to the inherent complexity of the robust average reward setting, stemming from the non-linearity of the Bellman operator and a more complicated high-dimensional solution space for the robust Bellman equation (Wang et al., 2023e).

8. Conclusion

In this work, we studied the fundamental connection between robust AMDPs and DMDPs. We reveal that obtaining an optimal policy for the robust average reward is equivalent to achieving a near-optimal policy under the discounted reward with a specific reduction discount factor, based on which we constructed a reduction-based framework that solves robust RL with average reward effectively. Our framework is adaptable to any method or oracle and versatile across uncertainty sets. It offers two key benefits: data efficiency and scalability, as illustrated by our design of both tabular and function approximation algorithms, along with their sample complexity analysis and experimental performance. Our results represent the first concrete solutions for robust RL in the average reward setting under the mild assumption (unichain), advancing the understanding of optimizing long-term RL performance under model uncertainty.

Acknowledgments

This work was supported by DARPA under Agreement No. HR0011-24-9-0427 and NSF under Award CCF-2106339. We would also like to show our appreciation to the reviewers of this work for their impactful insight during the revision of this paper.

Impact Statement

The work contained in this paper is to advance the field of robust reinforcement learning. As there are many different potential impacts to different fields and that we provide a theory-based general framework, we do not feel that there is a need to delve into specific potential impacts here.

References

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

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

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

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

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

Bartlett, P. L. and Tewari, A. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. *arXiv preprint arXiv:1205.2661*, 2012.

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

Bhatnagar, S., Precup, D., Silver, D., Sutton, R. S., Maei, H., and Szepesvári, C. Convergent temporal-difference learning with arbitrary smooth function approximation. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, volume 22, pp. 1204–1212, 2009.

Blanchet, J., Lu, M., Zhang, T., and Zhong, H. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *Advances in Neural Information Processing Systems*, 36, 2024.

Blanchet, J. H. and Glynn, P. W. Unbiased Monte Carlo for optimization and functions of expectations via multi-level randomization. In *2015 Winter Simulation Conference (WSC)*, pp. 3656–3667. IEEE, 2015.

Blanchet, J. H., Glynn, P. W., and Pei, Y. Unbiased multilevel Monte Carlo: Stochastic optimization, steady-state simulation, quantiles, and other applications. *arXiv preprint arXiv:1904.09929*, 2019.

Cai, Q., Yang, Z., Lee, J. D., and Wang, Z. Neural temporal-difference learning converges to global optima. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 11312–11322, 2019.

Chatterjee, K., Goharshady, E. K., Karrabi, M., Novotný, P., and Žikelić, u. Solving long-run average reward robust mdps via stochastic games. *arXiv preprint arXiv:2312.13912*, 2023.

Clavier, P., Pennec, E. L., and Geist, M. Towards minimax optimality of model-based robust reinforcement learning. *arXiv preprint arXiv:2302.05372*, 2023.

Dugac, P. Sur la correspondance de borel et le théorème de dirichlet-heine-weierstrass-borel-schoenflies-lebesgue. *Archives internationales d'histoire des sciences*, 39(122): 69–110, 1989.

Ghavamzadeh, M., Lazaric, A., Maillard, O., and Munos, R. LSTD with random projections. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2010.

Goyal, V. and Grand-Clement, J. Robust Markov decision process: Beyond rectangularity. *arXiv preprint arXiv:1811.00215*, 2018.

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

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

Ho, C. P., Petrik, M., and Wiesemann, W. Fast Bellman updates for robust MDPs. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1979–1988. PMLR, 2018.

Ho, C. P., Petrik, M., and Wiesemann, W. Partial policy iteration for L1-robust Markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46, 2021.

Hu, Z. and Hong, L. J. Kullback-Leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, pp. 1695–1724, 2013.

Huber, P. J. A robust version of the probability ratio test. *Ann. Math. Statist.*, 36:1753–1758, 1965.

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

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

Kaufman, D. L. and Schaefer, A. J. Robust modified policy iteration. *INFORMS Journal on Computing*, 25(3):396–410, 2013.

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

Konda, V. and Tsitsiklis, J. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Konidaris, G., Osentoski, S., and Thomas, P. Value function approximation in reinforcement learning using the fourier basis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 25, pp. 380–385, 2011.

Lan, G. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer Nature, 2020.

Levin, D. A. and Peres, Y. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in neural information processing systems*, 33:12861–12872, 2020.

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

Liu, R.-Z., Pang, Z.-J., Meng, Z.-Y., Wang, W., Yu, Y., and Lu, T. On efficient reinforcement learning for full-length game of starcraft ii. *Journal of Artificial Intelligence Research*, 75:213–260, 2022a.

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

Liu, Z., Bai, Q., Blanchet, J., Dong, P., Xu, W., Zhou, Z., and Zhou, Z. Distributionally robust Q-learning. In *Proc. International Conference on Machine Learning (ICML)*, pp. 13623–13643. PMLR, 2022b.

Lu, M., Zhong, H., Zhang, T., and Blanchet, J. Distributionally robust reinforcement learning with interactive data collection: Fundamental hardness and near-optimal algorithm. *arXiv preprint arXiv:2404.03578*, 2024.

Müller, A. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29 (2):429–443, 1997.

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

Panaganti, K. and Kalathil, D. Sample complexity of robust reinforcement learning with a generative model. *arXiv preprint arXiv:2112.01506*, 2021.

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

Panaganti, K., Xu, Z., Kalathil, D., and Ghavamzadeh, M. Robust reinforcement learning using offline data. *arXiv preprint arXiv:2208.05129*, 2022.

Puterman, M. L. Markov decision processes: Discrete stochastic dynamic programming, 1994.

Rahimian, H., Bayraksan, G., and De-Mello, T. H. Effective scenarios in multistage distributionally robust optimization with a focus on total variation distance. *SIAM Journal on Optimization*, 32(3):1698–1727, 2022.

Roy, A., Xu, H., and Pokutta, S. Reinforcement learning under model mismatch. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 3046–3055, 2017.

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

Shi, L., Li, G., Wei, Y., Chen, Y., Geist, M., and Chi, Y. The curious price of distributional robustness in reinforcement learning with a generative model. *arXiv preprint arXiv:2305.16589*, 2023.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Sobel, M. J. The variance of discounted markov decision processes. *Journal of Applied Probability*, 19(4):794–802, 1982.

Sutton, R. S. Generalization in reinforcement learning: Successful examples using sparse coarse coding. *Advances in neural information processing systems*, 8, 1995.

Tamar, A., Mannor, S., and Xu, H. Scaling up robust MDPs using function approximation. In *Proc. International Conference on Machine Learning (ICML)*, pp. 181–189. PMLR, 2014.

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

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

Tsitsiklis, J. N. and Van Roy, B. Average cost temporal-difference learning. *Automatica*, 35(11):1799–1808, 1999.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Wai, H.-T., Hong, M., Yang, Z., Wang, Z., and Tang, K. Variance reduced policy evaluation with smooth function approximation. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 5784–5795, 2019.

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

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

Wang, G. and Wang, T. Unbiased multilevel Monte Carlo methods for intractable distributions: Mlmc meets mcmc. *arXiv preprint arXiv:2204.04808*, 2022.

Wang, H., Shi, L., and Chi, Y. Sample complexity of offline distributionally robust linear markov decision processes. *arXiv preprint arXiv:2403.12946*, 2024a.

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

Wang, S., Blanchet, J., and Glynn, P. Optimal sample complexity for average reward Markov decision processes. *arXiv preprint arXiv:2310.08833*, 2023a.

Wang, S., Blanchet, J., and Glynn, P. Optimal sample complexity of reinforcement learning for mixing discounted Markov decision processes. *arXiv preprint arXiv:2302.07477*, 2023b.

Wang, S., Si, N., Blanchet, J., and Zhou, Z. A finite sample complexity bound for distributionally robust q-learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3370–3398. PMLR, 2023c.

Wang, Y. and Zou, S. Online robust reinforcement learning with model uncertainty. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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

Wang, Y., Velasquez, A., Atia, G., Prater-Bennette, A., and Zou, S. Robust average-reward Markov decision processes. In *Proc. Conference on Artificial Intelligence (AAAI)*, 2023d.

Wang, Y., Velasquez, A., Atia, G. K., Prater-Bennette, A., and Zou, S. Model-free robust average-reward reinforcement learning. In *International Conference on Machine Learning*, pp. 36431–36469. PMLR, 2023e.

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

Wang, Y., Zou, S., and Wang, Y. Model-free robust reinforcement learning with sample complexity analysis. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024c.

Wei, H., Chen, J., Ji, X., Qin, H., Deng, M., Li, S., Wang, L., Zhang, W., Yu, Y., Linc, L., et al. Honor of kings arena: an environment for generalization in competitive reinforcement learning. *Advances in Neural Information Processing Systems*, 35:11881–11892, 2022.

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

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

Yang, W., Zhang, L., and Zhang, Z. Towards theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics. *arXiv preprint arXiv:2105.03863*, 2021.

Yu, H. and Bertsekas, D. P. Convergence results for some temporal difference methods based on least squares. *IEEE Transactions on Automatic Control*, 54(7):1515–1531, 2009.

Zha, D., Xie, J., Ma, W., Zhang, S., Lian, X., Hu, X., and Liu, J. Douzero: Mastering doudizhu with self-play deep reinforcement learning. In *international conference on machine learning*, pp. 12333–12344. PMLR, 2021.

Zhang, H., Chen, H., Xiao, C., Li, B., Liu, M., Boning, D., and Hsieh, C.-J. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems*, 33: 21024–21037, 2020.

Zhang, S., Wan, Y., Sutton, R. S., and Whiteson, S. Average-reward off-policy policy evaluation with function approximation. In *Proc. International Conference on Machine Learning (ICML)*, pp. 12578–12588. PMLR, 2021a.

Zhang, S., Zhang, Z., and Maguluri, S. T. 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, 2021b.

Zhang, X., Chen, Y., Zhu, X., and Sun, W. Robust policy gradient against strong data corruption. *arXiv preprint arXiv:2102.05800*, 2021c.

Zhang, Y. and Ross, K. W. On-policy deep reinforcement learning for the average-reward criterion. In *Proc. International Conference on Machine Learning (ICML)*, pp. 12535–12545. PMLR, 2021.

Zhang, Z. and Xie, Q. 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.

Zhou, R., Liu, T., Cheng, M., Kalathil, D., Kumar, P., and Tian, C. Natural actor-critic for robust reinforcement learning with function approximation. *Advances in neural information processing systems*, 36, 2024.

Zhou, Z., Bai, Q., Zhou, Z., Qiu, L., Blanchet, J., and Glynn, P. 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.

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

Appendix

9. Additional Experiments for Tabular Settings

Building upon the results in 6, we aim to verify our theoretical findings by presenting concrete experimental validation using the true transition kernel as well as the empirical kernel in figures 7 and 8 respectively. In each setting, both the nominal transition kernel and the reward function are generated via a normal distribution for 20 states and 8 actions. Given this, we then construct the uncertainty set with either TV or CS divergence, both with a radius $R = 0.2$. For each fixed value of γ, we estimate π_γ for the associated robust DMDP using robust value iteration (Iyengar, 2005). We then use Algorithm 1 in (Wang et al., 2023d) on this policy to obtain its robust average reward $g_{\mathcal{P}}^{\pi_\gamma}$. Additionally, we plot the optimal robust average reward using Algorithm 2 from (Wang et al., 2023d). By iterating through discount factors and obtaining the robust average reward, our results show that as $\gamma \to 1$, the estimated policies obtained from the corresponding robust DMDPs yield an increasingly higher robust average reward, thus we converge to the optimal policy for the robust AMDP. To further solidify this point, given arbitrarily chosen values for ϵ, we can estimate the robust optimal bias span $\mathcal{H}$ using the work of (Wang et al., 2023e). With these values, we know the optimal discount factor necessary to obtain the optimal policy of the robust AMDP.

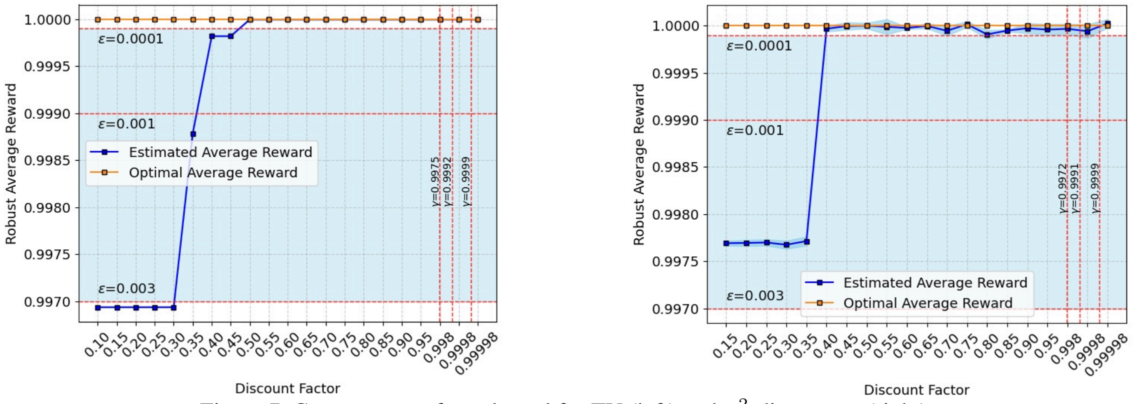

Figure 7. Convergence of true kernel for TV (left) and χ^2-divergence (right).

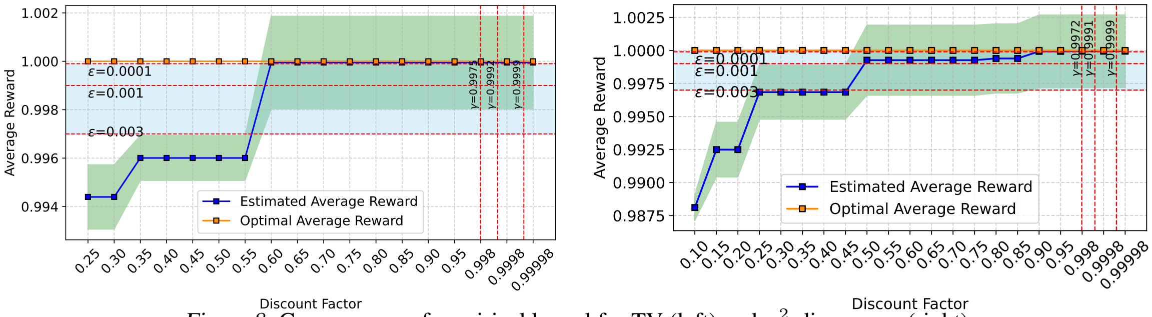

Figure 8. Convergence of empirical kernel for TV (left) and χ^2-divergence (right).

We also wish to compare other methods to our reduction framework. Due to the novelty of our work, we opted to compare our method to two baselines for robust AMDPs, robust value iteration (RVI) (Wang et al., 2023d), and robust relative value iteration (RRVI) (Wang et al., 2023e), under the tabular Garnet problem due to the baseline method's asymptotic convergence guarantees. We set the reduction factor to be 0.99 in our framework (corresponds to $\epsilon = 0.001$. As we show in Figure 9, our method obtains a better policy within the same number of steps, thus achieving state-of-the-art performance in robust average reward optimization.

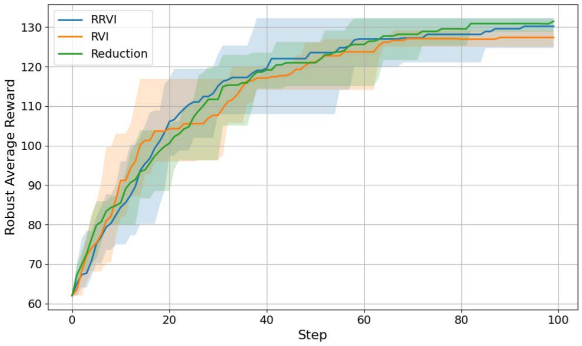

Figure 9. Comparison with AMDP baselines

10. Discussion on $\mathcal{H}$

For general uncertainty sets, we can obtain an upper bound of $\mathcal{H}$ with an additional assumption.

Lemma 10.1. *Assume there exists some positive integer $n > 0$ and some positive value ρ, such that*

$$(\mathsf{P}^n)_{s,x} \geq \rho, \forall \mathsf{P} \in \mathcal{P}, (s,x) \in \mathcal{S} \times \mathcal{S}. \tag{11}$$

Then,

$$\boldsymbol{Sp}(h^\pi_{\mathsf{P}}) \leq \frac{1}{1-\rho} \frac{1}{1-(1-\rho)^{\frac{1}{n}}}. \tag{12}$$

Proof. Without loss of generality, we prove the case $n = 1$. The results in other cases can be derived in a similar way. First, it holds that

$$\mathsf{P}^\pi_{x,y} \geq \rho \eta_{\mathsf{P}}(y), \tag{13}$$

where $\mathsf{P}^\pi_{\cdot,\cdot}$ is the reduced transition kernel by π, and η_{P} is the stationary distribution. Define a stochastic matrix Q through the equation

$$\mathsf{P} = \rho \mathsf{P}^* + (1-\rho) Q. \tag{14}$$

We claim that

$$(\mathsf{P})^k = (1-(1-\rho)^k)\mathsf{P}^* + (1-\rho)^k Q^k\ . \tag{15}$$

To show this, we use induction. Clearly (15) holds when $k = 1$. Assume that (15) holds when $k = n$:

$$(\mathsf{P})^n = (1-(1-\rho)^n)\mathsf{P}^* + (1-\rho)^n Q^n. \tag{16}$$

Then,

$$\begin{aligned}(\mathsf{P})^{1+n} &= (1-(1-\rho)^n)\mathsf{P}^*\mathsf{P} + (1-\rho)^n Q^n \mathsf{P} \\ &= (1-(1-\rho)^n)\mathsf{P}^* + (1-\rho)^n Q^n(\rho\mathsf{P}^* + (1-\rho)Q) \\ &= (1-(1-\rho)^{n+1})\mathsf{P}^* + (1-\rho)^{n+1} Q^{n+1},\end{aligned} \tag{17}$$

which proves the claim (15). Thus, rearranging terms implies

$$\|\mathsf{P}^k - \mathsf{P}^*\| \leq (1-\rho)^k \|Q^k - \mathsf{P}^*\| \leq \frac{1}{1-\rho}(1-\rho)^k. \tag{18}$$

Then, it holds that

$$\|h^\pi_{\mathsf{P}}\| = \|\lim_{T\to\infty}\sum_{t=0}^{T}\mathsf{P}^t r - T\mathsf{P}^* r\| \leq \lim_{T\to\infty}\|\sum_{t=0}^{T}\mathsf{P}^t r - T\mathsf{P}^* r\| \leq \sum_{t=0}^{\infty}(1-\rho)^{t-1} \leq \frac{1}{\rho(1-\rho)}, \tag{19}$$

which further implies that

$$\mathbf{Sp}(h^\pi_{\mathsf{P}}) \leq 2\|h^\pi_{\mathsf{P}}\| \leq \frac{2}{\rho(1-\rho)}. \tag{20}$$

□

Remark 10.2. *As a sufficient condition for the assumption in Lemma 10.1, it is commonly assumed that the Markov chains induced by any deterministic policy and transition kernels in the uncertainty set are aperiodic and irreducible (Levin & Peres, 2017). Since the aperiodicity can be ensured by applying the aperiodicity transformation (Puterman, 1994), the only additional assumption required is the irreducibility of the induced Markov chains.*

Remark 10.3. *Another well adopted assumption that can imply an upper bound on $\mathcal{H}$ is to assume uniformly Geometric ergodic chains. Namely, there exists constants m, ρ, such that for all deterministic policy and kernel,*

$$\|(\mathsf{P}^\pi)^n - \eta^\pi\| \leq m\rho^{-n}. \tag{21}$$

In this case, it holds that

$$\mathcal{H} \leq \max_{\mathsf{P}\in\mathcal{P}} \boldsymbol{Sp}\left(\lim_{T\to\infty}\sum_{i=1}^{T-1}((P^\pi)^i - \eta^\pi)r\right) \leq \frac{2m}{1-\rho}. \tag{22}$$

11. Proof of Theorem 3.4

We first prove some important lemmas.

Lemma 11.1. *Under Assumption 3.1, for any $\gamma \in (0,1)$ and any stationary policy π, it holds that*

$$\|g^\pi_{\mathcal{P}} - (1-\gamma)V^\pi_{\gamma,\mathcal{P}}\|_\infty \leq \boldsymbol{Sp}((1-\gamma)V^\pi_{\gamma,\mathcal{P}}).$$

Proof. Recalling the definition of g^π_{P} in (1), we have that for any kernel P and policy π,

$$g^\pi_{\mathsf{P}} = \mathsf{P}^*_\pi r_\pi,$$

where r_π is the induced immediate reward function: $r_\pi(s) = \mathbb{E}[r(s,a)|a\sim\pi(s)]$, and P^*_π is the Cesàro limit (Puterman, 1994) of the transition kernel P that follows the policy π.

On the other hand, we have that

$$V^\pi_{\gamma,\mathsf{P}} = (I - \gamma\mathsf{P}_\pi)^{-1} r_\pi. \tag{23}$$

Note that:

$$\mathsf{P}^*_\pi(I - \gamma\mathsf{P}_\pi) = \mathsf{P}^*_\pi - \gamma\mathsf{P}^*_\pi = (1-\gamma)\mathsf{P}^*_\pi, \tag{24}$$

where the above equation is from the fact that for every policy π, $\mathsf{P}_\pi\mathsf{P}^*_\pi = \mathsf{P}^*_\pi\mathsf{P}_\pi = \mathsf{P}^*_\pi\mathsf{P}^*_\pi = \mathsf{P}^*_\pi$. Thus,

$$\mathsf{P}^*_\pi = (1-\gamma)\mathsf{P}^*_\pi(I-\gamma\mathsf{P}_\pi)^{-1}. \tag{25}$$

Hence,

$$g^\pi_{\mathsf{P}} = \mathsf{P}^*_\pi r_\pi = (1-\gamma)\mathsf{P}^*_\pi(I-\gamma\mathsf{P}_\pi)^{-1} r_\pi = \mathsf{P}^*_\pi\cdot(1-\gamma)V^\pi_{\gamma,\mathsf{P}}. \tag{26}$$

Here, the last equation is from (23).

Since each row of P^*_π is a transition kernel of a unichain, $g^\pi_{\mathsf{P}}(s)$ is constant $\forall s$ and can be bounded as:

$$\min_s (1-\gamma)V^\pi_{\gamma,\mathsf{P}}(s) \le g^\pi_{\mathsf{P}} \le \max_s (1-\gamma)V^\pi_{\gamma,\mathsf{P}}(s). \tag{27}$$

Taking $\min\limits_{\mathsf{P}\in\mathcal{P}}$ on both sides of inequality (27) implies that

$$\min_{\mathsf{P}\in\mathcal{P}} \min_s (1-\gamma)V^\pi_{\gamma,\mathsf{P}}(s) \le \min_{\mathsf{P}\in\mathcal{P}} g^\pi_{\mathsf{P}} = g^\pi_{\mathcal{P}}. \tag{28}$$

By interchanging the order of two $\min$ operators, we have that

$$\min_s (1-\gamma) \min_{\mathsf{P}\in\mathcal{P}} V^\pi_{\gamma,\mathsf{P}}(s) = \min_s (1-\gamma)V^\pi_{\gamma,\mathcal{P}}(s) \le g^\pi_{\mathcal{P}}. \tag{29}$$

On the other hand, we have

$$g^\pi_{\mathcal{P}} = \min_{\mathsf{P}\in\mathcal{P}} \ g^\pi_{\mathsf{P}} = \min_{\mathsf{P}\in\mathcal{P}} \ \left[\mathsf{P}^*_\pi \cdot (1-\gamma)V^\pi_{\gamma,\mathsf{P}}\right] \ . \tag{30}$$

We denote by $Q_\gamma \in \mathcal{P}$ the worst case transition kernel of $V^\pi_{\gamma,\mathcal{P}}$. Then,

$$\begin{aligned} g^\pi_{\mathcal{P}} &\le Q^*_\gamma (1-\gamma) V^\pi_{\gamma,Q_\gamma} & (31)\\ &= Q^*_\gamma (1-\gamma) V^\pi_{\gamma,\mathcal{P}} & (32)\\ &\le \max_s (1-\gamma) V^\pi_{\gamma,\mathcal{P}}(s), & (33) \end{aligned}$$

where (31) is from (30), (32) is from the definition of Q_γ, and (33) is because for every distribution of Q_γ we have $\mathbb{E}_{Q_\gamma}[(1-\gamma)V^\pi_{\gamma,Q_\gamma}] \le \max\limits_s (1-\gamma)V^\pi_{\gamma,Q_\gamma}(s)$.

Combining (29) and (33), we get

$$\min_s (1-\gamma)V^\pi_{\gamma,\mathcal{P}}(s) \le g^\pi_{\mathcal{P}} \le \max_s (1-\gamma)V^\pi_{\gamma,\mathcal{P}}(s). \tag{34}$$

This implies that

$$\|g^\pi_{\mathcal{P}} - (1-\gamma)V^\pi_{\gamma,\mathcal{P}}\|_\infty \le \textbf{Sp}((1-\gamma)V^\pi_{\gamma,\mathcal{P}}), \tag{35}$$

which completes the proof. □

Lemma 11.2. *Under assumption 3.1, for any $\epsilon \in [0,\mathcal{H})$, if we set $\gamma := 1 - \frac{\epsilon}{\mathcal{H}}$, where $\mathcal{H} = \max\limits_{\mathsf{P}\in\mathcal{P}} \textbf{Sp}(h^{\pi^*}_{\mathsf{P}})$, then*

$$\textbf{Sp}((1-\gamma)V^{\pi_\gamma}_{\gamma,\mathcal{P}}) \le \epsilon.$$

Proof. Since the MDP is assumed to be a unichain, both $g^{\pi^*}_{\mathsf{P}}$ and $g^{\pi^*}_{\mathcal{P}}$ are constant. Moreover, $g^{\pi^*}_{\mathcal{P}}$ and $h^{\pi^*}_{\mathcal{P}}$ satisfy the robust Bellman optimality equation (Wang et al., 2023e):

$$(g^{\pi^*}_{\mathcal{P}} + h^{\pi^*}_{\mathcal{P}})(s) = \max_a \left\{ r(s,a) + \sigma_{\mathcal{P}^a_s}(h^{\pi^*}_{\mathcal{P}}) \right\}, \tag{36}$$

where $\sigma_{\mathcal{P}^a_s}(V) \triangleq \min\limits_{p\in\mathcal{P}^a_s} p^\top V$.

It is also known that for any $\gamma \in [0,1)$, the optimal robust discounted value function $V^{\pi_\gamma}_{\gamma,\mathcal{P}}$ satisfies the discounted Bellman equation:

$$V^{\pi_\gamma}_{\gamma,\mathcal{P}}(s) = \max_a \left\{ r(s,a) + \gamma\sigma_{\mathcal{P}^a_s}(V^{\pi_\gamma}_{\gamma,\mathcal{P}}) \right\}. \tag{37}$$

Next, we aim to rewrite the robust discounted Bellman equation to obtain a form similar to the Bellman equation for the average reward setting. First, we define

$$h^{\pi_\gamma,\pi^*}_{\gamma,\mathcal{P}} = V^{\pi_\gamma}_{\gamma,\mathcal{P}} - \frac{1}{1-\gamma} g^{\pi^*}_{\mathcal{P}}. \tag{38}$$

Note the fact that $h^{\pi_\gamma,\pi^*}_{\gamma,\mathcal{P}}$ and $V^{\pi_\gamma}_{\gamma,\mathcal{P}}$ have the same span because $g^{\pi^*}_{\mathcal{P}}$ is a constant:

$$\mathbf{Sp}(h^{\pi_\gamma,\pi^*}_{\gamma,\mathcal{P}}) = \mathbf{Sp}(V^{\pi^*}_{\gamma,\mathcal{P}}). \tag{39}$$

Now, we substitute (38) into (37) to obtain an equation similar to (36):

$$\begin{aligned} V^{\pi_\gamma}_{\gamma,\mathcal{P}}(s) = \left(h^{\pi_\gamma,\pi^*}_{\gamma,\mathcal{P}} + \frac{1}{1-\gamma} g^{\pi^*}_{\mathcal{P}}\right)(s) &= \max_a \left\{ r(s,a) + \gamma \sigma_{\mathcal{P}^a_s}\left(h^{\pi_\gamma,\pi^*}_{\gamma,\mathcal{P}} + \frac{1}{1-\gamma} g^{\pi^*}_{\mathcal{P}}\right)\right\} \\ &= \max_a \left\{ r(s,a) + \gamma \sigma_{\mathcal{P}^a_s}\left(h^{\pi_\gamma,\pi^*}_{\gamma,\mathcal{P}}\right) + \frac{\gamma}{1-\gamma} g^{\pi^*}_{\mathcal{P}}\right\}, \end{aligned} \tag{40}$$

where (40) is because $\frac{\gamma}{1-\gamma} g^{\pi^*}_{\mathcal{P}}$ is a constant and can be taken from the support function. The above equation can be written as

$$\left(h^{\pi_\gamma,\pi^*}_{\gamma,\mathcal{P}} + \frac{1}{1-\gamma} g^{\pi^*}_{\mathcal{P}}\right)(s) = \max_a \left\{ r(s,a) + \gamma \sigma_{\mathcal{P}^a_s}\left(h^{\pi_\gamma,\pi^*}_{\gamma,\mathcal{P}}\right)\right\} + \frac{\gamma}{1-\gamma} g^{\pi^*}_{\mathcal{P}}. \tag{41}$$

This implies that

$$\left(h^{\pi_\gamma,\pi^*}_{\gamma,\mathcal{P}} + g^{\pi^*}_{\mathcal{P}}\right)(s) = \max_a \left\{ r(s,a) + \gamma \sigma_{\mathcal{P}^a_s}\left(h^{\pi_\gamma,\pi^*}_{\gamma,\mathcal{P}}\right)\right\}. \tag{42}$$

Combine (36) and (42) and $\forall s \in S$ we have

$$\begin{aligned} |h^{\pi^*}_{\mathcal{P}} - h^{\pi_\gamma,\pi^*}_{\gamma,\mathcal{P}}(s)| &= |\max_a \left\{ r(s,a) + \sigma_{\mathcal{P}^a_s}(h^{\pi^*}_{\mathcal{P}})\right\} - \max_a \left\{ r(s,a) + \gamma \sigma_{\mathcal{P}^a_s}\left(h^{\pi_\gamma,\pi^*}_{\gamma,\mathcal{P}}\right)\right\}| & (43) \\ &\le |\max_a \{\sigma_{\mathcal{P}^a_s}(h^{\pi^*}_{\mathcal{P}}) - \gamma \sigma_{\mathcal{P}^a_s}\left(h^{\pi_\gamma,\pi^*}_{\gamma,\mathcal{P}}\right)\}| & (44) \\ &= |\max_a \{\sigma_{\mathcal{P}^a_s}(h^{\pi^*}_{\mathcal{P}}) - \gamma \sigma_{\mathcal{P}^a_s}\left(h^{\pi^*}_{\mathcal{P}}\right) + \gamma \sigma_{\mathcal{P}^a_s}\left(h^{\pi^*}_{\mathcal{P}}\right) - \gamma \sigma_{\mathcal{P}^a_s}\left(h^{\pi_\gamma,\pi^*}_{\gamma,\mathcal{P}}\right)\}| & (45) \\ &\le |\max_a \{\gamma \sigma_{\mathcal{P}^a_s}\left(h^{\pi^*}_{\mathcal{P}}\right) - \gamma \sigma_{\mathcal{P}^a_s}\left(h^{\pi_\gamma,\pi^*}_{\gamma,\mathcal{P}}\right)\}| + |\max_a \{(1-\gamma)\sigma_{\mathcal{P}^a_s}(h^{\pi^*}_{\mathcal{P}})\}|. & (46) \end{aligned}$$

And since $|\sigma_{\mathcal{P}^a_s}(V) - \sigma_{\mathcal{P}^a_s}(W)| \le \|V - W\|_\infty$, and $|\sigma_{\mathcal{P}^a_s}(V)| \le \|V\|_\infty$:

$$|h^{\pi^*}_{\mathcal{P}} - h^{\pi_\gamma,\pi^*}_{\gamma,\mathcal{P}}(s)| \le \gamma \|h^{\pi^*}_{\mathcal{P}} - h^{\pi_\gamma,\pi^*}_{\gamma,\mathcal{P}}\|_\infty + (1-\gamma)\|h^{\pi^*}_{\mathcal{P}}\|_\infty. \tag{47}$$

Thus, it follows that,

$$\|h^{\pi^*}_{\mathcal{P}} - h^{\pi_\gamma,\pi^*}_{\gamma,\mathcal{P}}\|_\infty \le \|h^{\pi^*}_{\mathcal{P}}\|_\infty. \tag{48}$$

Now, we combine (39) and (48):

$$\mathbf{Sp}(V^{\pi_\gamma}_{\gamma,\mathcal{P}}) = \mathbf{Sp}(h^{\pi_\gamma,\pi^*}_{\gamma,\mathcal{P}}) \le 2\|h^{\pi_\gamma,\pi^*}_{\gamma,\mathcal{P}}\|_\infty \le 4\|h^{\pi^*}_{\mathcal{P}}\|_\infty. \tag{49}$$

From Theorem 3.1 of (Wang et al., 2023e), for any policy π, there exists a transition kernel $\mathsf{P}_V \in \mathcal{P}_W$ such that $h^\pi_{\mathcal{P}} = h^\pi_{\mathsf{P}_V} + ce$ for $c \in \mathbb{R}$, where e denotes the vector $(1,1,1,..,1) \in \mathbb{R}^{|S|}$. Hence, we have that

$$\mathbf{Sp}(V^{\pi_\gamma}_{\gamma,\mathcal{P}}) \le 4\|h^{\pi^*}_{\mathcal{P}}\|_\infty = 4\|h^{\pi^*}_{\mathsf{P}_V}\|_\infty, \tag{50}$$

and

$$\mathbf{Sp}(V^{\pi_\gamma}_{\gamma,\mathcal{P}}) \le 4\|h^{\pi^*}_{\mathsf{P}_V}\|_\infty \le 4\mathbf{Sp}(h^{\pi^*}_{\mathsf{P}^{\pi^*}_V}) \le 4\mathcal{H}, \tag{51}$$

which completes the proof. □

Combine the above two results, then we direcly have the following result.

Corollary 11.3. *Under Assumption 3.1, for any $\gamma \in (0,1)$ and the policy π_γ, it holds that*

$$\|\frac{g_{\mathcal{P}}^{\pi_\gamma}}{1-\gamma} - V_{\gamma,\mathcal{P}}^{\pi_\gamma}\|_\infty \le \mathcal{H}.$$

Lemma 11.4. *Under assumption 3.1, for any $\epsilon \in [0, \mathcal{H})$, if we set $\gamma := 1 - \frac{\epsilon}{\mathcal{H}}$, then* **Sp**$((1-\gamma)V_{\gamma,\mathcal{P}}^{\pi^*}) \le \epsilon$.

Proof. First, we utilize the definitions of the finite time-horizon reward function $V_{T,\mathsf{P}}^{\pi}(s) \triangleq \mathbb{E}_{\pi,\mathsf{P}}\left[\sum_0^{T-1} r_t | S_0 = s\right]$ and the other definition of bias $h_{\mathsf{P}}^{\pi}(s) \triangleq \lim_{T\to\infty}\left[V_{T,\mathsf{P}}^{\pi} - Tg_{\mathsf{P}}^{\pi}\right]$. Note that:

$$\begin{aligned} V_{\gamma,\mathcal{P}}^{\pi^*} &= \min_{\mathsf{P}\in\mathcal{P}} V_{\gamma,\mathsf{P}}^{\pi^*} && (52)\\ &= \min_{\mathsf{P}\in\mathcal{P}} \left(\lim_{T\to\infty} \sum_{t=0}^{T-1} \gamma^t \mathsf{P}_{\pi^*}^t r_{\pi^*}\right) && (53)\\ &= \min_{\mathsf{P}\in\mathcal{P}} \left(\lim_{T\to\infty} V_{T,\mathsf{P}}^{\pi^*} - (1-\gamma)\sum_{t=1}^{T-1}\gamma^{t-1}\mathsf{P}_{\pi^*}^t V_{T-t,\mathsf{P}}^{\pi^*}\right). && (54)\end{aligned}$$

Recall the worst case transition kernel Q_γ where $Q_\gamma \in \mathcal{P}$:

$$V_{\gamma,\mathcal{P}}^{\pi^*} = V_{\gamma,Q_\gamma}^{\pi^*} = \lim_{T\to\infty} V_{T,Q_\gamma}^{\pi^*} - (1-\gamma)\sum_{t=1}^{T-1}\gamma^{t-1}Q_{\gamma,\pi^*}^t V_{T-t,Q_\gamma}^{\pi^*}. \tag{55}$$

We have:

$$\textbf{Sp}(V_{\gamma,\mathcal{P}}^{\pi^*}) = \textbf{Sp}(V_{\gamma,Q_\gamma}^{\pi^*}) \le \lim_{T\to\infty}\sup \textbf{Sp}(V_{T,Q_\gamma}^{\pi^*}) + (1-\gamma)\sum_{t=1}^{T-1}\gamma^{t-1}\textbf{Sp}(V_{T-t,Q_\gamma}^{\pi^*}) \tag{56}$$

Now the objective is to find an upper bound for **Sp**$(V_{T,Q_\gamma}^{\pi^*})$. From the definition of $h_{Q_\gamma}^{\pi}$, we have

$$\begin{aligned} h_{Q_\gamma}^{\pi} &= \lim_{t\to\infty}(V_{t,Q_\gamma}^{\pi} - tg_{Q_\gamma}^{\pi}) && (57)\\ &= \lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}(V_{t,Q_\gamma}^{\pi} - tg_{Q_\gamma}^{\pi}) && (58)\\ &= \lim_{N\to\infty}\frac{1}{N}\sum_{t=T+1}^{N}(V_{t,Q_\gamma}^{\pi} - tg_{Q_\gamma}^{\pi}). && (59)\end{aligned}$$

Thus, we have

$$\begin{aligned} h_{Q_\gamma}^{\pi} &= \lim_{N\to\infty}\frac{1}{N}\sum_{t=T+1}^{N}[(V_{T,Q_\gamma}^{\pi} - Tg_{Q_\gamma}^{\pi}) + Q_{\gamma,\pi}^T V_{t-T,Q_\gamma}^{\pi} - (t-T)g_{Q_\gamma}^{\pi}]\\ &= (V_{T,Q_\gamma}^{\pi} - Tg_{Q_\gamma}^{\pi}) + Q_{\gamma,\pi}^T \lim_{N\to\infty}\frac{1}{N}\sum_{t=T+1}^{N-T}(V_{t-T,Q_\gamma}^{\pi} - (t-T)g_{Q_\gamma}^{\pi}) && (60)\\ &= V_{T,Q_\gamma}^{\pi} - Tg_{Q_\gamma}^{\pi} + Q_{\gamma,\pi}^T h_{Q_\gamma}^{\pi}.\end{aligned}$$

It follows that

$$V_{T,Q_\gamma}^{\pi} = Tg_{Q_\gamma}^{\pi} + h_{Q_\gamma}^{\pi} - Q_{\gamma,\pi}^T h_{Q_\gamma}^{\pi}. \tag{61}$$

Thus, we have

$$\mathbf{Sp}(V^{\pi}_{T,Q_\gamma}) \leq 2\mathbf{Sp}(h^{\pi}_{Q_\gamma}). \tag{62}$$

Note that $\mathbf{Sp}(h^{\pi}_{\mathsf{P}})$ is continuous in P and $\mathcal{P}$ is a compact set, thus $\mathbf{Sp}(h^{\pi}_{\mathsf{P}}) \leq \mathcal{H}, \forall \mathsf{P} \in \mathcal{P}$. Now we can combine the result from (56) and (62):

$$\begin{aligned}\mathbf{Sp}(V^{\pi^*}_{\gamma,\mathcal{P}}) = \mathbf{Sp}(V^{\pi^*}_{\gamma,Q_\gamma}) &\leq \lim_{T\to\infty} \sup \mathbf{Sp}(V^{\pi^*}_{T,Q_\gamma}) + (1-\gamma)\sum_{t=1}^{T-1}\gamma^{t-1}\mathbf{Sp}(V^{\pi^*}_{T-t,Q_\gamma}) &(63)\\ &\leq 2\mathbf{Sp}(h^{\pi^*}_{\mathsf{P}})\left(1+(1-\gamma)\sum_{t=1}^{\infty}\gamma^{t-1}\right) &(64)\\ &= 4\mathbf{Sp}(h^{\pi^*}_{\mathsf{P}}). &(65)\end{aligned}$$

Thus,

$$\mathbf{Sp}(V^{\pi^*}_{\gamma,\mathcal{P}}) \leq 4\mathcal{H}. \tag{66}$$

□

As a direct corollary, it holds that

Corollary 11.5. *Under Assumption 3.1, for any $\gamma \in (0,1)$, it holds that*

$$\|\frac{g^{\pi^*}_{\mathcal{P}}}{1-\gamma} - V^{\pi^*}_{\gamma,\mathcal{P}}\|_\infty \leq 4\mathcal{H}.$$

Proof. From Lemma 11.1 and Equation (66),it holds that

$$\|g^{\pi^*}_{\mathcal{P}} - (1-\gamma)V^{\pi^*}_{\gamma,\mathcal{P}}\|_\infty \leq \mathbf{Sp}((1-\gamma)V^{\pi^*}_{\gamma,\mathcal{P}}) \leq 4(1-\gamma)\mathcal{H},$$

which completes the proof. □

We are now ready to prove the main theorem.

Theorem 11.6. *(Restatement of Theorem 3.4) Under Assumption 3.1, for any ϵ, if we set $\gamma := 1 - \frac{\epsilon}{\mathcal{H}}$, then any ϵ_γ-optimal policy $\hat{\pi}_\gamma$ for the robust γ-DMDP[8] is also an $\mathcal{O}(\epsilon)$-optimal policy for the robust AMDP, i.e.,*

$$g^{\pi^*}_{\mathcal{P}} - g^{\hat{\pi}_\gamma}_{\mathcal{P}} \leq \left(8 + \frac{5\epsilon_\gamma}{\mathcal{H}}\right)\epsilon. \tag{67}$$

Specifically, an $\mathcal{O}(\mathcal{H})$-optimal robust policy for the robust DMDP is an $\mathcal{O}(\epsilon)$-optimal robust policy under the average reward.

Proof. Under Assumption 3.1, for any $\epsilon \in (0,\mathcal{H}]$ and any $\delta \in (0,1]$, we consider $\gamma = 1 - \frac{\epsilon}{\mathcal{H}}$. Suppose π_γ is an optimal policy of the robust DMDP and $\hat{\pi}_\gamma$ is an ϵ_γ-optimal policy in the robust DMDP. Then,

$$\|(1-\gamma)V^{\pi_\gamma}_{\gamma,\mathcal{P}} - (1-\gamma)V^{\hat{\pi}_\gamma}_{\gamma,\mathcal{P}}\|_\infty \leq (1-\gamma)\epsilon_\gamma. \tag{68}$$

Considering (35), (51), and (66), we have that:

$$\begin{aligned}\mathbf{Sp}((1-\gamma)V^{\hat{\pi}_\gamma}_{\gamma,\mathcal{P}}) &\leq \mathbf{Sp}((1-\gamma)V^{\pi_\gamma}_{\gamma,\mathcal{P}}) + 2\|(1-\gamma)V^{\pi_\gamma}_{\gamma,\mathcal{P}} - (1-\gamma)V^{\hat{\pi}_\gamma}_{\gamma,\mathcal{P}}\|_\infty\\ &\leq 4\epsilon + 2(1-\gamma)\epsilon_\gamma. \end{aligned} \tag{69}$$

[8] For a robust DMDP, an ϵ-optimal policy is some policy π such that $V^{*}_{\gamma,\mathcal{P}}(s) - V^{\pi}_{\gamma,\mathcal{P}}(s) \leq \epsilon, \forall s \in \mathcal{S}$.

Moreover, because of the optimality of π_γ,

$$V^{\pi^*}_{\gamma,\mathcal{P}} \le V^{\pi_\gamma}_{\gamma,\mathcal{P}}, \tag{70}$$

recalling that π^* is an optimal policy of the robust AMDP.

By merging Lemma 11.1, (68), (69), and (70) we have:

$$\begin{aligned} g^{\pi^*}_{\mathcal{P}} &\le (1-\gamma)V^{\pi^*}_{\gamma,\mathcal{P}} + 4\epsilon + 2(1-\gamma)\epsilon_\gamma & (71)\\ &\le (1-\gamma)V^{\pi_\gamma}_{\gamma,\mathcal{P}} + 4\epsilon + 2(1-\gamma)\epsilon_\gamma & (72)\\ &\le (1-\gamma)V^{\hat{\pi}_\gamma}_{\gamma,\mathcal{P}} + 4\epsilon + 3(1-\gamma)\epsilon_\gamma & (73)\\ &\le g^{\hat{\pi}_\gamma}_{\mathcal{P}} + 8\epsilon + 5(1-\gamma)\epsilon_\gamma\ , & (74) \end{aligned}$$

which completes the proof. □

12. Proof of Theorem 4.4 Part 1

We first note that the result under the case $R \ge \frac{1}{\mathcal{H}}$ can be directly obtained from the existing result in (Shi et al., 2023). Specifically, it is shown in (Shi et al., 2023) that learning an ϵ_γ-optimal policy for a γ-discounted robust MDP requires samples of size

$$\mathcal{N}(\mathcal{S},\mathcal{A},\hat{\mathcal{P}},\gamma,\epsilon_\gamma) = \frac{CSA\log\frac{cSA}{\delta}}{(1-\gamma)^2\max\{R,1-\gamma\}\epsilon_\gamma^2}. \tag{75}$$

Thus, by setting $\gamma = 1-\frac{\epsilon}{\mathcal{H}}$ and $\epsilon_\gamma = \mathcal{H}$ as Theorem 4.1, we have that the complexity to learn an ϵ-optimal policy for the robust average reward MDP is

$$\frac{CSA\log\frac{cSA}{\delta}}{(1-\gamma)^2\max\{R,1-\gamma\}\epsilon_\gamma^2} = \frac{CSA\log\frac{cSA}{\delta}}{\epsilon^2 R} \le \frac{CSA\mathcal{H}\log\frac{cSA}{\delta}}{\epsilon^2}. \tag{76}$$

We hence mainly focus on the case of $R \le \frac{1}{\mathcal{H}}$. In the following proof, we denote ϵ_γ as ϵ.

We first introduce some notation. Let $\hat{\mathsf{P}}$ and $\hat{\mathcal{P}}$ be the estimated nominal kernel and estimated uncertainty set. For any policy π, let $\tilde{V}^\pi_{\gamma,\hat{\mathcal{P}}}, V^\pi_{\gamma,\hat{\mathcal{P}}}, V^\pi_{\gamma,\mathcal{P}}$ be the robust value function w.r.t. perturbed reward and estimated uncertainty set, unperturbed reward and estimated uncertainty set, unperturbed reward and the true uncertainty set. The optimal robust policy w.r.t. $\tilde{V}^\pi_{\hat{\mathcal{P}}}$ is denoted by $\tilde{\pi}^*$, and the corresponding optimal robust value functions are denoted by $\tilde{V}^*_{\gamma,\hat{\mathcal{P}}}$ and $\tilde{Q}^*_{\gamma,\hat{\mathcal{P}}}$.

We consider a general policy π, and we denote the worst-case kernel of any vector V under $\hat{\mathcal{P}}$ and $\mathcal{P}$ by $\hat{\mathsf{P}}^{\pi,V}_w$ and $\mathsf{P}^{\pi,V}_w$. Then, it holds that

$$\begin{aligned} V^\pi_{\gamma,\hat{\mathcal{P}}} - V^\pi_{\gamma,\mathcal{P}} &= r_\pi + \gamma\hat{\mathsf{P}}^{\pi,\widehat{V}}_w V^\pi_{\gamma,\hat{\mathcal{P}}} - \left(r_\pi + \gamma\mathsf{P}^{\pi,V}_w V^\pi_{\gamma,\mathcal{P}}\right)\\ &= \left(\gamma\hat{\mathsf{P}}^{\pi,\widehat{V}}_w V^\pi_{\gamma,\hat{\mathcal{P}}} - \gamma\mathsf{P}^{\pi,\widehat{V}}_w V^\pi_{\gamma,\hat{\mathcal{P}}}\right) + \left(\gamma\mathsf{P}^{\pi,\widehat{V}}_w V^\pi_{\gamma,\hat{\mathcal{P}}} - \gamma\mathsf{P}^{\pi,V}_w V^\pi_{\gamma,\mathcal{P}}\right)\\ &\overset{\text{(i)}}{\le} \gamma\Big(\mathsf{P}^{\pi,V}_w V^\pi_{\gamma,\hat{\mathcal{P}}} - \mathsf{P}^{\pi,V}_w V^\pi_{\gamma,\mathcal{P}}\Big) + \Big(\gamma\hat{\mathsf{P}}^{\pi,\widehat{V}}_w V^\pi_{\gamma,\hat{\mathcal{P}}} - \gamma\mathsf{P}^{\pi,\widehat{V}}_w V^\pi_{\gamma,\hat{\mathcal{P}}}\Big), \end{aligned}$$

where (i) holds by observing that

$$\mathsf{P}^{\pi,\widehat{V}}_w V^\pi_{\gamma,\hat{\mathcal{P}}} \le \mathsf{P}^{\pi,V}_w V^\pi_{\gamma,\hat{\mathcal{P}}}$$

due to the worst-case kernel. Rearranging terms leads to

$$V^\pi_{\gamma,\hat{\mathcal{P}}} - V^\pi_{\gamma,\mathcal{P}} \le \gamma\left(I - \gamma\mathsf{P}^{\pi,V}_w\right)^{-1}\Big(\hat{\mathsf{P}}^{\pi,\widehat{V}}_w V^\pi_{\gamma,\hat{\mathcal{P}}} - \mathsf{P}^{\pi,\widehat{V}}_w V^\pi_{\gamma,\hat{\mathcal{P}}}\Big). \tag{77}$$

Similarly, we can also deduce

$$\begin{aligned}
V^{\pi}_{\gamma,\hat{\mathcal{P}}} - V^{\pi}_{\gamma,\mathcal{P}} &= r_\pi + \gamma \hat{\mathsf{P}}^{\pi,\widehat{V}}_{w} V^{\pi}_{\gamma,\hat{\mathcal{P}}} - \left(r_\pi + \gamma \mathsf{P}^{\pi,V}_{w} V^{\pi}_{\gamma,\mathcal{P}}\right) \\
&= \left(\gamma \hat{\mathsf{P}}^{\pi,\widehat{V}}_{w} V^{\pi}_{\gamma,\hat{\mathcal{P}}} - \gamma \mathsf{P}^{\pi,\widehat{V}}_{w} V^{\pi}_{\gamma,\hat{\mathcal{P}}}\right) + \left(\gamma \mathsf{P}^{\pi,\widehat{V}}_{w} V^{\pi}_{\gamma,\hat{\mathcal{P}}} - \gamma \mathsf{P}^{\pi,V}_{w} V^{\pi}_{\gamma,\mathcal{P}}\right) \\
&\geq \gamma \left(\mathsf{P}^{\pi,\widehat{V}}_{w} V^{\pi}_{\gamma,\hat{\mathcal{P}}} - \mathsf{P}^{\pi,\widehat{V}}_{w} V^{\pi}_{\gamma,\mathcal{P}}\right) + \left(\gamma \hat{\mathsf{P}}^{\pi,\widehat{V}}_{w} V^{\pi}_{\gamma,\hat{\mathcal{P}}} - \gamma \mathsf{P}^{\pi,\widehat{V}}_{w} V^{\pi}_{\gamma,\hat{\mathcal{P}}}\right) \\
&\geq \gamma \left(I - \gamma \mathsf{P}^{\pi,\widehat{V}}_{w}\right)^{-1} \left(\hat{\mathsf{P}}^{\pi,\widehat{V}}_{w} V^{\pi}_{\gamma,\hat{\mathcal{P}}} - \mathsf{P}^{\pi,\widehat{V}}_{w} V^{\pi}_{\gamma,\hat{\mathcal{P}}}\right). \qquad (78)
\end{aligned}$$

Combining (77) and (78), we arrive at

$$\begin{aligned}
\|V^{\pi}_{\gamma,\hat{\mathcal{P}}} - V^{\pi}_{\gamma,\mathcal{P}}\|_\infty \leq \gamma \max \Big\{ &\Big\| \left(I - \gamma \mathsf{P}^{\pi,V}_{w}\right)^{-1} \left(\hat{\mathsf{P}}^{\pi,\widehat{V}}_{w} V^{\pi}_{\gamma,\hat{\mathcal{P}}} - \mathsf{P}^{\pi,\widehat{V}}_{w} V^{\pi}_{\gamma,\hat{\mathcal{P}}}\right) \Big\|_\infty, \\
&\Big\| \left(I - \gamma \mathsf{P}^{\pi,\widehat{V}}_{w}\right)^{-1} \left(\hat{\mathsf{P}}^{\pi,\widehat{V}}_{w} V^{\pi}_{\gamma,\hat{\mathcal{P}}} - \mathsf{P}^{\pi,\widehat{V}}_{w} V^{\pi}_{\gamma,\hat{\mathcal{P}}}\right) \Big\|_\infty \Big\}. \qquad (79)
\end{aligned}$$

By decomposing the error in a symmetric way, we can similarly obtain

$$\begin{aligned}
\|V^{\pi}_{\gamma,\hat{\mathcal{P}}} - V^{\pi}_{\gamma,\mathcal{P}}\|_\infty \leq \gamma \max \Big\{ &\Big\| \left(I - \gamma \hat{\mathsf{P}}^{\pi,\widehat{V}}_{w}\right)^{-1} \left(\hat{\mathsf{P}}^{\pi,V}_{w} V^{\pi}_{\gamma,\mathcal{P}} - \mathsf{P}^{\pi,V}_{w} V^{\pi}_{\gamma,\mathcal{P}}\right) \Big\|_\infty, \\
&\Big\| \left(I - \gamma \hat{\mathsf{P}}^{\pi,V}_{w}\right)^{-1} \left(\hat{\mathsf{P}}^{\pi,V}_{w} V^{\pi}_{\gamma,\mathcal{P}} - \mathsf{P}^{\pi,V}_{w} V^{\pi}_{\gamma,\mathcal{P}}\right) \Big\|_\infty \Big\}. \qquad (80)
\end{aligned}$$

12.1. Part A: $\|\tilde{V}^{\pi^*}_{\gamma,\hat{\mathcal{P}}} - V^{\pi^*}_{\gamma,\mathcal{P}}\|$

Consider $|\tilde{V}^{\pi^*}_{\gamma,\hat{\mathcal{P}}}(s) - V^{\pi^*}_{\gamma,\hat{\mathcal{P}}}(s)|$ for some state s. If $\tilde{V}^{\pi^*}_{\gamma,\hat{\mathcal{P}}}(s) - V^{\pi^*}_{\gamma,\hat{\mathcal{P}}}(s) > 0$, then

$$\begin{aligned}
|\tilde{V}^{\pi^*}_{\gamma,\hat{\mathcal{P}}}(s) - V^{\pi^*}_{\gamma,\hat{\mathcal{P}}}(s)| &= \tilde{V}^{\pi^*}_{\gamma,\hat{\mathcal{P}}}(s) - V^{\pi^*}_{\gamma,\hat{\mathcal{P}}}(s) \\
&= (I - \gamma \tilde{\mathsf{P}}^{\pi^*}_{w})^{-1} \tilde{r}^{\pi^*} - (I - \gamma \mathsf{P}^{\pi^*}_{w})^{-1} r^{\pi^*} \\
&\leq (I - \gamma \mathsf{P}^{\pi^*}_{w})^{-1} \tilde{r}^{\pi^*} - (I - \gamma \mathsf{P}^{\pi^*}_{w})^{-1} r^{\pi^*} \\
&\leq \frac{\epsilon}{6}, \qquad (81)
\end{aligned}$$

where $\tilde{\mathsf{P}}^{\pi^*}_{w}$ and $\mathsf{P}^{\pi^*}_{w}$ are the corresponding worst-case transition kernel.

A similar result can also be obtained for the other case, hence it holds that

$$\|\tilde{V}^{\pi^*}_{\gamma,\hat{\mathcal{P}}}(s) - V^{\pi^*}_{\gamma,\hat{\mathcal{P}}}\| \leq \frac{\epsilon}{6}. \qquad (82)$$

We thus have that

$$\|\tilde{V}^{\pi^*}_{\gamma,\hat{\mathcal{P}}} - V^{\pi^*}_{\gamma,\mathcal{P}}\| \leq \|\tilde{V}^{\pi^*}_{\gamma,\hat{\mathcal{P}}} - V^{\pi^*}_{\gamma,\hat{\mathcal{P}}}\| + \|V^{\pi^*}_{\gamma,\hat{\mathcal{P}}} - V^{\pi^*}_{\gamma,\mathcal{P}}\| \leq \frac{\epsilon}{6} + \|V^{\pi^*}_{\gamma,\hat{\mathcal{P}}} - V^{\pi^*}_{\gamma,\mathcal{P}}\|, \qquad (83)$$

and it suffices to study the second term $\|V^{\pi^*}_{\gamma,\hat{\mathcal{P}}} - V^{\pi^*}_{\gamma,\mathcal{P}}\|$. Applying (80) further implies that

$$\begin{aligned}
\|V^{\pi^*}_{\gamma,\hat{\mathcal{P}}} - V^{\pi^*}_{\gamma,\mathcal{P}}\| \leq \gamma \max \Big\{ &\Big\| \left(I - \gamma \hat{\mathsf{P}}^{\pi^*,\widehat{V}}_{w}\right)^{-1} \left(\hat{\mathsf{P}}^{\pi^*,V}_{w} V^{\pi^*}_{\gamma,\mathcal{P}} - \mathsf{P}^{\pi^*,V}_{w} V^{\pi^*}_{\gamma,\mathcal{P}}\right) \Big\|_\infty, \\
&\Big\| \left(I - \gamma \hat{\mathsf{P}}^{\pi^*,V}_{w}\right)^{-1} \left(\hat{\mathsf{P}}^{\pi^*,V}_{w} V^{\pi^*}_{\gamma,\mathcal{P}} - \mathsf{P}^{\pi^*,V}_{w} V^{\pi}_{\gamma,\mathcal{P}}\right) \Big\|_\infty \Big\}. \qquad (84)
\end{aligned}$$

We first consider the case that $\max\Big\{ \Big\| \left(I - \gamma \hat{\mathsf{P}}^{\pi^*,\widehat{V}}_{w}\right)^{-1} \left(\hat{\mathsf{P}}^{\pi^*,V}_{w} V^{\pi^*}_{\gamma,\mathcal{P}} - \mathsf{P}^{\pi^*,V}_{w} V^{\pi^*}_{\gamma,\mathcal{P}}\right) \Big\|_\infty, \Big\| \left(I - \gamma \hat{\mathsf{P}}^{\pi^*,V}_{w}\right)^{-1} \left(\hat{\mathsf{P}}^{\pi^*,V}_{w} V^{\pi^*}_{\gamma,\mathcal{P}} - \mathsf{P}^{\pi^*,V}_{w} V^{\pi}_{\gamma,\mathcal{P}}\right) \Big\|_\infty \Big\} = \Big\| \left(I - \gamma \hat{\mathsf{P}}^{\pi^*,\widehat{V}}_{w}\right)^{-1} \left(\hat{\mathsf{P}}^{\pi^*,V}_{w} V^{\pi^*}_{\gamma,\mathcal{P}} - \mathsf{P}^{\pi^*,V}_{w} V^{\pi^*}_{\gamma,\mathcal{P}}\right) \Big\|_\infty$.

We first apply the following lemma.

Lemma 12.1. *(Lemma 11 of (Shi et al., 2023)) Consider any $\delta \in (0,1)$. Setting $N \geq \log(\frac{18SAN}{\delta})$, with probability at least $1-\delta$, one has*

$$\left|\hat{\mathsf{P}}_w^{\pi^*,V} V_{\gamma,\mathcal{P}}^{\pi^*} - \mathsf{P}_w^{\pi^*,V} V_{\gamma,\mathcal{P}}^{\pi^*}\right| \leq 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\sqrt{\mathrm{Var}_{\mathsf{P}^{\pi^*}}(V_{\gamma,\mathcal{P}}^{\pi^*})} + \frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)}1. \tag{85}$$

Thus, it holds that

$$\begin{aligned}
&\left\| \left(I-\gamma\hat{\mathsf{P}}_w^{\pi^*,\hat{V}}\right)^{-1}\left(\hat{\mathsf{P}}_w^{\pi^*,V} V_{\gamma,\mathcal{P}}^{\pi^*} - \mathsf{P}_w^{\pi^*,V} V_{\gamma,\mathcal{P}}^{\pi^*}\right)\right\|_\infty \\
&\leq 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left\| \left(I-\gamma\hat{\mathsf{P}}_w^{\pi^*,\hat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\mathsf{P}^{\pi^*}}(V_{\gamma,\mathcal{P}}^{\pi^*})}\right\|_\infty + \frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)^2} \\
&\leq \underbrace{2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left\| \left(I-\gamma\hat{\mathsf{P}}_w^{\pi^*,\hat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\hat{\mathsf{P}}_w^{\pi^*,\hat{V}}}(V_{\gamma,\hat{\mathcal{P}}}^{\pi^*})}\right\|_\infty}_{A1} \\
&\quad + \underbrace{2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left\| \left(I-\gamma\hat{\mathsf{P}}_w^{\pi^*,\hat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\hat{\mathsf{P}}_w^{\pi^*,\hat{V}}}(V_{\gamma,\mathcal{P}}^{\pi^*}-V_{\gamma,\hat{\mathcal{P}}}^{\pi^*})}\right\|_\infty}_{A2} \\
&\quad + \underbrace{2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left\| \left(I-\gamma\hat{\mathsf{P}}_w^{\pi^*,\hat{V}}\right)^{-1}\sqrt{\left|\mathrm{Var}_{\hat{\mathsf{P}}_w^{\pi^*,\hat{V}}}(V_{\gamma,\mathcal{P}}^{\pi^*})-\mathrm{Var}_{\hat{\mathsf{P}}^{\pi^*}}(V_{\gamma,\mathcal{P}}^{\pi^*})\right|}\right\|_\infty}_{A3} \\
&\quad + \underbrace{2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left\| \left(I-\gamma\hat{\mathsf{P}}_w^{\pi^*,\hat{V}}\right)^{-1}\left(\sqrt{\mathrm{Var}_{\mathsf{P}^{\pi^*}}(V_{\gamma,\mathcal{P}}^{\pi^*})}-\sqrt{\mathrm{Var}_{\hat{\mathsf{P}}^{\pi^*}}(V_{\gamma,\mathcal{P}}^{\pi^*})}\right)\right\|_\infty}_{A4} + \frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)^2}.
\end{aligned} \tag{86}$$

Term $A1$. We note that $\mathsf{Q} \triangleq \hat{\mathsf{P}}_w^{\hat{V}} = \arg\min_{\mathsf{P}\in\hat{\mathcal{P}}} \mathsf{P}V_{\gamma,\hat{\mathcal{P}}}^{\pi^*}$, thus we have that

$$V_{\gamma,\hat{\mathcal{P}}}^{\pi^*} = r^{\pi^*} + \gamma\sigma_{\hat{\mathcal{P}}}^{\pi^*}(V_{\gamma,\hat{\mathcal{P}}}^{\pi^*}) = r^{\pi^*} + \gamma\mathsf{Q}^{\pi^*}V_{\gamma,\hat{\mathcal{P}}}^{\pi^*}; \tag{87}$$

On the other hand, we have that

$$V_{\gamma,\mathsf{Q}}^{\pi^*} = r^{\pi^*} + \gamma\mathsf{Q}^{\pi^*}V_{\gamma,\mathsf{Q}}^{\pi^*}, \tag{88}$$

hence both $V_{\gamma,\hat{\mathcal{P}}}^{\pi^*}$ and $V_{\gamma,\mathsf{Q}}^{\pi^*}$ are fixed points of the Bellman operator w.r.t. Q, which implies they are identical $V_{\gamma,\hat{\mathcal{P}}}^{\pi^*} = V_{\gamma,\mathsf{Q}}^{\pi^*}$.

Thus, the term $\left(I-\gamma\hat{\mathsf{P}}_w^{\pi^*,\hat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\hat{\mathsf{P}}_w^{\pi^*,\hat{V}}}(V_{\gamma,\hat{\mathcal{P}}}^{\pi^*})}$ can be rewritten as

$$\left(I-\gamma\hat{\mathsf{P}}_w^{\pi^*,\hat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\hat{\mathsf{P}}_w^{\pi^*,\hat{V}}}(V_{\gamma,\hat{\mathcal{P}}}^{\pi^*})} = \left(I-\gamma\mathsf{Q}^{\pi^*}\right)^{-1}\sqrt{\mathrm{Var}_{\mathsf{Q}^{\pi^*}}(V_{\gamma,\mathsf{Q}}^{\pi^*})}, \tag{89}$$

and the term $A1$ can be rewritten as

$$A1 = 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left\| \left(I-\gamma\mathsf{Q}^{\pi^*}\right)^{-1}\sqrt{\mathrm{Var}_{\mathsf{Q}^{\pi^*}}(V_{\gamma,\mathsf{Q}}^{\pi^*})}\right\|_\infty. \tag{90}$$

We then apply the following lemma to bound $A1$, which is the main result for the complexity improvement.

Lemma 12.2. *For any policy π, if $N \geq \mathcal{O}\left(\frac{\log\frac{SA}{(1-\gamma)\delta\epsilon}}{1-\gamma}\right)$, it holds with probability at least $1-\delta$ that*

$$\left\| \left(I-\gamma\mathsf{Q}^{\pi^*}\right)^{-1}\sqrt{\mathrm{Var}_{\mathsf{Q}^{\pi^*}}(V_{\gamma,\mathsf{Q}}^{\pi^*})}\right\|_\infty \leq \sqrt{\frac{c_1\mathcal{H}}{(1-\gamma)^2}}. \tag{91}$$

The proof of this lemma can be derived similarly to the ones in (Zurek & Chen, 2023). For completeness, we provide its proof in Section 12.4.

Thus, it holds that

$$A1 \leq \sqrt{\frac{2c_1 \log(\frac{18SAN}{\delta})\mathcal{H}}{(1-\gamma)^2 N}}. \tag{92}$$

Term $A2$. It holds that

$$\begin{aligned} A2 &= 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left\| \left(I - \gamma \hat{\mathsf{P}}_w^{\pi,\hat{V}}\right)^{-1} \sqrt{\mathrm{Var}_{\hat{\mathsf{P}}_w^{\pi,\hat{V}}}(V^{\pi^*}_{\gamma,\mathcal{P}} - V^{\pi^*}_{\gamma,\hat{\mathcal{P}}})}\right\|_\infty \\ &\leq 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)^2}}\left\|V^{\pi^*}_{\gamma,\mathcal{P}} - V^{\pi^*}_{\gamma,\hat{\mathcal{P}}}\right\|_\infty. \end{aligned} \tag{93}$$

Term $A3$. It holds that

$$\begin{aligned} &\left(I - \gamma \hat{\mathsf{P}}_w^{\pi,\hat{V}}\right)^{-1} \sqrt{\left|\mathrm{Var}_{\hat{\mathsf{P}}_w^{\pi,\hat{V}}}(V^{\pi^*}_{\gamma,\mathcal{P}}) - \mathrm{Var}_{\hat{\mathsf{P}}^{\pi^*}}(V^{\pi^*}_{\gamma,\mathcal{P}})\right|} \\ &= \left(I - \gamma \hat{\mathsf{P}}_w^{\pi,\hat{V}}\right)^{-1} \sqrt{\left|\prod^{\pi^*}(\mathrm{Var}_{\hat{\mathsf{P}}_w^{\pi,\hat{V}}}(V^{\pi^*}_{\gamma,\mathcal{P}}) - \mathrm{Var}_{\hat{\mathsf{P}}}(V^{\pi^*}_{\gamma,\mathcal{P}}))\right|} \\ &\leq \left(I - \gamma \hat{\mathsf{P}}_w^{\pi,\hat{V}}\right)^{-1} \sqrt{\left\|\mathrm{Var}_{\hat{\mathsf{P}}_w^{\pi,\hat{V}}}(V^{\pi^*}_{\gamma,\mathcal{P}}) - \mathrm{Var}_{\hat{\mathsf{P}}}(V^{\pi^*}_{\gamma,\mathcal{P}}))\right\|}. \end{aligned} \tag{94}$$

Note that both $\hat{\mathsf{P}}_w^{\pi,\hat{V}}$, $\hat{\mathsf{P}}$ belong to the uncertainty set $\hat{\mathcal{P}}$, hence $\|\hat{\mathsf{P}}_w^{\pi,\hat{V}} - \hat{\mathsf{P}}\|_1 \leq 2R$, which further implies that

$$\begin{aligned} &|\mathrm{Var}_{\hat{\mathsf{P}}_w^{\pi,\hat{V}}}(V^{\pi^*}_{\gamma,\mathcal{P}}) - \mathrm{Var}_{\hat{\mathsf{P}}}(V^{\pi^*}_{\gamma,\mathcal{P}}))|_{s,a} \\ &= |\mathrm{Var}_{\hat{\mathsf{P}}_w^{\pi,\hat{V}}}(V^{\pi^*}_{\gamma,\mathcal{P}} - \frac{1}{1-\gamma}g^{\pi^*}_{\mathcal{P}}) - \mathrm{Var}_{\hat{\mathsf{P}}}(V^{\pi^*}_{\gamma,\mathcal{P}}) - \frac{1}{1-\gamma}g^{\pi^*}_{\mathcal{P}})|_{s,a} \\ &\leq \|\hat{\mathsf{P}}_w^{\pi,\hat{V}} - \hat{\mathsf{P}}\|_1 \|V^{\pi^*}_{\gamma,\mathcal{P}} - \frac{1}{1-\gamma}g^{\pi^*}_{\mathcal{P}}\|^2 \\ &\leq 2R\mathcal{H}^2, \end{aligned} \tag{95}$$

where the last inequality is from Lemma 11.1 and Lemma 11.2.

Since $R \leq \frac{1}{\mathcal{H}}$, it holds that

$$A3 \leq 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left\| \left(I - \gamma \hat{\mathsf{P}}_w^{\pi,\hat{V}}\right)^{-1} \sqrt{2\mathcal{H}}\right\| \leq 2\sqrt{\frac{\log(\frac{18SAN}{\delta})\mathcal{H}}{N(1-\gamma)^2}}. \tag{96}$$

Term $A4$. We directly apply Lemma 6 of (Panaganti & Kalathil, 2022) and Lemma 11 of (Shi et al., 2023), and it implies that

$$A4 \leq \frac{4\log(\frac{18SAN}{\delta})}{N(1-\gamma)^2}. \tag{97}$$

We then plug (92), (93), (96) and (97) in (86), and we have that

$$\begin{aligned} &\|V^{\pi^*}_{\gamma,\hat{\mathcal{P}}} - V^{\pi^*}_{\gamma,\mathcal{P}}\| \\ &\leq \left\| \left(I - \gamma \hat{\mathsf{P}}_w^{\pi,\hat{V}}\right)^{-1} \left(\hat{\mathsf{P}}_w^{\pi,\hat{V}} V^{\pi}_{\gamma,\mathcal{P}} - \mathsf{P}_w^{\pi,V} V^{\pi}_{\gamma,\mathcal{P}}\right)\right\|_\infty \\ &\leq \frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)^2} + \sqrt{\frac{2c_1 \log(\frac{18SAN}{\delta})\mathcal{H}}{(1-\gamma)^2 N}} + 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)^2}}\left\|V^{\pi^*}_{\gamma,\mathcal{P}} - V^{\pi^*}_{\gamma,\hat{\mathcal{P}}}\right\|_\infty + 2\sqrt{\frac{\log(\frac{18SAN}{\delta})\mathcal{H}}{N(1-\gamma)^2}} + \frac{4\log(\frac{18SAN}{\delta})}{N(1-\gamma)^2}. \end{aligned} \tag{98}$$

We note that if we set $N \geq \frac{32\log(\frac{18SAN}{\delta})}{(1-\gamma)^2}$, it holds that

$$\|V^{\pi^*}_{\gamma,\hat{\mathcal{P}}} - V^{\pi^*}_{\gamma,\mathcal{P}}\| \leq \frac{C_1 \log(\frac{18SAN}{\delta})}{N(1-\gamma)^2} + \sqrt{\frac{2C_2\log(\frac{18SAN}{\delta})\mathcal{H}}{(1-\gamma)^2 N}}, \tag{99}$$

which completes the first term in (80).

To bound the second term in (80), following eq (69) in (Shi et al., 2023), we have that

$$\begin{aligned}&\left\| \left(I - \gamma \hat{\mathsf{P}}^{\pi^*,V}_w\right)^{-1} \left(\hat{\mathsf{P}}^{\pi^*,V}_w V^{\pi^*}_{\gamma,\mathcal{P}} - \mathsf{P}^{\pi^*,V}_w V^{\pi^*}_{\gamma,\mathcal{P}}\right)\right\|_\infty \\ &\leq 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}} \left\| \left(I - \gamma \hat{\mathsf{P}}^{\pi^*,V}_w\right)^{-1} \sqrt{\mathrm{Var}_{\mathsf{P}^{\pi^*}}(V^{\pi^*}_{\gamma,\mathcal{P}})}\right\|_\infty + \frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)^2}.\end{aligned} \tag{100}$$

Now applying Lemma 12.10,

$$\left\| \left(I - \gamma \hat{\mathsf{P}}^{\pi,V}_w\right)^{-1} \left(\hat{\mathsf{P}}^{\pi,V}_w V^{\pi}_{\gamma,\mathcal{P}} - \mathsf{P}^{\pi,V}_w V^{\pi}_{\gamma,\mathcal{P}}\right)\right\|_\infty \leq \frac{C_3 \log(\frac{18SAN}{\delta})}{N(1-\gamma)^2} + \sqrt{\frac{2C_4\log(\frac{18SAN}{\delta})\mathcal{H}^2}{(1-\gamma)^2 N}}. \tag{101}$$

We hence obtain the bound on $\|\tilde{V}^{\pi^*}_{\gamma,\hat{\mathcal{P}}} - V^{\pi^*}_{\gamma,\mathcal{P}}\|$ as follows:

$$\|\tilde{V}^{\pi^*}_{\gamma,\hat{\mathcal{P}}} - V^{\pi^*}_{\gamma,\mathcal{P}}\| \leq \frac{a_1 \log(\frac{18SAN}{\delta})}{N(1-\gamma)^2} + \sqrt{\frac{a_2\log(\frac{18SAN}{\delta})\mathcal{H}^2}{(1-\gamma)^2 N}} + \frac{\epsilon}{6}, \tag{102}$$

when $N \geq \frac{C\log(\frac{18SAN}{\delta})}{(1-\gamma)^2}$.

12.2. Part B: $\|\tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} - V^{\hat{\pi}}_{\gamma,\mathcal{P}}\|$

Similarly, we have that

$$\begin{aligned}\|\tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} - V^{\hat{\pi}}_{\gamma,\mathcal{P}}\| &\leq \|\tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} - \tilde{V}^{\hat{\pi}}_{\gamma,\mathcal{P}}\| + \|\tilde{V}^{\hat{\pi}}_{\gamma,\mathcal{P}} - V^{\hat{\pi}}_{\gamma,\mathcal{P}}\| \\ &\leq \frac{\epsilon}{6} + \|\tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} - \tilde{V}^{\hat{\pi}}_{\gamma,\mathcal{P}}\|,\end{aligned} \tag{103}$$

hence it suffices to bound the term $\|\tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} - \tilde{V}^{\hat{\pi}}_{\gamma,\mathcal{P}}\|$. By setting $\pi = \hat{\pi}$ in (79), we have that

$$\begin{aligned}\|\tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} - \tilde{V}^{\hat{\pi}}_{\gamma,\mathcal{P}}\|_\infty \leq \gamma \max\Big\{&\left\| \left(I - \gamma \mathsf{P}^{\hat{\pi},\tilde{\hat{V}}}_w\right)^{-1} \left(\hat{\mathsf{P}}^{\hat{\pi},\tilde{\hat{V}}}_w \tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} - \mathsf{P}^{\hat{\pi},\tilde{\hat{V}}}_w \tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}\right)\right\|_\infty, \\ &\left\| \left(I - \gamma \mathsf{P}^{\hat{\pi},\tilde{V}}_w\right)^{-1} \left(\hat{\mathsf{P}}^{\hat{\pi},\tilde{\hat{V}}}_w \tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} - \mathsf{P}^{\hat{\pi},\tilde{\hat{V}}}_w \tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}\right)\right\|_\infty\Big\}.\end{aligned} \tag{104}$$

We first bound the first term $\left\| \left(I - \gamma \mathsf{P}^{\hat{\pi},\tilde{\hat{V}}}_w\right)^{-1} \left(\hat{\mathsf{P}}^{\hat{\pi},\tilde{\hat{V}}}_w \tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} - \mathsf{P}^{\hat{\pi},\tilde{\hat{V}}}_w \tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}\right)\right\|_\infty$. To simplify notation, we rewrite $\mathsf{P}^{\hat{\pi},\tilde{\hat{V}}}_w$ as $\mathsf{P}^{\hat{\pi},\hat{V}}_w$ and $\hat{\mathsf{P}}^{\hat{\pi},\tilde{\hat{V}}}_w$ by $\hat{\mathsf{P}}^{\hat{\pi},\hat{V}}_w$.

We first introduce the following separation events:

$$\hat{\Omega}_\omega \triangleq \{\tilde{V}^*_{\gamma,\hat{\mathcal{P}}}(s) - \max_{a\neq\hat{\pi}^*(s)} \tilde{Q}^*_{\gamma,\hat{\mathcal{P}}}(s,a) \geq \omega, \forall s \in \mathcal{S}\}, \tag{105}$$

$$\Omega_\omega \triangleq \{\tilde{V}^*_{\gamma,\mathcal{P}}(s) - \max_{a\neq\pi^*(s)} \tilde{Q}^*_{\gamma,\mathcal{P}}(s,a) \geq \omega, \forall s \in \mathcal{S}\}. \tag{106}$$

These events indicate that there exists some threshold between the value functions of the optimal action and other actions, and there is no tie between the optimal robust value functions. It further implies that the optimal policy $\hat{\pi}^*$ and π^* are unique. As we shall show in Lemma 12.3, with a carefully chosen threshold, such events will occur with high probability.

Lemma 12.3. *Set $\omega = \frac{\xi\delta(1-\gamma)}{3SA^2}$, then both (105) and (106) occur with probability at least $1-\delta$.*

We then combine Lemma 14 from (Shi et al., 2023) and Lemma 9 in (Li et al., 2020) together to show the following result. Such a result allows us to decouple the dependence between $\hat{\pi}$ and other terms.

Lemma 12.4. *Consider any $\delta \in (0,1)$. Taking $N \geq \mathcal{O}\left(\frac{\log\left(\frac{54SAN^2}{(1-\gamma)\delta}\right)}{1-\gamma}\right)$, with probability at least $1-2\delta$, events (105) and (106) occur, and it holds that*

$$\left|\hat{\mathsf{P}}_w^{\widehat{\pi},\widehat{V}}\tilde{V}_{\gamma,\hat{\mathcal{P}}}^{\hat{\pi}} - \mathsf{P}_w^{\widehat{\pi},\widehat{V}}\tilde{V}_{\gamma,\hat{\mathcal{P}}}^{\hat{\pi}}\right| \leq 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\sqrt{\mathrm{Var}_{\mathsf{P}_{s,a}}(\tilde{V}_{\gamma,\hat{\mathcal{P}}}^{\hat{\pi}})}1 + \frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)}1. \tag{107}$$

With Lemma 12.4 in hand, we have

$$\begin{aligned}
&\left(I-\gamma\mathsf{P}_w^{\widehat{\pi},\widehat{V}}\right)^{-1}\left(\hat{\mathsf{P}}_w^{\widehat{\pi},\widehat{V}}\tilde{V}_{\gamma,\hat{\mathcal{P}}}^{\hat{\pi}} - \mathsf{P}_w^{\widehat{\pi},\widehat{V}}\tilde{V}_{\gamma,\hat{\mathcal{P}}}^{\hat{\pi}}\right)\\
&\overset{\text{(i)}}{\leq} \left(I-\gamma\mathsf{P}_w^{\widehat{\pi},\widehat{V}}\right)^{-1}\left|\hat{\mathsf{P}}_w^{\widehat{\pi},\widehat{V}}\tilde{V}_{\gamma,\hat{\mathcal{P}}}^{\hat{\pi}} - \mathsf{P}_w^{\widehat{\pi},\widehat{V}}\tilde{V}_{\gamma,\hat{\mathcal{P}}}^{\hat{\pi}}\right|\\
&\leq 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I-\gamma\mathsf{P}_w^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{P^{\widehat{\pi}}}(\tilde{V}_{\gamma,\hat{\mathcal{P}}}^{\hat{\pi}})} + \left(\frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2}\right)1\\
&\overset{\text{(ii)}}{\leq} \left(\frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2}\right)1 + \underbrace{2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I-\gamma\mathsf{P}_w^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\mathsf{P}_w^{\widehat{\pi},\widehat{V}}}(\tilde{V}_{\gamma,\hat{\mathcal{P}}}^{\hat{\pi}})}}_{=:B_1}\\
&\quad + \underbrace{2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I-\gamma\mathsf{P}_w^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\left|\mathrm{Var}_{P^{\widehat{\pi}}}(\tilde{V}_{\gamma,\hat{\mathcal{P}}}^{\hat{\pi}}) - \mathrm{Var}_{\mathsf{P}_w^{\widehat{\pi},\widehat{V}}}(\tilde{V}_{\gamma,\hat{\mathcal{P}}}^{\hat{\pi}})\right|}}_{=:B_2},
\end{aligned} \tag{108}$$

where (i) and (ii) hold by the fact that each row of $(1-\gamma)\left(I-\gamma\mathsf{P}_w^{\widehat{\pi},\widehat{V}}\right)^{-1}$ is a probability vector that falls into $\Delta(\mathcal{S})$.

Term $B1$. Similar to term $A1$, term $B1$ is equivalent to $2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}(I-\gamma\mathsf{P})^{-1}\sqrt{\mathrm{Var}_{\mathsf{P}}(V_{\gamma,\mathsf{P}})}$ with $\mathsf{P} = \mathsf{P}_w^{\widehat{\pi},\widehat{V}}$. Specifically, $\hat{\pi}$ can be viewed as the optimal policy for γ and the empirical uncertainty set. Thus, applying Corollary 11.3 implies that

$$\|\tilde{V}_{\gamma,\hat{\mathcal{P}}}^{\hat{\pi}} - \frac{\tilde{g}_{\hat{\mathcal{P}}}^*}{1-\gamma}\| \leq \mathcal{H}, \tag{109}$$

and hence

$$B1 \leq \sqrt{\frac{2d_1\log(\frac{18SAN}{\delta})\mathcal{H}^2}{(1-\gamma)^2N}}. \tag{110}$$

Term $B2$. Similar to term $A3$, term $B2$ can be bounded by noting that $R \leq \frac{1}{\mathcal{H}}$

$$B2 \leq 2\sqrt{\frac{\log(\frac{18SAN}{\delta})\mathcal{H}^2}{N(1-\gamma)^2}}. \tag{111}$$

Combine both bounds together, and we have that when $N \geq \frac{C\log\frac{SAN}{\delta}}{(1-\gamma)^2}$,

$$\left(I-\gamma\mathsf{P}_w^{\widehat{\pi},\widehat{V}}\right)^{-1}\left(\hat{\mathsf{P}}_w^{\widehat{\pi},\widehat{V}}\tilde{V}_{\gamma,\hat{\mathcal{P}}}^{\hat{\pi}} - \mathsf{P}_w^{\widehat{\pi},\widehat{V}}\tilde{V}_{\gamma,\hat{\mathcal{P}}}^{\hat{\pi}}\right) \leq \frac{D_1\log(\frac{18SAN}{\delta})}{N(1-\gamma)^2} + \sqrt{\frac{2D_2\log(\frac{18SAN}{\delta})\mathcal{H}^2}{(1-\gamma)^2N}} \tag{112}$$

with probability at least $1-\delta$. Similarly, we can get the bound on the second term of (104), which finally implies that

$$\left\|\tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}-\tilde{V}^{\hat{\pi}}_{\gamma,\mathcal{P}}\right\|_\infty \leq \frac{D_1\log(\frac{18SAN}{\delta})}{N(1-\gamma)^2}+\sqrt{\frac{2D_2\log(\frac{18SAN}{\delta})\mathcal{H}^2}{(1-\gamma)^2N}}. \tag{113}$$

12.3. Summing Up the Results

Combine the bounds obtained from both Part A and Part B, it holds that with probability at least $1-4\delta$,

$$V^*_{\gamma,\mathcal{P}}-V^{\hat{\pi}}_{\gamma,\mathcal{P}} \leq m_1\epsilon+m_2\frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)^2}+m_3\sqrt{\frac{\log(\frac{18SAN}{\delta})\mathcal{H}^2}{(1-\gamma)^2N}}, \tag{114}$$

when $N\geq \frac{C\log(\frac{18SAN}{\delta})}{(1-\gamma)^2}$.

Thus, to achieve an ϵ-optimal policy, it requires a total number of samples of

$$NSA=\frac{CSA\log(\frac{18SAN}{\delta})\mathcal{H}^2}{(1-\gamma)^2\epsilon^2}+\frac{CSA\log(\frac{18SAN}{\delta})}{(1-\gamma)^2}, \tag{115}$$

for some constant C. Then setting $\epsilon=\mathcal{H}$ and $1-\gamma=\frac{\epsilon}{\mathcal{H}}$ implies that

$$NSA\geq \frac{CSA\mathcal{H}^2\log\frac{SAN}{\delta}}{\epsilon^2} \tag{116}$$

samples are required to find an ϵ-optimal policy for robust average reward.

12.4. Proofs of Lemmas

Lemma 12.5. *(Lemma 6 of (Zurek & Chen, 2023)) For any deterministic stationary policy π, we have*

$$\gamma\left\|(I-\gamma\mathsf{P}^\pi)^{-1}\sqrt{\boldsymbol{Var}_{\mathsf{P}^\pi}\left[V^\pi_{\gamma,\mathsf{P}}\right]}\right\|_\infty \leq \sqrt{\frac{2}{1-\gamma}}\sqrt{\left\|\boldsymbol{Var}_{\mathsf{P}^\pi}\left[\sum_{t=0}^\infty\gamma^tR_t\right]\right\|_\infty} \tag{117}$$

Proof. The following variance Bellman equation holds from (Sobel, 1982):

$$\mathbf{Var}_{\mathsf{P}^\pi}\left[\sum_{t=0}^\infty\gamma^tR_t\right]=\gamma^2\mathbf{Var}_{\mathsf{P}^\pi}\left[V^\pi_{\gamma,\mathsf{P}}\right]+\gamma^2\mathsf{P}^\pi\mathbf{Var}_{\mathsf{P}^\pi}\left[\sum_{t=0}^\infty\gamma^tR_t\right]. \tag{118}$$

On the other hand, it holds that

$$\left|(1-\gamma)e_s^\top(I-\gamma\mathsf{P}^\pi)^{-1}\sqrt{\mathbf{Var}_{\mathsf{P}^\pi}\left[V^\pi_{\gamma,\mathsf{P}}\right]}\right|\leq\sqrt{\left|(1-\gamma)e_s^\top(I-\gamma\mathsf{P}^\pi)^{-1}\mathbf{Var}_{\mathsf{P}^\pi}\left[V^\pi_{\gamma,\mathsf{P}}\right]\right|}.$$

Denote that $v=\mathbf{Var}_{\mathsf{P}^\pi}\left[V^\pi_{\gamma,\mathsf{P}}\right]$, we then have that

$$\begin{aligned}\gamma\left\|(I-\gamma\mathsf{P}^\pi)^{-1}\sqrt{v}\right\|_\infty &=\gamma\frac{1}{1-\gamma}\left\|(1-\gamma)(I-\gamma\mathsf{P}^\pi)^{-1}\sqrt{v}\right\|_\infty && (119)\\ &\leq\gamma\frac{1}{1-\gamma}\sqrt{\left\|(1-\gamma)(I-\gamma\mathsf{P}^\pi)^{-1}v\right\|_\infty} && (120)\\ &=\gamma\frac{1}{\sqrt{1-\gamma}}\sqrt{\left\|(I-\gamma\mathsf{P}^\pi)^{-1}v\right\|_\infty}. && (121)\end{aligned}$$

Moreover,

$$\begin{aligned}
\left\|(I-\gamma \mathsf{P}^{\pi})^{-1}v\right\|_{\infty} &= \left\|(I-\gamma \mathsf{P}^{\pi})^{-1}(I-\gamma^2 \mathsf{P}^{\pi})(I-\gamma^2 \mathsf{P}^{\pi})^{-1}v\right\|_{\infty} && (122)\\
&= \left\|(I-\gamma \mathsf{P}^{\pi})^{-1}\left((1-\gamma)I+\gamma(I-\gamma \mathsf{P}^{\pi})\right)(I-\gamma^2 \mathsf{P}^{\pi})^{-1}v\right\|_{\infty} && (123)\\
&= \left\|\left((1-\gamma)(I-\gamma \mathsf{P}^{\pi})^{-1}+\gamma I\right)(I-\gamma^2 \mathsf{P}^{\pi})^{-1}v\right\|_{\infty} && (124)\\
&\le \left\|(1-\gamma)(I-\gamma \mathsf{P}^{\pi})^{-1}(I-\gamma^2 \mathsf{P}^{\pi})^{-1}v\right\|_{\infty}+\gamma\left\|(I-\gamma^2 \mathsf{P}^{\pi})^{-1}v\right\|_{\infty} && (125)\\
&\le (1-\gamma)\left\|(I-\gamma \mathsf{P}^{\pi})^{-1}\right\|_{\infty\to\infty}\left\|(I-\gamma^2 \mathsf{P}^{\pi})^{-1}v\right\|_{\infty}+\gamma\left\|(I-\gamma^2 \mathsf{P}^{\pi})^{-1}v\right\|_{\infty} && (126)\\
&\le (1+\gamma)\left\|(I-\gamma^2 \mathsf{P}^{\pi})^{-1}v\right\|_{\infty} && (127)\\
&\le 2\left\|(I-\gamma^2 \mathsf{P}^{\pi})^{-1}v\right\|_{\infty}. && (128)
\end{aligned}$$

Combining them with the variance Bellman equation (118), it holds that

$$\gamma\left\|(I-\gamma \mathsf{P}^{\pi})^{-1}\sqrt{v}\right\|_{\infty} \le \gamma\frac{1}{\sqrt{1-\gamma}}\sqrt{2\left\|(I-\gamma^2 \mathsf{P}^{\pi})^{-1}v\right\|_{\infty}} \le \sqrt{\frac{2}{1-\gamma}}\sqrt{\left\|\mathbf{Var}_{\mathsf{P}^{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t R_t\right]\right\|_{\infty}}. \tag{129}$$

□

Lemma 12.6. *(Lemma 7 of (Zurek & Chen, 2023)) For any integer $T \ge 1$, for any deterministic stationary policy π, we have*

$$\left\|\mathbf{Var}_{\mathsf{P}^{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t R_t\right]\right\|_{\infty} \le \frac{\left\|\mathbf{Var}_{\mathsf{P}^{\pi}}\left[\sum_{t=0}^{T-1}\gamma^t R_t+\gamma^T V_{\gamma}^{\pi}(S_T)\right]\right\|_{\infty}}{1-\gamma^{2T}}.$$

Lemma 12.7. *(Lemma 8 of (Zurek & Chen, 2023)) If $\gamma \ge 1-\frac{1}{\mathcal{H}}$ for some integer $\mathcal{H} \ge 1$, then*

$$\frac{1-\gamma^{2\mathcal{H}}}{1-\gamma} \ge \left(1-\frac{1}{e^2}\right)\mathcal{H} \ge \frac{4}{5}\mathcal{H}.$$

Lemma 12.8. *Letting π^* be the optimal policy for the robust DMDP $(\mathcal{S},\mathcal{A},\gamma,r,\mathcal{P})$, we have*

$$\left\|\mathbf{Var}_{\mathsf{P}^{\pi^*}}\left[\sum_{t=0}^{\infty}\gamma^t R_t\right]\right\|_{\infty} \le 5\frac{\mathcal{H}}{1-\gamma}.$$

Proof. By using Lemma 12.6, it suffices to bound $\left\|\mathbf{Var}_{\mathsf{P}^{\pi^*}}\left[\sum_{t=0}^{\mathcal{H}-1}\gamma^t R_t+\gamma^{\mathcal{H}} V_{\gamma,\mathsf{P}}^{\pi^*}(S_{\mathcal{H}})\right]\right\|_{\infty}$.

Fixing a state $s_0 \in \mathcal{S}$,

$$\begin{aligned}
\mathbf{Var}_{\mathsf{P}_{s_0}^{\pi^*}}\left[\sum_{t=0}^{\mathcal{H}-1}\gamma^t R_t+\gamma^{\mathcal{H}} V_{\gamma,\mathsf{P}}^{\pi^*}(S_{\mathcal{H}})\right] &= \mathbf{Var}_{\mathsf{P}_{s_0}^{\pi^*}}\left[\sum_{t=0}^{\mathcal{H}-1}\gamma^t R_t+\gamma^{\mathcal{H}}\left(V_{\gamma,\mathsf{P}}^{\pi^*}(S_{\mathcal{H}})-\frac{1}{1-\gamma}g_{\mathsf{P}}^{\pi^*}\right)\right]\\
&\le \mathbb{E}_{\mathsf{P}_{s_0}^{\pi^*}}\left|\sum_{t=0}^{\mathcal{H}-1}\gamma^t R_t+\gamma^{\mathcal{H}}\left(V_{\gamma,\mathsf{P}}^{\pi^*}(S_{\mathcal{H}})-\frac{1}{1-\gamma}g_{\mathsf{P}}^{\pi^*}\right)\right|^2\\
&\le 2\mathbb{E}_{\mathsf{P}_{s_0}^{\pi^*}}\left|\sum_{t=0}^{\mathcal{H}-1}\gamma^t R_t\right|^2+2\mathbb{E}_{\mathsf{P}_{s_0}^{\pi^*}}\left|\gamma^{\mathcal{H}}\left(V_{\gamma,\mathsf{P}}^{\pi^*}(S_{\mathcal{H}})-\frac{1}{1-\gamma}g_{\mathsf{P}}^{\pi^*}\right)\right|^2\\
&\le 2\mathcal{H}^2+2\sup_s\left(V_{\gamma,\mathsf{P}}^{\pi^*}(s)-\frac{1}{1-\gamma}g_{\mathsf{P}}^{\pi^*}\right)^2\\
&\le 4\mathcal{H}^2,
\end{aligned}$$

where the last inequality can be similarly derived as Lemma 11.1. We thus have that

$$\left\|\mathbf{Var}_{\mathsf{P}^{\pi^*}}\left[\sum_{t=0}^{\infty}\gamma^t R_t\right]\right\|_\infty \le \frac{4\mathcal{H}^2}{1-\gamma^{2H}}.$$

Together with Lemma 12.7, this completes the proof. □

Lemma 12.9. *(Lemma 12.2) For any policy π, if $N \ge \mathcal{O}\left(\frac{\log \frac{SA}{(1-\gamma)\delta\epsilon}}{1-\gamma}\right)$, it holds with probability at least $1-\delta$ that*

$$\left\|\left(I-\gamma \mathsf{Q}^{\pi^*}\right)^{-1}\sqrt{\mathrm{Var}_{\mathsf{Q}^{\pi^*}}(V^{\pi^*}_{\gamma,\mathsf{Q}})}\right\|_\infty \le \sqrt{\frac{c_1\mathcal{H}}{(1-\gamma)^2}}. \tag{130}$$

Proof. We prove a more general result: for any kernel P, it holds that

$$\left\|\left(I-\gamma \mathsf{P}^{\pi^*}\right)^{-1}\sqrt{\mathrm{Var}_{\mathsf{P}^{\pi^*}}(V^{\pi^*}_{\gamma,\mathsf{P}})}\right\|_\infty \le \sqrt{\frac{c_1\mathcal{H}}{(1-\gamma)^2}}. \tag{131}$$

By Lemma 12.5, we have that

$$\begin{aligned}
&\left\|\left(I-\gamma \mathsf{P}^{\pi^*}\right)^{-1}\sqrt{\mathrm{Var}_{\mathsf{P}^{\pi^*}}(V^{\pi^*}_{\gamma,\mathsf{P}})}\right\|_\infty\\
&\le \sqrt{\frac{2}{1-\gamma}}\sqrt{\left\|\mathbf{Var}_{\mathsf{P}^{\pi^*}}\left[\sum_{t=0}^{\infty}\gamma^t R_t\right]\right\|_\infty}\\
&\overset{(a)}{\le} \sqrt{\frac{2}{1-\gamma}}\sqrt{5\frac{\mathcal{H}}{1-\gamma}}\\
&= \sqrt{\frac{10\mathcal{H}}{(1-\gamma)^2}},
\end{aligned} \tag{132}$$

where (a) is due to Lemma 12.8. □

Lemma 12.10. *For any transition kernels q_1 and q_2, it holds that*

$$\left\|\left(I-\gamma q_1^{\pi^*}\right)^{-1}\sqrt{\mathrm{Var}_{q_2^{\pi^*}}(V^{\pi^*}_{\gamma,\mathcal{P}})}\right\|_\infty \le \sqrt{\frac{c_1\mathcal{H}^2}{(1-\gamma)^2}}. \tag{133}$$

Proof. Note that $\mathbf{Var}_q(V) = \mathbf{Var}_q(V-ke)$ for any k and $e=(1,...,1)$. Moreover, from Corollary 11.5, it holds that

$$\|V^{\pi^*}_{\gamma,\mathcal{P}} - \frac{1}{1-\gamma}g^{\pi^*}_{\mathcal{P}}\| \le 4\mathcal{H}. \tag{134}$$

Thus

$$\mathrm{Var}_{q_2^{\pi^*}}(V^{\pi^*}_{\gamma,\mathcal{P}}) = \mathrm{Var}_{q_2^{\pi^*}}(V^{\pi^*}_{\gamma,\mathcal{P}} - \frac{1}{1-\gamma}g^{\pi^*}_{\mathcal{P}}) \le \|V^{\pi^*}_{\gamma,\mathcal{P}} - \frac{1}{1-\gamma}g^{\pi^*}_{\mathcal{P}}\|^2 \le 16\mathcal{H}^2. \tag{135}$$

The proof is then completed. □

Lemma 12.11. *(Lemma 12.3) Set $\omega = \frac{\xi\delta(1-\gamma)}{3SA^2}$, then both (105) and (106) occur with probability at least $1-\delta$.*

Proof. The proof is similar for both events, hence we only present the proof for (106). We show a more general result, namely, with probability at least $1-\delta$, for any s and $a_1 \neq a_2$,

$$|Q^*_{\gamma,\mathcal{P}}(s,a_1) - Q^*_{\gamma,\mathcal{P}}(s,a_2)| > \frac{\xi\delta(1-\gamma)}{3SA^2}. \tag{136}$$

We further introduce the following notation:

$$r_\tau(s,a_1)=\tau, \tag{137}$$
$$r_\tau(s',a')=\tilde{r}(s',a'),\forall(s',a')\neq(s,a_1). \tag{138}$$

We denote the optimal robust value functions and the optimal policy w.r.t. r_τ as Q^*_τ, V^*_τ and π^*_τ.

We first prove the following claim: there exists some τ', such that

$$\pi^*_\tau(s)\neq a_1, \text{ for all } \tau<\tau', \tag{139}$$
$$\pi^*_\tau(s)=a_1, \text{ for all } \tau>\tau'. \tag{140}$$

Define

$$\tau'=\sup\{u:\pi^*_\tau(s)\neq a_1,\forall\tau<u\}, \tag{141}$$

then it suffices to show (140) for our choice, which exactly follows as the proofs of eq (95) in (Li et al., 2020). We then prove the lemma as follows.

First, define the following sets:

$$I_{0,\omega}\triangleq\{\tau:|Q^*_\tau(s,a_1)-Q^*_\tau(s,a_2)|<\omega\}, \tag{142}$$
$$I_{1,\omega}\triangleq\{\tau:\tau<\tau',|Q^*_\tau(s,a_1)-Q^*_\tau(s,a_2)|<\omega\}, \tag{143}$$
$$I_{2,\omega}\triangleq\{\tau:\tau\geq\tau',|Q^*_\tau(s,a_1)-Q^*_\tau(s,a_2)|<\omega\}. \tag{144}$$

Clearly, $I_{0,\omega}=I_{1,\omega}\cup I_{2,\omega}$, and we will show that the probability of these events is small.

Step 1. For $\tau\in I_{1,\omega}$, note that V^*_τ does not depend on τ, since $\pi^*_\tau(s)\neq a_1$, and τ is never active when calculating V^*_τ. Thus, the robust Bellman equation becomes

$$\begin{aligned}Q^*_\tau(s,a_1)&=\tau+\gamma\sigma_{\mathcal{P}^{a_1}_s}(V^*_\tau),\\ Q^*(s,a_2)&=\tilde{r}(s,a_2)+\gamma\sigma_{\mathcal{P}^{a_2}_s}(V^*_\tau).\end{aligned} \tag{145}$$

Thus, it holds that

$$I_{1,\omega}\subset\{\tau:|\tau+\gamma\sigma_{\mathcal{P}^{a_1}_s}(V^*_\tau)-Q^*(s,a_2)|<\omega\}. \tag{146}$$

Since both terms $\gamma\sigma_{\mathcal{P}^{a_1}_s}(V^*_\tau)$ and $Q^*(s,a_2)$ are independent from τ, the Lebesgue measure of $I_{1,\omega}$ is at most 2ω.

Step 2. We now consider $I_{2,\omega}$. First, note that

$$\begin{aligned}0\leq Q^*_{\tau_2}-Q^*_{\tau_1}&\leq r_{\tau_2}-r_{\tau_1}+\gamma(\sigma(V^*_{\tau_2})-\sigma(V^*_{\tau_1}))\\ &\leq r_{\tau_2}-r_{\tau_1}+\gamma\|V^*_{\tau_2}-V^*_{\tau_1}\|,\end{aligned} \tag{147}$$

for any $\tau_2>\tau_1>\tau'$, which is from the 1-Lipschitz of the support functions. Moreover, for any $(x,b)\neq(s,a_1)$, since $r_{\tau_2}(x,b)=r_{\tau_1}(x,b)$, it holds that

$$0\leq Q^*_{\tau_2}(x,b)-Q^*_{\tau_1}(x,b)\leq\gamma\|V^*_{\tau_2}-V^*_{\tau_1}\|. \tag{148}$$

On the other hand, note that

$$0\leq V^*_{\tau_2}-V^*_{\tau_1}=\max_a Q^*_{\tau_2}-\max_a Q^*_{\tau_1}\leq\|Q^*_{\tau_2}-Q^*_{\tau_1}\|, \tag{149}$$

and thus

$$\|V^*_{\tau_2}-V^*_{\tau_1}\|\leq\|Q^*_{\tau_2}-Q^*_{\tau_1}\|. \tag{150}$$

Note that (148) implies that

$$Q^*_{\tau_2}(x,b) - Q^*_{\tau_1}(x,b) \leq \gamma \|V^*_{\tau_2} - V^*_{\tau_1}\| < \|V^*_{\tau_2} - V^*_{\tau_1}\|, \forall (x,b) \neq (s,a_1), \tag{151}$$

thus

$$\|Q^*_{\tau_2} - Q^*_{\tau_1}\| = |Q^*_{\tau_2}(s,a_1) - Q^*_{\tau_1}(s,a_1)| \geq \|V^*_{\tau_2} - V^*_{\tau_1}\|. \tag{152}$$

Since $\tau_1, \tau_2 \geq \tau'$, $V^*_{\tau_2}(s) = Q^*\tau_2(s,a_1)$ and $V^*_{\tau_1}(s) = Q^*\tau_1(s,a_1)$, and we further have that

$$V^*_{\tau_2}(s) - V^*_{\tau_1}(s) = Q^*_{\tau_2}(s,a_1) - Q^*_{\tau_1}(s,a_1) \geq \|V^*_{\tau_2} - V^*_{\tau_1}\|, \tag{153}$$

and hence

$$\|V^*_{\tau_2} - V^*_{\tau_1}\| = Q^*_{\tau_2}(s,a_1) - Q^*_{\tau_1}(s,a_1). \tag{154}$$

Now from the robust Bellman equation, it holds that

$$\begin{aligned} & Q^*_{\tau_2}(s,a_1) - Q^*_{\tau_1}(s,a_1) \\ & = \|V^*_{\tau_2} - V^*_{\tau_1}\| \\ & = \|r_{\tau_2} - r_{\tau_1} + \gamma(\sigma(V^*_{\tau_2}) - \sigma(V^*_{\tau_1})\| \\ & \geq r_{\tau_2} - r_{\tau_1}, \end{aligned} \tag{155}$$

due to the monotonicity properties of the support functions.

We note that (148) and (155) exactly match eqs. (99) and (102) in (Li et al., 2020), and hence the rest of the proof follows similarly. □

Lemma 12.12. *(Lemma 12.4) Consider any $\delta \in (0,1)$. Taking $N \geq \mathcal{O}\left(\frac{\log\left(\frac{54SAN^2}{(1-\gamma)\delta}\right)}{1-\gamma}\right)$, with probability at least $1-2\delta$, events (105) and (106) occur, and it holds that*

$$\left| \hat{\mathsf{P}}^{\widehat{\pi},\widehat{V}}_w \tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} - \mathsf{P}^{\widehat{\pi},\widehat{V}}_w \tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} \right| \leq 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}} \sqrt{\mathrm{Var}_{\mathsf{P}_{s,a}}(\tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}})} 1 + \frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)} 1. \tag{156}$$

Proof. The proof is obtained similarly as Lemma 14 from (Shi et al., 2023), by only replace $r(s,a)$ therein by $\mathbb{E}[\tilde{r}(s,a)]$.

For any (s,a), by the duality we have that

$$\left| (\hat{\mathsf{P}}^{\widehat{\pi},\widehat{V}}_w)_{s,a} \tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} - (\mathsf{P}^{\widehat{\pi},\widehat{V}}_w)_{s,a} \tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} \right| \leq \max_{\alpha \in [\min_s \tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}(s), \max_s \tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}(s)]} \left| \left(\mathsf{P}_{s,a} - \hat{\mathsf{P}}_{s,a}\right) \left[\tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}\right]_\alpha \right|. \tag{157}$$

Construction of auxiliary RMDPs with deterministic empirical nominal transitions. Recall that we target the empirical infinite-horizon robust MDP with the nominal transition kernel $\hat{\mathsf{P}}$. We define the nominal transition kernel and reward function as $P^{s,u}$ and $r^{s,u}$, which are expressed as follows

$$\begin{cases} P^{s,u}(s'|s,a) = \mathbf{1}(s'=s) & \text{for all } (s',a) \in \mathcal{S}\times\mathcal{A}, \\ P^{s,u}(\cdot|\widetilde{s},a) = \hat{\mathsf{P}}(\cdot|\widetilde{s},a) & \text{for all } (\widetilde{s},a) \in \mathcal{S}\times\mathcal{A} \text{ and } \widetilde{s} \neq s, \end{cases} \tag{158}$$

and

$$\begin{cases} r^{s,u}(s,a) = u & \text{for all } a \in \mathcal{A}, \\ r^{s,u}(\widetilde{s},a) = \mathbb{E}[\tilde{r}(\widetilde{s},a)] & \text{for all } (\widetilde{s},a) \in \mathcal{S}\times\mathcal{A} \text{ and } \widetilde{s} \neq s. \end{cases} \tag{159}$$

Correspondingly, the associated robust Bellman operator is then

$$\forall (\tilde{s},a) \in \mathcal{S}\times\mathcal{A}: \quad \mathbf{T}_{s,u}(Q)(\tilde{s},a) = r^{s,u}(\tilde{s},a) + \gamma \inf_{\mathsf{P} \in \mathcal{P}(P^{s,u}_{\tilde{s},a})} \mathsf{P}V, \qquad \text{with } V(\tilde{s}) = \max_a Q(\tilde{s},a). \tag{160}$$

Fixed-point equivalence. Recall that $\tilde{Q}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}$ is the unique fixed point of the Bellman operator with the corresponding robust value $\tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}$. We assert that the corresponding robust value function $(\tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}})_{s,u^*}$ obtained from the fixed point of $\mathbf{T}_{s,u}(\cdot)$ aligns with the robust value function $\tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}$, as long as we choose u in the following manner:

$$u^* = u^*(s) = \tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}(s) - \gamma \inf_{\mathsf{P}\in\mathcal{P}(e_s)} \mathsf{P}\tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}. \tag{161}$$

where e_s is the s-th standard basis vector in $\mathbb{R}^S$. Towards verifying this, we shall break our arguments in two different cases.

- **For state** s: One has for any $a \in \mathcal{A}$:

  $$\begin{aligned} r^{s,u^*}(s,a) + \gamma \inf_{\mathsf{P}\in\mathcal{P}(P^{s,u^*}_{s,a})} \mathsf{P}\tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} &= u^* + \gamma \inf_{\mathsf{P}\in\mathcal{P}(e_s)} \mathsf{P}\tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} \\ &= \tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}(s) - \gamma \inf_{\mathsf{P}\in\mathcal{P}(e_s)} \mathsf{P}\tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} + \gamma \inf_{\mathsf{P}\in\mathcal{P}(e_s)} \mathsf{P}\tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} = \tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}(s), \end{aligned} \tag{162}$$

  where the first equality follows from the definition of $P^{s,u^*}_{s,a}$ in (158), and the second equality follows from plugging in the definition of u^* in (161).

- **For state** $s' \neq s$: It is easily verified that for all $a \in \mathcal{A}$,

  $$\begin{aligned} r^{s,u^*}(s',a) + \gamma \inf_{\mathsf{P}\in\mathcal{P}(P^{s,u^*}_{s',a})} \mathsf{P}\tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} &= r(s',a) + \gamma \inf_{\mathsf{P}\in\mathcal{P}(\hat{\mathsf{P}}_{s',a})} \mathsf{P}\tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} \\ &= \mathbf{T}(\tilde{Q}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}})(s',a) = \tilde{Q}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}(s',a), \end{aligned} \tag{163}$$

  where the first equality follows from the definitions in (159) and (158), and the last line arises from the definition of the robust Bellman operator, and that $\tilde{Q}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}$ is the fixed point of $\mathbf{T}(\cdot)$.

Combining the facts in the above two cases, we establish that there exists a fixed point $(\tilde{Q}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}})_{s,u^*}$ of the operator $\mathbf{T}_{s,u^*}(\cdot)$ by taking

$$\begin{cases} (\tilde{Q}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}})_{s,u^*}(s,a) = \tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}(s) & \text{for all } a \in \mathcal{A}, \\ (\tilde{Q}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}})_{s,u^*}(s',a) = \tilde{Q}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}(s',a) & \text{for all } s' \neq s \text{ and } a \in \mathcal{A}. \end{cases} \tag{164}$$

Consequently, we confirm the existence of a fixed point of the operator $\mathbf{T}_{s,u^*}(\cdot)$. In addition, its corresponding value function $(\tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}})_{s,u^*}$ also coincides with $\tilde{V}^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}$.

This equivalence exactly matches with Step 1 and Step 2 in Lemma 14 of (Shi et al., 2023), and hence the remaining part directly follows. □

13. Proof of Theorem 4.4 Part 2

The proof of Theorem 4.4 mainly follows a similar structure. We note that it is equivalent to show that

$$\left\| V^{\pi^*}_{\gamma,\mathcal{P}} - V^{\hat{\pi}}_{\gamma,\mathcal{P}} \right\|_\infty \leq 16\sqrt{\frac{2(1+R)\mathcal{H}^2 \log(\frac{36SAN^2}{\delta})}{(1-\gamma)^2 N}}. \tag{165}$$

In order to control the performance gap $\left\| V^{\pi^*}_{\gamma,\mathcal{P}} - V^{\hat{\pi}}_{\gamma,\mathcal{P}} \right\|_\infty$, note that

$$V^{\pi^*}_{\gamma,\mathcal{P}} - V^{\hat{\pi}}_{\gamma,\mathcal{P}} \leq V^{\pi^*}_{\gamma,\mathcal{P}} - V^{\pi^*}_{\gamma,\hat{\mathcal{P}}} + V^{\pi^*}_{\gamma,\hat{\mathcal{P}}} - V^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} + V^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} - V^{\hat{\pi}}_{\gamma,\mathcal{P}} \leq V^{\pi^*}_{\gamma,\mathcal{P}} - V^{\pi^*}_{\gamma,\hat{\mathcal{P}}} + V^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} - V^{\hat{\pi}}_{\gamma,\mathcal{P}}. \tag{166}$$

It is hence sufficient to bound the two terms on the RHS.

Part A: $\|V^{\pi^*}_{\gamma,\hat{\mathcal{P}}} - V^{\pi^*}_{\gamma,\mathcal{P}}\|_\infty$. Towards this, recall the bound in (80):

$$\|V^{\pi^*}_{\gamma,\hat{\mathcal{P}}} - V^{\pi^*}_{\gamma,\mathcal{P}}\|_\infty \le \gamma \max\Big\{\Big\|\Big(I-\gamma\hat{\mathsf{P}}_w^{\pi^\star,\widehat{V}}\Big)^{-1}\Big(\hat{\mathsf{P}}_w^{\pi^\star,V}V^{\pi^*}_{\gamma,\mathcal{P}} - \mathsf{P}_w^{\pi^\star,V}V^{\pi^*}_{\gamma,\mathcal{P}}\Big)\Big\|_\infty, \\ \Big\|\Big(I-\gamma\hat{\mathsf{P}}_w^{\pi^\star,V}\Big)^{-1}\Big(\hat{\mathsf{P}}_w^{\pi^\star,V}V^{\pi^*}_{\gamma,\mathcal{P}} - \mathsf{P}_w^{\pi^\star,V}V^{\pi^*}_{\gamma,\mathcal{P}}\Big)\Big\|_\infty\Big\}. \tag{167}$$

To control the main term $\hat{\mathsf{P}}_w^{\pi^\star,V}V^{\pi^*}_{\gamma,\mathcal{P}} - \mathsf{P}_w^{\pi^\star,V}V^{\pi^*}_{\gamma,\mathcal{P}}$ in (167), we first introduce the following lemma.

Lemma 13.1. *For any $\delta \in (0,1)$ and any fixed policy π, one has with probability at least $1-\delta$,*

$$\Big\|\hat{\mathsf{P}}_w^{\pi,V}V^{\pi}_{\gamma,\mathcal{P}} - \mathsf{P}_w^{\pi,V}V^{\pi}_{\gamma,\mathcal{P}}\Big\|_\infty \le 4\sqrt{\frac{2(1+R)\mathcal{H}^2\log(\frac{24SAN}{\delta})}{N}}. \tag{168}$$

Applying Lemma 13.1 by taking $\pi = \pi^\star$ gives

$$\Big\|\hat{\mathsf{P}}_w^{\pi^\star,V}V^{\pi^*}_{\gamma,\mathcal{P}} - \mathsf{P}_w^{\pi^\star,V}V^{\pi^*}_{\gamma,\mathcal{P}}\Big\|_\infty \le 4\sqrt{\frac{2(1+R)\mathcal{H}^2\log(\frac{24SAN}{\delta})}{N}}, \tag{169}$$

which directly leads to

$$\begin{aligned}&\Big\|\Big(I-\gamma\hat{\mathsf{P}}_w^{\pi^\star,\widehat{V}}\Big)^{-1}\Big(\hat{\mathsf{P}}_w^{\pi^\star,V}V^{\pi^*}_{\gamma,\mathcal{P}} - \mathsf{P}_w^{\pi^\star,V}V^{\pi^*}_{\gamma,\mathcal{P}}\Big)\Big\|_\infty \\ &\le \Big\|\hat{\mathsf{P}}_w^{\pi^\star,V}V^{\pi^*}_{\gamma,\mathcal{P}} - \mathsf{P}_w^{\pi^\star,V}V^{\pi^*}_{\gamma,\mathcal{P}}\Big\|_\infty \cdot \Big\|\Big(I-\gamma\hat{\mathsf{P}}_w^{\pi^\star,\widehat{V}}\Big)^{-1}1\Big\|_\infty \le 4\sqrt{\frac{2(1+R)\mathcal{H}^2\log(\frac{24SAN}{\delta})}{(1-\gamma)^2N}}.\end{aligned} \tag{170}$$

Similarly, we have

$$\Big\|\Big(I-\gamma\hat{\mathsf{P}}_w^{\pi^\star,V}\Big)^{-1}\Big(\hat{\mathsf{P}}_w^{\pi^\star,V}V^{\pi^*}_{\gamma,\mathcal{P}} - \mathsf{P}_w^{\pi^\star,V}V^{\pi^*}_{\gamma,\mathcal{P}}\Big)\Big\|_\infty \le 4\sqrt{\frac{2(1+R)\mathcal{H}^2\log(\frac{24SAN}{\delta})}{(1-\gamma)^2N}}. \tag{171}$$

Inserting (170) and (171) back to (167) yields

$$\|V^{\pi^*}_{\gamma,\hat{\mathcal{P}}} - V^{\pi^*}_{\gamma,\mathcal{P}}\|_\infty \le 4\sqrt{\frac{2(1+R)\mathcal{H}^2\log(\frac{24SAN}{\delta})}{(1-\gamma)^2N}}. \tag{172}$$

Part B: controlling $\Big\|V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}} - V^{\hat\pi}_{\gamma,\mathcal{P}}\Big\|_\infty$. Similarly, we have that

$$\|V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}} - V^{\hat\pi}_{\gamma,\mathcal{P}}\|_\infty \le \gamma \max\Big\{\Big\|\Big(I-\gamma\mathsf{P}_w^{\hat\pi,V}\Big)^{-1}\Big(\hat{\mathsf{P}}_w^{\hat\pi,\widehat{V}}V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}} - \mathsf{P}_w^{\hat\pi,\widehat{V}}V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}}\Big)\Big\|_\infty, \\ \Big\|\Big(I-\gamma\mathsf{P}_w^{\hat\pi,\widehat{V}}\Big)^{-1}\Big(\hat{\mathsf{P}}_w^{\hat\pi,\widehat{V}}V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}} - \mathsf{P}_w^{\hat\pi,\widehat{V}}V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}}\Big)\Big\|_\infty\Big\}. \tag{173}$$

We introduce the following lemma which controls $\hat{\mathsf{P}}_w^{\hat\pi,\widehat{V}}V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}} - \mathsf{P}_w^{\hat\pi,\widehat{V}}V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}}$ in (173);

Lemma 13.2. *With probability at least $1-\delta$, one has*

$$\Big\|\hat{\mathsf{P}}_w^{\hat\pi,\widehat{V}}V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}} - \mathsf{P}_w^{\hat\pi,\widehat{V}}V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}}\Big\|_\infty \le 12\sqrt{\frac{2(1+R)\mathcal{H}^2\log(\frac{36SAN^2}{\delta})}{N}}. \tag{174}$$

Repeating the arguments from (169) to (172) yields

$$\|V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}} - V^{\hat\pi}_{\gamma,\mathcal{P}}\|_\infty \le 12\sqrt{\frac{2(1+R)\mathcal{H}^2\log(\frac{36SAN^2}{\delta})}{(1-\gamma)^2N}}. \tag{175}$$

Finally, combining all bounds together implies that

$$\begin{aligned}
&\left\|V_{\gamma,\mathcal{P}}^{\pi^*} - V_{\gamma,\mathcal{P}}^{\hat{\pi}}\right\|_\infty \\
&\leq 4\sqrt{\frac{2(1+R)\mathcal{H}^2\log(\frac{24SAN}{\delta})}{(1-\gamma)^2 N}} + 12\sqrt{\frac{2(1+R)\mathcal{H}^2\log(\frac{36SAN^2}{\delta})}{(1-\gamma)^2 N}} \\
&\leq 16\sqrt{\frac{2(1+R)\mathcal{H}^2\log(\frac{36SAN^2}{\delta})}{(1-\gamma)^2 N}}.
\end{aligned} \tag{176}$$

13.1. Proofs of Lemmas

Lemma 13.3. *(Lemma 13.1) For any $\delta \in (0,1)$ and any fixed policy π, one has with probability at least $1-\delta$,*

$$\left\|\hat{\mathsf{P}}_w^{\pi,V} V_{\gamma,\mathcal{P}}^{\pi} - \mathsf{P}_w^{\pi,V} V_{\gamma,\mathcal{P}}^{\pi}\right\|_\infty \leq 4\sqrt{\frac{2(1+R)\mathcal{H}^2\log(\frac{24SAN}{\delta})}{N}}. \tag{177}$$

Proof. **Step 1: controlling the point-wise concentration.** Consider any fixed policy π and the corresponding robust value vector $V \triangleq V_{\gamma,\mathcal{P}}^{\pi} - \frac{g_{\mathcal{P}}^{\pi}}{1-\gamma}$ (independent from $\hat{\mathsf{P}}$). We note that $\|V\| \leq \mathcal{H}$ as showed before. By the duality of CS sets (Shi et al., 2023) it holds that

$$\begin{aligned}
&\left|(\hat{\mathsf{P}}_w^{\pi,V})_{s,a} V_{\gamma,\mathcal{P}}^{\pi} - (\mathsf{P}_w^{\pi,V})_{s,a} V_{\gamma,\mathcal{P}}^{\pi}\right| \\
&= \Bigg| \max_{\alpha\in[\min_s V(s),\max_s V(s)]} \left\{\mathsf{P}_{s,a}[V]_\alpha - \sqrt{R\mathsf{Var}_{\mathsf{P}_{s,a}}([V]_\alpha)}\right\} \\
&\qquad - \max_{\alpha\in[\min_s V(s),\max_s V(s)]} \left\{\hat{\mathsf{P}}_{s,a}[V]_\alpha - \sqrt{R\mathsf{Var}_{\hat{\mathsf{P}}_{s,a}}([V]_\alpha)}\right\} \Bigg| \\
&\leq \max_{\alpha\in[\min_s V(s),\max_s V(s)]} \left|\left(\mathsf{P}_{s,a} - \hat{\mathsf{P}}_{s,a}\right)[V]_\alpha + \sqrt{R\mathsf{Var}_{\hat{\mathsf{P}}_{s,a}}([V]_\alpha)} - \sqrt{R\mathsf{Var}_{\mathsf{P}_{s,a}}([V]_\alpha)}\right| \\
&\leq \max_{\alpha\in[\min_s V(s),\max_s V(s)]} \left|\left(\mathsf{P}_{s,a} - \hat{\mathsf{P}}_{s,a}\right)[V]_\alpha\right| + \\
&\quad + \max_{\alpha\in[\min_s V(s),\max_s V(s)]} \sqrt{R}\left|\sqrt{\mathsf{Var}_{\hat{\mathsf{P}}_{s,a}}([V]_\alpha)} - \sqrt{\mathsf{Var}_{\mathsf{P}_{s,a}}([V]_\alpha)}\right|,
\end{aligned} \tag{178}$$

where the first inequality follows by the maximum operator being 1-Lipschitz, and the second inequality follows from the triangle inequality.

The first term in (178) can be directly bounded through an ϵ-net technique and Hoeffding's inequality, which implies that with probability at least $1-\delta$,

$$\max_{\alpha\in[\min_s V(s),\max_s V(s)]} \left|\left(\mathsf{P}_{s,a} - \hat{\mathsf{P}}_{s,a}\right)[V]_\alpha\right| \leq 2\sqrt{\frac{\log(\frac{2SAN}{\delta})\mathcal{H}^2}{N}}, \tag{179}$$

holds for all (s,a).

Step 2: controlling the second term in (178). Consider a fixed $\alpha \in [0, \frac{1}{1-\gamma}]$, applying Lemma 6 of (Panaganti & Kalathil, 2021) with $\|[V]_\alpha\|_\infty \leq \mathcal{H}$, we get that

$$\left|\sqrt{\mathsf{Var}_{\hat{\mathsf{P}}_{s,a}}([V]_\alpha)} - \sqrt{\mathsf{Var}_{\mathsf{P}_{s,a}}([V]_\alpha)}\right| \leq \sqrt{\frac{2\log(\frac{2}{\delta})\mathcal{H}}{N}} \tag{180}$$

holds with probability at least $1-\delta$. We then introduce the following lemma, whose proof can be similarly derived as Lemma 18 of (Shi et al., 2023).

Lemma 13.4. *For any V obeying $\|V\|_\infty \leq \mathcal{H}$, the function $J_{s,a}(\alpha, V) := \left|\sqrt{\mathsf{Var}_{\hat{\mathsf{P}}_{s,a}}([V]_\alpha)} - \sqrt{\mathsf{Var}_{\mathsf{P}_{s,a}}([V]_\alpha)}\right|$ w.r.t. α obeys*

$$|J_{s,a}(\alpha_1, V) - J_{s,a}(\alpha_2, V)| \leq 4\sqrt{|\alpha_1 - \alpha_2|\mathcal{H}}.$$

We then construct an ϵ-net $\mathcal{N}$ over $[0, \mathcal{H}]$ with size $N_n \leq 3\epsilon\mathcal{H}$ (Vershynin, 2018), so that with probability at least $1 - \frac{\delta}{SA}$, it holds that for any (s, a),

$$\begin{aligned}
&\max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left| \sqrt{\mathsf{Var}_{\hat{\mathsf{P}}_{s,a}}\left([V]_\alpha\right)} - \sqrt{\mathsf{Var}_{\mathsf{P}_{s,a}}\left([V]_\alpha\right)} \right| \\
&\leq \max_{\alpha \in [0, 1/(1-\gamma)]} \left| \sqrt{\mathsf{Var}_{\hat{\mathsf{P}}_{s,a}}\left([V]_\alpha\right)} - \sqrt{\mathsf{Var}_{\mathsf{P}_{s,a}}\left([V]_\alpha\right)} \right| \\
&\overset{\text{(i)}}{\leq} 4\sqrt{\frac{\epsilon}{1-\gamma}} + \sup_{\alpha \in \mathcal{N}} \left| \sqrt{\mathsf{Var}_{\hat{\mathsf{P}}_{s,a}}\left([V]_\alpha\right)} - \sqrt{\mathsf{Var}_{\mathsf{P}_{s,a}}\left([V]_\alpha\right)} \right| \\
&\overset{\text{(ii)}}{\leq} 4\sqrt{\frac{\epsilon}{1-\gamma}} + \sqrt{\frac{2\mathcal{H}^2 \log(\frac{2SA|N_\epsilon|}{\delta})}{N}} \\
&\overset{\text{(iii)}}{\leq} 2\sqrt{\frac{2\mathcal{H}^2 \log(\frac{2SA|N_\epsilon|}{\delta})}{N}} \\
&\leq 2\sqrt{\frac{2\mathcal{H}^2 \log(\frac{24SAN}{\delta})}{N}},
\end{aligned} \tag{181}$$

where (i) holds by the property of the ϵ-net, (ii) follows from (180), (iii) follows from taking $\epsilon = \frac{\mathcal{H}\log(\frac{2SA|\mathcal{N}|}{\delta})}{8N}$.

Inserting (179) and (181) back to (178) and taking the union bound over (s, a), with probability at least $1 - \delta$,

$$\begin{aligned}
\left| (\hat{\mathsf{P}}^{\pi,V}_w)_{s,a} V - (\mathsf{P}^{\pi,V}_w)_{s,a} V \right| &\leq \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left| \left(\mathsf{P}_{s,a} - \hat{\mathsf{P}}_{s,a}\right) [V]_\alpha \right| + \\
&\quad + \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left| \sqrt{R\mathsf{Var}_{\hat{\mathsf{P}}_{s,a}}\left([V]_\alpha\right)} - \sqrt{R\mathsf{Var}_{\mathsf{P}_{s,a}}\left([V]_\alpha\right)} \right| \\
&\leq 4\sqrt{\frac{2(1+R)\mathcal{H}^2 \log(\frac{24SAN}{\delta})}{N}}.
\end{aligned}$$

Finally, we complete the proof by recalling the matrix form as below:

$$\left\| \hat{\mathsf{P}}^{\pi,V}_w V^\pi_{\gamma,\mathcal{P}} - \mathsf{P}^{\pi,V}_w V^\pi_{\gamma,\mathcal{P}} \right\|_\infty \leq \max_{(s,a)} \left| \hat{\mathsf{P}}^{\pi,V}_{s,a} V - (\mathsf{P}^{\pi,V}_w)_{s,a} V \right| \leq 4\sqrt{\frac{2(1+R)\mathcal{H}^2 \log(\frac{24SAN}{\delta})}{N}}.$$

□

Lemma 13.5. *(Lemma 13.2) With probability at least* $1 - \delta$*, one has*

$$\left\| \hat{\mathsf{P}}^{\hat{\pi},\hat{V}}_w V^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} - \mathsf{P}^{\hat{\pi},\hat{V}}_w V^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} \right\|_\infty \leq 12\sqrt{\frac{2(1+R)\mathcal{H}^2 \log(\frac{36SAN^2}{\delta})}{N}}. \tag{182}$$

Proof. For any (s, a), following the same arguments of (178) yields

$$\begin{aligned}
&\left| (\hat{\mathsf{P}}^{\hat{\pi},\hat{V}}_w)_{s,a} V^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} - (\mathsf{P}^{\hat{\pi},\hat{V}}_w)_{s,a} V^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} \right| \\
&= \left| (\hat{\mathsf{P}}^{\hat{\pi},\hat{V}}_w)_{s,a} (V^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} - g^{\hat{\pi}}_{\hat{\mathcal{P}}}) - (\mathsf{P}^{\hat{\pi},\hat{V}}_w)_{s,a} (V^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} - g^{\hat{\pi}}_{\hat{\mathcal{P}}}) \right| \\
&\leq \max_{\alpha \in \left[\min_s V^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}(s) - g^{\hat{\pi}}_{\hat{\mathcal{P}}}, \max_s V^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}(s) - g^{\hat{\pi}}_{\hat{\mathcal{P}}}\right]} \left| \left(\mathsf{P}_{s,a} - \hat{\mathsf{P}}_{s,a}\right) \left[V^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} - g^{\hat{\pi}}_{\hat{\mathcal{P}}}\right]_\alpha \right| + \\
&\quad + \max_{\alpha \in \left[\min_s V^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}(s) - g^{\hat{\pi}}_{\hat{\mathcal{P}}}, \max_s V^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}}(s) - g^{\hat{\pi}}_{\hat{\mathcal{P}}}\right]} \sqrt{R} \left| \sqrt{\mathsf{Var}_{\hat{\mathsf{P}}_{s,a}}\left(\left[V^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} - g^{\hat{\pi}}_{\hat{\mathcal{P}}}\right]_\alpha\right)} - \sqrt{\mathsf{Var}_{\mathsf{P}_{s,a}}\left(\left[V^{\hat{\pi}}_{\gamma,\hat{\mathcal{P}}} - g^{\hat{\pi}}_{\hat{\mathcal{P}}}\right]_\alpha\right)} \right|.
\end{aligned} \tag{183}$$

The first term in (183) can be bounded through Hoeffding's inequality as

$$\begin{aligned}
&\max_{\alpha\in\left[\min_s V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}}(s)-g^{\hat\pi}_{\hat{\mathcal{P}}},\max_s V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}}(s)-g^{\hat\pi}_{\hat{\mathcal{P}}}\right]}\left|\left(\mathsf{P}_{s,a}-\hat{\mathsf{P}}_{s,a}\right)\left[V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}}\right]_\alpha\right|\\
&\leq \max_{\alpha\in[-\mathcal{H},\mathcal{H}]}\left|\left(\mathsf{P}_{s,a}-\hat{\mathsf{P}}_{s,a}\right)\left[V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}}\right]_\alpha\right|\\
&\leq 4\sqrt{\mathcal{H}^2\frac{\log(\frac{3SAN^{3/2}}{(1-\gamma)\delta})}{N}}. \qquad (184)
\end{aligned}$$

We then consider the second term of (183). Towards this, we can construct an auxiliary robust MDP $(\mathcal{S},\mathcal{A},\mathcal{P}^{s,u},r_{s,u},\gamma)$ as in Section D.2.2. in (Shi et al., 2023), so that $V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}}-g^{\hat\pi}_{\hat{\mathcal{P}}}=V_{s,u^*}-g_{s,u^*}$ for some $u^*\in[-\mathcal{H},\mathcal{H}]$. We then construct an ϵ-net $\mathcal{N}$ over $[-\mathcal{H},\mathcal{H}]$, so that $|u^*-u|\leq\epsilon$ for some $u\in\mathcal{N}$. Following (Shi et al., 2023), it holds that

$$\begin{aligned}
&\max_{\alpha\in\left[\min_s V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}}(s)-g^{\hat\pi}_{\hat{\mathcal{P}}},\max_s V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}}(s)-g^{\hat\pi}_{\hat{\mathcal{P}}}\right]}\left|\sqrt{\mathsf{Var}_{\hat{\mathsf{P}}_{s,a}}\left(\left[V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}}-g^{\hat\pi}_{\hat{\mathcal{P}}}\right]_\alpha\right)}-\sqrt{\mathsf{Var}_{\mathsf{P}_{s,a}}\left(\left[V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}}-g^{\hat\pi}_{\hat{\mathcal{P}}}\right]_\alpha\right)}\right|\\
&\leq 6\sqrt{\frac{2R\mathcal{H}^2\log(\frac{36SAN^2|\mathcal{N}|}{\delta})}{N}}, \qquad (185)
\end{aligned}$$

with probability at least $1-\delta$.

Inserting (185) and (184) back to (183), we have that with probability at least $1-\delta$,

$$\left\|\hat{\mathsf{P}}^{\widehat\pi,\widehat V}_w V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}}-\mathsf{P}^{\widehat\pi,\widehat V}_w V^{\hat\pi}_{\gamma,\hat{\mathcal{P}}}\right\|_\infty\leq 12\sqrt{\frac{2(1+R)\mathcal{H}^2\log(\frac{36SAN^2}{\delta})}{N}}. \qquad (186)$$

□

14. Function Approximation for Robust AMDPs

The algorithm design part is almost identical to (Zhou et al., 2024). For completeness we provide a brief discussion.

Following (Zhou et al., 2024), we consider a class of Integral Probability Metric (IPM) Uncertainty Set. Given some function class $\mathcal{F}\subset\mathbb{R}^S$ including the zero function, the integral probability metric (IPM) is defined by $d_{\mathcal{F}}(p,q):=\sup_{f\in\mathcal{F}}\{p^\top f-q^\top f\}\geq 0$ (Müller, 1997). Many metrics such as Kantorovich metric, total variation, etc., are special cases of IPM under different function classes (Müller, 1997). The IPM uncertainty set is defined as $\mathcal{P}^a_s=\{q:d_{\mathcal{F}}(q,\mathsf{P}^a_s)\leq R\}$.

We also consider the linear function class as (Zhou et al., 2024). Denote $\Phi\in\mathbb{R}^{Sd}$ the feature matrix with rows $\phi(s)$, we set

$$\mathcal{F}:=\{s\mapsto\psi(s)^\top\xi:\xi\in\mathbb{R}^d,\|\xi\|\leq 1\}. \qquad (187)$$

Without loss of generality, assume Φ has full column rank, and let the first coordinate of $\phi(s)$ be 1 for any s.

We then detail the algorithm designs of the two updates steps in Algorithm 2.

Algorithm 3 **Robust Linear Temporal Difference (RLTD)**

```
Input: $\pi, K$
Initialization: $\theta_0$, $s_0$
for $k=0,1,\dots,K-1$ do
    Sample $a_k\sim\pi(\cdot|s_k)$, $y_{k+1}$ according to $\mathsf{P}_{s_k,a_k}$, and $s_{k+1}$ from $y_{k+1}$
    Update $w_{k+1}=\theta_k+\alpha_k\phi(s_k)\left[(\hat{\mathbf{T}}^\pi V_{\theta_k})(s_k,a_k,y_{k+1})-\phi(s_k)^\top\theta_k\right]$
end for
Output: $\theta_K$
```

We adopt the following standard assumption from robust RL studies.

Algorithm 4 Robust Q-Natural Policy Gradient (RQNPG)

```
Input: $\theta, \eta, w, N$
Initialization: $u_0, s_0$
for $n = 0, 1, \dots, N-1$ do
    Sample $a_n \sim \pi_w(\cdot|s_n)$, $y_{n+1}$ according to $\mathsf{P}_{s_k,a_k}$ and determine $s_{n+1}$ from $y_{n+1}$
    Update $u_{n+1} = u_n + \zeta_n \phi(s_n, a_n)\left[(\hat{\mathbf{T}}^\pi V_\theta)(s_n, a_n, y_{n+1}) - \phi(s_n, a_n)^\top u_n\right]$
end for
Output: $w + \eta u_N$
```

Assumption 14.1. There exists $\beta < 1$ such that

$$\gamma q^a_{s,s'} \leq \beta \mathsf{P}^a_{s,s'}, \forall q^a_s \in \mathcal{P}^a_s. \tag{188}$$

This assumption is widely adopted in function approximation studies in robust RL, e.g., (Tamar et al., 2014; Xu & Mannor, 2010; Zhou et al., 2024), to ensure the solvability of function approximation.

We then provide the formal statement of Theorem 5.1.

Theorem 14.2. *Set geometrically increasing step sizes $\eta^t \geq \frac{S^2\mathcal{H}^2}{\epsilon S\mathcal{H} - \epsilon^2}\eta^{t-1}$ for each $t = 1, 2, \dots, T$. Set $N, K = \mathcal{O}(\mathcal{H}\epsilon^{-2})$ in Algorithm 3 and Algorithm 4, and $T \geq \frac{\mathcal{H}\log\mathcal{H}}{\epsilon}$, then*

$$g^*_{\mathcal{P}} - \mathbb{E}[g^{\pi_T}_{\mathcal{P}}] \leq C\left(1 - \frac{\epsilon}{\mathcal{H}S}\right)^{T-1} + \frac{M\mathcal{H}^2}{\epsilon^2}\epsilon_e^2,$$

where C is some constant, and $\epsilon_e = \epsilon_{stat} + \epsilon_{bias}$ measures the approximation error .

Proof. To apply Theorem 1 of (Zhou et al., 2024), it is assumed that for initial state distribution ρ, there exists M such that $\sup_{\kappa \in \mathcal{P}} \|\frac{d^{*,\kappa}_\rho}{\rho}\|_\infty \leq M < \infty$. We note that since $g^\pi_{\mathcal{P}}$ is a constant and does not depend on the initial distribution, we can simply set uniform initial distribution $\rho = (\frac{1}{S}, ..., \frac{1}{S})$, in which case $M = S$.

By applying Theorem 1 of (Zhou et al., 2024) with $\gamma = 1 - \frac{\epsilon}{\mathcal{H}}$, it holds that

$$V^*_{\gamma,\mathcal{P}} - \mathbb{E}[V^{\pi_T}_{\gamma,\mathcal{P}}] \leq \left(1 - \frac{\epsilon}{\mathcal{H}S}\right)^{T-1} C + \frac{S\mathcal{H}^2}{\epsilon}\epsilon_e^2. \tag{189}$$

Note that $\epsilon_e = \tilde{O}(\frac{1}{\sqrt{K}} + \frac{1}{\sqrt{N}})$ when omitting the critic error, thus setting $N, K = \mathcal{O}(S\mathcal{H}\epsilon^{-2})$ implies that

$$V^*_{\gamma,\mathcal{P}} - \mathbb{E}[V^{\pi_T}_{\gamma,\mathcal{P}}] \leq \left(1 - \frac{\epsilon}{\mathcal{H}S}\right)^{T-1} C + S\mathcal{H}. \tag{190}$$

Note that if we set T large enough so that $\left(1 - \frac{\epsilon}{\mathcal{H}S}\right)^{T-1} \leq \mathcal{H}$, which can be satisfied if $T \geq \frac{\mathcal{H}\log\mathcal{H}}{\epsilon}$, then π_T is an $\mathcal{H}$-optimal policy for the robust DMDP. Combining the above with Theorem 3.4 completes the proof. □

14.1. Details for Experiments

To expand on our results in Section 6, we provide additional detail here. In our experiments, for different discount factors γ, we run the corresponding RNAC algorithm (Zhou et al., 2024) to learn a robust policy, and then we estimate the robust average reward of the learned policy by calculating the average reward under environment perturbations. Experiments were performed in two custom perturbed MuJoCo environments, Walker2d-v3 and Hopper-v3, using the integral probability metric (IPM) uncertainty set (Zhou et al., 2024). We evaluate the performance of our algorithm with different perturbation levels, and the results, including one standard deviation from the mean, are shown below in Figure 10 for varying perturbation levels. These experimental results align with our theoretical results, and thus verify the scalability of combining our framework with function approximation.

To show that our framework does not result in high computational costs regarding varying reduction factors, we show the execution time of our method using these factors. As can be seen in Tables 1 and 2, even with a very large value for γ, the

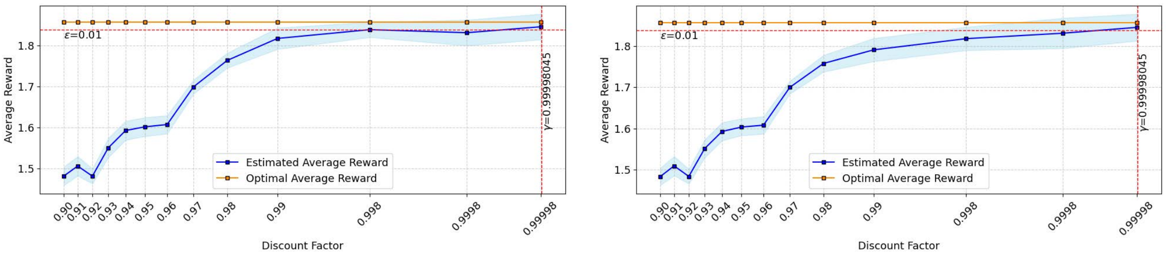

Figure 10. Scalability under Walker2d-v3.

execution time is similar, thus showing that our method is also computationally efficient. Similar to how we showed in Section 6, we show additional results comparing our robust reduction method to that of the non-robust reduction method in Figures 11 and 12 for neural network approximation in the Walker2d-v3 and Hopper-v3 environments, respectively. Additionally, while the focus of this work is on distributional robustness, we hypothesize that our reduction framework *should* work under the adversarial robustness formulation. Due to this idea and inspired by (Zhang et al., 2020), we conducted a preliminary experiment on (discounted) adversarial robust RL in the Humanoid-v4 environment using increasing factors of γ. As our result in Figure 13 shows, the reward under attack increases as γ increases, indicating the potential to develop a similar reduction framework for this setting.

Table 1. Computational efficiency under the Walker2d-v3 environment.

PHASE	γ=[0.9, 0.99, 0.999, 0.9999, 0.99998, 0.999980448383733]
TRAINING	603.92, 630.72, 690.33, 706.32, 711.70, 676.36
EVALUATION	5.93, 25.60, 10.03, 9.16, 10.44, 6.05

Table 2. Computational efficiency under the Hopper-v3 environment.

PHASE	γ=[0.9, 0.99, 0.999, 0.9999, 0.99998, 0.9999925319260162]
TRAINING	579.41, 634.98, 678.29, 682.53, 686.92, 683.56
EVALUATION	6.42, 37.80, 7.15, 5.80, 7.10, 6.84

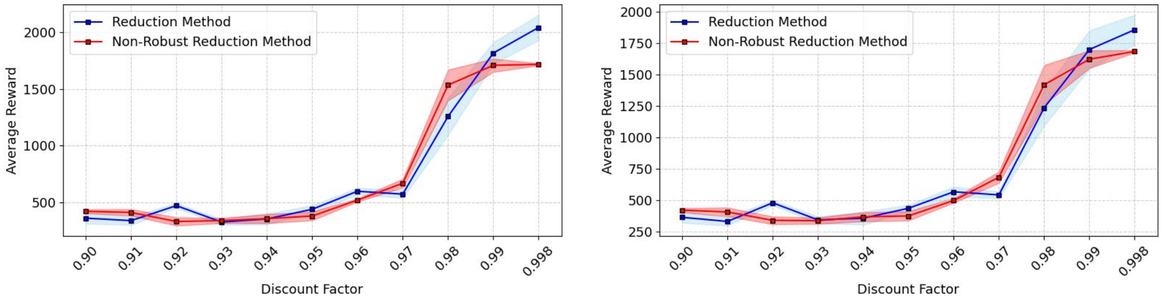

Figure 11. Neural network approximation under Walker2d-v3.

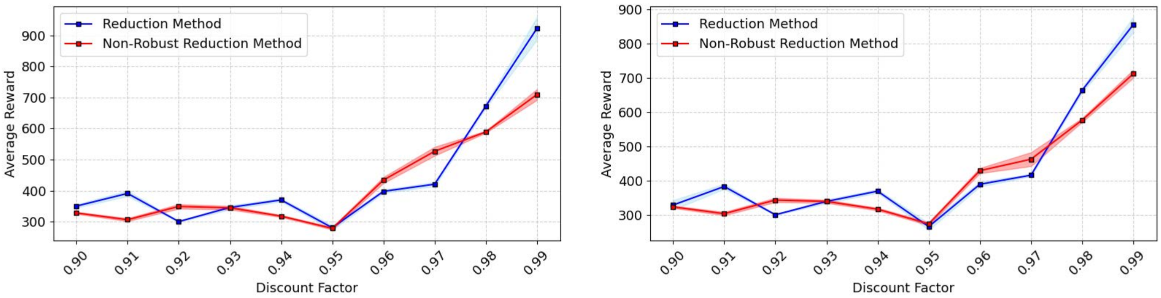

Figure 12. Neural network approximation under Hopper-v3.

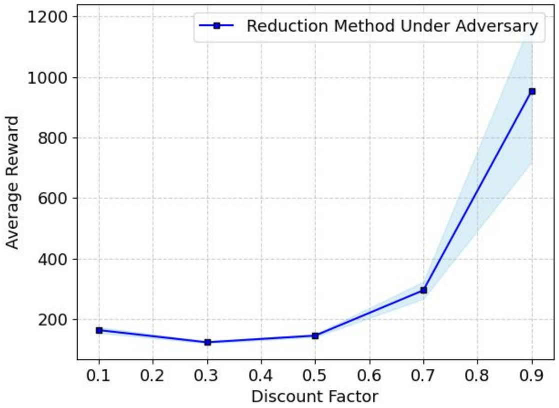

Figure 13. Preliminary discounted adversarial robust RL under Humanoid-v4.

15. Model-Free RL for Robust AMDPs

15.1. Multi-Level Monte Carlo Robust Q-Learning

One potential way to improve scalability is to design model-free algorithms. In contrast to the model-based method in the last section, model-free methods do not store the transition kernels and aim to learn the optimal policy directly. To illustrate the applicability of our framework and ensure it's scalability, we develop two model-free algorithms that can be applied efficiently for large-scale problems. First, we introduce an algorithm that combines our reduction framework with the multi-level Monte Carlo (MLMC) technique (Liu et al., 2022b; Wang et al., 2023e; Blanchet & Glynn, 2015; Blanchet et al., 2019; Wang & Wang, 2022), and provide a theoretical analysis of its sample complexity. We further design another mini-batch model-free robust Q-learning algorithm for robust AMDP in Section 15.2. Both algorithms do not require model estimation and storage, making them scalable for robust RL with average reward.

The MLMC approach is widely used to construct an unbiased estimator of the robust Bellman operator, which is challenging due to its non-linear dependence on the nominal kernel[9] (Wang et al., 2023e). It relies on a geometric distribution with parameter $\Psi \in (0, 1)$, and requires 2^{N+1} samples at each step, with $N \sim \textbf{Geom}(\Psi)$, to construct an unbiased estimator of $\sigma_{\mathcal{P}_s^a}(V)$.

By adapting the MLMC estimator with our reduction-based framework, we propose the MLMC robust Q-learning algorithm for robust AMDPs, given in Algorithm 5.

Our algorithm is hence the first model-free one for robust RL under average reward, along with finite sample analysis. These results demonstrate the broad applicability of our reduction-based framework, enabling the direct integration of any model-free algorithm designed for robust DMDPs to yield algorithms for robust AMDPs with sample complexity guarantees.

[9] Specifically, the support function $\sigma_{\mathcal{P}_s^a}(V)$ is not linear in P_s^a.

Algorithm 5 MLMC robust Q-learning for robust AMDPs

```
Input: A generative model of $(\mathcal{S}, \mathcal{A}, \mathsf{P}, r)$, uncertainty radius $\sigma$, robust bias span $\mathcal{H}$, accuracy $\epsilon$, $\Psi = 0.5$, threshold $N_m$
Initialization:$Q_1 \leftarrow 0$, $s_0$
$\gamma \leftarrow 1 - \frac{\epsilon}{\mathcal{H}}, T \leftarrow \frac{1}{\mathcal{H}^2(1-\gamma)^5}$
for $t < T$ do
    for all $s \in \mathcal{S}, a \in \mathcal{A}$ do
        $V(s) \leftarrow \max_a Q(s,a)$
        Sample $N \sim$ Geom($\Psi$)
        $N \leftarrow \min\{N_m, N\}$
        Sample $2^{N+1}$ samples following $\mathsf{P}_s^a$
        Obtain the multi-level estimator according to (191)
        $Q(s,a) \leftarrow r(s,a) + \gamma \hat{\sigma}_{\mathcal{P}_s^a}(V)$
    end for
end for
$\hat{\pi}_\gamma(s) \leftarrow \arg\max_a Q(s,a), \forall s$
Output: $\hat{\pi}_\gamma$
```

As a result, our framework is scalable to large-scale problems while maintaining high data efficiency.

The MLMC operator is based on a geometric distribution with parameter $\Psi \in (0,1)$. For any s, a, we first generate a number N from a geometric distribution with parameter $\Psi \in (0,1)$. Then, we take action a at state s for 2^{N+1} times, and observe $r(s,a)$ and the subsequent state $\{s_i'\}, i = 1, \ldots, 2^{N+1}$. We divide these 2^{N+1} samples into two groups: samples with odd indices, and samples with even indices. We then individually calculate the empirical distribution of s' using the even-index samples, the odd-index samples, all the samples, and the first sample: $\hat{\mathsf{P}}^{a,E}_{s,N+1} = \frac{1}{2^N}\sum_{i=1}^{2^N} \mathbf{1}_{s'_{2i}}$, $\hat{\mathsf{P}}^{a,O}_{s,N+1} = \frac{1}{2^N}\sum_{i=1}^{2^N} \mathbf{1}_{s'_{2i-1}}$, $\hat{\mathsf{P}}^{a}_{s,N+1} = \frac{1}{2^{N+1}}\sum_{i=1}^{2^{N+1}} \mathbf{1}_{s'_{i}}$, $\hat{\mathsf{P}}^{a,1}_{s,N+1} = \mathbf{1}_{s'_1}$. Then, we use these estimated transition kernels as nominal kernels to construct four estimated uncertainty sets (with the same uncertainty radius): $\hat{\mathcal{P}}^{a,E}_{s,N+1}, \hat{\mathcal{P}}^{a,O}_{s,N+1}, \hat{\mathcal{P}}^{a}_{s,N+1}, \hat{\mathcal{P}}^{a,1}_{s,N+1}$. The multi-level estimator is then defined as

$$\hat{\sigma}_{\mathcal{P}_s^a}(V) \triangleq \sigma_{\hat{\mathcal{P}}^{a,1}_{s,N+1}}(V) + \frac{\Delta_N(V)}{p_N}, \tag{191}$$

where $p_N = \Psi(1-\Psi)^N$ and

$$\Delta_N(V) \triangleq \sigma_{\hat{\mathcal{P}}^{a}_{s,N+1}}(V) - \frac{\sigma_{\hat{\mathcal{P}}^{a,E}_{s,N+1}}(V) + \sigma_{\hat{\mathcal{P}}^{a,O}_{s,N+1}}(V)}{2}.$$

Our threshold-MLMC estimator is constructed as follows

$$\hat{\sigma}_{\mathcal{P}_s^a}(V) \triangleq \sigma_{\hat{\mathcal{P}}^{a,1}_{s,\min\{N_m+1,N+1\}}}(V) + \frac{\Delta_{\min\{N,N_m\}}(V)}{p_{\min\{N,N_m\}}}.$$

The formal statements on the sample complexity are as follows. The proofs are straightforward by combining our framework with the results in (Wang et al., 2024c).

Theorem 15.1. *(1). For the TV-defined uncertainty set, set $N_m = \frac{2\log T}{\log 2}$ and the step size as $\beta_t = \frac{2\log T}{(1-\gamma)T}$. Then, the output of Algorithm 5 is an ϵ-optimal policy for the robust AMDP if*

$$N \geq \widetilde{\mathcal{O}}\left(\frac{SA\mathcal{H}^3}{\epsilon^5}\right). \tag{192}$$

(2). For the Chi-Square-divergence-defined uncertainty set, set $N_m = \frac{2\log T}{\log 2}$ and step size as $\beta_t = \frac{2\log T}{(1-\gamma)T}$. Then the output of Algorithm 5 is an ϵ-optimal policy for the robust AMDP if

$$N \geq \widetilde{\mathcal{O}}\left(\frac{SA\mathcal{H}^3}{\epsilon^5}\right). \tag{193}$$

(3). For the KL-divergence-defined uncertainty set, set

$$N_m = \max\left\{\frac{2\log T}{\log 2}, \frac{\log(1+p_\wedge^2\log(2S)\log T)}{\log 2}\right\},$$

where $p_\wedge$ is the minimal positive entry of the nominal kernel P. *Then, the output of Algorithm 5 is an ϵ-optimal policy for the robust AMDP if*

$$N \geq \widetilde{\mathcal{O}}\left(\frac{SA\mathcal{H}^3}{p_\wedge^2\epsilon^5}\right). \tag{194}$$

15.2. Mini-Batch Robust Q-Learning with Variance Reduction

In this section, we present a model-free mini-batch Q-learning algorithm. The algorithm is derived from the robust DMDP algorithm (Wang et al., 2023c), which employs a variance reduction technique to improve the sample complexity. When combined with our framework, we can also achieve improved sample complexity for robust AMDPs. The details can be found in Algorithm 6.

Algorithm 6 Mini-batch robust Q-learning for robust AMDPs

```
Input: A generative model of $(\mathcal{S}, \mathcal{A}, \mathsf{P}, r)$, uncertainty radius $\sigma$, robust bias span $\mathcal{H}$, accuracy $\epsilon$, batch size $n$, and behavior policy $\pi$
Initialization: $Q_1 \leftarrow 0$, $s_0$
Set $\gamma \leftarrow 1 - \frac{\epsilon}{\mathcal{H}}$
for $t < T$ do
    $V_t(s) \leftarrow \max_a Q_t(s,a), \forall s$
    Sample $a_t \sim \pi(\cdot|s_t)$
    Sample $n$ samples from $\mathsf{P}_{s_t}^{a_t}$ and obtain empirical uncertainty set $\hat{\mathcal{P}}_{s_t}^{a_t}$
    $\lambda_t \leftarrow \frac{1}{1+(1-\gamma)t}$
    $Q_{t+1}(s_t,a_t) \leftarrow (1-\lambda_t)Q_t(s_t,a_t) + \lambda_t(r(s_t,a_t) + \gamma\sigma_{\hat{\mathcal{P}}_{s_t}^{a_t}}(V_t))$
    Sample the next state $s_{t+1} \sim \mathsf{P}_{s_t}^{a_t}$ from the generative model
end for
$\hat{\pi}_\gamma(s) \leftarrow \arg\max_a Q_T(s,a), \forall s$
Output: $\hat{\pi}_\gamma$
```

We then derive the sample complexity of the mini-batch robust Q-learning algorithm, as stated in the following theorem.

Theorem 15.2. *Consider the uncertainty set defined by the KL divergence. If we set*[10]

$$T = \tilde{\mathcal{O}}\left(\frac{SA\mathcal{H}^2}{\epsilon^2}\right), n = \tilde{\mathcal{O}}\left(\frac{\mathcal{H}}{\epsilon}\right) \tag{195}$$

in Algorithm 6, then the output policy $\hat{\pi}_\gamma$ of Algorithm 6 is an ϵ-optimal policy for robust AMDP.

The result shows that the mini-batch robust Q-learning requires $\tilde{\mathcal{O}}\left(\frac{SA\mathcal{H}^3}{\epsilon^3}\right)$ samples to obtain an ϵ-optimal policy for the robust AMDP. This improves the sample complexity for our MLMC robust Q-learning algorithm and represents the state-of-the-art in robust AMDP model-free sample complexity.

15.3. Numerical Experiments

To exemplify the scalability of our framework for robust AMDPs, we present concrete proof of convergence of our model-free *MLMC robust Q-learning* algorithm in figure 14. We first created the uncertainty set using total variation. Then for each γ value we wanted to test, we generate a number N from a geometric distribution with $\Psi \in (0,1)$ and take action a at state

[10] We omit the complex dependencies of T and n on other parameters, including the uncertainty set radius, to highlight the potential of our framework.

s for 2^{N+1} times in order to learn the optimal policy under the robust discounted DMDP. Once we have the ϵ_γ-optimal policy, $\hat{\pi}_\gamma$, we then applied algorithm 1 from (Wang et al., 2023d) to evaluate the robust average reward for each policy. This experiment was independently conducted 5 times, where we plot the mean of the estimated robust average reward for every γ factor along with 1 standard deviation above and below the mean in figure 14. Following this, we then implemented algorithm 2 from (Wang et al., 2023d) in order to find the optimal robust average reward. By combining our reduction-based framework with the multi-level Monte Carlo (MLMC) technique (Liu et al., 2022b; Wang et al., 2023e; Blanchet & Glynn, 2015; Blanchet et al., 2019; Wang & Wang, 2022), we are able to prove our theoretical results in showing that our model-free algorithm converges to the optimal robust average reward.

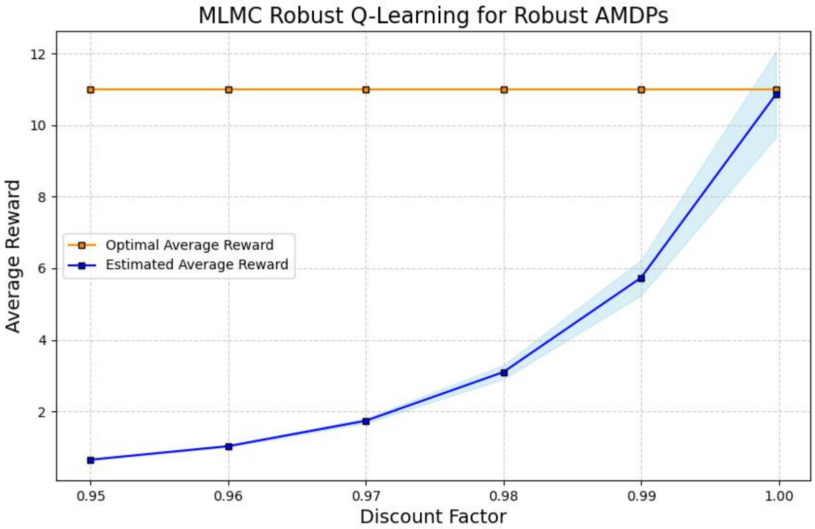

Figure 14. Model-Free Experimental Results Under Total Variation