Model-Free Robust Average-Reward Reinforcement Learning

Yue Wang [1] **Alvaro Velasquez** [2] **George Atia** [3] **Ashley Prater-Bennette** [4] **Shaofeng Zou** [1]

Abstract

Robust Markov decision processes (MDPs) address the challenge of model uncertainty by optimizing the worst-case performance over an uncertainty set of MDPs. In this paper, we focus on the robust average-reward MDPs under the model-free setting. We first theoretically characterize the structure of solutions to the robust average-reward Bellman equation, which is essential for our later convergence analysis. We then design two model-free algorithms, robust relative value iteration (RVI) TD and robust RVI Q-learning, and theoretically prove their convergence to the optimal solution. We provide several widely used uncertainty sets as examples, including those defined by the contamination model, total variation, Chi-squared divergence, Kullback-Leibler (KL) divergence and Wasserstein distance.

1. Introduction

Reinforcement learning (RL) has demonstrated remarkable success in applications like robotics, finance, and games. However, RL often suffers from a severe performance degradation when deployed in real-world environments, which is often due to the model mismatch between the training and testing environments. Such a model mismatch could be a result of non-stationary conditions, modeling errors between simulated and real systems, external perturbations and adversarial attacks. To address this challenge, a framework of robust MDPs and robust RL was developed in (Nilim & El Ghaoui, 2004; Iyengar, 2005; Bagnell et al., 2001), where an uncertainty set of MDPs is constructed, and a pessimistic approach is adopted to optimize the worst-case performance over the uncertainty set. Such a minimax approach guarantees performance for all MDPs within the uncertainty set, making it robust to model mismatch.

[1]University at Buffalo [2]University of Colorado Boulder [3]University of Central Florida [4]Air Force Research Laboratory. Correspondence to: Yue Wang <ywang294@buffalo.edu>, Shaofeng Zou <szou3@buffalo.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Two performance criteria are commonly used for infinite-horizon MDPs: 1) the discounted-reward setting, where the reward is discounted exponentially with time; and 2) the average-reward setting, where the long-term average-reward over time is of interest. For systems that operate for an extended period of time, e.g., queue control, inventory management in supply chains, or communication networks, it is more important to optimize the average-reward since policies obtained from the discounted-reward setting may be myopic and have poor long-term performance (Kazemi et al., 2022). However, much of the work on robust MDPs has focused on the discounted-reward setting, and robust average-reward MDPs remain largely unexplored.

Robust MDPs under the two criteria are fundamentally different. Compared to the discounted-reward setting, the average-reward setting places equal weight on immediate and future rewards, thus depends on the limiting state-action frequencies of the underlying MDPs. This makes it more challenging to investigate than the discounted-reward setting. For example, to establish a contraction in the discounted-reward setting, it generally suffices to have a discount factor strictly less than one, whereas in the average-reward setting, convergence depends on the structure of the MDP. Studies on robust average-reward MDPs are quite limited in the literature, e.g., (Tewari & Bartlett, 2007; Lim et al., 2013; Wang et al., 2023), and are model-based, where the uncertainty set is fully known to the learner. However, the more practical model-free setting, where only samples from the nominal MDP (the centroid of the uncertainty set) can be observed, has yet to be explored. Algorithms and fundamental convergence guarantees in this setting have not been established yet, which is the focus of this paper.

1.1. Challenges and Contributions

In this paper, we develop a fundamental characterization of solutions to the robust average-reward Bellman equation, design model-free algorithms with provable optimality guarantees, and construct an *unbiased* estimator of the robust average-reward Bellman operator for various types of uncertainty sets. Our results fill the gap in model-free algorithms for robust average-reward RL, and are fundamentally different from existing studies on robust discounted RL and robust and non-robust average-reward MDPs. In the following, we summarize the major challenges and our key contributions.

We characterize the fundamental structure of solutions to the robust average-reward Bellman equation. Value-based approaches typically converge to a solution to the (robust) Bellman equation. For example, TD and Q-learning converge to the *unique* fixed-point that yields an optimal solution to the Bellman equation in the discounted setting (Sutton & Barto, 2018). However, in the (robust) average-reward setting, solutions to the (robust) average-reward Bellman equation may not be unique (Puterman, 1994; Wang et al., 2023), and the fundamental structure of the solution space usually plays an important role in proving convergence (e.g., (Wan et al., 2021; Abounadi et al., 2001; Tsitsiklis & Van Roy, 1999)). In the non-robust average-reward setting, the solution to the average-reward Bellman equation is unique up to an additive constant (Puterman, 1994). This result is largely due to the nice linear structure of the non-robust average-reward Bellman equation. However, in the robust setting, where the worst-case performance is of interest, the robust average-reward Bellman equation is nonlinear, which poses a significant challenge in characterizing its solution. We demonstrate that the worst-case transition kernel may not be unique and then show that any solution to the robust average-reward Bellman equation is the relative value function for some worst-case transition kernel (up to an additive constant). We note that this structure is of key importance later in our convergence analysis.

We develop model-free approaches for robust policy evaluation and optimal control. We then design model-free algorithms for robust average-reward RL. To ensure stability of the iterates, we adopt a similar idea as in the non-robust average-reward setting (Puterman, 1994; Bertsekas, 2011), which is to subtract an offset function and design robust RVI TD and Q-learning algorithms. Unlike the convergence proof for model-free algorithms for non-robust average-reward MDPs, e.g., (Wan et al., 2021; Abounadi et al., 2001) and robust discounted MDPs, e.g., (Wang & Zou, 2021; Liu et al., 2022), where the Bellman operator is either linear or a contraction, our robust average-reward Bellman operator is neither linear nor a contraction, which makes the convergence analysis challenging. Nevertheless, based on the solution structure of the robust average-reward Bellman equation discussed above, we prove that our algorithms converge to solutions to the robust average-reward Bellman equations. Specifically, for robust RVI TD, its output converges to the worst-case average-reward; and for the robust RVI Q-learning, the greedy policy w.r.t. the Q-function converges to an optimal robust policy.

We construct unbiased estimators with bounded variance for robust Bellman operators under various uncertainty set models. One popular approach in RL is bootstrapping, i.e., use $\mathbf{H}Q$ as an estimate of the true Q-function, where $\mathbf{H}$ is the (robust) Bellman operator. In the non-robust setting, an unbiased estimate of $\mathbf{H}$ can be easily derived because $\mathbf{H}$ is linear in the transition kernel. This linearity, however, does not hold in the robust setting since we minimize over the transition kernel to account for the worst-case and, therefore, directly plugging in the empirical transition kernel results in a biased estimate. In this paper, we employ the multi-level Monte-Carlo method (Blanchet & Glynn, 2015) and construct unbiased estimators for five popular uncertainty models, including those defined by contamination models, total variation, Chi-squared divergence, Kullback-Leibler (KL) divergence, and Wasserstein distance. We then prove that our estimator is unbiased and has a bounded variance. These properties play an important role in establishing the convergence of our algorithms.

1.2. Related Work

Robust average-reward MDPs. Studies on robust average-reward MDPs are quite limited in the literature. Model-based robust average-reward MDPs were first studied in (Tewari & Bartlett, 2007), where the authors showed the existence of the Blackwell optimal policy for a specific uncertainty set of bounded parameters. Results for general uncertainty sets were developed in recent work (Wang et al., 2023), following a fundamentally different approach from (Tewari & Bartlett, 2007). However, these two methods are model-based. The work of (Lim et al., 2013) designed a model-free algorithm for the uncertainty set defined by total variation and characterized its regret bound. However, their method cannot be generalized to other uncertainty sets. In this paper, we design the first model-free method for general uncertainty sets and provide fundamental insights into the model-free robust average-reward RL problem.

Robust discounted-reward MDPs. Model-based methods for robust discounted MDPs were studied in (Iyengar, 2005; Nilim & El Ghaoui, 2004; Bagnell et al., 2001; Satia & Lave Jr, 1973; Wiesemann et al., 2013; Lim & Autef, 2019; Xu & Mannor, 2010; Yu & Xu, 2015; Lim et al., 2013; Tamar et al., 2014; Neufeld & Sester, 2022), where the uncertainty set is assumed to be known, and the problem can be solved using robust dynamic programming. Later, the studies were generalized to the model-free setting (Roy et al., 2017; Badrinath & Kalathil, 2021; Wang & Zou, 2021; 2022; Tessler et al., 2019; Liu et al., 2022; Zhou et al., 2021; Yang et al., 2021; Panaganti & Kalathil, 2021; Goyal & Grand-Clement, 2018; Kaufman & Schaefer, 2013; Ho et al., 2018; 2021). Our work focuses on the average-reward setting. First, the robust average-reward Bellman operator does not come with a simple contraction property like the one for the discounted setting, and the average-reward depends on the limiting behavior of the underlying MDP. Moreover, the average-reward Bellman function is a function of two variables: the average-reward and the relative value function, whereas its discounted-reward counterpart is a function of only the value function. These challenges make the robust

average-reward problem more intricate. Furthermore, existing studies mostly focus on a certain type of uncertainty set, e.g., contamination (Wang & Zou, 2021) and total variation (Panaganti et al., 2022); in this paper, our method can be used to solve a wide range of uncertainty sets.

Non-robust average-reward MDPs. Early contributions to non-robust average-reward MDPs include a fundamental characterization of the problem and model-based methods (Puterman, 1994; Bertsekas, 2011). Model-free methods in the tabular setting, e.g., RVI Q-learning (Abounadi et al., 2001) and differential Q-learning (Wan et al., 2021; Wan & Sutton, 2022), were developed recently and are both shown to converge to the optimal average-reward. There is also work on average-reward RL with function approximation, e.g., (Zhang et al., 2021b; Tsitsiklis & Van Roy, 1999; Zhang et al., 2021a; Yu & Bertsekas, 2009). In this paper, we focus on the robust setting, where the key challenge lies in the non-linearity of the robust average-reward Bellman equation, whereas it is linear in the non-robust setting.

2. Preliminaries and Problem Formulation

Average-reward MDPs. An MDP $(\mathcal{S}, \mathcal{A}, \mathsf{P}, r)$ 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, and a reward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$. At each time step t, the agent at state s_t takes an action a_t, the environment then transitions to the next state $s_{t+1} \sim \mathsf{P}_{s_t}^{a_t}$, and provides a reward signal $r_t \in [0, 1]$.

A stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ maps a state to a distribution over $\mathcal{A}$, following which the agent takes action a at state s with probability $\pi(a|s)$. Under a transition kernel P, the average-reward of π starting from $s \in \mathcal{S}$ is defined as

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

The relative value function is defined to measure the cumulative difference between the reward and g_{P}^{π}:

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

Then $(g_{\mathsf{P}}^{\pi}, V_{\mathsf{P}}^{\pi})$ satisfies the following Bellman equation (Puterman, 1994):

$$V_{\mathsf{P}}^{\pi}(s) = \mathbb{E}_{\pi,\mathsf{P}}\left[r(s, A) - g_{\mathsf{P}}^{\pi}(s) + \sum_{s'\in\mathcal{S}} p_{s,s'}^{A} V_{\mathsf{P}}^{\pi}(s')\right]. \tag{3}$$

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

Robust average-reward MDPs. For robust MDPs, the transition kernel is assumed to be in some uncertainty set $\mathcal{P}$. At each time step, the environment transits to the next state according to an arbitrary transition kernel $\mathsf{P} \in \mathcal{P}$. In this paper, we focus on the (s, a)-rectangular uncertainty set (Nilim & El Ghaoui, 2004; Iyengar, 2005), i.e., $\mathcal{P} = \bigotimes_{s,a} \mathcal{P}_s^a$, where $\mathcal{P}_s^a \subseteq \Delta(\mathcal{S})$. Popular uncertainty sets include those defined by the contamination model (Huber, 1965; Wang & Zou, 2022), total variation (Lim et al., 2013), Chi-squared divergence (Iyengar, 2005), Kullback-Leibler (KL) divergence (Hu & Hong, 2013) and Wasserstein distance (Gao & Kleywegt, 2022). We will investigate these uncertainty sets in detail in Section 5.

We investigate the worst-case average-reward over the uncertainty set of MDPs. Specifically, define the robust average-reward of a policy π as

$$g_{\mathcal{P}}^{\pi}(s) \triangleq \min_{\kappa\in\bigotimes_{n\geq 0}\mathcal{P}} \lim_{T\to\infty} \mathbb{E}_{\pi,\kappa}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_t | S_0 = s\right], \tag{4}$$

where $\kappa = (\mathsf{P}_0, \mathsf{P}_1 ...) \in \bigotimes_{n\geq 0} \mathcal{P}$. It was shown in (Wang et al., 2023) that the worst case under the time-varying model is equivalent to the one under the stationary model:

$$g_{\mathcal{P}}^{\pi}(s) = \min_{\mathsf{P}\in\mathcal{P}} \lim_{T\to\infty} \mathbb{E}_{\pi,\mathsf{P}}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_t | S_0 = s\right]. \tag{5}$$

Therefore, we limit our focus to the stationary model. We refer to the minimizers of (5) as the worst-case transition kernels for the policy π, and denote the set of all possible worst-case transition kernels by Ω_g^{π}, i.e., $\Omega_g^{\pi} \triangleq \{\mathsf{P} \in \mathcal{P} : g_{\mathsf{P}}^{\pi} = g_{\mathcal{P}}^{\pi}\}$. As shown in Example A.1 in the appendix, the worst-case transition kernel may not be unique.

We focus on the model-free setting, where only samples from the nominal MDP (centroid of the uncertainty set) are available. We investigate two problems: 1) given a policy π, estimate its robust average-reward $g_{\mathcal{P}}^{\pi}$, and 2) find an optimal robust policy that optimizes the robust average-reward:

$$\max_{\pi} g_{\mathcal{P}}^{\pi}(s), \text{ for any } s \in \mathcal{S}. \tag{6}$$

We denote by $g_{\mathcal{P}}^{*}(s) \triangleq \max_{\pi} g_{\mathcal{P}}^{\pi}(s)$ the optimal robust average-reward.

3. Robust RVI TD for Policy Evaluation

In this section, we study the problem of policy evaluation, which aims to estimate the robust average-reward $g_{\mathcal{P}}^{\pi}$ for a fixed policy π.

For technical convenience, we make the following assumption to guarantee that the average-reward is independent of the initial state (Abounadi et al., 2001; Wan et al., 2021; Zhang et al., 2021a; Zhang & Ross, 2021; Chen et al., 2022; Wang et al., 2023).

Assumption 1. The Markov chain induced by π is a unichain for all $\mathsf{P} \in \mathcal{P}$.

In general, the average-reward depends on the initial state. For example, imagine a policy that induces a multichain in an MDP with two closed communicating classes. A learning algorithm would be able to learn the average-reward for each communicating class; however, the average-rewards for the two classes may be different. To remove this complexity, it is common and convenient to rule out this possibility. Under Assumption 1, the average-reward w.r.t. any $\mathsf{P} \in \mathcal{P}$ is identical for any start state, i.e., $g^\pi_\mathsf{P}(s) = g^\pi_\mathsf{P}(s'), \forall s, s' \in \mathcal{S}$.

We first revisit the robust average-reward Bellman equation in (Wang et al., 2023), and further characterize the structure of its solutions. For $V \in \mathbb{R}^{|\mathcal{S}|}$, denote by $\sigma_{\mathcal{P}^a_s}(V) \triangleq \min_{p \in \mathcal{P}^a_s} pV$, $\mathsf{P}_V(s,a) \triangleq \arg\min_{p \in \mathcal{P}^a_s} pV$ and let $\mathsf{P}_V = \{\mathsf{P}_V(s,a), s \in \mathcal{S}, a \in \mathcal{A}\}$.

Theorem 3.1 (Robust Bellman equation)**.** *If (g, V) is a solution to the robust Bellman equation*

$$V(s) = \sum_a \pi(a|s)(r(s,a) - g + \sigma_{\mathcal{P}^a_s}(V)), \forall s, \quad (7)$$

then 1) $g = g^\pi_\mathcal{P}$ (Wang et al., 2023); 2) $\mathsf{P}_V \in \Omega^\pi_g$; 3) $V = V^\pi_{\mathsf{P}_V} + ce$ for some $c \in \mathbb{R}$, where e denotes the vector $(1, 1, ..., 1) \in \mathbb{R}^{|\mathcal{S}|}$.

The robust Bellman equation in (7) was initially developed in (Wang et al., 2023). The authors also defined a robust relative value function:

$$V^\pi_\mathcal{P}(s) \triangleq \min_{\mathsf{P} \in \mathcal{P}} \mathbb{E}_{\pi,\mathsf{P}} \left[\sum_{n=0}^\infty (r_n - g^\pi_\mathcal{P}) | S_0 = s \right], \quad (8)$$

which is the worst-case relative value function, and showed that $(g^\pi_\mathcal{P}, V^\pi_\mathcal{P})$ is a solution to (7). However, conversely, it may not be the case that any solution (g, V) to (7) can be written as $V = V^\pi_\mathcal{P} + ce$ for some $c \in \mathbb{R}$. This is in contrast to the results for the non-robust average-reward setting, where the set of solutions to the non-robust average-reward Bellman equation can be written as $\{(g^\pi_\mathsf{P}, V^\pi_\mathsf{P} + ce) : c \in \mathbb{R}\}$ (Puterman, 1994).

In Theorem 3.1, we show that for any solution (g, V) to (7), the transition kernel $P_V \in \Omega^\pi_g$, i.e., it is a worst-case transition kernel for $g^\pi_\mathcal{P}$. Moreover, any solution to (7) can be written as $V = V^\pi_{\mathsf{P}_V} + ce$ for some constant c. This is different from the non-robust setting, as $V^\pi_{\mathsf{P}_V}$ actually depends on V. As will be seen later, this result is crucial to establish the convergence of our robust RVI TD.

Theorem 3.1 also reveals the fundamental difference between the robust and the non-robust average-reward settings. Under the non-robust setting, the solution set to the Bellman equation can be written as $\{(g^\pi_\mathsf{P}, V^\pi_\mathsf{P} + ce) : c \in \mathbb{R}\}$. The solution is uniquely determined by the transition kernel (up to some constant vector ke). In contrast, in the robust setting, the robust Bellman equation is no longer linear. Any solution V to (7) is a relative value function w.r.t. *some* worst-case transition kernel $\mathsf{P} \in \Omega^\pi_g$ (up to some additive constant vector), i.e., $V \in \{V^\pi_\mathsf{P} + ce : \mathsf{P} \in \Omega^\pi_g, c \in \mathbb{R}\}$.

A natural question that arises is whether, for any $\mathsf{P} \in \Omega^\pi_g$, $(g^\pi_\mathcal{P}, V^\pi_\mathsf{P})$ is a solution to (7)? Lemma 3.1 refutes this.

Lemma 3.1. *There exists a robust MDP such that for some $\mathsf{P} \in \Omega^\pi_g$, $(g^\pi_\mathcal{P}, V^\pi_\mathsf{P})$ is not a solution to* (7)*.*

Lemma 3.1 implies that the solution set to (7) is a subset of $\{V^\pi_\mathsf{P} + ce, \mathsf{P} \in \Omega^\pi_g, c \in \mathbb{R}\}$. Note that explicit characterization of the solution set to (7) is challenging due to its non-linear structure; however, result 3 in Theorem 3.1 suffices for the convergence proof (see appendix for proofs).

Motivated by the robust Bellman equation in (7), we propose a model-free robust RVI TD algorithm in Algorithm 1, where $\hat{\mathbf{T}}$ and function f will be discussed later.

Algorithm 1 Robust RVI TD

Input: $V_0, \alpha_n, n = 0, 1, ..., N-1$

```
for n = 0, 1, ..., N − 1 do
    for all s ∈ S do
        V_{n+1}(s) ← V_n(s) + α_n(T̂V_n(s) − f(V_n) − V_n(s))
    end for
end for
```

Note that (7) can be written as $V = \mathbf{T}V - g$, where $\mathbf{T}$ is the robust average-reward Bellman operator. Since in the model-free setting $\mathcal{P}$ is unknown, in Algorithm 1, we construct $\hat{\mathbf{T}}V$ as an estimate of $\mathbf{T}V$ satisfying

$$\mathbb{E}[\hat{\mathbf{T}}V] = \mathbf{T}V, \quad \text{Var}[\hat{\mathbf{T}}V(s)] \le C(1 + \|V\|^2), \quad (9)$$

for some constant $C > 0$. In this paper, if not specified, $\|\cdot\|$ denotes the infinity norm $\|\cdot\|_\infty$.

It is challenging to construct such $\hat{\mathbf{T}}$ as $\mathbf{T}$ is non-linear in the nominal transition kernel from which samples are generated. In Section 5, we will present in detail how to construct such $\hat{\mathbf{T}}$ for various uncertainty set models.

To make the iterates stable, we follow the idea of RVI in the non-robust setting and introduce an offset function f satisfying the following assumption (Puterman, 1994).

Assumption 2. $f : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}$ is L_f-Lipschitz and satisfies $f(e) = 1, f(x + ce) = f(x) + c, f(cx) = cf(x), \forall c \in \mathbb{R}$.

Assumption 2 can be easily satisfied, e.g., $f(V) = V(s_0)$ for some reference state $s_0 \in \mathcal{S}$, and $f(V) = \frac{\sum_s V(s)}{|\mathcal{S}|}$ (Abounadi et al., 2001). Compared with the discounted setting, f is critical here. As we discussed above, in the average-reward setting, the solution to the Bellman equation

$V + ce$ can be arbitrarily large because c can be any real number. This may lead to a non-convergent sequence V_n (see, e.g., example 8.5.2 of (Puterman, 1994)). Hence, a function f is introduced to "offset" V_n and keep the iterates stable. Also, $f(V_n)$ serves as an estimator of the average-reward $g^\pi_\mathcal{P}$, as we shall see later.

We then assume the Robbins-Monro condition on the stepsize, and further show the convergence of robust RVI TD.

Assumption 3. The stepsize $\{\alpha_n\}_{n=0}^\infty$ satisfies the Robbins-Monro condition, i.e., $\sum_{n=0}^\infty \alpha_n = \infty, \quad \sum_{n=0}^\infty \alpha_n^2 < \infty$.

Theorem 3.2 (Convergence of robust RVI TD)**.** *Under Assumptions 1,2,3, and if* $\hat{\mathbf{T}}$ *satisfies* (9)*, then almost surely,* $(f(V_n), V_n)$ *converges to a solution to* (7) *which may depend on the initialization.*

The result implies that $f(V_n) \to g^\pi_\mathcal{P}$ a.s., which means our robust RVI TD converges to the worst-case average-reward for the given policy π.

Remark 3.1. Our robust RVI TD algorithm is shown to converge to a solution to (7). This model-free result is the same as the model-based results in (Wang et al., 2023). Though it was shown in (Wang et al., 2023) that $(g^\pi_\mathcal{P}, V^\pi_\mathcal{P})$ (defined in (8)) is a solution to (7), it is not guaranteed that the convergence is to $(g^\pi_\mathcal{P}, V^\pi_\mathcal{P} + ce)$ for some c.

Remark 3.2. As we discussed following the statement of Theorem 3.1, result 3) of Theorem 3.1 is crucial to the convergence proof of Theorem 3.2. Specifically, it is necessary in order to characterize the equilibrium of the associated ODE, and thus the limit of the iterates $f(V_n) \to g^\pi_\mathcal{P}$.

4. Robust RVI Q-Learning for Control

In this section, we study the problem of optimal control, which aims to find a policy that optimizes the robust average-reward: $\pi^* = \arg\max_\pi g^\pi_\mathcal{P}$.

Similar to Assumption 1, we make the following assumption to guarantee that the average-reward is independent of the initial state (Abounadi et al., 2001; Li et al., 2022).

Assumption 4. The Markov chain induced by any $\mathsf{P} \in \mathcal{P}$ and any π is a unichain.

We first revisit the optimal robust Bellman equation in (Wang et al., 2023) and further present a characterization of its solutions. Consider a Q-function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and define $V_Q(s) = \max_a Q(s,a), \forall s \in \mathcal{S}$.

Lemma 4.1 (Optimal robust Bellman equation)**.** *If* (g, Q) *is a solution to the optimal robust Bellman equation*

$$Q(s,a) = r(s,a) - g + \sigma_{\mathcal{P}^a_s}(V_Q), \forall s, a, \qquad (10)$$

then 1) $g = g^*_\mathcal{P}$ *(Wang et al., 2023); 2) the greedy policy w.r.t.* Q*:* $\pi_Q(s) = \arg\max_a Q(s,a)$ *is an optimal robust policy (Wang et al., 2023); 3)* $V_Q = V^{\pi_Q}_\mathsf{P} + ce$ *for some* $\mathsf{P} \in \Omega^{\pi_Q}_g, c \in \mathbb{R}$.

According to result 2 in Lemma 4.1, finding a solution to (10) is sufficient to get the optimal robust average-reward and to derive the optimal robust policy. We note that a complete characterization of the solution set to (10) can be obtained similarly to the one in result 3 of Theorem 3.1. Here, we only provide its structure to simplify the presentation and avoid cumbersome notation. This structure as outlined in Lemma 4.1 is sufficient for our convergence analysis.

We hence present the following model-free robust RVI Q-learning algorithm.

Algorithm 2 Robust RVI Q-learning

Input: Q_0, α_n

for $n = 0, ..., N-1$ **do**
  for all $s \in \mathcal{S}, a \in \mathcal{A}$ **do**
    $Q_{n+1}(s,a) \leftarrow Q_n(s,a) + \alpha_n\big(\hat{\mathbf{H}}Q_n(s,a) - f(Q_n) - Q_n(s,a)\big)$
  end for
end for

Similar to the robust RVI TD algorithm, denote the optimal robust Bellman operator by $\mathbf{H}Q(s,a) \triangleq r(s,a) + \sigma_{\mathcal{P}^a_s}(V_Q)$, and we construct an estimate $\hat{\mathbf{H}}$ such that for some finite constant C,

$$\mathbb{E}[\hat{\mathbf{H}}Q] = \mathbf{H}Q, \quad \text{Var}[\hat{\mathbf{H}}Q(s,a)] \le C(1 + \|Q\|^2). \quad (11)$$

In Section 5, we will present in detail how to construct such $\hat{\mathbf{H}}$ for various uncertainty set models.

The following theorem shows the convergence of the robust RVI Q-learning to the optimal robust average-reward $g^*_\mathcal{P}$ and the optimal robust policy π^*.

Theorem 4.1 (Convergence of robust RVI Q-learning)**.** *Under Assumptions 2, 3 and 4, and if* $\hat{\mathbf{H}}$ *satisfies* (11)*, then almost surely,* $(f(Q_n), Q_n)$ *converges to a solution to* (10)*, i.e.,* $f(Q_n)$ *converges to* $g^*_\mathcal{P}$*, and the greedy policy* $\pi_{Q_n}(s) \triangleq \arg\max_a Q_n(s,a)$ *converges to an optimal robust average-reward* π^*.

5. Case Studies

In the previous two sections, we showed that if an unbiased estimator with bounded variance is available for the robust Bellman operator, then both robust algorithms proposed converge to the optimum. In this section, we present the design of these estimators for various uncertainty set models.

The major challenge in designing the estimated operators satisfying (9) and (11) lies in estimating the support function $\sigma_{\mathcal{P}^a_s}(V)$ unbiasedly with bounded variance using samples from the nominal transition kernel. However, the nominal transition kernel P in general is different from the worst-case

transition kernel, and the straightforward estimator is shown to be biased. Specifically, if we centered at the empirical transition kernel $\hat{\mathsf{P}}$ and construct the uncertainty set $\hat{\mathcal{P}}$, then the estimator is biased: $\mathbb{E}[\sigma_{\hat{\mathcal{P}}_s^a}(V)] \neq \sigma_{\mathcal{P}_s^a}(V)$.

We consider several widely-used uncertainty models including the contamination model, the total variation model, the Chi-square model, the KL-divergence model, and the Wasserstein distance model. We show that our estimators are unbiased and have bounded variance in the following theorem. We will present the design in later sections.

Theorem 5.1. *For each uncertainty set, denote its corresponding estimators by $\hat{\mathbf{T}}$ and $\hat{\mathbf{H}}$ as in Sections 5.1 and 5.2. Then, there exists some constant C, such that* (9) *and* (11) *hold.*

In the following sections, we construct an operator $\hat{\sigma}_{\mathcal{P}_s^a}$ to estimate the support function $\sigma_{\mathcal{P}_s^a}, \forall s \in \mathcal{S}, a \in \mathcal{A}$ for each uncertainty set. We further define the estimated robust Bellman operators as $\hat{\mathbf{T}}V(s) \triangleq \sum_a \pi(a|s)(r(s,a) + \hat{\sigma}_{\mathcal{P}_s^a}(V))$ and $\hat{\mathbf{H}}Q(s,a) \triangleq r(s,a) + \hat{\sigma}_{\mathcal{P}_s^a}(V_Q)$.

5.1. Linear Model: Contamination Uncertainty Set

The δ-contamination uncertainty set is $\mathcal{P}_s^a = \{(1-\delta)\mathsf{P}_s^a + \delta q : q \in \Delta(\mathcal{S})\}$, where $0 < \delta < 1$ is the radius. Under this uncertainty set, the support function can be computed as $\sigma_{\mathcal{P}_s^a}(V) = (1-\delta)\mathsf{P}_s^a V + \delta \min_s V(s)$, and this is linear in the nominal transition kernel P_s^a. We hence use the transition to the subsequent state to construct our estimator:

$$\hat{\sigma}_{\mathcal{P}_s^a}(V) \triangleq (1-\delta)\gamma V(s') + \delta \min_x V(x), \tag{12}$$

where s' is a subsequent state sample after (s,a).

5.2. Non-Linear Models

Unlike the δ-contamination model, most uncertainty sets result in a non-linear support function of the nominal transition kernel. We will employ the approach of multi-level Monte-Carlo which is widely used in quantile estimation under stochastic environments (Blanchet & Glynn, 2015; Blanchet et al., 2019; Wang & Wang, 2022) to construct an unbiased estimator with bounded variance.

For any s, a, we first generate N according to 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 $\{s_i'\}, i = 1, ..., 2^{N+1}$. 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, odd-index ones, all the samples, and the first sample: $\hat{\mathsf{P}}_{s,N+1}^{a,E} = \frac{1}{2^N}\sum_{i=1}^{2^N} \mathbb{1}_{s'_{2i}}$, $\hat{\mathsf{P}}_{s,N+1}^{a,O} = \frac{1}{2^N}\sum_{i=1}^{2^N} \mathbb{1}_{s'_{2i-1}}$, $\hat{\mathsf{P}}_{s,N+1}^{a} = \frac{1}{2^{N+1}}\sum_{i=1}^{2^{N+1}} \mathbb{1}_{s'_i}$, $\hat{\mathsf{P}}_{s,N+1}^{a,1} = \mathbb{1}_{s'_1}$. Then, we use these estimated transition kernels as nominal kernels to construct four estimated uncertainty sets $\hat{\mathcal{P}}_{s,N+1}^{a,E}, \hat{\mathcal{P}}_{s,N+1}^{a,O}, \hat{\mathcal{P}}_{s,N+1}^{a}, \hat{\mathcal{P}}_{s,N+1}^{a,1}$. The multi-level estimator is then defined as

$$\hat{\sigma}_{\mathcal{P}_s^a}(V) \triangleq \sigma_{\hat{\mathcal{P}}_{s,N+1}^{a,1}}(V) + \frac{\Delta_N(V)}{p_N}, \tag{13}$$

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

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

We note that in previous results of the multi-level Monte-Carlo estimator (Blanchet & Glynn, 2015; Blanchet et al., 2019; Wang & Wang, 2022), several assumptions are needed to show that the estimator is unbiased. These assumptions, however, do not hold in our cases. For example, the function $\sigma_{\mathcal{P}}(V)$ is not continuously differentiable. Hence, their analysis cannot be directly applied here.

We then present four examples of non-linear uncertainty sets. Under each example, a solution to the support function $\sigma_{\mathcal{P}}(V)$ is given, and by plugging it into (13) the unbiased estimator can then be constructed. More details can be found in Appendix D and Appendix E.

Total Variation Uncertainty Set. The total variation uncertainty set is $\mathcal{P}_s^a = \{q \in \Delta(|\mathcal{S}|) : \frac{1}{2}\|q - \mathsf{P}_s^a\|_1 \leq \delta\}$, and the support function can be computed using its dual function (Iyengar, 2005):

$$\sigma_{\mathcal{P}_s^a}(V) = \max_{\mu \geq 0} \left(\mathsf{P}_s^a(V-\mu) - \delta \mathrm{Span}(V-\mu)\right). \tag{14}$$

Chi-square Uncertainty Set. The Chi-square uncertainty set is $\mathcal{P}_s^a = \{q \in \Delta(|\mathcal{S}|) : d_c(\mathsf{P}_s^a, q) \leq \delta\}$, where $d_c(q,p) = \sum_s \frac{(p(s)-q(s))^2}{p(s)}$. Its support function can be computed using its dual function (Iyengar, 2005):

$$\sigma_{\mathcal{P}_s^a}(V) = \max_{\mu \geq 0} \left(\mathsf{P}_s^a(V-\mu) - \sqrt{\delta \mathrm{Var}_{\mathsf{P}_s^a}(V-\mu)}\right). \tag{15}$$

Kullback–Leibler (KL) Divergence Uncertainty Set. The KL-divergence between two distributions p, q is defined as $D_{KL}(q||p) = \sum_s q(s)\log\frac{q(s)}{p(s)}$, and the uncertainty set defined via KL divergence is

$$\mathcal{P}_s^a = \{q : D_{KL}(q||\mathsf{P}_s^a) \leq \delta\}, \forall s \in \mathcal{S}, a \in \mathcal{A}. \tag{16}$$

Its support function can be efficiently solved using the duality result in (Hu & Hong, 2013):

$$\sigma_{\mathcal{P}_s^a}(V) = -\min_{\alpha \geq 0}\left(\delta\alpha + \alpha \log\left(\mathbb{E}_{\mathsf{P}_s^a}\left[e^{\frac{-V}{\alpha}}\right]\right)\right). \tag{17}$$

The above estimator for the KL-divergence uncertainty set has also been developed in (Liu et al., 2022) for robust discounted MDPs. Its extension to our average-reward setting is similar.

Wasserstein Distance Uncertainty Sets. Consider the metric space $(\mathcal{S}, d)$ by defining some distance metric d. For some parameter $l \in [1, \infty)$ and two distributions $p, q \in \Delta(\mathcal{S})$, define the l-Wasserstein distance between them as $W_l(q,p) = \inf_{\mu \in \Gamma(p,q)} \|d\|_{\mu,l}$, where $\Gamma(p,q)$ denotes the distributions over $\mathcal{S} \times \mathcal{S}$ with marginal distributions p, q, and $\|d\|_{\mu,l} = \left(\mathbb{E}_{(X,Y)\sim\mu}\left[d(X,Y)^l\right]\right)^{1/l}$. The Wasserstein distance uncertainty set is then defined as

$$\mathcal{P}_s^a = \{q \in \Delta(|\mathcal{S}|) : W_l(\mathsf{P}_s^a, q) \leq \delta\}. \tag{18}$$

To solve the support function w.r.t. the Wasserstein distance set, we first prove the following duality lemma.

Lemma 5.1. *It holds that*

$$\sigma_{\mathcal{P}_s^a}(V) = \sup_{\lambda \geq 0} \left(-\lambda \delta^l + \mathbb{E}_{\mathsf{P}_s^a}\left[\inf_y \left(V(y) + \lambda d(S,y)^l\right)\right] \right). \tag{19}$$

Thus, the support function can be solved using its dual form, and the estimator can then be constructed following (13).

6. Experiments

We numerically verify our previous convergence results and demonstrate the robustness of our algorithms. Additional experiments can be found in Appendix G.

6.1. Convergence of Robust RVI TD and Q-Learning

We first verify the convergence of our robust RVI TD and Q-learning algorithms under a Garnet problem $\mathcal{G}(30, 20)$ (Archibald et al., 1995). There are 30 states and 20 actions. The nominal transition kernel $\mathsf{P} = \{\mathsf{P}_s^a, s \in \mathcal{S}, a \in \mathcal{A}\}$ is randomly generated by a normal distribution: $\mathsf{P}_s^a \sim \mathcal{N}(1, \sigma_s^a)$ and then normalized, and the reward function $r(s,a) \sim \mathcal{N}(1, \mu_s^a)$, where $\mu_s^a, \sigma_s^a \sim \textbf{Uniform}[0, 100]$. We set $\delta = 0.4$, $\alpha_n = 0.01$, $f(V) = \frac{\sum_s V(s)}{|\mathcal{S}|}$ and $f(Q) = \frac{\sum_{s,a} Q(s,a)}{|\mathcal{S}||\mathcal{A}|}$. Due to the space limit, we only show the results under the Chi-square and Wasserstein Distance models. The results under the other three uncertainty sets are presented in Appendix G.

For policy evaluation, we evaluate the robust average-reward of the uniform policy $\pi(a|s) = \frac{1}{|\mathcal{A}|}$. We implement our robust RVI TD algorithm under different uncertainty models. We run the algorithm independently for 30 times and plot the average value of $f(V)$ over all 30 trajectories. We also plot the 95th and 5th percentiles of the 30 curves as the upper and lower envelopes of the curves. To compare, we plot the true robust average-reward computed using the model-based robust value iteration method in (Wang et al., 2023). It can be seen from the results in Figure 1 that our robust RVI TD algorithm converges to the true robust average-reward value.

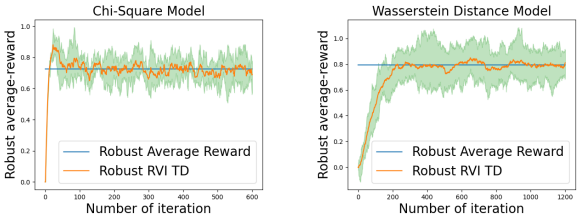


Figure 1: Robust RVI TD Algorithm.

We then consider policy optimization. We run our robust RVI Q-learning independently for 30 times. The curves in Figure 2 show the average value of $f(Q)$ over 30 trajectories, and the upper/lower envelopes are the 95/5 percentiles. We also plot the optimal robust average-reward $g_{\mathcal{P}}^*$ computed by the model-based RVI method in (Wang et al., 2023). Our robust RVI Q-learning converges to the optimal robust average-reward $g_{\mathcal{P}}^*$ under each uncertainty set, which verifies our theoretical results.

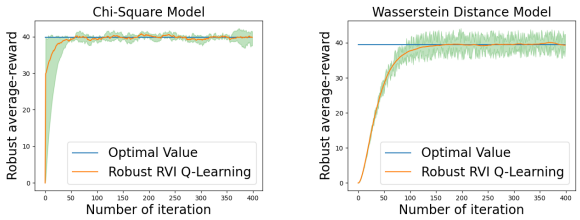


Figure 2: Robust RVI Q-Learning Algorithm.

6.2. Robustness of Robust RVI Q-Learning

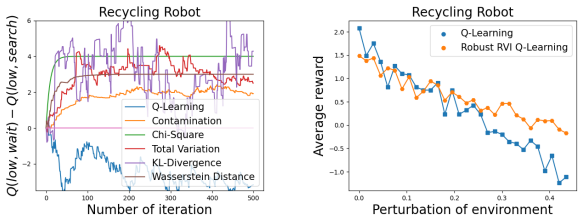
We then demonstrate the robustness of our robust RVI Q-learning by showing that our method achieves a higher average-reward when there is model deviation between training and evaluation.

6.2.1. Recycling Robot

We first consider the recycling robot problem (Example 3.3 (Sutton & Barto, 2018)). A mobile robot running on a rechargeable battery aims to collect empty soda cans. It has 2 battery levels: low and high. The robot can either 1) search for empty cans; 2) remain stationary and wait for someone to bring it a can; 3) go back to its home base to recharge. Under low (high) battery level, the robot finds an empty can with probabilities α (β), and remains at the same battery level. If the robot goes out to search but finds nothing, it will run out of its battery and can only be carried back by human. More details can be found in (Sutton & Barto, 2018).

In this experiment, the uncertainty lies in the probabilities α, β of finding an empty can if the robot chooses the action 'search'. We set $\delta = 0.4$ and implement our algorithms and vanilla Q-learning under the nominal environment ($\alpha = \beta = 0.5$) with stepsize 0.01. To show the

difference among the policies that the algorithms learned, we plot the difference of Q values at low battery level in Figure 3(a). In the low battery level, the robust algorithms find conservative policies which choose to wait instead of search, whereas the vanilla Q-learning finds a policy that chooses to search. To test the robustness of the obtained policies, we evaluate the average-reward of the learned policies in perturbed environments. Specifically, let x denote the amplitude of the perturbation. Then, we estimate the worst performance of the two policies over the testing uncertainty set $(0.5 - x, 0.5 + x)$, and plot them in Figure 3(b). It can be seen that when the perturbation is small, the true worst-case kernels (w.r.t. δ during training) are far from the testing environment, and hence the vanilla Q-learning has a higher reward; however, as the perturbation level becomes larger, the testing environment gets closer to the worst-case kernels, and then our robust algorithms perform better. It can be seen that the performance of Q-learning decreases rapidly while our robust algorithm is stable and outperforms the non-robust Q-learning. This implies that our algorithm is robust to the model uncertainty.



(a) Q(low,wait)-Q(low,search) (b) Perturbed environment

Figure 3: Recycling Robot.

6.2.2. Inventory Control Problem

We now consider the supply chain problem (Giannoccaro & Pontrandolfo, 2002; Kemmer et al., 2018; Liu et al., 2022). At the beginning of each month, the manager of a warehouse inspects the current inventory of a product. Based on the current stock, the manager decides whether or not to order additional stock from a supplier. During this month, if the customer demand is satisfied, the warehouse can make a sale and obtain profits; but if not, the warehouse will obtain a penalty associated with being unable to satisfy customer demand for the product. The warehouse also needs to pay the holding cost for the remaining stock and new items ordered. The goal is to maximize the average profit.

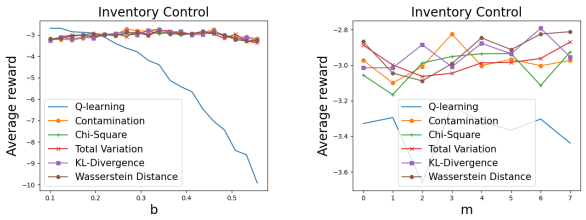
We let s_t denote the inventory at the beginning of the t-th month, D_t be a random demand during this month, and a_t be the number of units ordered by the manager. We assume that D_t follows some distribution and is independent over time. When the agent takes action a_t, the order cost is a_t, and the holding cost is $3 \cdot (s_t + a_t)$. If the demand $D_t \leq s_t + a_t$, then selling the item brings $5 \cdot D_t$ in total; but if the demand $D_t > s_t + a_t$, then there will not be any sale and a penalty of -15 will be received. We set $\mathcal{S} = \{0, 1, ..., 16\}$ and $\mathcal{A} = \{0, ..., 8\}$.

We first set $\delta = 0.4$ and $\alpha_t = 0.01$, and implement our algorithms and vanilla Q-learning under the nominal environment where $D_t \sim$ **Uniform**$(0, 16)$ is generated following the uniform distribution. To verify the robustness, we test the obtained policies under different perturbed environments. More specifically, we perturb the distribution of the demand to $D_t \sim U_{(m,b)}$, where

$$U_{(m,b)}(x) = \begin{cases} \frac{1}{|\mathcal{S}|} + b\frac{|\mathcal{S}|-2}{2|\mathcal{S}|}, & \text{if } x \in \{m, m+1\}, \\ \frac{1-b}{|\mathcal{S}|}, & \text{else.} \end{cases}$$

The results are plotted in Figure 4. We first fix $m = 0$ and plot the performance under different values of b in Figure 4(a), then we fix $b = 0.25$ and plot the performance under different values of m in Figure 4(b).

As the results show, when b is small, i.e., the perturbation of the environment is small, the non-robust Q-learning obtains higher reward than our robust methods; as b becomes larger, the performance of the non-robust method decreases rapidly, while our robust methods are more robust and outperform the non-robust one. When b is fixed, our robust methods outperform the non-robust Q-learning, which also demonstrates the robustness of our methods.



(a) $m = 0$ (b) $b = 0.25$

Figure 4: Inventory Control.

7. Conclusion

In this paper, we developed the first model-free algorithms with provable convergence and optimality guarantee for robust average-reward RL under a broad range of uncertainty set models. We characterized the fundamental structure of solutions to the robust average-reward Bellman equation, which is crucial for the convergence analysis. We designed model-free robust algorithms based on the ideas of relative value iteration for non-robust average-reward MDPs and the robust average-reward Bellman equation. We developed concrete solutions to five popular uncertainty sets, where we generalized the idea of multi-level Monte-Carlo and constructed an unbiased estimate of the non-linear robust average-reward Bellman operator.

8. Acknowledgments

This work is supported by the National Science Foundation under Grants CCF-2106560, CCF-2007783, CCF-2106339, and CCF-1552497. This material is based upon work supported under the AI Research Institutes program by National Science Foundation and the Institute of Education Sciences, U.S. Department of Education through Award # 2229873 - National AI Institute for Exceptional Education. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

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.

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. 2001.

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

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.

Borkar, V. S. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

Borkar, V. S. and Soumyanatha, K. An analog scheme for fixed point computation. i. theory. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 44(4):351–355, 1997.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016.

Chen, L., Jain, R., and Luo, H. Learning infinite-horizon average-reward Markov decision processes with constraints. *arXiv preprint arXiv:2202.00150*, 2022.

Gao, R. and Kleywegt, A. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 2022.

Giannoccaro, I. and Pontrandolfo, P. Inventory management in supply chains: a reinforcement learning approach. *International Journal of Production Economics*, 78(2): 153–161, 2002.

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

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.

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

Kazemi, M., Perez, M., Somenzi, F., Soudjani, S., Trivedi, A., and Velasquez, A. Translating omega-regular specifications to average objectives for model-free reinforcement learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 732–741, 2022.

Kemmer, L., von Kleist, H., de Rochebouët, D., Tziortziotis, N., and Read, J. Reinforcement learning for supply chain optimization. In *European Workshop on Reinforcement Learning*, volume 14, 2018.

Li, T., Wu, F., and Lan, G. Stochastic first-order methods for average-reward Markov decision processes. *arXiv preprint arXiv:2205.05800*, 2022.

Lim, S. H. and Autef, A. Kernel-based reinforcement learning in robust Markov decision processes. In *Proc. International Conference on Machine Learning (ICML)*, pp. 3973–3981. PMLR, 2019.

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, 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, 2022.

Neufeld, A. and Sester, J. Robust Q-learning algorithm for Markov decision processes under Wasserstein uncertainty. *arXiv preprint arXiv:2210.00898*, 2022.

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., 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.

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.

Satia, J. K. and Lave Jr, R. E. Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740, 1973.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, Massachusetts, 2018.

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.

Tessler, C., Efroni, Y., and Mannor, S. Action robust reinforcement learning and applications in continuous control. In *Proc. International Conference on Machine Learning (ICML)*, pp. 6215–6224. PMLR, 2019.

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.

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

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, 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)*, 2023.

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.

Yu, P. and Xu, H. Distributionally robust counterpart in Markov decision processes. *IEEE Transactions on Automatic Control*, 61(9):2538–2543, 2015.

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, 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.

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.

A. Proof of Lemma 3.1

We construct the following example.

Example A.1. Consider an MDP with 3 states (1,2,3) and only one action a, and set a (s,a)-rectangular uncertainty set $\mathcal{P} = \mathcal{P}_1^a \bigotimes \mathcal{P}_2^a \bigotimes \mathcal{P}_3^a$ where $\mathcal{P}_1^a = \{\mathsf{P}_{11}^a, \mathsf{P}_{12}^a\}$, $\mathcal{P}_2^a = \{(0,0,1)^\top\}$ and $\mathcal{P}_3^a = \{(0,1,0)^\top\}$, where $\mathsf{P}_{11}^a = (0,1,0)^\top, \mathsf{P}_{12}^a = (0,0,1)^\top$. Hence, the uncertainty set contains two transition kernels $\mathcal{P} = \{\mathsf{P}_1, \mathsf{P}_2\}$. The reward of each state is set to be $r = (r_1, r_2, r_3)$. The only stationary policy π in this example is $\pi(i) = a, \forall i$.

Note that this robust MDP is a unichain and hence satisfies Assumption 1 with $g_{\mathsf{P}_1}^\pi(1) = g_{\mathsf{P}_1}^\pi(2) = g_{\mathsf{P}_1}^\pi(3), g_{\mathsf{P}_2}^\pi(1) = g_{\mathsf{P}_2}^\pi(2) = g_{\mathsf{P}_2}^\pi(3)$.

Under both transition kernels $\mathsf{P}_1, \mathsf{P}_2$, the average-reward are identical: $g_{\mathsf{P}_1}^\pi = g_{\mathsf{P}_2}^\pi = 0.5r_2 + 0.5r_3$. Hence, both $\mathsf{P}_1, \mathsf{P}_2$ are the worst-case transition kernels.

According to Section A.5 of (Puterman, 1994), the relative value functions w.r.t. $\mathsf{P}_1, \mathsf{P}_2$ can be computed as

$$V_{\mathsf{P}_1}^\pi = \left(r_1 - \frac{1}{4}r_2 - \frac{3}{4}r_3, \frac{1}{4}r_2 - \frac{1}{4}r_3, -\frac{1}{4}r_2 + \frac{1}{4}r_3\right)^\top,$$

$$V_{\mathsf{P}_2}^\pi = \left(r_1 - \frac{3}{4}r_2 - \frac{1}{4}r_3, \frac{1}{4}r_2 - \frac{1}{4}r_3, -\frac{1}{4}r_2 + \frac{1}{4}r_3\right)^\top.$$

When $r_3 > r_2$, only $V_{\mathsf{P}_1}^\pi$ is the solution to (7); and when $r_2 > r_3$, only $V_{\mathsf{P}_2}^\pi$ is the solution to (7). Hence, this proves Lemma 3.1 and implies that not any relative value function w.r.t. a worst-case transition kernel is a solution to (7).

B. Robust RVI TD Method for Policy Evaluation

We define the following notation:

$$r_\pi(s) \triangleq \sum_a \pi(a|s) r(s,a),$$
$$\sigma_{\mathcal{P}_s}(V) \triangleq \sum_a \pi(a|s) \sigma_{\mathcal{P}_s^a}(V),$$
$$\sigma_{\mathcal{P}}(V) \triangleq (\sigma_{\mathcal{P}_{s_1}}(V), \sigma_{\mathcal{P}_{s_2}}(V), ..., \sigma_{\mathcal{P}_{s_{|\mathcal{S}|}}}(V)) \in \mathbb{R}^{|\mathcal{S}|}.$$

B.1. Proof of Theorem 3.1

Theorem B.1 (Restatement of Theorem 3.1)**.** *If (g, V) is a solution to the robust Bellman equation*

$$V(s) = \sum_a \pi(a|s)(r(s,a) - g + \sigma_{\mathcal{P}_s^a}(V)), \forall s, \tag{20}$$

then 1) $g = g_{\mathcal{P}}^\pi$ (Wang et al., 2023); 2) $\mathsf{P}_V \in \Omega_g^\pi$; 3) $V = V_{\mathsf{P}_V}^\pi + ce$ for some $c \in \mathbb{R}$.

Proof. 1). The robust Bellman equation in (20) can be rewritten as

$$g + V(s) - r_\pi(s) = \sigma_{\mathcal{P}_s}(V), \forall s \in \mathcal{S}. \tag{21}$$

From the definition, it follows that

$$\sigma_{\mathcal{P}_s}(V) = \sum_a \pi(a|s) \min_{\mathsf{P}_s^a \in \mathcal{P}_s^a} \mathsf{P}_s^a V. \tag{22}$$

Hence, for any transition kernel $\mathsf{P} = (\mathsf{P}_s^a) \in \bigotimes_{s,a} \mathcal{P}_s^a$,

$$g + V(s) - r_\pi(s) - \sum_a \pi(a|s) \mathsf{P}_s^a V \leq 0, \forall s. \tag{23}$$

It can be further rewritten in matrix form as:

$$ge \leq r_\pi + (\mathsf{P}^\pi - I)V, \tag{24}$$

where P^π is the state transition matrix induced by π and P, i.e., the s-th row of P^π is

$$\sum_a \pi(a|s)\mathsf{P}_s^a. \tag{25}$$

Note that P^π has non-negative components since it is a transition matrix. Multiplying by P^π on both sides, we have that

$$\begin{aligned}
\mathsf{P}^\pi ge = ge &\leq \mathsf{P}^\pi r_\pi + \mathsf{P}^\pi(\mathsf{P}^\pi - I)V, \\
ge &\leq (\mathsf{P}^\pi)^2 r_\pi + (\mathsf{P}^\pi)^2(\mathsf{P}^\pi - I)V, \\
&... \\
ge &\leq (\mathsf{P}^\pi)^{n-1} r_\pi + (\mathsf{P}^\pi)^{n-1}(\mathsf{P}^\pi - I)V.
\end{aligned} \tag{26}$$

Now, by summing up all these inequalities in (24) and (26), we have that

$$nge \leq \sum_{i=0}^{n-1}(\mathsf{P}^\pi)^i r_\pi + ((\mathsf{P}^\pi)^n - I)V, \tag{27}$$

and hence,

$$ge \leq \frac{\sum_{i=0}^{n-1}(\mathsf{P}^\pi)^i r_\pi}{n} + \frac{((\mathsf{P}^\pi)^n - I)V}{n}. \tag{28}$$

Let $n \to \infty$, and we have that

$$\begin{aligned}
ge &\leq \lim_{n\to\infty} \frac{\sum_{i=0}^{n-1}(\mathsf{P}^\pi)^i r_\pi}{n} + \lim_{n\to\infty} \frac{((\mathsf{P}^\pi)^n - I)V}{n} \\
&= g_{\mathsf{P}}^\pi e,
\end{aligned} \tag{29}$$

where the last inequality is from the definition of g_{P}^π and the fact that $\lim_{n\to\infty} \frac{((\mathsf{P}^\pi)^n - I)V}{n} = 0$. Hence, $g \leq g_{\mathsf{P}}^\pi$ for any $\mathsf{P} \in \bigotimes_{s,a} \mathcal{P}_s^a$.

Consider the worst-case transition kernel P_V of V. The robust Bellman equation can be equivalently rewritten as

$$ge = r_\pi - V + \mathsf{P}_V^\pi V. \tag{30}$$

This means that (g, V) is a solution to the non-robust Bellman equation for transition kernel P_V and policy π:

$$xe = r_\pi - Y + \mathsf{P}_V^\pi Y. \tag{31}$$

Thus, by Thm 8.2.6 from (Puterman, 1994),

$$g = g_{\mathsf{P}_V}^\pi, \tag{32}$$

$$V = V_{\mathsf{P}_V}^\pi + ce, \text{ for some } c \in \mathbb{R}. \tag{33}$$

However, note that

$$g_{\mathsf{P}_V}^\pi = g \leq g_{\mathcal{P}}^\pi = \min_{\mathsf{P}\in\mathcal{P}} g_{\mathsf{P}}^\pi \leq g_{\mathsf{P}_V}^\pi, \tag{34}$$

thus,

$$g_{\mathsf{P}_V}^\pi = g = g_{\mathcal{P}}^\pi. \tag{35}$$

2). From (35),

$$g_{\mathsf{P}_V}^\pi = g_{\mathcal{P}}^\pi. \tag{36}$$

It then follows from the definition of Ω_g^π that $\mathsf{P}_V \in \Omega_g^\pi$.

3). Since (g, V) is a solution to the non-robust Bellman equation

$$xe = r_\pi - Y + \mathsf{P}_V^\pi Y, \tag{37}$$

the claim then follows from Theorem 8.2.6 in (Puterman, 1994). □

B.2. Proof of Theorem 3.2

Theorem B.2. *(Restatement of Theorem 3.2) Under Assumptions 1,2,3, $(f(V_n), V_n)$ converges to a (possible sample path dependent) solution to* (7) *a.s..*

We first show the stability of the robust RVI TD algorithm in the following lemma.

Lemma B.1. *Algorithm 1 remains bounded during the update, i.e.,*

$$\sup_n \|V_n\| < \infty, a.s.. \tag{38}$$

Proof. Denote by

$$h(V) \triangleq r_\pi + \sigma_{\mathcal{P}}(V) - f(V)e - V. \tag{39}$$

Then the update of robust RVI TD can be rewritten as

$$V_{n+1} = V_n + \alpha_n(h(V_n) + M_{n+1}), \tag{40}$$

where $M_{n+1} \triangleq \hat{\mathbf{T}}V_n - r_\pi - \sigma_{\mathcal{P}}(V)$ is the noise term.

Further, define the limit function h_∞:

$$h_\infty(V) \triangleq \lim_{c\to\infty} \frac{h(cV)}{c}. \tag{41}$$

Then, from $\sigma_{\mathcal{P}_s^a}(cV) = c\sigma_{\mathcal{P}_s^a}(V)$ and $f(cV) = cf(V)$, it follows that

$$h_\infty(V) = \lim_{c\to\infty} \frac{r_\pi}{c} + \sigma_{\mathcal{P}}(V) - f(V)e - V = \sigma_{\mathcal{P}}(V) - f(V)e - V. \tag{42}$$

According to Section 2.1 and Section 3.2 of (Borkar, 2009), it suffices to verify the following assumptions:

(1). h is Lipschitz;

(2). Stepsize α_n satisfies Assumption 3;

(3). Denoting by $\mathcal{F}_n$ the σ-algebra generated by $V_0, M_1, ..., M_n$, then $\mathbb{E}[M_{n+1}|\mathcal{F}_n] = 0$, $\mathbb{E}[\|M_{n+1}\|^2|\mathcal{F}_n] \le K(1 + \|V_n\|^2)$ for some constant $K > 0$.

(4). h_∞ has the origin as its unique globally asymptotically stable equilibrium.

First, note that

$$\begin{aligned}
\|h(V_1) - h(V_2)\| &= \max_s \left| \sum_a \pi(a|s)(\sigma_{\mathcal{P}_s^a}(V_1) - \sigma_{\mathcal{P}_s^a}(V_2)) - (f(V_1) - f(V_2)) - (V_1(s) - V_2(s)) \right| \\
&\le \max_s \left\{ \left| \sum_a \pi(a|s)(\sigma_{\mathcal{P}_s^a}(V_1) - \sigma_{\mathcal{P}_s^a}(V_2)) \right| + |(f(V_1) - f(V_2))| + |(V_1(s) - V_2(s))| \right\} \\
&\le (2 + L_f)\|V_1 - V_2\|,
\end{aligned} \tag{43}$$

where the last inequality follows from the fact that the support function $\sigma_{\mathcal{P}}(\cdot)$ is 1-Lipschitz and the assumptions on f in Assumption 2. Thus, h is Lipschitz, which verifies (1).

It is straightforward that (3) is satisfied if $\mathbb{E}[\hat{\mathbf{T}}V_n|\mathcal{F}_n] = r_\pi + \sigma_{\mathcal{P}}(V_n)$ and $\text{Var}[\hat{\mathbf{T}}V_n|\mathcal{F}_n] \le K(1 + \|V_n\|^2)$. As discussed in Section 3, we assume the existence of an unbiased estimator $\hat{\mathbf{T}}$ with bounded variance here, and we will construct the estimator in Section 5.

Then, it suffices to verify condition (4), i.e., to show that the ODE

$$\dot{x}(t) = h_\infty(x(t)) \tag{44}$$

has 0 as its unique globally asymptotically stable equilibrium.

Define an operator $\mathbf{T}_0(V)(s) \triangleq \sum_a \pi(a|s)\sigma_{\mathcal{P}_s^a}(V)$. Then, any equilibrium W of (44) satisfies

$$\mathbf{T}_0(W) - f(W)e - W = 0. \tag{45}$$

This equation can be further rewritten as a set of equations:

$$\begin{cases} W = \mathbf{T}_0(W) - ge, \\ g = f(W). \end{cases} \tag{46}$$

The equation in (46) is the robust Bellman equation for a zero-reward robust MDP. Hence, from Theorem 3.1, any solution (g, W) to (46) satisfies

$$g = g_{\mathcal{P}}^{\pi}, W = V_{\mathsf{P}}^{\pi} + ce, \tag{47}$$

where V_{P}^{π} is the relative value function w.r.t. some worst-case transition kernel P (i.e., $g_{\mathsf{P}}^{\pi} = \min_{\mathsf{P}\in\mathcal{P}} g_{\mathsf{P}}^{\pi}$), and some $c \in \mathbb{R}$.

Hence, any equilibrium of (44) satisfies

$$W = V_{\mathsf{P}}^{\pi} + ce, f(W) = g_{\mathcal{P}}^{\pi}. \tag{48}$$

However, note that this robust Bellman equation is for a zero-reward robust MDP, hence for any P,

$$g_{\mathsf{P}}^{\pi} = \lim_{T\to\infty} \mathbb{E}_{\mathsf{P}}\left[\sum_{t=0}^{T-1} \frac{r_t}{T}\right] = 0, \tag{49}$$

$$V_{\mathsf{P}}^{\pi} = \mathbb{E}_{\mathsf{P}}\left[\sum_{t=0}^{\infty} (r_t - g_{\mathsf{P}}^{\pi})\right] = 0, \tag{50}$$

thus $g_{\mathcal{P}}^{\pi} = 0$ and $W = ce$ for some $c \in \mathbb{R}$. From (48), it follows that $f(W) = f(ce) = 0$, for any equilibrium W. From Assumption 2, we have that $f(ce) = cf(e) = c = 0$. This further implies that

$$W = V_{\mathcal{P}}^{\pi} + ce = 0. \tag{51}$$

Thus, the only equilibrium of (44) is 0.

We then show that 0 is globally asymptotically stable. Recall that the zero-reward robust Bellman operator

$$\mathbf{T}_0 V(s) = \sum_a \pi(a|s)(\sigma_{\mathcal{P}_s^a}(V)). \tag{52}$$

We further introduce two operators:

$$\mathbf{T}_0' V \triangleq \mathbf{T}_0 V - f(V)e, \tag{53}$$

$$\tilde{\mathbf{T}}_0 V \triangleq \mathbf{T}_0 V - g_{\mathcal{P}}^{\pi} e. \tag{54}$$

Note that in the zero-reward robust MDP, $g_{\mathcal{P}}^{\pi} = 0$ and $\tilde{\mathbf{T}}_0 = \mathbf{T}_0$, but we introduce this notation for future use.

Consider the ODEs w.r.t. these two operators:

$$\dot{x} = \mathbf{T}_0' x - x, \tag{55}$$

$$\dot{y} = \tilde{\mathbf{T}}_0 y - y. \tag{56}$$

First, it can be easily shown that both $\mathbf{T}_0'$ and $\tilde{\mathbf{T}}_0$ are Lipschitz with constants $1 + L_f$ and 1, respectively. Hence, both two ODEs are well-posed. Also, it can be seen that (55) is the same as the ODE in (44).

Since the second equation (56) is a non-expansion (Lipschitz with parameter no larger than 1), Theorem 3.1 of (Borkar & Soumyanatha, 1997) implies that any solution $y(t)$ to (56) converges to the set of equilibrium points, i.e.,

$$y(t) \to \left\{W : W = \tilde{\mathbf{T}}_0 W\right\}, a.s.. \tag{57}$$

Similar to the discussion for $\mathbf{T}_0$, our Theorem 3.1 implies that the set of equilibrium points of (56) is $\{W = ce : c \in \mathbb{R}\}$. Hence, for any solution $y(t)$ to (56), $y(t) \to ce$ for some constant k that may depend on the initial value of $y(t)$.

Now, consider the solution $x(t)$ to (55). According to Lemma F.1 (note that $\mathbf{T}_0$ here is a special case of $\mathbf{T}$ in Lemma F.1 with $r = 0$), if the solutions $x(t), y(t)$ have the same initial value $x(0) = y(0)$, then

$$x(t) = y(t) + r(t)e, \tag{58}$$

where $r(t)$ is a solution to $\dot{r}(t) = -r(t) + g_{\mathcal{P}}^{\pi} - f(y(t)), r(0) = 0$.

Note that the solution $r(t)$ with $r(0) = 0$ can be written as

$$r(t) = \int_0^t e^{-(t-s)}(g_{\mathcal{P}}^{\pi} - f(y(s)))ds \tag{59}$$

by variation of constants formula (Abounadi et al., 2001). If we denote the limit of $y(t)$ by $y^* = ce$, then $\lim_{t\to\infty} r(t) = g_{\mathcal{P}}^{\pi} - f(y^*)$ (Lemma B.4 in (Wan et al., 2021), Theorem 3.4 in (Abounadi et al., 2001)). Hence, $x(t) = y(t) + r(t)e$ converges to $y^* + (g_{\mathcal{P}}^{\pi} - f(y^*))e$, i.e.,

$$x(t) \to ce - f(ce)e = 0. \tag{60}$$

Hence, any solution $x(t)$ to (55) converges to 0, which is its unique equilibrium. This thus implies that 0 is the unique globally asymptotically stable equilibrium. Together with Theorem 3.7 in (Borkar, 2009), it further implies the boundedness of V_n, which completes the proof. □

We can readily prove Theorem B.2.

Proof. In Lemma B.1, we have shown that

$$\sup_n \|V_n\| < \infty, a.s.. \tag{61}$$

Thus, we have verified that conditions (A1-A3) and (A5) in (Borkar, 2009) are satisfied. Lemma 2.1 in (Borkar, 2009) thus implies that it suffices to study the solution to the ODE $\dot{x}(t) = h(x(t))$.

For the robust Bellman operator $\mathbf{T}V = r_\pi + \sigma_{\mathcal{P}}(V)$, define

$$\mathbf{T}'V \triangleq \mathbf{T}V - f(V)e, \tag{62}$$

$$\tilde{\mathbf{T}}V \triangleq \mathbf{T}V - g_{\mathcal{P}}^{\pi}e. \tag{63}$$

From Lemma F.1, we know that if $x(t), y(t)$ are the solutions to equations

$$\dot{x} = \mathbf{T}'x - x, \tag{64}$$

$$\dot{y} = \tilde{\mathbf{T}}y - y, \tag{65}$$

with the same initial value $x(0) = y(0)$, then

$$x(t) = y(t) + r(t)e, \tag{66}$$

where $r(t)$ satisfies

$$\dot{r}(t) = -r(t) + g_{\mathcal{P}}^{\pi} - f(y(t)), r(0) = 0. \tag{67}$$

Thus, by the variation of constants formula,

$$r(t) = \int_0^t e^{-(t-s)}(g_{\mathcal{P}}^{\pi} - f(y(s)))ds. \tag{68}$$

Note that $\tilde{\mathbf{T}}$ is also non-expansive, hence $y(t)$ converges to some equilibrium of (65) (Theorem 3.1 of (Borkar & Soumyanatha, 1997)). The set of equilibrium points of (65) can be characterized as

$$\left\{W : \tilde{\mathbf{T}}W = W\right\} = \{W : W = TW - g^\pi_{\mathcal{P}} e\} = \left\{W : W(s) = \sum_a \pi(a|s)(r(s,a) - g^\pi_{\mathcal{P}} + \sigma_{\mathcal{P}^a_s}(W)), \forall s \in \mathcal{S}\right\}. \quad (69)$$

From Theorem 3.1, any equilibrium of (65) can be rewritten as

$$W = V^\pi_{\mathsf{P}} + ce, \text{ for some } \mathsf{P} \in \Omega^\pi_g, c \in \mathbb{R}. \quad (70)$$

Thus, $y(t)$ converges to an equilibrium denoted by y^*:

$$y(t) \to y^* \triangleq V^\pi_{\mathsf{P}^*} + c^* e, \text{ for some } \mathsf{P}^* \in \Omega^\pi_g, c^* \in \mathbb{R}. \quad (71)$$

Similar to Lemma B.1, it can be showed that $r(t) \to g^\pi_{\mathcal{P}} - f(y^*)$ (Lemma B.4 in (Wan et al., 2021), Theorem 3.4 in (Abounadi et al., 2001)). This further implies that

$$x(t) \to y^* + (g^\pi_{\mathcal{P}} - f(y^*))e = V^\pi_{\mathsf{P}^*} + (c^* + g^\pi_{\mathcal{P}} - f(y^*))e, \quad (72)$$

and we denote $m^* = c^* + g^\pi_{\mathcal{P}} - f(y^*)$. Moreover, since f is continuous (because it is Lipschitz), we have that

$$\begin{aligned} f(x(t)) &\to f(V^\pi_{\mathsf{P}^*} + (c^* + g^\pi_{\mathcal{P}} - f(y^*))e) \\ &= f(V^\pi_{\mathsf{P}^*}) + c^* + g^\pi_{\mathcal{P}} - f(y^*) \\ &= f(V^\pi_{\mathsf{P}^*}) + c^* + g^\pi_{\mathcal{P}} - f(V^\pi_{\mathsf{P}^*} + c^* e) \\ &= f(V^\pi_{\mathsf{P}^*}) + c^* + g^\pi_{\mathcal{P}} - f(V^\pi_{\mathsf{P}^*}) - c^* \\ &= g^\pi_{\mathcal{P}}. \end{aligned} \quad (73)$$

Hence, we show that

$$x(t) \to V^\pi_{\mathsf{P}^*} + m^* e, \quad (74)$$
$$f(x(t)) \to g^\pi_{\mathcal{P}}. \quad (75)$$

Following Lemma 2.1 from (Borkar, 2009), we conclude that a.s.,

$$V_n \to V^\pi_{\mathsf{P}^*} + m^* e, \quad (76)$$
$$f(V_n) \to g^\pi_{\mathcal{P}}, \quad (77)$$

which completes the proof. □

C. Robust RVI Q-Learning

C.1. Proof of Lemma 4.1

Part of the following theorem is proved in (Wang et al., 2023), but we include the proof for completeness.

Theorem C.1 (Restatement of Lemma 4.1)**.** *If (g, Q) is a solution to the optimal robust Bellman equation*

$$Q(s,a) = r(s,a) - g + \sigma_{\mathcal{P}^a_s}(V_Q), \forall s, a, \quad (78)$$

then 1) $g = g^*_{\mathcal{P}}$ (Wang et al., 2023); 2) the greedy policy w.r.t. Q: $\pi_Q(s) = \arg\max_a Q(s,a)$ is an optimal robust policy (Wang et al., 2023); 3) $V_Q = V^{\pi_Q}_{\mathsf{P}} + ce$ for some $\mathsf{P} \in \Omega^{\pi_Q}_g, c \in \mathbb{R}$.

Proof. Taking the maximum on both sides of (78) w.r.t. a, we have that

$$\max_a Q(s,a) = \max_a \{r(s,a) - g + \sigma_{\mathcal{P}^a_s}(V_Q)\}, \forall s \in \mathcal{S}. \quad (79)$$

This is equivalent to

$$V_Q(s) = \max_a \{r(s,a) - g + \sigma_{\mathcal{P}_s^a}(V_Q)\}, \forall s \in \mathcal{S}. \tag{80}$$

By Theorem 7 in (Wang et al., 2023), we can show that $g = g^*_{\mathcal{P}}$, which proves claim (1).

Recall that $V_Q(s) = \max_a Q(s,a)$. It can be also written as

$$V_Q(s) = \sum_a \pi_Q(a|s) Q(s,a). \tag{81}$$

Here, we slightly abuse the notation of π_Q, and use $\pi_Q(s)$ and $\pi_Q(a|s)$ interchangeably.

Then, the optimal robust Bellman equation in (79) can be rewritten as

$$Q(s, \pi_Q(s)) = r(s, \pi_Q(s)) - g + \sigma_{\mathcal{P}_s^{\pi_Q(s)}} \bigg(\sum_a \pi_Q(a|\cdot) Q(\cdot, a) \bigg). \tag{82}$$

Moreover, if we denote by $W(s) = Q(s,a) = Q(s, \pi_Q(s)) = \max_a Q(s,a)$, then the equation above is equivalent to

$$W(s) = \sum_a \pi_Q(a|s)(r(s,a) - g + \sigma_{\mathcal{P}_s^a}(W)). \tag{83}$$

Therefore, (W, g) is a solution to the robust Bellman equation for the policy π_Q in Theorem 3.1. By Theorem 3.1, we have that

$$g = g^{\pi_Q}_{\mathcal{P}}, \tag{84}$$

$$W = V^{\pi_Q}_{\mathsf{P}} + ce, \tag{85}$$

for some $\mathsf{P} \in \Omega_g^{\pi_Q}$ and $c \in \mathbb{R}$.

Combining this with the claim (1) implies that π_Q is an optimal robust policy. Claims (2) and (3) are thus proved. □

C.2. Proof of Theorem 4.1

Lemma C.1. *If $\hat{\mathbf{H}}$ satisfies that for any $Q, s \in \mathcal{S}, a \in \mathcal{A}$, $\mathbb{E}[\hat{\mathbf{H}}Q(s,a)] = \mathbf{H}Q(s,a)$ and Var$(\hat{\mathbf{H}}Q(s,a)) \le C(1 + \|Q\|^2)$ for some constant C, then under Assumptions 2, 4 and 3, Algorithm 2 remains bounded during the update almost surely, i.e.,*

$$\sup_n \|Q_n\| < \infty, a.s.. \tag{86}$$

Proof. Denote by

$$h(Q) \triangleq r_\pi + \sigma_{\mathcal{P}}(V_Q) - f(Q)e - Q. \tag{87}$$

Then, the update of robust RVI Q-learning can be rewritten as

$$Q_{n+1} = Q_n + \alpha_n (h(Q_n) + M_{n+1}), \tag{88}$$

where $M_{n+1} \triangleq \hat{\mathbf{H}}Q_n - r_\pi - \sigma_{\mathcal{P}}(V_Q)$ is the noise term.

Further, define the limit function h_∞:

$$h_\infty(Q) \triangleq \lim_{c \to \infty} \frac{h(cQ)}{c}. \tag{89}$$

Then, note that $\sigma_{\mathcal{P}_s^a}(V_{cQ}) = \sigma_{\mathcal{P}_s^a}(cV_Q) = c\sigma_{\mathcal{P}_s^a}(V_Q)$ for $c > 0$ and $f(cQ) = cf(Q)$. It then follows that

$$h_\infty(Q) = \lim_{c \to \infty} \frac{r_\pi}{c} + \sigma_{\mathcal{P}}(V_Q) - f(Q)e - Q = \sigma_{\mathcal{P}}(V_Q) - f(Q)e - Q. \tag{90}$$

Similar to the proof of Lemma B.1, it suffices to verify the following conditions:

(1). h is Lipschitz;

(2). Stepsize α_n satisfies Assumption 3;

(3). $\mathbb{E}[M_{n+1}|\mathcal{F}_n] = 0$, and $\mathbb{E}[\|M_{n+1}\|^2|\mathcal{F}_n] \leq K(1 + \|Q_n\|^2)$ for some constant K.

(4). h_∞ has the origin as its unique globally asymptotically stable equilibrium.

Clearly, (2) and (3) can be verified similarly to Lemma B.1. We then verify (1) and (4).

Firstly, it can be shown that

$$\begin{aligned}|h(Q_1)(s,a) - h(Q_2)(s,a)| &= \left|\sigma_{\mathcal{P}_s^a}(V_{Q_1}) - f(Q_1) - Q_1(s,a) - \sigma_{\mathcal{P}_s^a}(V_{Q_2}) - f(Q_2) - Q_2(s,a)\right| \\ &\leq \left|\sigma_{\mathcal{P}_s^a}(V_{Q_1}) - \sigma_{\mathcal{P}_s^a}(V_{Q_2})\right| + |f(Q_1) - f(Q_2)| + |Q_1(s,a) - Q_2(s,a)| \\ &\leq \|V_{Q_1} - V_{Q_2}\| + L_f\|Q_1 - Q_2\| + \|Q_1 - Q_2\| \\ &\leq (2 + L_f)\|Q_1 - Q_2\|, \end{aligned} \tag{91}$$

where the last inequality is from the fact that $\|V_{Q_1} - V_{Q_2}\| \leq \|Q_1 - Q_2\|$. This implies that h is Lipschitz.

To verify (4), note that the stability equation is

$$\dot{X}(t) = h_\infty(X(t)) = \sigma_{\mathcal{P}}(V_X(t)) - f(X(t))e - X(t), \tag{92}$$

where $V_X(t)$ is a $|\mathcal{S}|$-dimensional vector with $V_X(t)(s) = \max_a X(t)(s,a)$.

Any equilibrium Q of the stability equation (92) satisfies that

$$Q(s,a) = \sigma_{\mathcal{P}_s^a}(V_Q) - f(Q)e, \tag{93}$$

which can be viewed as an optimal robust Bellman equation (10) with zero reward. Hence, by Lemma 4.1, it implies that

$$f(Q) = g_{\mathcal{P}}^* = 0, \tag{94}$$

$$V_Q = V_{\mathsf{P}}^{\pi_Q} + ce \text{ for some } \mathsf{P} \in \Omega_g^{\pi_Q}, c \in \mathbb{R}. \tag{95}$$

In the zero-reward MDP, we have that $V_{\mathsf{P}}^\pi = 0$ for any π, P, thus $V_Q(s) = \max_a Q(s,a) = c$ for any $s \in \mathcal{S}$.

Note that from (93), Q satisfies that

$$Q(s,a) = \sigma_{\mathcal{P}_s^a}(V_Q) = \sigma_{\mathcal{P}_s^a}(ce) = c. \tag{96}$$

Since $f(Q) = 0$, it implies that

$$f(Q) = f(ce) = c = 0. \tag{97}$$

Therefore,

$$c = 0, \tag{98}$$

$$Q = 0. \tag{99}$$

Thus, 0 is the unique equilibrium of the stability equation.

We then show that 0 is globally asymptotically stable. Define the zero-reward optimal robust Bellman operator

$$\mathbf{H}_0 Q(s,a) = \sigma_{\mathcal{P}_s^a}(V_Q), \tag{100}$$

and further introduce two operators

$$\mathbf{H}_0' Q(s,a) = \sigma_{\mathcal{P}_s^a}(V_Q) - f(Q), \tag{101}$$

$$\tilde{\mathbf{H}}_0 Q(s,a) = \sigma_{\mathcal{P}_s^a}(V_Q) - g_{\mathcal{P}}^*. \tag{102}$$

It is straightforward to verify that $\tilde{\mathbf{H}}_0$ is non-expansive. Hence by (Borkar & Soumyanatha, 1997), the solution $y(t)$ to equation

$$\dot{y} = \tilde{\mathbf{H}}_0 y - y \tag{103}$$

converges to the set of equilibrium points

$$\{W : W(s,a) = \sigma_{\mathcal{P}_s^a}(V_W) - g_{\mathcal{P}}^*\}, a.s.. \tag{104}$$

This again can be viewed as an optimal robust Bellman equation with zero-reward. Hence, any equilibrium W of (103) satisfies

$$\max_a W(s,a) = c, \forall s. \tag{105}$$

This together with (104) further implies that the equilibrium W of (103) satisfies

$$W(s,a) = \sigma_{\mathcal{P}_s^a}(V_W) = \sigma_{\mathcal{P}_s^a}(ce) = c, \tag{106}$$

and hence $y(t)$ converges to $\{ce : c \in \mathbb{R}\}$. We denote its limit by $y^* = c^* e$.

Lemma F.6 implies the solution $x(t)$ to the ODE $\dot{x} = \mathbf{H}_0'(x) - x$ can be decomposed as $x(t) = y(t) + r(t)e$, where $r(t)$ satisfies $\dot{r}(t) = -r(t) + g_{\mathcal{P}}^* - f(y(t)), r(0) = 0$.

Then, similar to Lemma B.1, Lemma B.4 in (Wan et al., 2021) and Theorem 3.4 in (Abounadi et al., 2001), it can be shown that $r(t) \to g_{\mathcal{P}}^* - f(y(t)) = -c^*$. Hence,

$$x(t) \to 0, \tag{107}$$

which proves the asymptotic stability.

Thus, we conclude that 0 is the unique globally asymptotically stable equilibrium of the stability equation, which implies the boundedness of $\{Q_n\}$ together with results from Section 2.1 and 3.2 from (Borkar, 2009). □

Theorem C.2 (Restatement of Theorem 4.1)**.** *The sequence $\{Q_n\}$ generated by Algorithm 2 converges to a solution Q^* to the optimal robust Bellman equation a.s., and $f(Q_n)$ converges to the optimal robust average-reward $g_{\mathcal{P}}^*$ a.s..*

Proof. According to Lemma 1 from (Borkar, 2009) and Theorem 3.5 from (Abounadi et al., 2001), the sequence $\{Q_n\}$ converge to the same limit as the solution $x(t)$ to the ODE $\dot{x} = \mathbf{H}'x - x$. Hence the proof can be completed by showing convergence of $x(t)$ and $f(x(t))$.

For the optimal robust Bellman operator,

$$\mathbf{H}Q(s,a) = r(s,a) + \sigma_{\mathcal{P}_s^a}(V_Q), \tag{108}$$

define two operators

$$\mathbf{H}'Q \triangleq \mathbf{H}Q - f(Q)e, \tag{109}$$

$$\tilde{\mathbf{H}}Q \triangleq \mathbf{H}Q - g_{\mathcal{P}}^* e. \tag{110}$$

From Lemma F.6, we know that if $x(t), y(t)$ are the solutions to equations

$$\dot{x} = \mathbf{H}'x - x, \tag{111}$$

$$\dot{y} = \tilde{\mathbf{H}}y - y, \tag{112}$$

with the same initial value $x(0) = y(0)$, then

$$x(t) = y(t) + r(t)e, \tag{113}$$

where $r(t)$ satisfies

$$\dot{r}(t) = -r(t) + g_{\mathcal{P}}^* - f(y(t)), r(0) = 0. \tag{114}$$

It can be easily verified that $\tilde{\mathbf{H}}$ is non-expansive. Hence $y(t)$ converges to the set of equilibrium points of of (112) (Theorem 3.1 of (Borkar & Soumyanatha, 1997)), which can be characterized as

$$\left\{W : \tilde{\mathbf{H}}W = W\right\} = \{W : W = \mathbf{H}W - g^*_{\mathcal{P}}e\} = \left\{W : W(s,a) = r(s,a) - g^*_{\mathcal{P}} + \sigma_{\mathcal{P}^a_s}(V_W), \forall s, a\right\}. \tag{115}$$

From Lemma 4.1, any equilibrium W satisfies

$$V_W = V^{\pi_W}_{\mathsf{P}} + ce, \text{ for some } \mathsf{P} \in \Omega^{\pi_W}_g, c \in \mathbb{R}, \tag{116}$$

and π_W is robust optimal. We denote the limit of $y(t)$ by W^*.

Similar to (72) to (76), it can be shown that $r(t) \to g^*_{\mathcal{P}} - f(W^*)$. This further implies that

$$x(t) \to W^* + (g^*_{\mathcal{P}} - f(W^*))e \triangleq W^* + m^*e, \tag{117}$$

where $m^* = g^*_{\mathcal{P}} - f(W^*)$. Note that $W^* + m^*e$ is a solution to the optimal robust Bellman equation, hence $x(t)$ converges to a solution to (10). Moreover, since f is continuous (because it is Lipschitz), we have that

$$\begin{aligned} f(x(t)) &\to f(W^* + m^*e) \\ &= f(W^*) + g^*_{\mathcal{P}} - f(W^*) \\ &= g^*_{\mathcal{P}}. \end{aligned} \tag{118}$$

This completes the proof. □

D. Case Studies for Robust RVI TD

In this section, we provide the proof of the first part of Theorem 5.1, i.e., that $\hat{\mathbf{T}}$ is unbiased and has bounded variance under each uncertainty model.

We first show a lemma, by which the problem can be reduced to investigating whether $\hat{\sigma}_{\mathcal{P}^a_s}$ is unbiased and has bounded variance.

Lemma D.1. *If*

$$\mathbb{E}[\hat{\sigma}_{\mathcal{P}^a_s} V] = \sigma_{\mathcal{P}^a_s}(V), \forall s, a, \tag{119}$$

and moreover, there exists a constant C, such that

$$\mathit{Var}(\hat{\sigma}_{\mathcal{P}^a_s} V) \le C(1 + \|V\|^2), \forall s, a, \tag{120}$$

then

$$\mathbb{E}[\hat{\mathbf{T}}V(s)] = \mathbf{T}V(s), \forall s, \tag{121}$$

and

$$\mathit{Var}(\hat{\mathbf{T}}V(s)) \le |\mathcal{A}|C(1 + \|V\|^2), \forall s. \tag{122}$$

Proof. From the definition, $\hat{\mathbf{T}}V(s) = \sum_a \pi(a|s)(r(s,a) + \hat{\sigma}_{\mathcal{P}^a_s} V)$. Thus,

$$\begin{aligned} \mathbb{E}[\hat{\mathbf{T}}V(s)] &= \mathbb{E}\left[\sum_a \pi(a|s)(r(s,a) + \hat{\sigma}_{\mathcal{P}^a_s} V)\right] \\ &= \sum_a \pi(a|s)(r(s,a) + \mathbb{E}[\hat{\sigma}_{\mathcal{P}^a_s} V]) \\ &= \sum_a \pi(a|s)(r(s,a) + \sigma_{\mathcal{P}^a_s}(V)) = \mathbf{T}V(s), \end{aligned} \tag{123}$$

which shows that $\hat{\mathbf{T}}$ is unbiased. On the other hand, we have that

$$\begin{aligned}
\text{Var}(\hat{\mathbf{T}}V(s)) &= \mathbb{E}\bigg[\bigg(\sum_a \pi(a|s)(r(s,a)+\hat{\sigma}_{\mathcal{P}_s^a}V) - \mathbb{E}\bigg[\sum_a \pi(a|s)(r(s,a)+\hat{\sigma}_{\mathcal{P}_s^a}V)\bigg]\bigg)^2\bigg] \\
&= \mathbb{E}\bigg[\bigg(\sum_a \pi(a|s)(r(s,a)+\hat{\sigma}_{\mathcal{P}_s^a}V) - \sum_a \pi(a|s)(r(s,a)+\mathbb{E}\big[\hat{\sigma}_{\mathcal{P}_s^a}V\big]\bigg)^2\bigg] \\
&= \mathbb{E}\bigg[\bigg(\sum_a \pi(a|s)(\hat{\sigma}_{\mathcal{P}_s^a}V) - \mathbb{E}\big[\hat{\sigma}_{\mathcal{P}_s^a}V\big]\bigg)^2\bigg] \\
&\overset{(a)}{\leq} \mathbb{E}\bigg[\sum_a \pi(a|s)(\hat{\sigma}_{\mathcal{P}_s^a}V - \mathbb{E}\big[\hat{\sigma}_{\mathcal{P}_s^a}V\big])^2\bigg] \\
&= \sum_a \pi(a|s)\mathbb{E}\bigg[(\hat{\sigma}_{\mathcal{P}_s^a}V - \mathbb{E}\big[\hat{\sigma}_{\mathcal{P}_s^a}V^2\big])^2\bigg] \\
&\leq \sum_a \pi(a|s)\text{Var}(\hat{\sigma}_{\mathcal{P}_s^a}V) \\
&\leq |\mathcal{A}|C(1+\|V\|^2),
\end{aligned} \tag{124}$$

where (a) is because $(\mathbb{E}[X])^2 \leq \mathbb{E}[X^2]$, which completes the proof. □

This lemma implies that to prove Theorem 5.1, it suffices to show that $\hat{\sigma}_{\mathcal{P}_s^a}$ is unbiased and has bounded variance.

D.1. Contamination Uncertainty Set

Theorem D.1. $\hat{\mathbf{T}}$ *defined in* (12) *is unbiased and has bounded variance.*

Proof. First, note that

$$\begin{aligned}
V_{n+1}(s) &= V_n(s) + \alpha_n(r(s,a) + ((1-\delta)V_n(s') + \delta \min_x V_n(x) - f(V_n) - V_n(s)) \\
&= V_n(s) + \alpha_n(\mathbf{T}V_n(s) - f(V_n) - V_n(s) + M_n(s)),
\end{aligned} \tag{125}$$

where

$$M_n(s) = r(s,a) + (1-\delta)V_n(s') + \delta \min_x V_n(x) - \mathbf{T}V_n(s), \tag{126}$$

and

$$\mathbf{T}V_n(s) = \sum_a \pi(a|s)\bigg(r(s,a) + (1-\delta)\sum_{s'} \mathsf{P}^a_{s,s'}V_n(s') + \delta \min_x V_n(x)\bigg). \tag{127}$$

Thus,

$$\begin{aligned}
\mathbb{E}[M_n(s)] &= \mathbb{E}\big[r(s,a) + (1-\delta)V_n(s') + \delta \min_x V_n(x)\big] - \sum_a \pi(a|s)\bigg(r(s,a) + (1-\delta)\sum_{s'} \mathsf{P}^a_{s,s'}V_n(s') + \delta \min_x V_n(x)\bigg) \\
&= \sum_a \pi(a|s)\bigg(r(s,a) + (1-\delta)\sum_{s'} \mathsf{P}^a_{s,s'}V_n(s') + \delta \min_x V_n(x)\bigg) \\
&\qquad - \sum_a \pi(a|s)\bigg(r(s,a) + (1-\delta)\sum_{s'} \mathsf{P}^a_{s,s'}V_n(s') + \delta \min_x V_n(x))\bigg) \\
&= 0.
\end{aligned} \tag{128}$$

Hence, the operator is unbiased.

We also have that

$$\begin{aligned}\mathbb{E}[|M_n(s)|^2] &= \mathbb{E}\bigg[\bigg(r(s,a)+(1-\delta)V_n(s')+\delta\min_x V_n(x)-\mathbf{T}V_n(s)\bigg)^2\bigg]\\ &\le 2\mathbb{E}\bigg[\bigg(r(s,a)+(1-\delta)V_n(s')+\delta\min_x V_n(x)\bigg)^2\bigg]+2\mathbb{E}[(\mathbf{T}V_n(s))^2]\\ &\overset{(a)}{\le} 8+8\|V_n\|^2\\ &\le 8(1+\|V_n\|^2),\end{aligned} \tag{129}$$

where (a) is from the fact that $\mathbb{E}\big[\big((1-\delta)V_n(s')+\delta\min_x V_n(x)\big)^2\big]=\mathbb{E}\big[\big|(1-\delta)V_n(s')+\delta\min_x V_n(x)\big|^2\big]\le\mathbb{E}\big[\big(\big|(1-\delta)V_n(s')\big|+\big|\delta\min_x V_n(x)\big|\big)^2\big]\le\mathbb{E}\big[\big((1-\delta)\|V_n\|+\big(\delta\|V_n\|\big)^2\big]\le\|V_n\|^2$.

The proof is completed. □

D.2. Total Variation Uncertainty Set

The estimator under the total variation uncertainty set can be written as

$$\hat{\sigma}_{\mathcal{P}_s^a}(V)=\max_{\mu\ge 0}\big(\hat{\mathsf{P}}^{a,1}_{s,N+1}(V-\mu)-\delta\text{Span}(V-\mu)\big)+\frac{\Delta_N(V)}{p_N}, \tag{130}$$

where

$$\begin{aligned}\Delta_N(V)=&\max_{\mu\ge 0}\big(\hat{\mathsf{P}}^{a}_{s,N+1}(V-\mu)-\delta\text{Span}(V-\mu)\big)\\ &-\frac{1}{2}\max_{\mu\ge 0}\big(\hat{\mathsf{P}}^{a,O}_{s,N+1}(V-\mu)-\delta\text{Span}(V-\mu)\big)\\ &-\frac{1}{2}\max_{\mu\ge 0}\big(\hat{\mathsf{P}}^{a,E}_{s,N+1}(V-\mu)-\delta\text{Span}(V-\mu)\big).\end{aligned} \tag{131}$$

Theorem D.2. *The estimated operator $\hat{\sigma}_{\mathcal{P}_s^a}$ defined in* (130) *is unbiased, i.e.,*

$$\mathbb{E}[\hat{\sigma}_{\mathcal{P}_s^a}V]=\sigma_{\mathcal{P}_s^a}(V). \tag{132}$$

Proof. First, denote the dual function (14) by g:

$$g^V_{s,a}(\mu)=\mathsf{P}^a_s(V-\mu)-\delta\text{Span}(V-\mu), \tag{133}$$

and denote its optimal solution by $\mu^V_{s,a}$:

$$\mu^V_{s,a}=\arg\max_{\mu\ge 0}\big(\mathsf{P}^a_s(V-\mu)-\delta\text{Span}(V-\mu)\big). \tag{134}$$

Then, the support function $\sigma_{\mathcal{P}_s^a}(V)=g^V_{s,a}(\mu^V_{s,a})$. Similarly, define the empirical function

$$g^V_{s,a,N+1}(\mu)=\hat{\mathsf{P}}^a_{s,N+1}(V-\mu)-\delta\text{Span}(V-\mu), \tag{135}$$

$$g^V_{s,a,N+1,O}(\mu)=\hat{\mathsf{P}}^{a,O}_{s,N+1}(V-\mu)-\delta\text{Span}(V-\mu), \tag{136}$$

$$g^V_{s,a,N+1,E}(\mu)=\hat{\mathsf{P}}^{a,E}_{s,N+1}(V-\mu)-\delta\text{Span}(V-\mu), \tag{137}$$

and their optimal solutions are denoted by $\mu^V_{s,a,N+1},\mu^V_{s,a,N+1,O},\mu^V_{s,a,N+1,E}$. We have that

$$\begin{aligned}\mathbb{E}[\hat{\sigma}_{\mathcal{P}_s^a}V]&=\mathbb{E}\bigg[\max_{\mu\ge 0}\big(\hat{\mathsf{P}}^{a,1}_{s,N+1}(V-\mu)-\delta\text{Span}(V-\mu)\big)+\frac{\Delta_N(V)}{p_N}\bigg]\\ &=\mathbb{E}[g^V_{s,a,0}(\mu^V_{s,a,0})]+\mathbb{E}\bigg[\frac{\Delta_N(V)}{p_N}\bigg]\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}[g^V_{s,a,0}(\mu^V_{s,a,0})] + \sum_{n=0}^{\infty} p(N=n)\mathbb{E}\left[\frac{\Delta_N(V)}{p_N}|N=n\right] \\
&= \mathbb{E}[g^V_{s,a,0}(\mu^V_{s,a,0})] + \sum_{n=0}^{\infty} \mathbb{E}[\Delta_n(V)] \\
&= \mathbb{E}[g^V_{s,a,0}(\mu^V_{s,a,0})] + \sum_{n=0}^{\infty} \mathbb{E}\left[g^V_{s,a,n+1}(\mu^V_{s,a,n+1}) - \frac{g^V_{s,a,n+1,O}(\mu^V_{s,a,n+1,O}) + g^V_{s,a,n+1,E}(\mu^V_{s,a,n+1,E})}{2}\right] \\
&= \mathbb{E}[g^V_{s,a,0}(\mu^V_{s,a,0})] + \sum_{n=0}^{\infty} \mathbb{E}\left[g^V_{s,a,n+1}(\mu^V_{s,a,n+1}) - g^V_{s,a,n}(\mu^V_{s,a,n})\right], \qquad (138)
\end{aligned}
$$

where the last inequality is from Lemma F.2. The last equation can be further rewritten as

$$
\begin{aligned}
\mathbb{E}[\hat{\sigma}_{\mathcal{P}^a_s} V] &= \mathbb{E}[g^V_{s,a,0}(\mu^V_{s,a,0})] + \sum_{n=0}^{\infty} \mathbb{E}\left[g^V_{s,a,n+1}(\mu^V_{s,a,n+1}) - g^V_{s,a,n}(\mu^V_{s,a,n})\right] \\
&= \lim_{n\to\infty} \mathbb{E}\left[g^V_{s,a,n}(\mu^V_{s,a,n})\right]. \qquad (139)
\end{aligned}
$$

To show that $\hat{\sigma}_{\mathcal{P}^a_s}$ is unbiased, it suffices to prove that

$$
\lim_{n\to\infty} \mathbb{E}\left[g^V_{s,a,n}(\mu^V_{s,a,n})\right] = g^V_{s,a}(\mu^V_{s,a}). \qquad (140)
$$

For any arbitrary i.i.d. samples $\{X_i\}$ and its corresponding function $g^V_{s,a,n}$, together with Lemma F.3, we have that

$$
\begin{aligned}
&|g^V_{s,a,n}(\mu^V_{s,a,n}) - g^V_{s,a}(\mu^V_{s,a})| \\
&= |\max_{0\le\mu\le V+\|V\|e} g^V_{s,a}(\mu) - \max_{0\le\mu\le V+\|V\|e} g^V_{s,a,n}(\mu)| \\
&\le \max_{0\le\mu\le V+\|V\|e} |g^V_{s,a}(\mu) - g^V_{s,a,n}(\mu)| \\
&= \max_{0\le\mu\le V+\|V\|e} |\mathsf{P}^a_s(V-\mu) - \delta\text{Span}(V-\mu) - \hat{\mathsf{P}}^a_{s,n}(V-\mu) + \delta\text{Span}(V-\mu)| \\
&= \max_{0\le\mu\le V+\|V\|e} |\mathsf{P}^a_s(V-\mu) - \hat{\mathsf{P}}^a_{s,n}(V-\mu)| \\
&\le \max_{0\le\mu\le V+\|V\|e} \|V-\mu\|\|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n}\|_1 \\
&\le 3\|V\|\|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n}\|_1. \qquad (141)
\end{aligned}
$$

Thus, by Hoeffding's inequality and Theorem 3.7 from (Liu et al., 2022),

$$
\mathbb{E}[|g^V_{s,a,n}(\mu^V_{s,a,n}) - g^V_{s,a}(\mu^V_{s,a})|] \le 3\|V\|\frac{|\mathcal{S}|^2\sqrt{\pi}}{2^{\frac{n+1}{2}}}, \qquad (142)
$$

which implies that

$$
\lim_{n\to\infty} \mathbb{E}\left[g^V_{s,a,n}(\mu^V_{s,a,n})\right] = g^V_{s,a}(\mu^V_{s,a}), \qquad (143)
$$

completing the proof. □

Theorem D.3. *The estimated operator $\hat{\sigma}_{\mathcal{P}^a_s}$ defined in* (130) *has bounded variance, i.e., there exists a constant C_0, such that*

$$
\textit{Var}(\hat{\sigma}_{\mathcal{P}^a_s} V) \le (1 + 18(1+2\delta)^2 + 2C_0)\|V\|^2. \qquad (144)
$$

Proof. Similar to Theorem D.2, we have that

$$
\text{Var}(\hat{\sigma}_{\mathcal{P}^a_s} V)
$$

$$
\begin{aligned}
&= \mathbb{E}[(\hat{\sigma}_{\mathcal{P}_s^a} V)^2] - \sigma_{\mathcal{P}_s^a}(V)^2 \\
&\leq \mathbb{E}\left[\left(g^V_{s,a,0}(\mu^V_{s,a,0}) + \frac{\Delta_N(V)}{p_N}\right)^2\right] + (\sigma_{\mathcal{P}_s^a}(V))^2 \\
&\leq 2\mathbb{E}\left[\left(g^V_{s,a,0}(\mu^V_{s,a,0})\right)^2\right] + 2\mathbb{E}\left[\left(\frac{\Delta_N(V)}{p_N}\right)^2\right] + (\sigma_{\mathcal{P}_s^a}(V))^2 \\
&\leq (1 + 18(1+2\delta)^2)\|V\|^2 + 2\sum_{i=0}^{\infty} \frac{\mathbb{E}[(\Delta_i(V))^2]}{p_i},
\end{aligned} \tag{145}
$$

where the last inequality is from Lemma F.3. For any $n \geq 1$, we have that

$$
\begin{aligned}
\mathbb{E}[(\Delta_n(V))^2] &= \mathbb{E}\left[\left(g^V_{s,a,n}(\mu^V_{s,a,n}) - \frac{g^V_{s,a,n,E}(\mu^V_{s,a,n,E}) + g^V_{s,a,n,O}(\mu^V_{s,a,n,O})}{2}\right)^2\right] \\
&= \mathbb{E}\left[\left(g^V_{s,a,n}(\mu^V_{s,a,n}) - g^V_{s,a}(\mu^V_{s,a}) + g^V_{s,a}(\mu^V_{s,a}) - \frac{g^V_{s,a,n,E}(\mu^V_{s,a,n,E}) + g^V_{s,a,n,O}(\mu^V_{s,a,n,O})}{2}\right)^2\right] \\
&\leq 2\mathbb{E}[(g^V_{s,a,n}(\mu^V_{s,a,n}) - g^V_{s,a}(\mu^V_{s,a}))^2] + 2\mathbb{E}\left[\left(g^V_{s,a}(\mu^V_{s,a}) - \frac{g^V_{s,a,n,E}(\mu^V_{s,a,n,E}) + g^V_{s,a,n,O}(\mu^V_{s,a,n,O})}{2}\right)^2\right] \\
&\overset{(a)}{=} 2\mathbb{E}[(g^V_{s,a,n}(\mu^V_{s,a,n}) - g^V_{s,a}(\mu^V_{s,a}))^2] + 2\mathbb{E}[(g^V_{s,a,n-1}(\mu^V_{s,a,n-1}) - g^V_{s,a}(\mu^V_{s,a}))^2] \\
&\leq 18\|V\|^2\mathbb{E}[\|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n}\|_1^2] + 18\|V\|^2\mathbb{E}[\|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n-1}\|_1^2],
\end{aligned} \tag{146}
$$

where (a) is due to Lemma F.2 and the last inequality follows a similar argument to (141). Note that $p_n = \Psi(1-\Psi)^n$ for $\Psi \in (0, 0.5)$, thus similar to Theorem 3.7 of (Liu et al., 2022), we can show that there exists a constant C_0, such that

$$
\sum_{i=0}^{\infty} \frac{\mathbb{E}[(\Delta_i(V))^2]}{p_i} \leq C_0\|V\|^2. \tag{147}
$$

Thus,

$$
\mathrm{Var}(\hat{\sigma}_{\mathcal{P}_s^a} V) \leq (1 + 18(1+2\delta)^2)\|V\|^2 + 2C_0\|V\|^2 = (1 + 18(1+2\delta)^2 + 2C_0)\|V\|^2 . \tag{148}
$$

□

D.3. Chi-Square Uncertainty Set

The estimator under the Chi-square uncertainty set can be written as

$$
\begin{aligned}
\hat{\sigma}_{\mathcal{P}_s^a} V = \max_{\mu \geq 0} &\left(\hat{\mathsf{P}}^{a,1}_{s,N+1}(V-\mu) - \sqrt{\delta \mathrm{Var}_{\hat{\mathsf{P}}^{a,1}_{s,N+1}}(V-\mu)}\right) \\
&+ \frac{\Delta_N(V)}{p_N},
\end{aligned} \tag{149}
$$

where

$$
\begin{aligned}
\Delta_N(V) = &\max_{\mu \geq 0} \left(\mathbb{E}_{\hat{\mathsf{P}}^a_{s,N+1}}[V-\mu] - \sqrt{\delta \mathrm{Var}_{\hat{\mathsf{P}}^a_{s,N+1}}(V-\mu)}\right) \\
&- \frac{1}{2}\max_{\mu \geq 0} \left(\mathbb{E}_{\hat{\mathsf{P}}^{a,O}_{s,N+1}}[V-\mu] - \sqrt{\delta \mathrm{Var}_{\hat{\mathsf{P}}^{a,O}_{s,N+1}}(V-\mu)}\right) \\
&- \frac{1}{2}\max_{\mu \geq 0} \left(\mathbb{E}_{\hat{\mathsf{P}}^{a,E}_{s,N+1}}[V-\mu] - \sqrt{\delta \mathrm{Var}_{\hat{\mathsf{P}}^{a,E}_{s,N+1}}(V-\mu)}\right).
\end{aligned}
$$

Theorem D.4. *The estimated operator defined in* (149) *is unbiased, i.e.,*

$$
\mathbb{E}[\hat{\sigma}_{\mathcal{P}_s^a} V] = \sigma_{\mathcal{P}_s^a}(V). \tag{150}
$$

Proof. Denote the dual function (15) by g:

$$g^V_{s,a}(\mu) = \mathsf{P}^a_s(V-\mu) - \sqrt{\delta \mathrm{Var}_{\mathsf{P}^a_s}(V-\mu)}, \tag{151}$$

and denote its optimal solution by $\mu^V_{s,a}$:

$$\mu^V_{s,a} = \arg\max_{\mu\geq 0} \left(\mathsf{P}^a_s(V-\mu) - \sqrt{\delta \mathrm{Var}_{\mathsf{P}^a_s}(V-\mu)}\right). \tag{152}$$

Then, the support function $\sigma_{\mathcal{P}^a_s}(V) = g^V_{s,a}(\mu^V_{s,a})$. Similarly, define the empirical function

$$g^V_{s,a,N+1}(\mu) = \hat{\mathsf{P}}^a_{s,N+1}(V-\mu) - \sqrt{\delta \mathrm{Var}_{\hat{\mathsf{P}}^a_{s,N+1}}(V-\mu)}, \tag{153}$$

$$g^V_{s,a,N+1,O}(\mu) = \hat{\mathsf{P}}^{a,O}_{s,N+1}(V-\mu) - \sqrt{\delta \mathrm{Var}_{\hat{\mathsf{P}}^{a,O}_{s,N+1}}(V-\mu)}, \tag{154}$$

$$g^V_{s,a,N+1,E}(\mu) = \hat{\mathsf{P}}^{a,E}_{s,N+1}(V-\mu) - \sqrt{\delta \mathrm{Var}_{\hat{\mathsf{P}}^{a,E}_{s,N+1}}(V-\mu)}, \tag{155}$$

and their optimal solutions are denoted by $\mu^V_{s,a,N+1}, \mu^V_{s,a,N+1,O}, \mu^V_{s,a,N+1,E}$. We have that

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}_{\mathcal{P}^a_s} V] &= \mathbb{E}[g^V_{s,a,0}(\mu^V_{s,a,0})] + \mathbb{E}\left[\frac{\Delta_N(V)}{p_N}\right] \\
&= \mathbb{E}[g^V_{s,a,0}(\mu^V_{s,a,0})] + \sum_{n=0}^{\infty} p(N=n)\mathbb{E}\left[\frac{\Delta_N(V)}{p_N}|N=n\right] \\
&= \mathbb{E}[g^V_{s,a,0}(\mu^V_{s,a,0})] + \sum_{n=0}^{\infty} \mathbb{E}[\Delta_n] \\
&= \mathbb{E}[g^V_{s,a,0}(\mu^V_{s,a,0})] + \sum_{n=0}^{\infty} \mathbb{E}\left[g^V_{s,a,n+1}(\mu^V_{s,a,n+1}) - \frac{g^V_{s,a,n+1,O}(\mu^V_{s,a,n+1,O}) + g^V_{s,a,n+1,E}(\mu^V_{s,a,n+1,E})}{2}\right] \\
&= \mathbb{E}[g^V_{s,a,0}(\mu^V_{s,a,0})] + \sum_{n=0}^{\infty} \mathbb{E}\left[g^V_{s,a,n+1}(\mu^V_{s,a,n+1}) - g^V_{s,a,n}(\mu^V_{s,a,n})\right],
\end{aligned} \tag{156}$$

where the last inequality is from Lemma F.2. The last equation can be further rewritten as

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}_{\mathcal{P}^a_s} V] &= \mathbb{E}[g^V_{s,a,0}(\mu^V_{s,a,0})] + \sum_{n=0}^{\infty} \mathbb{E}\left[g^V_{s,a,n+1}(\mu^V_{s,a,n+1}) - g^V_{s,a,n}(\mu^V_{s,a,n})\right] \\
&= \lim_{n\to\infty} \mathbb{E}\left[g^V_{s,a,n}(\mu^V_{s,a,n})\right].
\end{aligned} \tag{157}$$

To show that $\hat{\sigma}_{\mathcal{P}^a_s}$ is unbiased, it suffices to prove that

$$\lim_{n\to\infty} \mathbb{E}\left[g^V_{s,a,n}(\mu^V_{s,a,n})\right] = g^V_{s,a}(\mu^V_{s,a}). \tag{158}$$

For any arbitrary i.i.d. samples $\{X_i\}$ and its corresponding function $g^V_{s,a,n}$, together with Lemma F.4, we have that

$$\begin{aligned}
&|g^V_{s,a,n}(\mu^V_{s,a,n}) - g^V_{s,a}(\mu^V_{s,a})| \\
&= |\max_{0\leq\mu\leq V+\|V\|e} g^V_{s,a}(\mu) - \max_{0\leq\mu\leq V+\|V\|e} g^V_{s,a,n}(\mu)| \\
&\leq \max_{0\leq\mu\leq V+\|V\|e} |g^V_{s,a}(\mu) - g^V_{s,a,n}(\mu)| \\
&= \max_{0\leq\mu\leq V+\|V\|e} \left|\mathsf{P}^a_s(V-\mu) - \hat{\mathsf{P}}^a_{s,n}(V-\mu) - \left(\sqrt{\delta \mathrm{Var}_{\mathsf{P}^a_s}(V-\mu)} - \sqrt{\delta \mathrm{Var}_{\hat{\mathsf{P}}^a_{s,n}}(V-\mu)}\right)\right|
\end{aligned}$$

$$\leq \max_{0\leq\mu\leq V+\|V\|e} |\mathsf{P}_s^a(V-\mu) - \hat{\mathsf{P}}_{s,n}^a(V-\mu)| + \max_{0\leq\mu\leq V+\|V\|e} \left| \left(\sqrt{\delta \text{Var}_{\mathsf{P}_s^a}(V-\mu)} - \sqrt{\delta \text{Var}_{\hat{\mathsf{P}}_{s,n}^a}(V-\mu)} \right) \right|$$
$$\overset{(a)}{\leq} \max_{0\leq\mu\leq V+\|V\|e} \|V-\mu\| \|\mathsf{P}_s^a - \hat{\mathsf{P}}_{s,n}^a\|_1 + \max_{0\leq\mu\leq V+\|V\|e} \sqrt{|\delta \text{Var}_{\mathsf{P}_s^a}(V-\mu) - \delta \text{Var}_{\hat{\mathsf{P}}_{s,n}^a}(V-\mu)|}, \tag{159}$$

where (a) is due to $|\sqrt{x} - \sqrt{y}| \leq \sqrt{|x-y|}$. Note that for any distribution $p, q \in \Delta(|\mathcal{S}|)$ and any random variable X,

$$\begin{aligned} |\text{Var}_p[X] - \text{Var}_q[X]| &= |\mathbb{E}_p[X^2] - \mathbb{E}_p[X]^2 - \mathbb{E}_q[X^2] + \mathbb{E}_q[X]^2| \\ &\leq |\mathbb{E}_p[X^2] - \mathbb{E}_q[X^2]| + |(\mathbb{E}_p[X] + \mathbb{E}_q[X])(\mathbb{E}_p[X] - \mathbb{E}_q[X])| \\ &\leq \sup|X^2| \|p-q\|_1 + 2(\sup|X|)^2 \|p-q\|_1. \end{aligned} \tag{160}$$

Hence,

$$\sqrt{|\delta \text{Var}_{\mathsf{P}_s^a}(V-\mu) - \delta \text{Var}_{\hat{\mathsf{P}}_{s,n}^a}(V-\mu)|} \leq \sqrt{3\delta \|V-\mu\|^2 \|\mathsf{P}_s^a - \hat{\mathsf{P}}_{s,n}^a\|_1}. \tag{161}$$

Thus, by Hoeffding's inequality and Theorem 3.7 from (Liu et al., 2022),

$$\mathbb{E}[|g_{s,a,n}^V(\mu_{s,a,n}^V) - g_{s,a}^V(\mu_{s,a}^V)|] \leq 3\|V\| \left(\frac{|\mathcal{S}|^2\sqrt{\pi}}{2^{\frac{n+1}{2}}} + \sqrt{\frac{3\delta|\mathcal{S}|^2\sqrt{\pi}}{2^{\frac{n+1}{2}}}} \right), \tag{162}$$

which implies that

$$\lim_{n\to\infty} \mathbb{E}\left[g_{s,a,n}^V(\mu_{s,a,n}^V) \right] = g_{s,a}^V(\mu_{s,a}^V), \tag{163}$$

which completes the proof. □

Theorem D.5. *The estimated operator $\hat{\sigma}_{\mathcal{P}_s^a}$ defined in* (149) *has bounded variance, i.e., there exists a constant C_0, such that*

$$\textit{Var}(\hat{\sigma}_{\mathcal{P}_s^a} V) \leq (1 + 18(1+\sqrt{2\delta})^2 + 2C_0)\|V\|^2. \tag{164}$$

Proof. We have that

$$\begin{aligned} &\text{Var}(\hat{\sigma}_{\mathcal{P}_s^a} V) \\ &= \mathbb{E}[(\hat{\sigma}_{\mathcal{P}_s^a} V)^2] - \sigma_{\mathcal{P}_s^a}(V)^2 \\ &\leq \mathbb{E}\left[\left(g_{s,a,0}^V(\mu_{s,a,0}^V) + \frac{\Delta_N(V)}{p_N} \right)^2 \right] + (\sigma_{\mathcal{P}_s^a}(V))^2 \\ &\leq 2\mathbb{E}\left[\left(g_{s,a,0}^V(\mu_{s,a,0}^V) \right)^2 \right] + 2\mathbb{E}\left[\left(\frac{\Delta_N(V)}{p_N} \right)^2 \right] + (\sigma_{\mathcal{P}_s^a}(V))^2 \\ &\leq (1 + 18(1+\sqrt{2\delta})^2)\|V\|^2 + 2\sum_{i=0}^{\infty} \frac{\mathbb{E}[(\Delta_i(V))^2]}{p_i}, \end{aligned} \tag{165}$$

where the last inequality is from Lemma F.4. For any $n \geq 1$, we have that

$$\begin{aligned} \mathbb{E}[(\Delta_n(V))^2] &= \mathbb{E}\left[\left(g_{s,a,n}^V(\mu_{s,a,n}^V) - \frac{g_{s,a,n,E}^V(\mu_{s,a,n,E}^V) + g_{s,a,n,O}^V(\mu_{s,a,n,O}^V)}{2} \right)^2 \right] \\ &= \mathbb{E}\left[\left(g_{s,a,n}^V(\mu_{s,a,n}^V) - g_{s,a}^V(\mu_{s,a}^V) + g_{s,a}^V(\mu_{s,a}^V) - \frac{g_{s,a,n,E}^V(\mu_{s,a,n,E}^V) + g_{s,a,n,O}^V(\mu_{s,a,n,O}^V)}{2} \right)^2 \right] \\ &\leq 2\mathbb{E}[(g_{s,a,n}^V(\mu_{s,a,n}^V) - g_{s,a}^V(\mu_{s,a}^V))^2] + 2\mathbb{E}\left[\left(g_{s,a}^V(\mu_{s,a}^V) - \frac{g_{s,a,n,E}^V(\mu_{s,a,n,E}^V) + g_{s,a,n,O}^V(\mu_{s,a,n,O}^V)}{2} \right)^2 \right] \end{aligned}$$

$$\begin{aligned}
&\overset{(a)}{=} 2\mathbb{E}[(g^V_{s,a,n}(\mu^V_{s,a,n}) - g^V_{s,a}(\mu^V_{s,a}))^2] + 2\mathbb{E}[(g^V_{s,a,n-1}(\mu^V_{s,a,n-1}) - g^V_{s,a}(\mu^V_{s,a}))^2] \\
&\leq 18(1+\sqrt{3\delta})^2\|V\|^2\mathbb{E}[\|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n}\|_1^2 + \|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n}\|_1] \\
&\quad + 18(1+\sqrt{3\delta})^2\|V\|^2\mathbb{E}[\|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n-1}\|_1^2 + \|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n-1}\|_1], \qquad (166)
\end{aligned}$$

where (a) is due to Lemma F.2 and the last inequality follows a similar argument to (159). Note that $p_n = \Psi(1-\Psi)^n$ for $\Psi \in \left(0, 1-\frac{\sqrt{2}}{2}\right)$. Thus, similar to Theorem 3.7 of (Liu et al., 2022), we can show that there exists a constant C_0, such that

$$\sum_{i=0}^{\infty} \frac{\mathbb{E}[(\Delta_i(V))^2]}{p_i} \leq C_0\|V\|^2. \qquad (167)$$

Thus,

$$\text{Var}(\hat{\sigma}_{\mathcal{P}^a_s} V) \leq (1 + 18(1+\sqrt{2\delta})^2)\|V\|^2 + 2C_0\|V\|^2 = (1 + 18(1+\sqrt{2\delta})^2 + 2C_0)\|V\|^2. \qquad (168)$$

□

D.4. KL-Divergence Uncertainty Sets

The estimator under the KL-Divergence uncertainty set can be written as

$$\hat{\sigma}_{\mathcal{P}^a_s} V \triangleq -\min_{\alpha\geq 0}\left(\delta\alpha + \alpha\log\left(e^{\frac{-V(s'_1)}{\alpha}}\right)\right) + \frac{\Delta_N(V)}{p_N},$$

where

$$\begin{aligned}
\Delta_N(V) = &-\min_{\alpha\geq 0}\left(\delta\alpha + \alpha\log\left(\mathbb{E}_{\hat{\mathsf{P}}^a_{s,N+1}}\left[e^{\frac{-V}{\alpha}}\right]\right)\right) \\
&+ \frac{1}{2}\min_{\alpha\geq 0}\left(\delta\alpha + \alpha\log\left(\mathbb{E}_{\hat{\mathsf{P}}^{a,O}_{s,N+1}}\left[e^{\frac{-V}{\alpha}}\right]\right)\right) \\
&+ \frac{1}{2}\min_{\alpha\geq 0}\left(\delta\alpha + \alpha\log\left(\mathbb{E}_{\hat{\mathsf{P}}^{a,E}_{s,N+1}}\left[e^{\frac{-V}{\alpha}}\right]\right)\right). \qquad (169)
\end{aligned}$$

Theorem D.6. *(Liu et al., 2022) The estimated operator $\hat{\sigma}_{\mathcal{P}^a_s}$ is unbiased and has bounded variance, i.e., there exists a constant C_0, such that Var$(\hat{\sigma}_{\mathcal{P}^a_s} V) \leq C_0(1+\|V\|^2)$.*

D.5. Wasserstein Distance Uncertainty Sets

To study the support function w.r.t. this uncertainty model, we first introduce some notation.

Definition D.1. For any function $f : \mathcal{Z} \to \mathbb{R}$, $\lambda \geq 0$ and $x \in \mathcal{Z}$, define the regularization operator

$$\Phi(\lambda, x) \triangleq \inf_{y\in\mathcal{Z}}(\lambda d(x,y)^l + f(y)). \qquad (170)$$

The growth rate κ of function f and any distribution q over $\mathcal{Z}$ is defined as

$$\kappa_q \triangleq \inf\left(\lambda \geq 0 : \sum_{x\in\mathcal{Z}} q(x)\Phi(\lambda, x) > -\infty\right). \qquad (171)$$

Lemma D.2. *(Gao & Kleywegt, 2022) Consider the distributional robust optimization of a function f:*

$$\inf_{W_l(q,p)\leq\delta} \mathbb{E}_{x\sim q}[f(x)], \qquad (172)$$

and define its dual problem as

$$\sup_{\lambda\geq 0}(-\lambda\delta^l + \sum_{x\in\mathcal{Z}} p(x)\inf_{y\in\mathcal{Z}}(f(y) + \lambda d(x,y)^l)). \qquad (173)$$

If $\kappa_p < \infty$, then the strong duality holds, i.e.,

$$\inf_{W_l(q,p)\leq\delta} \mathbb{E}_{x\sim q}[f(x)] = \sup_{\lambda\geq 0}(-\lambda\delta^l + \sum_{x\in\mathcal{Z}} p(x)\inf_{y\in\mathcal{Z}}(f(y) + \lambda d(x,y)^l)). \qquad (174)$$

We first verify that this strong duality holds for our support function.

Lemma D.3. *(Restatement of Equation* (19)*) It holds that*

$$\sigma_{\mathcal{P}_s^a}(V) = \sup_{\lambda \geq 0} \bigg(-\lambda\delta^l + \sum_x \mathsf{P}_{s,x}^a \inf_y (V(y) + \lambda d(x,y)^l) \bigg). \tag{175}$$

Proof. In our case, the regularization operator is

$$\Phi(\lambda, x) = \inf_{s \in \mathcal{S}} (\lambda d(s,x)^l + V(s)). \tag{176}$$

Note that for any $\lambda \geq 0$,

$$\sum_{x \in \mathcal{S}} \mathsf{P}_s^a(x) \Phi(\lambda, x) = \sum_{x \in \mathcal{S}} \mathsf{P}_s^a(x) \inf_{s \in \mathcal{S}} (\lambda d(s,x)^l + V(s)) \geq -\|V\| > -\infty. \tag{177}$$

Hence, the growth rate $\kappa_{\mathsf{P}_s^a} = 0 < \infty$. Thus, the strong duality holds. □

Then, the estimator under the Wasserstein distance uncertainty set can be constructed as

$$\hat{\sigma}_{\mathcal{P}_s^a} V \triangleq \sup_{\lambda \geq 0} \bigg(-\lambda\delta^l + \inf_y (V(y) + \lambda d(s_1', y)^l) \bigg) + \frac{\Delta_N(V)}{p_N} + r(s,a), \tag{178}$$

where

$$\begin{aligned}
&\Delta_N(V) \\
&= \sup_{\lambda \geq 0} \bigg(-\lambda\delta^l + \mathbb{E}_{\hat{\mathsf{P}}_{s,N+1}^a} \bigg[\inf_y (V(y) + \lambda d(S,y)^l) \bigg] \bigg) \\
&\quad - \sup_{\lambda \geq 0} \bigg(-\lambda\delta^l + \mathbb{E}_{\hat{\mathsf{P}}_{s,N+1}^{a,O}} \bigg[\inf_y (V(y) + \lambda d(S,y)^l) \bigg] \bigg) \\
&\quad - \sup_{\lambda \geq 0} \bigg(-\lambda\delta^l + \mathbb{E}_{\hat{\mathsf{P}}_{s,N+1}^{a,E}} \bigg[\inf_y (V(y) + \lambda d(S,y)^l) \bigg] \bigg).
\end{aligned}$$

Theorem D.7. *The estimated operator defined in* (178) *is unbiased, i.e.,*

$$\mathbb{E}[\hat{\sigma}_{\mathcal{P}_s^a} V] = \sigma_{\mathcal{P}_s^a}(V). \tag{179}$$

Proof. Denote the dual function (19) by g:

$$g_{s,a}^V(\lambda) = -\lambda\delta^l + \mathbb{E}_{S \sim \mathsf{P}_s^a}[\inf_{x \in \mathcal{S}} (V(x) + \lambda d(S,x)^l)], \tag{180}$$

and denote its optimal solution by $\lambda_{s,a}^V$:

$$\lambda_{s,a}^V = \arg\max_{\lambda \geq 0} \bigg(-\lambda\delta^l + \mathbb{E}_{S \sim \mathsf{P}_s^a}[\inf_{x \in \mathcal{S}} (V(x) + \lambda d(S,x)^l)] \bigg). \tag{181}$$

Then, the support function $\sigma_{\mathcal{P}_s^a}(V) = g_{s,a}^V(\lambda_{s,a}^V)$. Similarly, define the empirical function $g_{s,a,N+1}^V, g_{s,a,N+1,O}^V, g_{s,a,N+1,E}^V$, and denote their optimal solutions by $\lambda_{s,a,N+1}^V, \lambda_{s,a,N+1,O}^V, \lambda_{s,a,N+1,E}^V$. We have that

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}_{\mathcal{P}_s^a} V] &= \mathbb{E}[g_{s,a,0}^V(\lambda_{s,a,0}^V)] + \mathbb{E}\bigg[\frac{\Delta_N(V)}{p_N}\bigg] \\
&= \mathbb{E}[g_{s,a,0}^V(\lambda_{s,a,0}^V)] + \sum_{n=0}^{\infty} p(N=n) \mathbb{E}\bigg[\frac{\Delta_N(V)}{p_N} | N = n\bigg]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}[g^V_{s,a,0}(\lambda^V_{s,a,0})] + \sum_{n=0}^{\infty} \mathbb{E}[\Delta_n] \\
&= \mathbb{E}[g^V_{s,a,0}(\lambda^V_{s,a,0})] + \sum_{n=0}^{\infty} \mathbb{E}\left[g^V_{s,a,n+1}(\lambda^V_{s,a,n+1}) - \frac{g^V_{s,a,n+1,O}(\lambda^V_{s,a,n+1,O}) + g^V_{s,a,n+1,E}(\lambda^V_{s,a,n+1,E})}{2}\right] \\
&= \mathbb{E}[g^V_{s,a,0}(\lambda^V_{s,a,0})] + \sum_{n=0}^{\infty} \mathbb{E}\left[g^V_{s,a,n+1}(\lambda^V_{s,a,n+1}) - g^V_{s,a,n}(\lambda^V_{s,a,n})\right], \qquad (182)
\end{aligned}
$$

where the last inequality is from Lemma F.2. The last equation can be further rewritten as

$$
\begin{aligned}
\mathbb{E}[\hat{\sigma}_{\mathcal{P}^a_s} V] &= \mathbb{E}[g^V_{s,a,0}(\lambda^V_{s,a,0})] + \sum_{n=0}^{\infty} \mathbb{E}\left[g^V_{s,a,n+1}(\lambda^V_{s,a,n+1}) - g^V_{s,a,n}(\lambda^V_{s,a,n})\right] \\
&= \lim_{n\to\infty} \mathbb{E}\left[g^V_{s,a,n}(\lambda^V_{s,a,n})\right]. \qquad (183)
\end{aligned}
$$

To show that $\hat{\sigma}_{\mathcal{P}^a_s}$ is unbiased, it suffices to prove that

$$
\lim_{n\to\infty} \mathbb{E}\left[g^V_{s,a,n}(\lambda^V_{s,a,n})\right] = g^V_{s,a}(\lambda^V_{s,a}). \qquad (184)
$$

For any arbitrary i.i.d. samples $\{X_i\}$ and its corresponding function $g^V_{s,a,n}$, together with Lemma F.5, we have that

$$
\begin{aligned}
&|g^V_{s,a,n}(\lambda^V_{s,a,n}) - g^V_{s,a}(\lambda^V_{s,a})| \\
&= |\max_{0\le\lambda\le\frac{2\|V\|}{\delta^l}} g^V_{s,a}(\lambda) - \max_{0\le\lambda\le\frac{2\|V\|}{\delta^l}} g^V_{s,a,n}(\lambda)| \\
&\le \max_{0\le\lambda\le\frac{2\|V\|}{\delta^l}} |g^V_{s,a}(\lambda) - g^V_{s,a,n}(\lambda)| \\
&= \max_{0\le\lambda\le\frac{2\|V\|}{\delta^l}} \left|\mathbb{E}_{S\sim\mathsf{P}^a_s}[\inf_{x\in\mathcal{S}}(V(x) + \lambda d(S,x)^l)] - \mathbb{E}_{S\sim\hat{\mathsf{P}}^a_{s,n}}[\inf_{x\in\mathcal{S}}(V(x) + \lambda d(S,x)^l)]\right| \\
&\le \max_{0\le\lambda\le\frac{2\|V\|}{\delta^l}} \|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n}\|_1 \sup_{x,S\in\mathcal{S}} (|V(x) + \lambda d(S,x)^l|) \\
&\le \left(1 + \frac{2D^l}{\delta^l}\right)\|V\|\|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n}\|_1, \qquad (185)
\end{aligned}
$$

where the last inequality is from the bound on λ and D is the diameter of the metric space $(\mathcal{S}, d)$.

By Hoeffding's inequality and similar to the previous proofs, we have that

$$
\mathbb{E}[|g^V_{s,a,n}(\lambda^V_{s,a,n}) - g^V_{s,a}(\lambda^V_{s,a})|] \le \left(1 + \frac{2D^l}{\delta^l}\right)\left(\frac{|\mathcal{S}|^2\sqrt{\pi}}{2^{\frac{n+1}{2}}}\right)\|V\|, \qquad (186)
$$

which implies that

$$
\lim_{n\to\infty} \mathbb{E}\left[g^V_{s,a,n}(\lambda^V_{s,a,n})\right] = g^V_{s,a}(\lambda^V_{s,a}). \qquad (187)
$$

This completes the proof. □

Theorem D.8. *The estimated operator $\hat{\sigma}_{\mathcal{P}^a_s}$ defined in* (178) *has bounded variance, i.e., there exists a constant C_0, such that*

$$
\mathit{Var}(\hat{\sigma}_{\mathcal{P}^a_s} V) \le (3 + 2C_0)\|V\|^2. \qquad (188)
$$

Proof. We first have that

$$
\mathrm{Var}(\hat{\sigma}_{\mathcal{P}^a_s} V)
$$

$$
\begin{aligned}
&= \mathbb{E}[(\hat{\sigma}_{\mathcal{P}_s^a} V)^2] - \sigma_{\mathcal{P}_s^a}(V)^2 \\
&\leq \mathbb{E}\left[\left(g^V_{s,a,0}(\lambda^V_{s,a,0}) + \frac{\Delta_N(V)}{p_N}\right)^2\right] + (\sigma_{\mathcal{P}_s^a}(V))^2 \\
&\leq 2\mathbb{E}\left[\left(g^V_{s,a,0}(\lambda^V_{s,a,0})\right)^2\right] + 2\mathbb{E}\left[\left(\frac{\Delta_N(V)}{p_N}\right)^2\right] + (\sigma_{\mathcal{P}_s^a}(V))^2 \\
&\leq 3\|V\|^2 + 2\sum_{i=0}^{\infty} \frac{\mathbb{E}[(\Delta_i(V))^2]}{p_i},
\end{aligned} \tag{189}
$$

where the last inequality is from Lemma F.4. For any $n \geq 1$, we have that

$$
\begin{aligned}
\mathbb{E}[(\Delta_n(V))^2] &= \mathbb{E}\left[\left(g^V_{s,a,n}(\lambda^V_{s,a,n}) - \frac{g^V_{s,a,n,E}(\lambda^V_{s,a,n,E}) + g^V_{s,a,n,O}(\lambda^V_{s,a,n,O})}{2}\right)^2\right] \\
&= \mathbb{E}\left[\left(g^V_{s,a,n}(\lambda^V_{s,a,n}) - g^V_{s,a}(\lambda^V_{s,a}) + g^V_{s,a}(\lambda^V_{s,a}) - \frac{g^V_{s,a,n,E}(\lambda^V_{s,a,n,E}) + g^V_{s,a,n,O}(\lambda^V_{s,a,n,O})}{2}\right)^2\right] \\
&\leq 2\mathbb{E}[(g^V_{s,a,n}(\lambda^V_{s,a,n}) - g^V_{s,a}(\lambda^V_{s,a}))^2] + 2\mathbb{E}\left[\left(g^V_{s,a}(\lambda^V_{s,a}) - \frac{g^V_{s,a,n,E}(\lambda^V_{s,a,n,E}) + g^V_{s,a,n,O}(\lambda^V_{s,a,n,O})}{2}\right)^2\right] \\
&\overset{(a)}{=} 2\mathbb{E}[(g^V_{s,a,n}(\lambda^V_{s,a,n}) - g^V_{s,a}(\lambda^V_{s,a}))^2] + 2\mathbb{E}[(g^V_{s,a,n-1}(\lambda^V_{s,a,n-1}) - g^V_{s,a}(\lambda^V_{s,a}))^2] \\
&\leq 2\left(1 + \frac{2D^l}{\delta^l}\right)^2 \|V\|^2 \mathbb{E}[\|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n}\|_1^2] + 2\left(1 + \frac{2D^l}{\delta^l}\right)^2 \|V\|^2 \mathbb{E}[\|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n-1}\|_1^2],
\end{aligned} \tag{190}
$$

where (a) is due to Lemma F.2 and the last inequality follows a similar argument to (186). Note that $p_n = \Psi(1-\Psi)^n$ for $\Psi \in \big(0, 0.5\big)$, thus similar to Theorem 3.7 of (Liu et al., 2022), we can show that there exists a constant C_0, such that

$$
\sum_{i=0}^{\infty} \frac{\mathbb{E}[(\Delta_i(V))^2]}{p_i} \leq C_0 \|V\|^2. \tag{191}
$$

Thus, we have that

$$
\mathrm{Var}(\hat{\sigma}_{\mathcal{P}_s^a} V) \leq 3\|V\|^2 + 2C_0\|V\|^2 = (3 + 2C_0)\|V\|^2. \tag{192}
$$

□

E. Case Studies for Robust RVI Q-Learning

In this section, we provide the proof of the second part of Theorem 5.1, i.e., $\hat{\mathbf{H}}$ is bounded and unbiased under each uncertainty model. We note that the proofs in this part can be easily derived by following the ones in Appendix D.

We first prove a lemma necessary to the proofs in this section.

Lemma E.1. *It holds that*

$$
\|V_Q\| \leq \|Q\|. \tag{193}
$$

Proof. From the definition of V_Q, we have that

$$
\|V_Q\| = \max_s |V_Q(s)| = \max_s |\max_a Q(s,a)| \triangleq |Q(s^*, a^*)|. \tag{194}
$$

Clearly, $|Q(s^*, a^*)| \leq \max_{s,a} |Q(s,a)|$, hence

$$
\|V_Q\| \leq \|Q\|. \tag{195}
$$

□

Similar to Appendix D, the propositions of $\hat{\mathbf{H}}$ can be reduced to the ones of $\hat{\sigma}_{\mathcal{P}_s^a}$.

Lemma E.2. *If $\mathbb{E}[\hat{\sigma}_{\mathcal{P}_s^a} V] = \sigma_{\mathcal{P}_s^a}(V)$, and moreover there exists a constant C, such that for any s, a, Var$(\hat{\sigma}_{\mathcal{P}_s^a} V) \leq C(1 + \|V\|^2)$, then $\mathbb{E}[\hat{\mathbf{H}}Q(s,a)] = \mathbf{H}Q(s,a)$, and Var$(\hat{\mathbf{H}}Q(s,a)) \leq C(1 + \|Q\|^2)$.*

Proof. First, we have that

$$\mathbb{E}[\hat{\mathbf{H}}Q(s,a)] = \mathbb{E}[r(s,a) + \hat{\sigma}_{\mathcal{P}_s^a} V_Q(s)] = r(s,a) + \sigma_{\mathcal{P}_s^a}(V_Q) = \mathbf{H}Q(s,a). \tag{196}$$

For boundedness, note that

$$\begin{aligned}
\text{Var}(\hat{\mathbf{H}}Q(s,a)) &= \mathbb{E}\left[(\hat{\mathbf{H}}Q(s,a) - \mathbf{H}Q(s,a))^2\right] \\
&= \mathbb{E}\left[\left(\hat{\sigma}_{\mathcal{P}_s^a} V_Q(s) - \sigma_{\mathcal{P}_s^a}(V_Q)\right)^2\right] \\
&\leq C(1 + \|V_Q\|^2) \\
&\leq C(1 + \|Q\|^2),
\end{aligned} \tag{197}$$

where the last inequality is from Lemma E.1. □

This implies that the problem is reduced to verifying whether $\hat{\sigma}_{\mathcal{P}_s^a}$ is unbiased and has bounded variance, which is identical to the results in Appendix D. We thus omit the proofs for this part.

F. Technical Lemmas

Lemma F.1. *For a robust Bellman operator $\mathbf{T}$, define*

$$\mathbf{T}'V \triangleq \mathbf{T}V - f(V)e, \tag{198}$$

$$\tilde{\mathbf{T}}V \triangleq \mathbf{T}V - g_{\mathcal{P}}^{\pi} e. \tag{199}$$

Assume that $x(t), y(t)$ are the solutions to equations

$$\dot{x} = \mathbf{T}'x - x, \tag{200}$$

$$\dot{y} = \tilde{\mathbf{T}}y - y, \tag{201}$$

with the same initial value $x(0) = y(0) = x_0$. Then,

$$x(t) = y(t) + r(t)e, \tag{202}$$

where $r(t)$ satisfies

$$\dot{r}(t) = -r(t) + g_{\mathcal{P}}^{\pi} - f(y(t)). \tag{203}$$

Proof. Note that $\mathbf{T}'V = \tilde{\mathbf{T}}V + (g_{\mathcal{P}}^{\pi} - f(V))e$, then from the variation of constants formula, we have that

$$x(t) = x_0 e^{-t} + \int_0^t e^{-(t-s)} \tilde{\mathbf{T}}(x(s)) ds + \left(\int_0^t e^{-(t-s)} (g_{\mathcal{P}}^{\pi} - f(x(s))) ds \right) e, \tag{204}$$

$$y(t) = x_0 e^{-t} + \int_0^t e^{-(t-s)} \tilde{\mathbf{T}}(y(s)) ds. \tag{205}$$

Hence, the maximal and minimal components of $x(t) - y(t)$ can be bounded as:

$$\begin{aligned}
\max_i (x_i(t) - y_i(t)) &\leq \int_0^t e^{-(t-s)} \max_i (\tilde{\mathbf{T}}_i(x(s)) - \tilde{\mathbf{T}}_i(y(s))) ds + \int_0^t e^{-(t-s)} (g_{\mathcal{P}}^{\pi} - f(x(s))) ds, \\
\min_i (x_i(t) - y_i(t)) &\geq \int_0^t e^{-(t-s)} \min_i (\tilde{\mathbf{T}}_i(x(s)) - \tilde{\mathbf{T}}_i(y(s))) ds + \int_0^t e^{-(t-s)} (g_{\mathcal{P}}^{\pi} - f(x(s))) ds.
\end{aligned} \tag{206}$$

This hence implies that

$$\begin{aligned}\text{Span}(x(t)-y(t)) &\leq \int_0^t e^{-(t-s)}\text{Span}(\tilde{\mathbf{T}}(x(s))-\tilde{\mathbf{T}}(y(s)))ds \\ &\leq \int_0^t e^{-(t-s)}\text{Span}(x(s)-y(s))ds, \end{aligned} \tag{207}$$

where the last inequality is because $\tilde{\mathbf{T}}$ is non-expansive w.r.t. the span semi-norm (Wang et al., 2023).

Gronwall's inequality implies that $\text{Span}(x(t)-y(t)) \leq 0 \cdot \int_0^t e^{-(t-s)}ds = 0$ for any $t \geq 0$. However, since Span is non-negative, then $\text{Span}(x(t)-y(t)) = 0$. Hence, we have that $x(t) = y(t) + r(t)e$ for some $r(t)$ satisfying $r(0) = 0$.

Also note that the differential of $r(t)$ can be written as

$$\begin{aligned}\dot{r}(t)e &= \dot{x}(t) - \dot{y}(t) \\ &= \tilde{\mathbf{T}}x(t) + (g_{\mathcal{P}}^{\pi} - f(x(t)))e - x(t) - \tilde{\mathbf{T}}y(t) + y(t) \\ &= (-r(t) + g_{\mathcal{P}}^{\pi} - f(y(t)))e, \end{aligned} \tag{208}$$

where the last equation is because

$$\tilde{\mathbf{T}}x(t) = \tilde{\mathbf{T}}(y(t)+r(t)e) = \tilde{\mathbf{T}}(y(t)) + r(t)e, \tag{209}$$

$$f(x(t)) = f(y(t)+r(t)e) = f(y(t)) + r(t). \tag{210}$$

This completes the proof. □

Lemma F.2. *For any function $g : \Delta(|\mathcal{S}|) \to \mathbb{R}$, assume there are 2^{n+1} i.i.d. samples $X_i \sim q$. Denote the empirical distributions from samples $\{X_i : i = 1, ..., 2^{n+1}\}, \{X_{2i-1} : i = 1, ..., 2^n\}, \{X_{2i} : i = 1, ..., 2^n\}$ by $\hat{q}_{n+1}, \hat{q}_{n+1,O}, \hat{q}_{n+1,E}$. Then,*

$$\mathbb{E}[g(\hat{q}_{n+1,O})] = \mathbb{E}[g(\hat{q}_{n+1,E})] = \mathbb{E}[g(\hat{q}_n)]. \tag{211}$$

Proof. Note that

$$\hat{q}_{n+1,O}(s) = \frac{\sum_{i=1}^{2^n} \mathbb{1}_{X_{2i-1}=s}}{2^n}, \tag{212}$$

hence,

$$\begin{aligned}\mathbb{E}[g(\hat{q}_{n+1,O})] &= \sum_{p=(p_1,...,p_{|\mathcal{S}|})\in\Delta(|\mathcal{S}|)} g(p)\mathbb{P}(\hat{q}_{n+1,O} = p) \\ &= \sum_{p=(p_1,...,p_{|\mathcal{S}|})\in\Delta(|\mathcal{S}|)} \mathbb{P}\left(\frac{\sum_{i=1}^{2^n}\mathbb{1}_{X_{2i-1}=s_1}}{2^n} = p_1, ..., \frac{\sum_{i=1}^{2^n}\mathbb{1}_{X_{2i-1}=s_{|\mathcal{S}|}}}{2^n} = p_{|\mathcal{S}|}\middle| q\right) g(p), \end{aligned} \tag{213}$$

where $2^n p_i \in \mathbb{N}$ and $\sum_{i=1}^{|\mathcal{S}|} p_i = 1$. On the other hand,

$$\begin{aligned}\mathbb{E}[g(\hat{q}_n)] &= \sum_{p=(p_1,...,p_{|\mathcal{S}|})\in\Delta(|\mathcal{S}|)} g(p)\mathbb{P}(\hat{q}_n = p) \\ &= \sum_{p=(p_1,...,p_{|\mathcal{S}|})\in\Delta(|\mathcal{S}|)} \mathbb{P}\left(\frac{\sum_{i=1}^{2^n}\mathbb{1}_{X_i=s_1}}{2^n} = p_1, ..., \frac{\sum_{i=1}^{2^n}\mathbb{1}_{X_i=s_{|\mathcal{S}|}}}{2^n} = p_{|\mathcal{S}|}\middle| q\right) g(p). \end{aligned} \tag{214}$$

Note that X_i are i.i.d., hence,

$$\mathbb{P}\left(\frac{\sum_{i=1}^{2^n}\mathbb{1}_{X_i=s_1}}{2^n} = p_1, ..., \frac{\sum_{i=1}^{2^n}\mathbb{1}_{X_i=s_{|\mathcal{S}|}}}{2^n} = p_{|\mathcal{S}|}\middle| q\right) = \mathbb{P}\left(\frac{\sum_{i=1}^{2^n}\mathbb{1}_{X_{2i-1}=s_1}}{2^n} = p_1, ..., \frac{\sum_{i=1}^{2^n}\mathbb{1}_{X_{2i-1}=s_{|\mathcal{S}|}}}{2^n} = p_{|\mathcal{S}|}\middle| q\right).$$

Thus,

$$\mathbb{E}[g(\hat{q}_{n+1,O})] = \mathbb{E}[g(\hat{q}_n)]. \tag{215}$$

Similarly, $\mathbb{E}[g(\hat{q}_{n+1,E})] = \mathbb{E}[g(\hat{q}_n)]$ and hence it completes the proof. □

Lemma F.3. *Under the total variation uncertainty model, the optimal solution and optimal value for $g^V_{s,a}, g^V_{s,a,N+1}, g^V_{s,a,N+1,E}, g^V_{s,a,N+1,O}$ are bounded. Specifically,*

$$\mu^V_{s,a}, \mu^V_{s,a,N+1}, \mu^V_{s,a,N+1,E}, \mu^V_{s,a,N+1,O} \le V + \|V\|e, \tag{216}$$

$$\|\mu^V_{s,a}\|, \|\mu^V_{s,a,N+1}\|, \|\mu^V_{s,a,N+1,E}\|, \|\mu^V_{s,a,N+1,O}\| \le 2\|V\|, \tag{217}$$

$$|g^V_{s,a}(\mu^V_{s,a})|, |g^V_{s,a,N+1}(\mu^V_{s,a,N+1})|, |g^V_{s,a,N+1,O}(\mu^V_{s,a,N+1,O})|, |g^V_{s,a,N+1,E}(\mu^V_{s,a,N+1,E})| \le 3(1+2\delta)\|V\|. \tag{218}$$

Proof. First we show the bounds on the optimal solutions. If we denote the minimal entry of V by w: $w = \min_s V(s)$, then $W \triangleq V - we \ge 0$. Note that,

$$\begin{aligned}\mu^W_{s,a} &= \arg\max_{\mu\ge 0} \big(\mathsf{P}^a_s(W-\mu) - \delta\text{Span}(W-\mu)\big)\\ &= \arg\max_{\mu\ge 0} \big(-w + \mathsf{P}^a_s(V-\mu) - \delta\text{Span}(V-\mu)\big),\end{aligned} \tag{219}$$

which is because $\text{Span}(V+ke) = \text{Span}(V)$ and $\mathsf{P}^a_s(V+ke) = k + \mathsf{P}^a_s V$. Hence, $\mu^W_{s,a} = \mu^V_{s,a}$. Moreover note that $W \ge 0$, hence $\mu^W_{s,a}$ is bounded: $\mu^W_{s,a} \le W$, this further implies that

$$\|\mu^V_{s,a}\| = \|\mu^W_{s,a}\| \le \|W\| \le 2\|V\|. \tag{220}$$

The bounds on $\mu^V_{s,a,N+1}, \mu^V_{s,a,N+1,O}, \mu^V_{s,a,N+1,E}$ can be similarly derived.

We then consider the optimal value. Note that,

$$\begin{aligned}g^V_{s,a}(\mu^V_{s,a}) &= \mathsf{P}^a_s(V-\mu^V_{s,a}) - \delta\text{Span}(V-\mu^V_{s,a})\\ &\le \|V\| + \|\mu^V_{s,a}\| + \delta|\max_i(V(i)-\mu^V_{s,a}(i))| + \delta|\min_i(V(i)-\mu^V_{s,a}(i))|\\ &\le 3\|V\| + 2\delta(\|V\| + \|\mu^V_{s,a}\|)\\ &\le 3(1+2\delta)\|V\|.\end{aligned} \tag{221}$$

On the other hand,

$$g^V_{s,a}(\mu^V_{s,a}) \ge g^V_{s,a}(0) = \mathsf{P}^a_s V - \delta\text{Span}(V) = \mathsf{P}^a_s V - \delta\max_i V(i) + \delta\min_i V(i). \tag{222}$$

Denote the maximal and minimal entries of V by $V(M)$ and $V(m)$, then we have that

$$\begin{aligned}&\mathsf{P}^a_s V - \delta\max_i V(i) + \delta\min_i V(i)\\ &= \sum_x \mathsf{P}^a_{s,x} V(x) - \delta V(M) + \delta V(m)\\ &\ge -\|V\| - 2\delta\|V\|,\end{aligned} \tag{223}$$

where the last inequality is from $\|V\| \ge V(i) \ge -\|V\|$ for any entry i. Thus, combining (222) and (223) implies that

$$-(1+2\delta)\|V\| \le g^V_{s,a}(\mu^V_{s,a}) \le 3(1+2\delta)\|V\|. \tag{224}$$

Similarly, the bounds on $g^V_{s,a,N+1}(\mu^V_{s,a,N+1}), g^V_{s,a,N+1,O}(\mu^V_{s,a,N+1,O}), g^V_{s,a,N+1,E}(\mu^V_{s,a,N+1,E})$ can be derived. □

Lemma F.4. *Under the chi-square uncertainty model, the optimal solution and optimal value for $g^V_{s,a}, g^V_{s,a,N+1}, g^V_{s,a,N+1,E}, g^V_{s,a,N+1,O}$ are bounded. Specifically,*

$$\mu^V_{s,a}, \mu^V_{s,a,N+1}, \mu^V_{s,a,N+1,E}, \mu^V_{s,a,N+1,O} \le V + \|V\|e, \tag{225}$$

$$\|\mu^V_{s,a}\|, \|\mu^V_{s,a,N+1}\|, \|\mu^V_{s,a,N+1,E}\|, \|\mu^V_{s,a,N+1,O}\| \le 2\|V\|, \tag{226}$$

$$|g^V_{s,a}(\mu^V_{s,a})|, |g^V_{s,a,N+1}(\mu^V_{s,a,N+1})|, |g^V_{s,a,N+1,O}(\mu^V_{s,a,N+1,O})|, |g^V_{s,a,N+1,E}(\mu^V_{s,a,N+1,E})| \le 3(1+\sqrt{2\delta})\|V\|. \tag{227}$$

Proof. First, we show the bounds on the optimal solutions. If we denote the minimal entry of V by w: $w = \min_s V(s)$, then $W \triangleq V - we \geq 0$. Note that,

$$\begin{aligned}\mu^W_{s,a} &= \arg\max_{W\geq\mu\geq 0} \big(\mathsf{P}^a_s(W-\mu) - \sqrt{\delta \mathrm{Var}_{\mathsf{P}^a_s}(W-\mu)}\big)\\ &= \arg\max_{W\geq\mu\geq 0} \big(-w + \mathsf{P}^a_s(V-\mu) - \sqrt{\delta \mathrm{Var}_{\mathsf{P}^a_s}(V-\mu)}\big),\end{aligned} \tag{228}$$

which is because $\mathrm{Var}_{\mathsf{P}^a_s}(V-\mu-we) = \mathrm{Var}_{\mathsf{P}^a_s}(V-\mu) + \mathrm{Var}_{\mathsf{P}^a_s}(we) - 2\mathrm{Cov}_{\mathsf{P}^a_s}(V-\mu, we) = \mathrm{Var}_{\mathsf{P}^a_s}(V-\mu)$. Hence $\mu^W_{s,a} = \mu^V_{s,a}$. Moreover note that $W \geq 0$, hence $\mu^W_{s,a}$ is bounded: $\mu^W_{s,a} \leq W$, this further implies that

$$\|\mu^V_{s,a}\| = \|\mu^W_{s,a}\| \leq \|W\| \leq 2\|V\|. \tag{229}$$

The bounds on $\mu^V_{s,a,N+1}, \mu^V_{s,a,N+1,O}, \mu^V_{s,a,N+1,E}$ can be similarly derived.

We then consider the optimal value. Note that,

$$\begin{aligned}|g^V_{s,a}(\mu^V_{s,a})| &= |\mathsf{P}^a_s(V-\mu^V_{s,a}) - \sqrt{\delta \mathrm{Var}_{\mathsf{P}^a_s}(V-\mu^V_{s,a})}|\\ &\leq \|V\| + \|\mu^V_{s,a}\| + \sqrt{2\delta\|V-\mu^V_{s,a}\|^2}\\ &\leq 3\|V\| + \sqrt{2\delta}(\|V\| + \|\mu^V_{s,a}\|)\\ &\leq 3(1+\sqrt{2\delta})\|V\|.\end{aligned} \tag{230}$$

Similarly, the bounds on $g^V_{s,a,N+1}(\mu^V_{s,a,N+1}), g^V_{s,a,N+1,O}(\mu^V_{s,a,N+1,O}), g^V_{s,a,N+1,E}(\mu^V_{s,a,N+1,E})$ can be derived. □

Lemma F.5. *Under the Wasserstein distance uncertainty model, the optimal solution and optimal value for $g^V_{s,a}, g^V_{s,a,N+1}, g^V_{s,a,N+1,E}, g^V_{s,a,N+1,O}$ are bounded. Specifically,*

$$\lambda^V_{s,a}, \lambda^V_{s,a,n}, \lambda^V_{s,a,n,O}, \lambda^V_{s,a,n,E} \leq \frac{2\|V\|}{\delta^l}, \tag{231}$$

$$|g^V_{s,a}(\lambda^V_{s,a})|, |g^V_{s,a,n}(\lambda^V_{s,a,n})|, |g^V_{s,a,n,O}(\lambda^V_{s,a,n,O})|, |g^V_{s,a,n,E}(\lambda^V_{s,a,n,E})| \leq \|V\|. \tag{232}$$

Proof. First, we show the bounds on the optimal solutions. Denote the optimal solution to $\max_{\lambda\geq 0} g^V_{s,a}(\lambda)$ by $\lambda^V_{s,a}$. Moreover, for each state $y \in \mathcal{S}$ and any $\lambda \geq 0$, denote $s^y_\lambda \triangleq \arg\min_{x\in\mathcal{S}}\{\lambda d(x,y)^l + V(x)\}$. Hence,

$$g^V_{s,a}(\lambda) = -\lambda\delta^l + \mathbb{E}_{S\sim\mathsf{P}^a_s}[\lambda d(S, s^S_\lambda)^l + V(s^S_\lambda)]. \tag{233}$$

Moreover, note that $g^V_{s,a}(\lambda^V_{s,a}) = \max_{\lambda\geq 0} g^V_{s,a}(\lambda)$, hence,

$$-\lambda^V_{s,a}\delta^l + \mathbb{E}_{S\sim\mathsf{P}^a_s}[\lambda^V_{s,a} d(S, s^S_{\lambda^V_{s,a}})^l + V(s^S_{\lambda^V_{s,a}})] \geq g^V_{s,a}(0) = \mathbb{E}_{S\sim\mathsf{P}^a_s}[V(s^S_0)] = \min_x V(x), \tag{234}$$

where the last equation is due to the fact that $s^S_0 = \arg\min_{x\in\mathcal{S}}\{V(x)\} = \min_x V(x)$. Now consider the inner problem $\mathbb{E}_{S\sim\mathsf{P}^a_s}[\lambda^V_{s,a} d(S, s^S_{\lambda^V_{s,a}})^l + V(s^S_{\lambda^V_{s,a}})]$. Note that,

$$\begin{aligned}&\mathbb{E}_{S\sim\mathsf{P}^a_s}[\lambda^V_{s,a} d(S, s^S_{\lambda^V_{s,a}})^l + V(s^S_{\lambda^V_{s,a}})]\\ &= \sum_x \mathsf{P}^a_{s,x}\big(\lambda^V_{s,a} d(x, s^x_{\lambda^V_{s,a}})^l + V(s^x_{\lambda^V_{s,a}})\big)\\ &\overset{(a)}{\leq} \sum_x \mathsf{P}^a_{s,x}\big(\lambda^V_{s,a} d(x,x)^l + V(x)\big)\\ &= \mathbb{E}_{\mathsf{P}^a_s}[V(S)],\end{aligned} \tag{235}$$

where (a) is because $s^x_{\lambda^V_{s,a}} = \arg\min_{y\in\mathcal{S}}\{\lambda^V_{s,a} d(x,y)^l + V(y)\}$ and hence $\lambda^V_{s,a} d(x, s^x_{\lambda^V_{s,a}})^l + V(s^x_{\lambda^V_{s,a}}) \leq \lambda^V_{s,a} d(x,x)^l + V(x)$.

Combine (234) and (235), then we further have that

$$\min_x V(x) \leq -\lambda_{s,a}^V \delta^l + \mathbb{E}_{S\sim \mathsf{P}_s^a}[\lambda_{s,a}^V d(S, s^S_{\lambda^V_{s,a}})^l + V(s^S_{\lambda^V_{s,a}})] \leq -\lambda_{s,a}^V \delta^l + \mathbb{E}_{\mathsf{P}_s^a}[V(S)]. \tag{236}$$

This implies that

$$\lambda_{s,a}^V \leq \frac{\mathbb{E}_{\mathsf{P}_s^a}[V(S)] - \min_x V(x)}{\delta^l} \leq \frac{2\|V\|}{\delta^l}, \tag{237}$$

and hence $\lambda_{s,a}^V$ is bounded.

On the other hand, note that $g_{s,a}^V(\lambda_{s,a}^V) = \sigma_{\mathcal{P}_s^a}[V(S)]$, hence,

$$\left|g_{s,a}^V(\lambda_{s,a}^V)\right| \leq \|V\|. \tag{238}$$

Same bound can be similarly derived for $\lambda^V_{s,a,n}, \lambda^V_{s,a,n,O}, \lambda^V_{s,a,n,E}, g^V_{s,a,n}(\lambda^V_{s,a,n}), g^V_{s,a,n,O}(\lambda^V_{s,a,n,O}), g^V_{s,a,n,E}(\lambda^V_{s,a,n,E})$. □

Lemma F.6. *For an optimal robust Bellman operator:* $\mathbf{H}Q(s,a) = r(s,a) + \sigma_{\mathcal{P}_s^a}(V_Q)$, *define*

$$\mathbf{H}'Q \triangleq \mathbf{H}Q - f(Q)e, \tag{239}$$

$$\tilde{\mathbf{H}}Q \triangleq \mathbf{H}Q - g^*_{\mathcal{P}}e. \tag{240}$$

Assume that $x(t), y(t)$ *are the solutions to equations*

$$\dot{x} = \mathbf{H}'x - x, \tag{241}$$

$$\dot{y} = \tilde{\mathbf{H}}y - y, \tag{242}$$

with the same initial value $x(0) = y(0)$. *Then* $x(t) = y(t) + r(t)e$, *where* $r(t)$ *satisfies* $\dot{r}(t) = -r(t) + g^*_{\mathcal{P}} - f(y(t))$.

Proof. The proof follows exactly that of Lemma F.1 if we show that $\tilde{\mathbf{H}}$ is non-expansion w.r.t. the span semi-norm.

Following Theorem 17 of (Wang et al., 2023), it can be shown that

$$\text{Span}(\tilde{\mathbf{H}}(Q_1) - \tilde{\mathbf{H}}(Q_2)) \leq \text{Span}(V_{Q_1} - V_{Q_2}). \tag{243}$$

Let

$$s = \arg\max_i \{\max_a Q_1(i,a) - \max_a Q_2(i,a)\}, \tag{244}$$

$$t = \arg\min_i \{\max_a Q_1(i,a) - \max_a Q_2(i,a)\}. \tag{245}$$

Then,

$$\begin{aligned}
\text{Span}(V_{Q_1} - V_{Q_2}) &= (\max_a Q_1(s,a) - \max_a Q_2(s,a)) - (\max_a Q_1(t,a) - \max_a Q_2(t,a)) \\
&\leq Q_1(s,a_s) - Q_2(s,a_s) - (Q_1(t,a_t) - Q_2(t,a_t)) \\
&\leq \max_{x,b}(Q_1(x,b) - Q_2(x,b)) - \min_{x,b}(Q_1(x,b) - Q_2(x,b)) \\
&= \text{Span}(Q_1 - Q_2).
\end{aligned} \tag{246}$$

where $a_s = \arg\max_a Q_1(S,a)$ and $a_t = \arg\max_a Q_2(t,a)$. This completes the proof. □

G. Additional Experiments

In this section, we first show the additional experiments on the Garnet problem in Section 6.1. Then, we further verify our theoretical results using some additional experiments.

G.1. Garnet Problem

We first verify the convergence of our robust RVI TD and robust RVI Q-learning under the Garnet problem with the same setting as in Section 6.1. Our results show that both our algorithms converge to the (optimal) robust average-reward under the other three uncertainty sets.

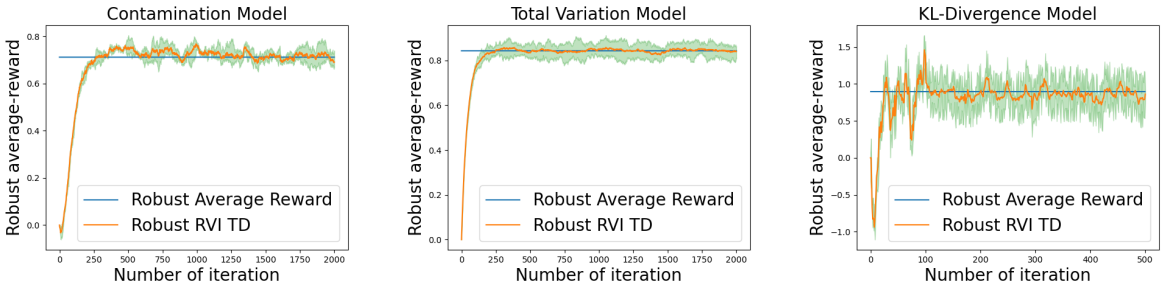


Figure 5: Robust RVI TD Algorithm under Garnet Problem.

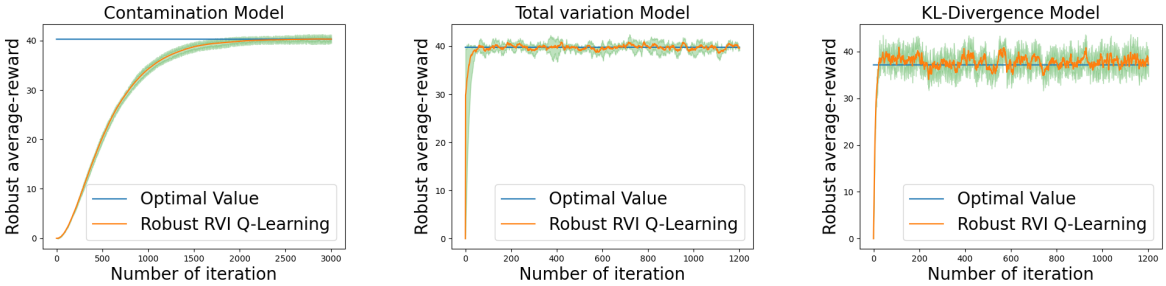


Figure 6: Robust RVI Q-Learning Algorithm under Garnet Problem.

G.2. Frozen-Lake Problem

We first verify our robust RVI TD algorithm and robust RVI Q-learning under the Frozen-Lake environment of OpenAI (Brockman et al., 2016). We set the uncertainty radius $\delta = 0.4$, $\alpha_n = 0.01$ and plot the (optimal) robust average-reward computed using model-based methods in (Wang et al., 2023) as the baseline. We evaluate the uniform policy for the policy evaluation problem, plot the average value of $f(V_t)$ of 30 trajectories and plot the 95/5 percentile as the upper/lower envelope. For the optimal control problem, we plot the average value of $f(Q_t)$ of 30 trajectories and plot the 95/5 percentile as the upper/lower envelope. The results show that both algorithms converge to the (optimal) robust average-reward.

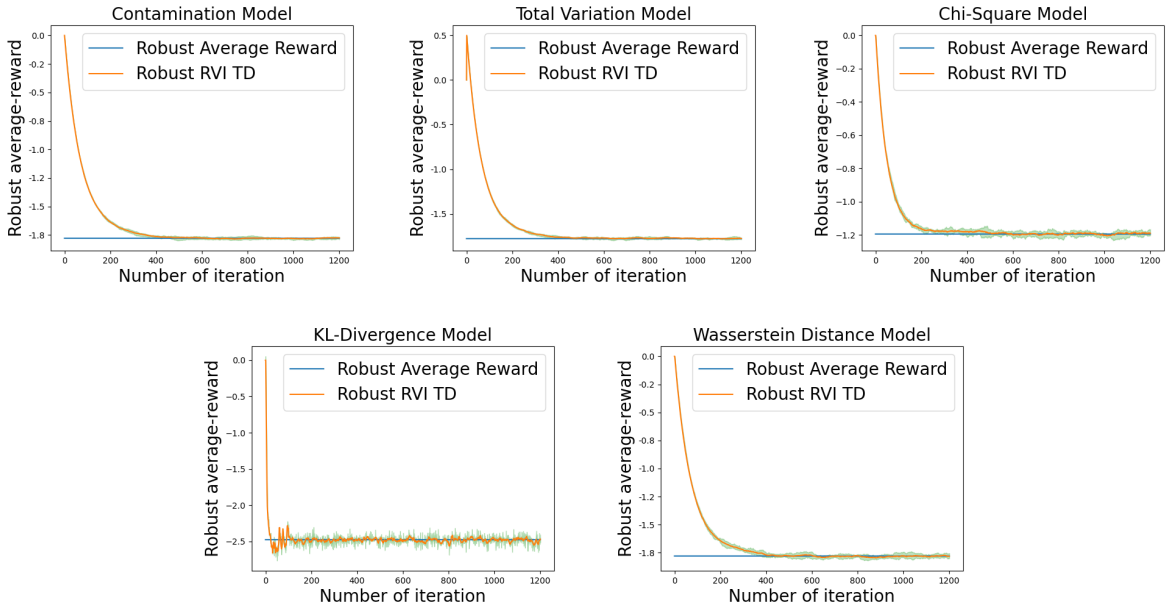


Figure 7: Robust RVI TD Algorithm under Frozen-Lake environment.

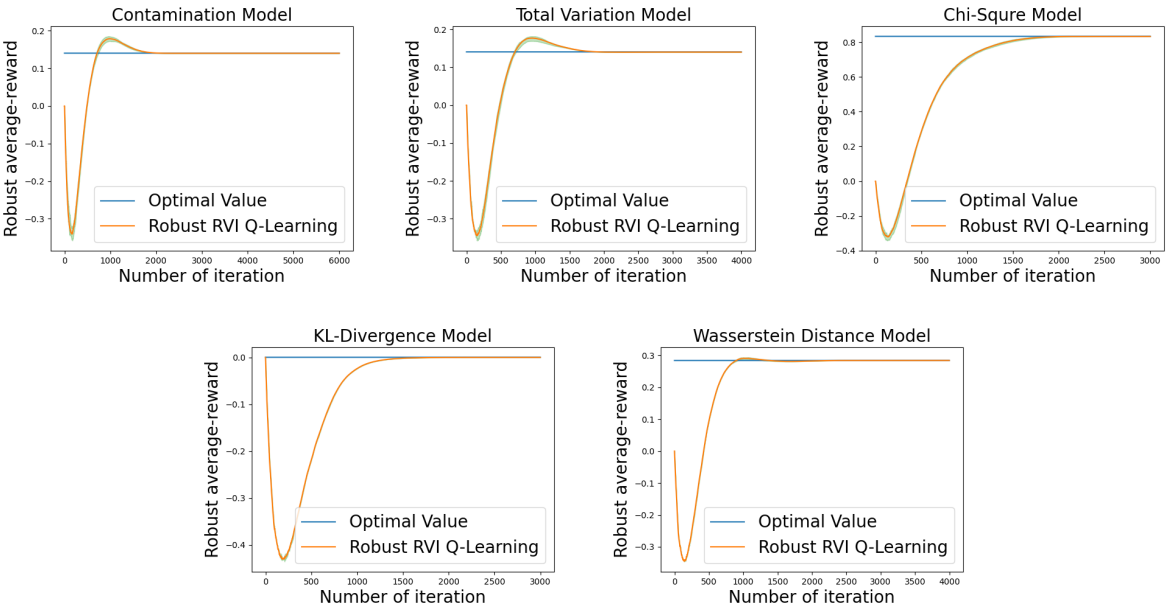


Figure 8: Robust RVI Q-learning Algorithm under Frozen-Lake environment.

G.3. Robustness of Robust RVI Q-Learning

We further use the simple, yet widely-used problem, referred to as the one-loop task problem (Panaganti & Kalathil, 2021), to verify the robustness of our robust RVI Q-learning. This environment is widely used to demonstrate that robust methods can learn different optimal polices from the non-robust methods, which are more robust to model uncertainty. The one-loop MDP contains 2 states s_1, s_2, and 2 actions a_l, a_r indicating going left or right.

The nominal environment is shown in the left of Figure 9, where at state s_1, going left and right will result in a transition to s_1 or s_2; and at s_2, going left and right will result in a transition to s_1 or s_2.

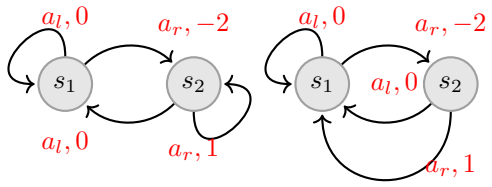
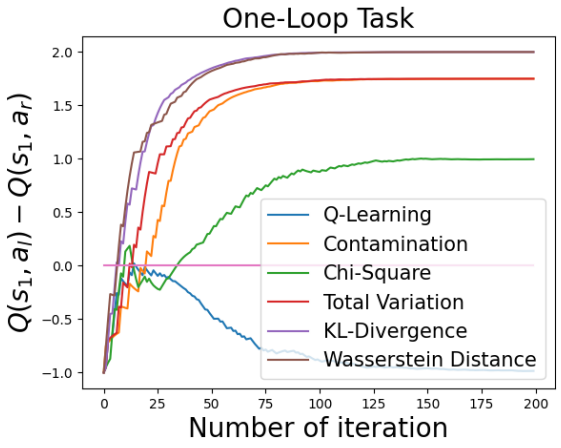


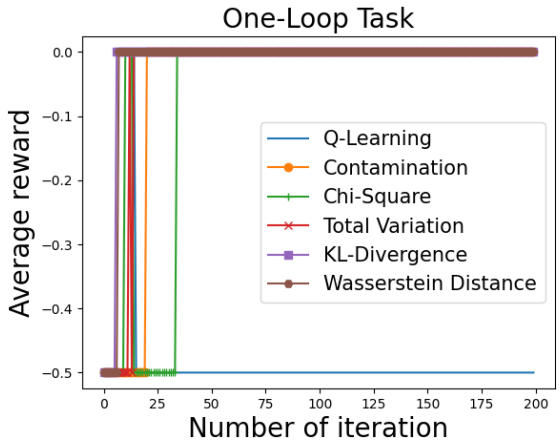
Figure 9: One-Loop Task.

We implement our robust RVI Q-learning and vanilla non-robust Q-learning as the baseline in this environment. At each time step t, we plot the difference between $Q_t(s_1, a_l)$ and $Q_t(s_1, a_r)$ in Figure 10(a). If $Q_t(s_1, a_l) - Q_t(s_1, a_r) < 0$, the greedy policy will be going right; and if $Q_t(s_1, a_l) - Q_t(s_1, a_r) > 0$, the policy will be going left. As the results show, the vanilla Q-learning will finally learn a policy $\pi(s_1) = a_r$, while our algorithms learn a policy $\pi(s_1) = a_l$.

To verify the robustness of our method, we test the learned policies under a perturbed testing environment, shown on the right of Figure 9. We plot the average-reward of policies π_t under this perturbed environment. The results are shown in Figure 10(b).



(a) $Q(s_1, a_l) - Q(s_2, a_r)$



(b) Average-Reward under Testing MDP

Figure 10: One-Loop Task.

As the results show, our robust RVI Q-learning learns a more robust policy under the nominal environment, which obtains a higher reward in the perturbed environment; whereas the non-robust Q-learning learns a policy that is optimal w.r.t. the nominal environment, but less robust when the environment is perturbed. This verifies that our algorithm is more robust than the vanilla method.