Journal of Artificial Intelligence Research 80 (2024) 719-803 Submitted 09/2023; published 06/2024

Robust Average-Reward Reinforcement Learning

Yue Wang YUE.WANG@UCF.EDU
University of Central Florida

Alvaro Velasquez ALVARO.VELASQUEZ@COLORADO.EDU
University of Colorado Boulder

George Atia GEORGE.ATIA@UCF.EDU
University of Central Florida

Ashley Prater-Bennette ASHLEY.PRATER-BENNETTE@US.AF.MIL
Air Force Research Laboratory

Shaofeng Zou SZOU3@BUFFALO.EDU
University at Buffalo, The State University of New York

Abstract

Robust Markov decision processes (MDPs) aim to find a policy that optimizes the worst-case performance over an uncertainty set of MDPs. Existing studies mostly have focused on the robust MDPs under the discounted reward criterion, leaving the ones under the average-reward criterion largely unexplored. In this paper, we develop the first comprehensive and systematic study of robust average-reward MDPs, where the goal is to optimize the long-term average performance under the worst case. Our contributions are four-folds: (1) we prove the uniform convergence of the robust discounted value function to the robust average-reward function as the discount factor γ goes to 1; (2) we derive the robust average-reward Bellman equation, characterize the structure of its solution set, and prove the equivalence between solving the robust Bellman equation and finding the optimal robust policy; (3) we design robust dynamic programming algorithms, and theoretically characterize their convergence to the optimal policy; and (4) we design two model-free algorithms unitizing the multi-level Monte-Carlo approach, and prove their asymptotic convergence.

1. Introduction

The Markov decision process (MDP) is an effective mathematical tool for modeling sequential decision-making problems in stochastic environments (Derman, 1970; Puterman, 1994). Solving an MDP problem entails finding an optimal policy that maximizes a cumulative reward according to a given criterion. However, due to reasons including non-stationarity of the environment, modeling error, exogenous perturbation, partial observability, and adversarial attacks, a model mismatch exists between the assumed MDP model and the underlying environment and can result in solution policies with poor performance. To solve this issue, a framework of robust MDPs, e.g., (Bagnell et al., 2001; Nilim & El Ghaoui, 2004; Iyengar, 2005), has been proposed. Rather than adopting a fixed MDP model, in robust MDP, one seeks to optimize the worst-case performance over an uncertainty set of possible MDP models. The solution provides performance guarantees for all MDP models in the uncertainty set, and is thus robust to the model mismatch.

Robust MDPs falling under different reward optimality criteria are fundamentally different. In robust discounted MDPs, the goal is to find a policy that maximizes the cumulative

discounted reward in the worst case, where the reward received diminishes exponentially over time. Much of the prior work in the robust setting has focused on the discounted reward criterion (see Sec. 1.2). Although discounted MDPs induce an elegant contractive Bellman operator which is mathematically convenient, the policy obtained may have a poor long-term performance when a system operates for an extended period of time. When the discount factor is close to 1, the agent may prefer to compare policies on the basis of their average expected reward instead of their expected total discounted reward, e.g., queueing control, inventory management in supply chains, scheduling automatic guided vehicles and applications in communication networks (Kober et al., 2013). Therefore, it is of great importance to optimize the long-term average performance of a system, especially in the present of environment uncertainty or model mismatch.

Nevertheless, robust MDPs under the average-reward criterion are largely understudied. Compared to the discounted reward, the average-reward depends on the limiting behavior of the underlying stochastic process and is markedly more intricate. A recognized instance of such intricacy concerns the one-to-one correspondence between the stationary policies and the limit points of state-action frequencies, which while true for discounted MDPs, breaks down under the average-reward criterion even in the non-robust setting except in some very special cases (Puterman, 1994; Atia et al., 2021). This is largely due to the dependence on the necessary conditions for establishing a contraction in average-reward settings on the graph structure of the MDP, versus the discounted-reward setting where it simply suffices to have a discount factor that is strictly less than one (Kazemi, Perez, Somenzi, Soudjani, Trivedi, & Velasquez, 2022). Heretofore, only a handful of studies have considered average-reward MDPs in the robust setting. The first work by (Tewari & Bartlett, 2007) considers robust average-reward MDPs under a specific finite interval uncertainty set, but their method is not easily applicable to other uncertainty sets. More recently, (Lim et al., 2013) proposed an algorithm for robust average-reward MDPs under the ℓ_1 uncertainty set, and in (Grand-Clément & Petrik, 2023), a characterization of the similarity between the optimal robust policy for average-reward and discounted reward MDPs is studied for a class of uncertainty models. Beyond these works, however, obtaining fundamental characterizations of the problem and convergence guarantees remains elusive.

On the other hand, model-free approaches for robust MDPs, even for the discounted reward setting, are still far from well-established. Recent work (Roy et al., 2017) developed the first model-free algorithm for robust discounted RL, where they develop a relaxation to the uncertainty set which can be over-pessimistic due to the deviation from the nominal kernel. Moreover, a strong assumption is made on the discount factor in order to guarantee the convergence. To solve these issues, recent works (Wang & Zou, 2021; Liu et al., 2022; Liang et al., 2023; Wang et al., 2023) developed robust Q-learning algorithm for various uncertainty sets with convergence guarantee and sample complexity analyses. However, these approaches are for the discounted reward setting, and there is still a gap in developing model-free approaches for the average-reward setting.

1.1 Challenges and Contributions

In this paper, we derive characterizations of robust average-reward MDPs with general uncertainty sets, and develop model-based and model-free approaches with provable the-

oretical guarantees. Our approach is fundamentally different from prior work on robust discounted MDPs, and robust and non-robust average-reward MDPs. In particular, the key challenges and the main contributions are summarized below.

We characterize the limiting behavior of robust discounted value function as the discount factor $\gamma \to 1$. For the standard non-robust MDP and for a fixed transition kernel, the discounted non-robust value function converges to the average-reward non-robust value function as $\gamma \to 1$ (Puterman, 1994). However, in the robust setting, we need to consider the worst-case limiting behavior under all possible transition kernels in the uncertainty set. Hence, the previous point-wise convergence result (Puterman, 1994) cannot be directly applied. In (Tewari & Bartlett, 2007), a finite interval uncertainty set is studied, where due to its special structure, the number of possible worst-case transition kernels of robust discounted MDPs is finite, and hence the order of min (over transition kernel) and $\lim_{\gamma\to 1}$ can be exchanged, and therefore, the robust discounted value function converges to the robust average-reward value function. This result, however, does not hold for general uncertainty sets investigated in this paper. We first prove the *uniform* convergence of discounted non-robust value function to average-reward w.r.t. the transition kernels and policies. Based on this uniform convergence, we show the convergence of the robust discounted value function to the robust average-reward. This uniform convergence result is the first in the literature and is of key importance to motivate our algorithm design and to guarantee convergence to the optimal robust policy in the average-reward setting.

We design algorithms for robust policy evaluation and optimal control based on the limit method. Based on the uniform convergence, we then use robust discounted MDPs to approximate robust average-reward MDPs. We show that when γ is large, any optimal policy of the robust discounted MDP is also an optimal policy of the robust average-reward, and hence solves the robust optimal control problem in the average-reward setting. This result is similar to the Blackwell optimality (Blackwell, 1962; Hordijk & Yushkevich, 2002) for the non-robust setting. However, our proof is fundamentally different. Technically, the proof in (Blackwell, 1962; Hordijk & Yushkevich, 2002) is based on the fact that the difference between the discounted value functions of two policies is a rational function of the discount factor, which has a finite number of zeros. However, in the robust setting with a general uncertainty set, the difference is no longer a rational function due to the min over the transition kernel. We construct a novel proof based on the limiting behavior of robust discounted MDPs, and show that the (optimal) robust discounted value function converges to the (optimal) robust average-reward as $\gamma \to 1$. Motivated by these insights, we then design our algorithms by applying a sequence of robust discounted Bellman operators with an increasing discount factor. We prove that our method can (i) evaluate the robust average-reward for a given policy and (ii) find the optimal robust value function and, in turn, the optimal robust policy for general uncertainty sets.

We derive the robust average-reward Bellman equation and design a direct model-based algorithm with convergence guarantee. The fundamental structure of MDPs is usually characterized by a bootstrap-type equation, namely, the Bellman equation, which reveals the relation among the value function, reward function, and the transition kernels. It is of great importance in value-based approaches, since finding the optimal policy and solving the Bellman equation are equivalent. We derive a robust Bellman equation for robust average-reward MDPs, and show that the pair of robust relative value functions

and robust average-reward is a solution to the robust Bellman equation under the average-reward setting. We further prove the equivalence between finding its solution and optimizing the robust average-reward. We then design a robust value iteration method that provably converges to the solution of the robust Bellman equation, i.e., solves the optimal policy for the robust average-reward MDP problem.

We design model-free algorithms for robust policy evaluation and optimal control. We then design model-free algorithms for robust average-reward RL. The major challenges are two-fold: 1) Constructing an unbiased estimator of the robust Bellman operator that captures the worst-case performance using the samples from the nominal transition kernel; and 2) Deriving the convergence of the stochastic algorithms. Regarding the first problem, one plausible approach is to define the estimator using the empirical transition kernel. However, due to the non-linear dependence of the robust Bellman operator on the nominal transition kernel, this plug-in estimator is biased. We hence employ the multi-level Monte-Carlo method (Blanchet & Glynn, 2015) and construct unbiased estimators. We then utilize them to design robust RVI TD and Q-learning algorithms. We leverage the stochastic approximation approach and the characterization of the Bellman equation to show the global asymptotic stability of our algorithms and further derive the convergence of our stochastic algorithms. Specifically, robust RVI TD 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.

1.2 Related Work

Robust discounted 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), 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 where stochastic samples from the nominal MDP of the uncertainty set are available in an online fashion (Roy et al., 2017; Badrinath & Kalathil, 2021; Wang & Zou, 2021, 2022; Tessler et al., 2019) and an offline fashion (Zhou et al., 2021; Yang et al., 2022; Panaganti & Kalathil, 2022; Goyal & Grand-Clement, 2018; Kaufman & Schaefer, 2013; Ho et al., 2018, 2021; Si et al., 2020). There are also empirical studies on robust RL, e.g., (Vinitsky et al., 2020; Pinto et al., 2017; Abdullah et al., 2019; Hou et al., 2020; Rajeswaran et al., 2017; Huang et al., 2017; Kos & Song, 2017; Lin et al., 2017; Pattanaik et al., 2018; Mandlekar et al., 2017). For discounted MDPs, the robust Bellman operator is a contraction, based on which robust dynamic programming and value-based methods can be designed. In this paper, we focus on robust average-reward MDPs, where the robust Bellman operator for average-reward MDPs is not a contraction, and its fixed point may not be unique. Moreover, the average-reward setting depends on the limiting behavior of the underlying stochastic process, which is thus more intricate.

Robust average-reward MDPs. Studies on robust average-reward MDPs are quite limited in the literature. Robust average-reward MDPs under a specific finite interval uncertainty set was studied in (Tewari & Bartlett, 2007), where the authors showed the existence of a robust Blackwell optimality constant, i.e., there exists some $\delta \in [0, 1)$, such that the

optimal robust policy for the robust average-reward MDP exists and remains unchanged for the robust discounted reward ones with any discount factor $\gamma \in [\delta, 1)$. However, this result depends on the structure of the uncertainty set, where the number of possible worst-case transition kernels is assumed finite. Under the similar assumptions, a recent work (Grand-Clément & Petrik, 2023) derived a lower bound on the robust Blackwell optimality constant δ; Under a similar polytopic assumption, (Chatterjee et al., 2023) design a policy iteration algorithm with convergence and computational complexity analysis. For more general uncertainty sets, the studies of robust average-reward MDPs, are not well-explored. Another work (Lim et al., 2013) designed a model-free algorithm for a specific ℓ_1-norm uncertainty set and characterized its regret bound. However, their method also relies on the structure of the ℓ_1-norm uncertainty set, and may not be generalizable to other types of uncertainty sets. In this paper, our results can be applied to various types of uncertainty sets, and thus is more general.

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 Model

In this section, we introduce some preliminaries on discounted MDPs, average-reward MDPs, and robust MDPs.

Discounted MDPs. A discounted MDP $(\mathcal{S}, \mathcal{A}, \mathsf{P}, r, \gamma)$ is specified by: a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, a transition kernel $\mathsf{P} = \{p_s^a \in \Delta(\mathcal{S}), a \in \mathcal{A}, s \in \mathcal{S}\}$[1], where p_s^a is the distribution of the next state over $\mathcal{S}$ upon taking action a in state s (with $p_{s,s'}^a$ denoting the probability of transitioning to s'), a reward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$, and a discount factor $\gamma \in [0, 1)$. At each time step t, the agent at state s_t takes an action a_t, the environment then transitions to the next state s_{t+1} according to $p_{s_t}^{a_t}$, and produces a reward signal $r(s_t, a_t) \in [0, 1]$ to the agent. In this paper, we also write $r_t = r(s_t, a_t)$ for convenience.

A stationary policy[2] $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is a distribution over $\mathcal{A}$ for any given state s, and the agent takes action a at state s with probability $\pi(a|s)$. The discounted value function of a stationary policy π starting from $s \in \mathcal{S}$ is defined as the expected discounted cumulative reward by following policy π: $V_{\mathsf{P},\gamma}^{\pi}(s) \triangleq \mathbb{E}_{\pi,\mathsf{P}}\left[\sum_{t=0}^{\infty} \gamma^t r_t | S_0 = s\right]$.

Average-Reward MDPs. Different from discounted MDPs, average-reward MDPs do not discount the reward over time, and consider the behavior of the underlying Markov

1. $\Delta(\mathcal{S})$: the $(|\mathcal{S}| - 1)$-dimensional probability simplex on $\mathcal{S}$.
2. In this paper, we focus on the stationary policies. The studies under the history-dependent policies are left for future exploration.

process under the steady-state distribution. More specifically, under a specific transition kernel P, the average-reward of a policy π starting from $s \in \mathcal{S}$ is defined as

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

which we also refer to in this paper as the average-reward value function for convenience.

The average-reward value function can also be equivalently written as follows: $g^{\pi}_{\mathsf{P}} = \lim_{n\to\infty}\frac{1}{n}\sum_{t=0}^{n-1}(\mathsf{P}^{\pi})^t r_\pi \triangleq \mathsf{P}^{\pi}_* r_\pi$, where $(\mathsf{P}^{\pi})_{s,s'} \triangleq \sum_a \pi(a|s)p^a_{s,s'}$ and $r_\pi(s) \triangleq \sum_a \pi(a|s)r(s,a)$ are the transition matrix and reward function induced by π, and $\mathsf{P}^{\pi}_* \triangleq \lim_{n\to\infty}\frac{1}{n}\sum_{t=0}^{n-1}(\mathsf{P}^{\pi})^t$ is the limit matrix of P^{π}.

In the average-reward setting, we also define the following relative value function

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

which is the cumulative difference over time between the reward and the average value g^{π}_{P}. It has been shown that (Puterman, 1994): $V^{\pi}_{\mathsf{P}} = H^{\pi}_{\mathsf{P}} r_\pi$, where $H^{\pi}_{\mathsf{P}} \triangleq (I - \mathsf{P}^{\pi} + \mathsf{P}^{\pi}_*)^{-1}(I - \mathsf{P}^{\pi}_*)$ is defined as the deviation matrix of P^{π}.

The relationship between the average-reward and the relative value functions can be characterized by the following Bellman equation (Puterman, 1994):

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

Robust discounted and average-reward MDPs. For robust MDPs, the transition kernel is not fixed but belongs to some uncertainty set $\mathcal{P}$. After the agent takes an action, 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., $\mathcal{P} = \bigotimes_{s,a} \mathcal{P}^a_s$, where $\mathcal{P}^a_s \subseteq \Delta(\mathcal{S})$. We note that there are also studies on relaxing the (s,a)-rectangular uncertainty set to s-rectangular uncertainty set, which is not the focus of this paper.

Under the robust setting, we consider the worst-case performance over the uncertainty set of MDPs. More specifically, the robust discounted value function of a policy π for a discounted MDP is defined as

$$V^{\pi}_{\mathcal{P},\gamma}(s) \triangleq \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\pi,\mathsf{P}}\left[\sum_{t=0}^{\infty}\gamma^t r_t | S_0 = s\right]. \tag{4}$$

In this paper, we focus on the following worst-case average-reward for a policy π:

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

to which, for convenience, we refer as the robust average-reward value function[3].

3. Here we consider the worst case performance among the stationary model, i.e., the transition kernels are identical at each time step. However, as shown later in this paper, the worst case performance under the dynamic uncertain model is the same as the one under the stationary model. Hence it is sufficient to consider the stationary model.

For robust discounted MDPs, it has been shown that the robust discounted value function is the unique fixed-point of the robust discounted Bellman operator (Nilim & El Ghaoui, 2004; Iyengar, 2005; Puterman, 1994):

$$\mathbf{T}_\pi V(s) \triangleq \sum_{a\in\mathcal{A}} \pi(a|s)\left(r(s,a) + \gamma\sigma_{\mathcal{P}_s^a}(V)\right), \tag{6}$$

where $\sigma_{\mathcal{P}_s^a}(V) \triangleq \min_{p\in\mathcal{P}_s^a} p^\top V$ is the support function of V on $\mathcal{P}_s^a$. Based on the contraction of $\mathbf{T}_\pi$, robust dynamic programming approaches, e.g., robust value iteration, can be designed (Nilim & El Ghaoui, 2004; Iyengar, 2005) (see Appendix B for a review of these methods). However, there is no such contraction result for robust average-reward MDPs. In this paper, our goal is to find a policy that optimizes the robust average-reward value function:

$$\max_{\pi\in\Pi} g_{\mathcal{P}}^\pi(s), \text{ for any } s\in\mathcal{S}, \tag{7}$$

where Π is the set of all stationary policies, and we denote by $g_{\mathcal{P}}^*(s) \triangleq \max_\pi g_{\mathcal{P}}^\pi(s)$ the optimal robust average-reward.

3. Limit Approach

We first take a limit approach to solve the problem of robust average-reward MDPs in (7). It is known that under the non-robust setting, for any fixed π and P, the discounted value function converges to the average-reward value function as the discount factor γ approaches 1 (Puterman, 1994), i.e.,

$$\lim_{\gamma\to 1}(1-\gamma)V_{\mathsf{P},\gamma}^\pi = g_{\mathsf{P}}^\pi. \tag{8}$$

Note that the term $(1-\gamma)$ is necessary to ensure the finiteness of the limit, otherwise $V_{\mathsf{P},\gamma}^\pi \to \infty$ if $\gamma \to 1$.

We take a similar idea, and show that the same result holds in the robust case: $\lim_{\gamma\to 1}(1-\gamma)V_{\mathcal{P},\gamma}^\pi = g_{\mathcal{P}}^\pi$ under a mild assumption. This result further enables us to draw numerous characterizations of the fundamental structure of the robust MDPs under the average-reward setting. Moreover, we design algorithms (Algorithms 1 and 2) to solve robust MDPs under the average-reward criterion based on this result, and further prove its convergence and optimality.

3.1 Uniform Convergence of Robust Discounted Value Functions

In this section, we first show that the convergence $\lim_{\gamma\to 1}(1-\gamma)V_{\mathsf{P},\gamma}^\pi = g_{\mathsf{P}}^\pi$ is uniform on the set $\Pi\times\mathcal{P}$. In studies of average-reward MDPs, it is usually the case that a certain class of MDPs is considered, e.g., unichain and communicating (Wei et al., 2020; Zhang & Ross, 2021; Chen et al., 2022; Wan et al., 2021). In this paper, we focus on the unichain setting to highlight the major technical novelty to achieve robustness.

Assumption 1. *For any $s\in\mathcal{S}, a\in\mathcal{A}$, the uncertainty set $\mathcal{P}_s^a$ is a compact subset of $\Delta(\mathcal{S})$. And for any $\pi\in\Pi, \mathsf{P}\in\mathcal{P}$, the induced MDP is a unichain.*

The first part of Assumption 1 amounts to assuming that the uncertainty set is closed. We remark that many standard uncertainty sets satisfy this assumption, e.g., those defined by ϵ-contamination (Huber, 1965), finite interval (Tewari & Bartlett, 2007), total-variation (Rahimian et al., 2022) and KL-divergence (Hu & Hong, 2013). The unichain assumption is also widely used in studies of average-reward MDPs, e.g., (Puterman, 1994; Wan et al., 2021; Zhang & Ross, 2021; Lan, 2020; Zhang et al., 2021b). Also, it is worth noting that under the unichain assumption, the average-reward is identical for every starting state, i.e., $g^\pi_{\mathsf{P}}(s_1) = g^\pi_{\mathsf{P}}(s_2), \forall s_1, s_2 \in \mathcal{S}$ (Bertsekas, 2011).

Remark 1. *The results in this section actually only require the uniform boundedness of* $\|H^\pi_{\mathsf{P}}\|, \forall \pi \in \Pi, \mathsf{P} \in \mathcal{P}$ *(Lemma 4 in the appendix). Assumption 1 is one sufficient condition.*

In (Puterman, 1994), the convergence $\lim_{\gamma\to 1}(1-\gamma)V^\pi_{\mathsf{P},\gamma} = g^\pi_{\mathsf{P}}$ for a fixed policy π and a fixed transition kernel P (non-robust setting) is point-wise. However, such point-wise convergence does not provide any convergence guarantee on the robust discounted value function, as the robust value function measures the worst-case performance over the uncertainty set and the order of lim and min may not be exchangeable in general. In the following theorem, we prove the uniform convergence of the discounted value function under the foregoing assumption.

Theorem 1 (Uniform convergence)**.** *Under Assumption 1, the discounted value function converges uniformly to the average-reward value function on* $\Pi \times \mathcal{P}$ *as* $\gamma \to 1$*, i.e.,*

$$\lim_{\gamma\to 1}(1-\gamma)V^\pi_{\mathsf{P},\gamma} = g^\pi_{\mathsf{P}} \text{ uniformly on } \Pi \times \mathcal{P}. \tag{9}$$

With uniform convergence in Theorem 1, the order of the limit $\gamma \to 1$ and $\min_{\mathsf{P}}$ can be interchanged. Then, the following convergence of the robust discounted value function can be established.[4]

Theorem 2. *The robust discounted value function in* (4) *converges to the robust average-reward uniformly on* Π*:*

$$\lim_{\gamma\to 1}(1-\gamma)V^\pi_{\mathcal{P},\gamma} = g^\pi_{\mathcal{P}} \text{ uniformly on } \Pi. \tag{10}$$

We note that the convergence can also be derived under some other assumptions or for specific uncertainty sets. For example, a similar convergence result is shown in (Tewari & Bartlett, 2007) for a special uncertainty set of finite interval type. This result is further generalized in (Grand-Clément & Petrik, 2023; Goyal & Grand-Clement, 2018), where the convergence is obtained under the assumption that the number of the possible worst-case transition kernels is finite, i.e., $\{\mathsf{P} \in \mathcal{P} : g^\pi_{\mathsf{P}} = g^\pi_{\mathcal{P}}\}$ is a finite set for any policy π. Besides the finite interval uncertainty set, it is shown that the uncertainty sets defined by l_p-norm (Grand-Clément & Petrik, 2023; Goyal & Grand-Clement, 2018) also satisfy this assumption. Under this assumption, the lim and max are interchangeable and hence the convergence can be obtained. Our Theorem 2 holds for general compact uncertainty sets. Moreover, it is worth highlighting that our proof technique is fundamentally different from

4. During the preparation of our manuscript, a recent work (Grand-Clement, Petrik, & Vieille, 2023) develops a similar result without the unichian assumption.

the one in (Tewari & Bartlett, 2007; Grand-Clément & Petrik, 2023), where the worst-case transition kernels are from a finite set, i.e., $V^{\pi}_{\mathcal{P},\gamma} = \min_{\mathsf{P}\in\mathcal{M}} V^{\pi}_{\mathsf{P},\gamma}$ for a finite set $\mathcal{M} \subseteq \mathcal{P}$. This hence implies the interchangeability of lim and min. However, for general uncertainty sets, the number of worst-case transition kernels may not be finite. We demonstrate the interchangeability via our uniform convergence result in Theorem 1.

The preceding uniform convergence result enables us to exchange the order of the operators $\lim_{\gamma\to 1}$, $\min_{\mathsf{P}\in\mathcal{P}}$ and $\max_{\pi\in\Pi}$. This connects robust MDPs under the discounted reward with average-reward settings. In the following theorem, we show that we can also restrict ourselves to deterministic polices when optimizing the robust MDPs under the average-reward criterion.

Theorem 3. *(Deterministic Optimality) There exists a deterministic optimal robust policy, i.e., $\exists \pi \in \Pi_D$, such that $g^{\pi}_{\mathcal{P}} = g^{*}_{\mathcal{P}}$.*

The uniform convergence result in Theorem 2 motivates the use of robust discounted MDPs with $\gamma \to 1$ to approximate robust average-reward MDPs, which we refer to as the limit method. As discussed in (Blackwell, 1962; Hordijk & Yushkevich, 2002), the average-reward criterion is insensitive and under selective since it is only interested in the performance under the steady-state distribution. For example, two policies providing rewards: $100 + 0 + 0 + \cdots$ and $0 + 0 + 0 + \cdots$ are equally good/bad. For the non-robust setting, a more sensitive term of optimality was introduced by Blackwell (Blackwell, 1962). More specifically, a policy is said to be Blackwell optimal if it optimizes the discounted value function for any discount factor $\gamma \in (\delta, 1)$ for some $\delta \in (0, 1)$. Together with (8), the optimal policy obtained by taking $\gamma \to 1$ is optimal not only for the average-reward criterion, but also for the discounted criterion with large γ. Intuitively, it is optimal under the average-reward setting, and is sensitive to early rewards.

Following a similar idea, we justify that the optimal robust policy for the robust average-reward MDPs is also sensitive to early rewards. Denote by Π^{*}_{D} the set of all the deterministic optimal policies for robust average-reward (proved to exist in Lemma 9), i.e. $\Pi^{*}_{D} = \{\pi \in \Pi_D : g^{\pi}_{\mathcal{P}} = g^{*}_{\mathcal{P}}\}$.

Theorem 4 (Blackwell optimality)**.** *There exists $0 < \delta < 1$, such that for any $\gamma > \delta$, the deterministic optimal robust policy for robust discounted value function $V^{*}_{\mathcal{P},\gamma}$ is also optimal under the average-reward criterion. Moreover, when Π^{*}_{D} is a singleton, there exists a unique Blackwell optimal policy.*

This result implies that the optimal robust policy for average-reward MDPs has an additional advantage that the policy it finds not only optimizes the average-reward in steady state, but also is sensitive to early rewards.

It is worth highlighting the distinction of our results from the technique used in the proof of Blackwell optimality (Blackwell, 1962). In the non-robust setting, the existence of a stationary Blackwell optimal policy is proved via contradiction, where a difference function of two policies π and ν: $f_{\pi,\nu}(\gamma) \triangleq V^{\pi}_{\mathsf{P},\gamma} - V^{\nu}_{\mathsf{P},\gamma}$ is used in the proof. It was shown by contradiction that f has infinitely many zeros, which however contradicts with the fact that f is a rational function of γ with a finite number of zeros. A similar technique was also used in (Tewari & Bartlett, 2007) for the finite interval uncertainty set. Specifically, in (Tewari & Bartlett, 2007), it was shown that the worst-case transition kernels for any π, γ

are from a finite set $\mathcal{M}$, hence $f_{\pi,\nu}(\gamma) \triangleq \min_{\mathsf{P}\in\mathcal{M}} V^{\pi}_{\mathsf{P},\gamma} - \min_{\mathsf{P}\in\mathcal{M}} V^{\nu}_{\mathsf{P},\gamma}$ can also be shown to be a rational function with a finite number of zeroes. For a general uncertainty set $\mathcal{P}$, the difference function $f_{\pi,\nu}(\gamma)$, however, may not be rational. This makes the method in (Blackwell, 1962; Tewari & Bartlett, 2007) inapplicable to our problem.

3.2 Limit Method: Robust Value Iteration

Results in Section 3.1 play a fundamental role in developing the limit method for robust average-reward MDPs, and are of key importance to motivate the design of the following two algorithms. The basic idea is to apply a sequence of robust discounted Bellman operators on an arbitrary initialization while increasing the discount factor at a certain rate.

We first consider the robust policy evaluation problem, which aims to estimate the robust average-reward $g^{\pi}_{\mathcal{P}}$ for a fixed policy π. This problem for robust discounted MDPs is well studied in the literature. However, results for robust average-reward MDPs are quite limited except for the one in (Tewari & Bartlett, 2007; Goyal & Grand-Clement, 2018) for specific uncertainty sets. We present a robust value iteration (robust VI) algorithm for evaluating the robust average-reward with general uncertainty sets in Algorithm 1. At each time step

Algorithm 1 Robust VI: Policy Evaluation

Input: $\pi, V_0(s) = 0, \forall s, T$

1: **for** $t = 0, 1, ..., T-1$ **do**
2: $\quad\gamma_t \leftarrow \frac{t+1}{t+2}$
3: $\quad$**for** all $s \in \mathcal{S}$ **do**
4: $\quad\quad V_{t+1}(s) \leftarrow \mathbb{E}_{\pi}[(1-\gamma_t) r(s,A) + \gamma_t \sigma_{\mathcal{P}^A_s}(V_t)] = \mathbb{E}_{\pi}[(1-\gamma_t) r(s,A) + \gamma_t \min_{\mathsf{P}\in\mathcal{P}^A_s}(\mathsf{P}V_t)]$
5: $\quad$**end for**
6: **end for**
7: **return** V_T

t, the discount factor γ_t is set to be $\frac{t+1}{t+2}$, which converges to 1 as $t \to \infty$. Subsequently, a robust Bellman operator w.r.t discount factor γ_t is applied on the current estimate V_t of the robust discounted value function $(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}$. As the discount factor approaches 1, the estimated robust discounted value function converges to the robust average-reward $g^{\pi}_{\mathcal{P}}$ by Theorem 2. The following result shows that the output of Algorithm 1 converges to the robust average-reward.

Theorem 5. *Algorithm 1 converges to the robust average-reward, i.e.,* $\lim_{T\to\infty} V_T = g^{\pi}_{\mathcal{P}}$.

Besides the robust policy evaluation problem, it is also of great practical importance to find an optimal policy that maximizes the worst-case average-reward, i.e., to solve (7). Based on a similar idea as the one of Algorithm 1, we extend our limit approach to solve the robust optimal control problem in Algorithm 2.

Algorithm 2 Robust VI: Optimal Control

Input: $V_0(s) = 0, \forall s, T$

for $t = 0, 1, ..., T-1$ **do**
  $\gamma_t \leftarrow \frac{t+1}{t+2}$
  for all $s \in \mathcal{S}$ **do**
    $V_{t+1}(s) \leftarrow \max_{a \in \mathcal{A}} \left\{(1-\gamma_t) r(s,a) + \gamma_t \sigma_{\mathcal{P}_s^a}(V_t)\right\}$
  end for
end for
for $s \in \mathcal{S}$ **do**
  $\pi_T(s) \leftarrow \arg\max_{a \in \mathcal{A}} \left\{(1-\gamma_t) r(s,a) + \gamma_t \sigma_{\mathcal{P}_s^a}(V_T)\right\}$
end for
return V_T, π_T

The discount factor γ_t is set similarly as in Algorithm 1, and a one-step robust discounted Bellman operator (for optimal control) w.r.t. γ_t is applied to the current estimate V_t. The following theorem establishes that V_T in Algorithm 2 converges to the optimal robust value function, and hence can find the optimal robust policy.

Theorem 6. *The output V_T in Algorithm 2 converges to the optimal robust average-reward $g^*_{\mathcal{P}}$, i.e., $V_T \to g^*_{\mathcal{P}}$ as $T \to \infty$.*

4. Direct Approach

The limit approach in Section 3 is based on the uniform convergence of the robust discounted value function, and uses discounted MDPs to approximate average-reward MDPs. In this section, we develop a direct approach to solving the robust average-reward MDPs that does not adopt discounted MDPs as intermediate steps.

4.1 Robust Bellman Equation

One of the most important results which enable the dynamic programming approach for solving MDPs is the Bellman equation. Such results have been generalized to robust discounted MDPs (Nilim & El Ghaoui, 2004; Iyengar, 2005), and we develop an analog result for robust average-reward MDPs as follows.

We first generalize the relative value function in (2) to the robust relative value function. The robust relative value function measures the difference between the worst-case cumulative reward and the worst-case average-reward for a policy π.

Definition 1. *The robust relative value function is defined as*

$$V^\pi_{\mathcal{P}}(s) \triangleq \min_{\kappa \in \bigotimes_{t \geq 0} \mathcal{P}} \mathbb{E}_{\kappa,\pi}\left[\sum_{t=0}^{\infty} (r_t - g^\pi_{\mathcal{P}}) | S_0 = s\right], \tag{11}$$

where $g^\pi_{\mathcal{P}}$ is the worst-case average-reward defined in (5).

We further introduce several notations. For $V \in \mathbb{R}^{|\mathcal{S}|}$, denote by $\mathsf{P}_V(s,a) \triangleq \arg\min_{p \in \mathcal{P}_s^a} pV$ and let $\mathsf{P}_V = \{\mathsf{P}_V(s,a), s \in \mathcal{S}, a \in \mathcal{A}\}$. Moreover, denote the set of the worst-case transition kernels by Ω^π_g, i.e., $\Omega^\pi_g = \{\mathsf{P} \in \mathcal{P} : g^\pi_{\mathsf{P}} = g^\pi_{\mathcal{P}}\}$.

Theorem 7. *For any π, $(V_{\mathcal{P}}^{\pi}, g_{\mathcal{P}}^{\pi})$ is a solution to the following robust Bellman equation:*

$$V(s) + g = \sum_{a} \pi(a|s) \left(r(s,a) + \sigma_{\mathcal{P}_s^a}(V) \right), \forall s. \tag{12}$$

Moreover, if (g, V) is a solution to it, then

- *1) $g = g_{\mathcal{P}}^{\pi}$;*
- *2) $\mathsf{P}_V \in \Omega_g^{\pi}$;*
- *3) $V = V_{\mathsf{P}_V}^{\pi} + ce$ for some $c \in \mathbb{R}$, where e denotes the vector $(1, 1, ..., 1) \in \mathbb{R}^{|\mathcal{S}|}$.*

It can be seen that the robust Bellman equation for average-reward MDPs has a similar structure to the one for discounted MDPs in (6) except for a discount factor. This actually reveals a fundamental difference between the robust Bellman operator of the discounted MDPs and the average-reward ones. For a discounted MDP, its robust Bellman operator is a contraction with constant γ (Nilim & El Ghaoui, 2004; Iyengar, 2005), and hence the fixed point is unique. Based on this, the robust value function can be found by recursively applying the robust Bellman operator (see Appendix B for a review). In sharp contrast, in the average-reward setting, the robust Bellman operator is not necessarily a contraction, and the fixed point may not be unique. Therefore, repeatedly applying the robust Bellman operator in the average-reward setting may not even converge, which underscores that the two problem settings are fundamentally different.

The second part of Theorem 7 provides a characterization of the solutions to the Bellman equation, where we show that for any solution (g, V) to (12), the transition kernel $P_V \in \Omega_g^{\pi}$, i.e., it is a worst-case transition kernel for $g_{\mathcal{P}}^{\pi}$. This result also distinguishes the structure of the robust Bellman equation from the non-robust one. Under the non-robust setting, the solution set to the Bellman equation can be written as $\{(g_{\mathsf{P}}^{\pi}, V_{\mathsf{P}}^{\pi} + ce) : c \in \mathbb{R}\}$. The solution is uniquely determined by the transition kernel (up to some constant vector ce). In contrast, in the robust setting, the robust Bellman equation is no longer linear. Any solution V to (12) is a relative value function w.r.t. *some* worst-case transition kernel $\mathsf{P} \in \Omega_g^{\pi}$ (up to some additive constant vector), i.e., $V \in \{V_{\mathsf{P}}^{\pi} + ce : \mathsf{P} \in \Omega_g^{\pi}, c \in \mathbb{R}\}$. A natural question that arises is whether, for any $\mathsf{P} \in \Omega_g^{\pi}$, $(g_{\mathcal{P}}^{\pi}, V_{\mathsf{P}}^{\pi})$ is a solution to (12)? Lemma 1 refutes this.

Lemma 1. *There exists a robust MDP such that for some $\mathsf{P} \in \Omega_g^{\pi}$, $(g_{\mathcal{P}}^{\pi}, V_{\mathsf{P}}^{\pi})$ is not a solution to* (12).

Lemma 1 implies that the solution set to (12) is a subset of $\{V_{\mathsf{P}}^{\pi} + ce, \mathsf{P} \in \Omega_g^{\pi}, c \in \mathbb{R}\}$. Note that an explicit characterization of the solution set to (12) is challenging due to its non-linear structure; however, result 3 in Theorem 7 suffices to establish the convergence of our model-free algorithms (shown later in Section 5).

Theorem 7 characterizes the robust average-reward for a fixed policy π, which plays an essential role in policy evaluation problems, i.e., to estimate the robust average-reward for π. To find the optimal robust policy, we similarly derive the following optimality condition for robust average-reward MDPs. It is generally useful to consider the action-value functions in optimal control problems, hence we consider a Q-function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and define $V_Q(s) = \max_a Q(s,a), \forall s \in \mathcal{S}$.

Theorem 8 (Optimal robust Bellman equation)**.** *If (g, Q) is a solution to the optimal robust Bellman equation*

$$Q(s,a) = r(s,a) - g + \sigma_{\mathcal{P}_s^a}(V_Q), \forall s, a, \tag{13}$$

then

- *1) $g = g_{\mathcal{P}}^*$;*
- *2) the greedy policy w.r.t. Q: $\pi_Q(s) = \arg\max_a Q(s,a)$ is an optimal robust policy;*
- *3) $V_Q = V_{\mathsf{P}}^{\pi_Q} + ce$ for some $\mathsf{P} \in \Omega_g^{\pi_Q}, c \in \mathbb{R}$.*

Note that the theorem is presented using the action-value function Q; Similar results can be easily adapted using the V function as follows.

Corollary 1. *For any (g, V) that is a solution to*

$$\max_a \left\{ r(s,a) - g + \sigma_{\mathcal{P}_s^a}(V) - V(s) \right\} = 0, \forall s, \tag{14}$$

$g = g_{\mathcal{P}}^*$. If we further set

$$\pi^*(s) = \arg\max_a \left\{ r(s,a) + \sigma_{\mathcal{P}_s^a}(V) \right\}, \forall s \in \mathcal{S}, \tag{15}$$

then π^* is an optimal robust policy.

According to Theorem 8, finding a solution to (13) is sufficient to get the optimal robust average-reward and to derive the optimal robust policy. Similarly to Theorem 7, we omit the explicated study and characterization of the solution set to (13), but the above results are sufficient for the convergence proof of our direct methods and algorithms.

Results in this section provide a comprehensive characterization of the fundamental structure of robust MDPs under the average-reward criterion, and indicates the equivalence between solving them and find the solutions to the robust Bellman equations. However, as discussed, the solution set to the robust Bellman equations can be complicated, and are not straightforward to solve. In the next section, we develop a model-based algorithm to solve the equations and find the optimal robust policy.

4.2 Direct Method: Robust Relative Value Iteration

In the following, we generalize the RVI approach to the robust setting, and design a robust RVI algorithm in Algorithm 3. We will further show that the output of this algorithm converges to a solution to (14), and further the optimal policy could be obtained by (15).

Here, *sp* denotes the span semi-norm: $sp(w) = \max_s w(s) - \min_s w(s)$, and $f : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}$ is an offset function introduced to stabilize the algorithm updating. We adopt the following assumption from (Puterman, 1994).

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

Algorithm 3 Robust RVI

Input: V_0, ϵ

1: $w_0 \leftarrow V_0 - f(V_0)$
2: **while** $sp(w_t - w_{t+1}) \geq \epsilon$ **do**
3: **for** all $s \in \mathcal{S}$ **do**
4: $V_{t+1}(s) \leftarrow \max_a(r(s,a) + \sigma_{\mathcal{P}_s^a}(w_t))$
5: $w_{t+1}(s) \leftarrow V_{t+1}(s) - f(V_{t+1})$
6: **end for**
7: **end while**
8: **return** w_t, V_t

Assumption 2 can be easily satisfied, e.g., $f(V) = V(s_0)$ for some reference state $s_0 \in \mathcal{S}$, or $f(V) = \frac{\sum_s V(s)}{|\mathcal{S}|}$ (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.

Different from Algorithm 2, in Algorithm 3, we do not apply the robust discounted Bellman operator. The method directly solves the robust optimal control problem for average-reward robust MDPs. We note that in the previous studies of non-robust average-reward, a stronger assumption is made to guarantee the convergence of the non-robust relative value iteration (see, e.g., (Puterman, 1994)). However, it can be weakened using the aperiodic transform technique. In this paper, we further generalize such technique to the robust setting, and show the convergence of our robust RVI under Assumption 1.

In the following theorem, we show that our Algorithm 3 converges to a solution of (14). Then according to Theorem 8, the optimal robust policy can be obtained by setting π according to (15) from the limit of the algorithm.

Theorem 9. *(w_t, V_t) converges to a solution (w, V) to (14) as $\epsilon \to 0$.*

Remark 2. *In this section, we present the robust RVI algorithm for the robust optimal control problem, and its asymptotic convergence and optimality guarantee. A robust RVI algorithm for robust policy evaluation can be similarly designed by replacing the* max *in line 4, Algorithm 3 with an expectation w.r.t. π. The convergence results in Theorem 9 can also be similarly derived. Our algorithms are expected to converge linearly, as we show in Theorem 9 that it is a multi-step contraction. However, the exact number of steps and the contraction coefficients are involved dependent on the underlying MDP, and hence we leave the exact characterization of its convergence rate as a future research interest.*

5. Model-Free Approaches for Robust Average-Reward MDPs

In the previous sections, we developed the fundamental characterizations of the robust average-reward MDPs and designed two algorithms under the **model-based** setting, where we assume full knowledge of the uncertainty set $\mathcal{P}$. However, in practice the learner may

not know exactly the nominal MDP and thus the uncertainty set, instead, it can obtain samples from the nominal MDP.

A natural idea for the model-free setting is to replace the robust Bellman operators in Algorithms 2 and 3 using some estimators obtained from the samples. However, such an extension is not straightforward due to two reasons: 1) The convergence results above are not guaranteed with stochastic estimators; 2) the construction of an unbiased estimator can be difficult, due to the distribution shift between the nominal kernel and the worst-case kernel. In this section, we first construct unbiased estimators of the robust Bellman operator for five uncertainty set models, including the contamination model, the total variation model, the Chi-square model, the KL-divergence model, and the Wasserstein distance model. We then design stochastic algorithms using these unbiased estimators and show their convergence.

5.1 Robust RVI TD for Policy Evaluation

In this section, we first study the problem of robust 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).

Assumption 3. *The Markov chain induced by π is a unichain for all $\mathsf{P} \in \mathcal{P}$.*

Note that this assumption is weaker than Assumption 1, which is because the policy evaluation problem only considers a fixed policy. 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 3, the average-reward w.r.t. any $\mathsf{P} \in \mathcal{P}$ is identical for any start state, i.e., $g_{\mathsf{P}}^{\pi}(s) = g_{\mathsf{P}}^{\pi}(s'), \forall s, s' \in \mathcal{S}$.

Motivated by the robust Bellman equation in (12), we propose a model-free robust RVI TD algorithm in Algorithm 4, where $\hat{\mathbf{T}}$ is some estimator of the robust Bellman operator and will be discussed later.

Algorithm 4 Robust RVI TD

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

for $n = 0, 1, ..., N-1$ **do**
  for all $s \in \mathcal{S}$ **do**
    $V_{n+1}(s) \leftarrow V_n(s) + \alpha_n(\hat{\mathbf{T}}V_n(s) - f(V_n) - V_n(s))$
  end for
end for

Note that (12) 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 4, we construct $\hat{\mathbf{T}}V$ as an estimate of $\mathbf{T}V$ satisfying

$$\mathbb{E}[\hat{\mathbf{T}}V] = \mathbf{T}V, \quad \mathrm{Var}[\hat{\mathbf{T}}V(s)] \leq C(1 + \|V\|^2), \tag{16}$$

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 5.3, we will present in detail how to construct such $\hat{\mathbf{T}}$ for various uncertainty set models.

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

Assumption 4. *The stepsize $\{\alpha_n\}_{n=0}^\infty$ satisfies the Robbins-Monro condition, i.e., $\sum_{n=0}^\infty \alpha_n = \infty, \sum_{n=0}^\infty \alpha_n^2 < \infty$.*

Theorem 10 (Convergence of robust RVI TD)**.** *Under Assumptions 2, 3, 4, and if $\hat{\mathbf{T}}$ satisfies* (16)*, then almost surely, $(f(V_n), V_n)$ converges to a solution to* (12) *which may depend on the initialization.*

The result implies that $f(V_n) \to g_{\mathcal{P}}^\pi$ a.s., which means our robust RVI TD converges to the worst-case average-reward for the given policy π. Our robust RVI TD algorithm is hence shown to converge to a solution to (12), which solves the problem under the model-free setting.

To show the convergence of the stochastic model-free algorithm, we utilize the stochastic approximation approach. We study the associated ODE of the algorithm, tackle the stochastic noise and show the algorithm converge to the equilibrium of the ODE. As we discussed after Theorem 7, result 3 of Theorem 7 is crucial to the convergence proof of Theorem 10. Specifically, it is necessary in order to characterize the equilibrium of the associated ODE, and thus the limit of the iterates $f(V_n) \to g_{\mathcal{P}}^\pi$.

Remark 3. *Although our algorithm is presented in a synchronous fashion, the similar convergence result is also expected to hold for asynchronous version of algorithm, under the assumption that all state-action pairs are visited for infinite number of times. Such an extension is standard, see, e.g., (Wan et al., 2021).*

5.2 Robust RVI Q-Learning for Control

In this section, we aim to find the optimal robust policy under the model-free setting, i.e., find $\pi^* = \arg\max_\pi g_{\mathcal{P}}^\pi$.

Inspired by the model-based methods and the previous section, We hence present the following model-free robust RVI Q-learning algorithm.

Algorithm 5 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 $\mathbf{H}Q(s,a) \triangleq r(s,a) + \sigma_{\mathcal{P}_s^a}(V_Q)$, and we construct an estimate $\hat{\mathbf{H}}$ such that for some finite

constant C,

$$\mathbb{E}[\hat{\mathbf{H}}Q] = \mathbf{H}Q, \quad \mathrm{Var}[\hat{\mathbf{H}}Q(s,a)] \le C(1 + \|Q\|^2). \tag{17}$$

In Section 5.3, 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 11 (Convergence of robust RVI Q-learning)**.** *Under Assumptions 1, 2, 4, and if $\hat{\mathbf{H}}$ satisfies* (17)*, then almost surely, $(f(Q_n), Q_n)$ converges to a solution to* (13)*, 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^*.*

The above results imply that our robust RVI Q-learning algorithm finds the optimal robust average-reward function and the optimal robust policy, under the model-free setting. The proof technique is similar to the one of Theorem 10, where we first characterize the equilibrium of the associated ODE, and prove the global stability and convergence of our algorithm.

5.3 Construction of Estimated Operator: 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 (16) and (17) lies in estimating the support function $\sigma_{\mathcal{P}^a_s}(V)$ using samples from the nominal transition kernel P^a_s, which in general is different from the worst-case transition kernel. The function $\sigma_{\mathcal{P}^a_s}(V)$ is non-linear in the nominal kernel, which makes it challenging to construct such an estimator. For instance, the most widely-used MLE plugging-in estimator is shown to be biased. If we use the empirical transition kernel $\hat{\mathsf{P}}$ as the centroid to construct an uncertainty set $\hat{\mathcal{P}}$, then the estimator is biased: $\mathbb{E}[\sigma_{\hat{\mathcal{P}}^a_s}(V)] \neq \sigma_{\mathcal{P}^a_s}(V)$.

To solve this issue, we hence utilize the multi-level Monte-Carlo approach (Blanchet & Glynn, 2015), and construct an unbiased estimator for several widely-used non-linear uncertainty models including 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 12. *For each uncertainty set, denote its corresponding estimators by $\hat{\mathbf{T}}$ and $\hat{\mathbf{H}}$ as in Sections 5.3.1 and 5.3.2. Then, there exists some constant C, such that* (16) *and* (17) *hold.*

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

5.3.1 Linear Model: Contamination Uncertainty Set

The ζ-contamination uncertainty set is $\mathcal{P}_s^a = \{(1-\zeta)\mathsf{P}_s^a + \zeta q : q \in \zeta(\mathcal{S})\}$, where $0 < \zeta < 1$ is the radius. Under this uncertainty set, the support function can be computed as $\sigma_{\mathcal{P}_s^a}(V) = (1-\zeta)\mathsf{P}_s^a V + \zeta \min_s V(s)$, 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:

$$\hat{\sigma}_{\mathcal{P}_s^a}(V) \triangleq (1-\zeta)\gamma V(s') + \zeta \min_x V(x), \tag{18}$$

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

5.3.2 Non-Linear Models

Unlike the 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: $\hat{\mathsf{P}}_{s,N+1}^{a,E} = \frac{1}{2^N}\sum_{i=1}^{2^N} \mathbf{1}_{s_{2i}'}$, $\hat{\mathsf{P}}_{s,N+1}^{a,O} = \frac{1}{2^N}\sum_{i=1}^{2^N} \mathbf{1}_{s_{2i-1}'}$, $\hat{\mathsf{P}}_{s,N+1}^{a} = \frac{1}{2^{N+1}}\sum_{i=1}^{2^{N+1}} \mathbf{1}_{s_i'}$, $\hat{\mathsf{P}}_{s,N+1}^{a,1} = \mathbf{1}_{s_1'}$. 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

$$\hat{\sigma}_{\mathcal{P}_s^a}(V) \triangleq \sigma_{\hat{\mathcal{P}}_{s,N+1}^{a,1}}(V) + \frac{\Delta_N(V)}{p_N}, \tag{19}$$

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 (19) the unbiased estimator can then be constructed. More details can be found in Section H and Section I.

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

$$\sigma_{\mathcal{P}_s^a}(V) = \max_{\mu \ge 0} \big(\mathsf{P}_s^a(V-\mu) - \zeta \mathrm{Span}(V-\mu)\big). \tag{20}$$

Chi-square Uncertainty Set. The Chi-square uncertainty set is $\mathcal{P}_s^a = \{q \in \Delta(|\mathcal{S}|) : d_c(\mathsf{P}_s^a, q) \leq \zeta\}$, 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):

$$\sigma_{\mathcal{P}_s^a}(V) = \max_{\mu \geq 0} \left(\mathsf{P}_s^a(V-\mu) - \sqrt{\zeta \mathrm{Var}_{\mathsf{P}_s^a}(V-\mu)} \right). \tag{21}$$

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

$$\mathcal{P}_s^a = \{q : D_{KL}(q||\mathsf{P}_s^a) \leq \zeta\}, \forall s \in \mathcal{S}, a \in \mathcal{A}. \tag{22}$$

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

$$\sigma_{\mathcal{P}_s^a}(V) = -\min_{\alpha \geq 0} \left(\zeta\alpha + \alpha \log \left(\mathbb{E}_{\mathsf{P}_s^a} \left[e^{\frac{-V}{\alpha}} \right] \right) \right). \tag{23}$$

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 $\|d\|_{\mu,l} = \left(\mathbb{E}_{(X,Y)\sim\mu}[d(X,Y)^l] \right)^{1/l}$. The Wasserstein distance uncertainty set is then defined as

$$\mathcal{P}_s^a = \{q \in \Delta(|\mathcal{S}|) : W_l(\mathsf{P}_s^a, q) \leq \zeta\}. \tag{24}$$

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

Lemma 2. *It holds that*

$$\sigma_{\mathcal{P}_s^a}(V) = \sup_{\lambda \geq 0} \left(-\lambda\zeta^l + \mathbb{E}_{\mathsf{P}_s^a} \left[\inf_y \left(V(y) + \lambda d(S,y)^l \right) \right] \right). \tag{25}$$

Thus, the support function can be solved using its dual form, and the estimator can then be constructed following (19).

Remark 4. *We note that in the construction of MLMC estimators for the non-linear uncertainty sets, we do not need to estimate and store any transition models. When updating asynchronously, we only need a memory space of $|\mathcal{S}| \times |\mathcal{A}|$ to store the nominal kernel of the specific state-action pair, instead of the whole model. From this aspect, we refer to our algorithms with MLMC estimators as model-free approaches. It is left for further exploration of model-free algorithms with less memory space.*

6. Numerical Results

In this section, we numerically verify our theoretical results. We aim to verify two aspects of our methods: the convergence of the algorithms, and the robustness of them. Additional experiments can be found in Appendix A.

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). The problem can be characterized as a MDP $(30, 20, \mathcal{P}, r)$, where 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. 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 the radius of the uncertainty set $\zeta = 0.4$, $\alpha_n = 0.01$, $f(V) = \frac{\sum_s V(s)}{|\mathcal{S}|}$ and $f(Q) = \frac{\sum_{s,a} Q(s,a)}{|\mathcal{S}||\mathcal{A}|}$. We show the results under the Chi-square and Wasserstein Distance models. The results under the other three uncertainty sets are presented in Appendix A.

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. It can be seen from the results in Figure 1 that our robust RVI TD algorithm converges to the true robust average-reward value.

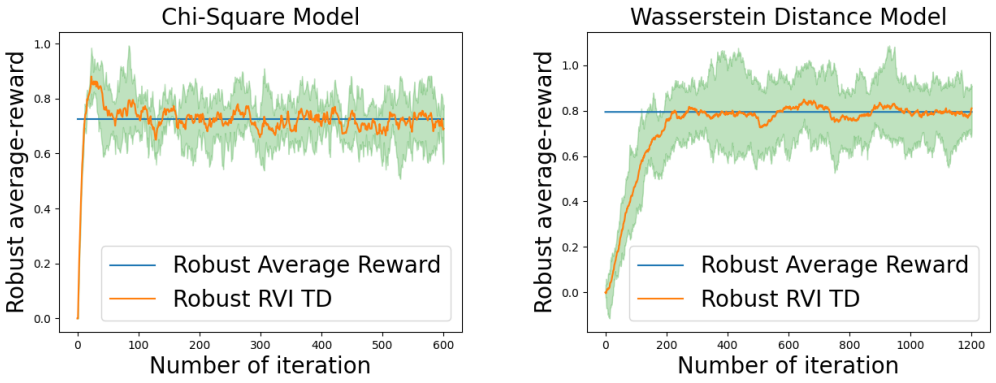


Figure 1: Robust RVI TD Algorithm.

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. Our robust RVI Q-learning converges to the optimal robust average-reward $g_{\mathcal{P}}^*$ under each uncertainty set, which verifies our theoretical results.

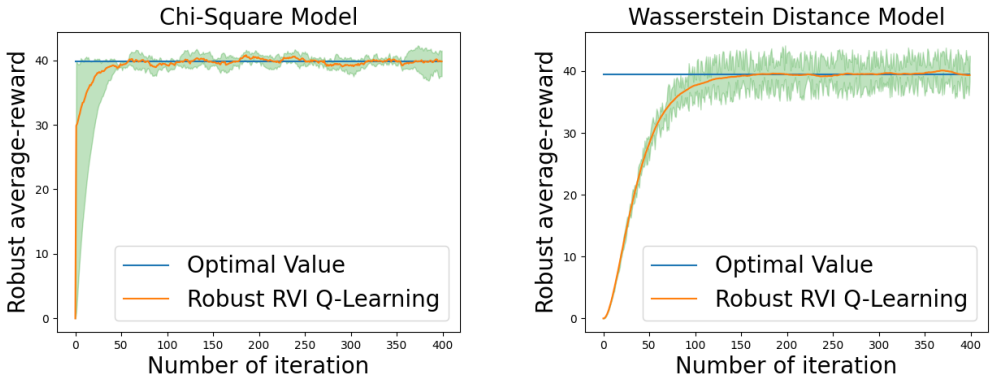


Figure 2: Robust RVI Q-Learning Algorithm.

6.2 Robustness of Our Robust Approaches

In this section, we aim to demonstrate the robustness of our approaches, including both model-based and model-free ones. We first compare our model-based approaches with the non-robust model-based ones under the Garnet problem (Archibald et al., 1995). Then we verify the robustness of our model-free robust RVI Q-learning under two problems, namely, the Recycling Robot and the inventory control problem.

6.2.1 Robustness of Model-Based Approaches

We study several commonly used uncertainty set models, including contamination model, Kullback-Lerbler (KL) divergence, and total-variation defined model. As can be observed from Algorithm 1, 2, and 3 for different uncertainty sets, the only difference lies in how the support function $\sigma_{\mathcal{P}_s^a}(V)$ is calculated. In the sequel, we discuss how to efficiently calculate the support function for various uncertainty sets.

We numerically compare our robust (relative) value iteration v.s. non-robust (relative) value iteration methods on different uncertainty sets. Our experiments are based on the same Garnet problem $\mathcal{G}(20, 40)$ considered in 6.1, with the same nominal transition kernel, reward functions, and uncertainty set structures. We run different algorithms, i.e., (robust) value iteration and (robust) relative value iteration, and obtain the greedy policies at each time step. Then, we use robust average-reward policy evaluation (Algorithm 1) to evaluate the robust average-reward of these policies. We plot the robust average-reward against the number of iterations.

Contamination model. Our experimental results under the contamination model are shown in Figure 3.

Total variation. Our experimental results under the total variation model are shown in Figure 4.

Kullback-Lerbler (KL) divergence. Our experimental results under the KL-divergence model are shown in Figure 5.

It can be seen that our robust methods can obtain policies that achieve higher worst-case rewards. Also, both our limit-based robust value iteration and our direct method of

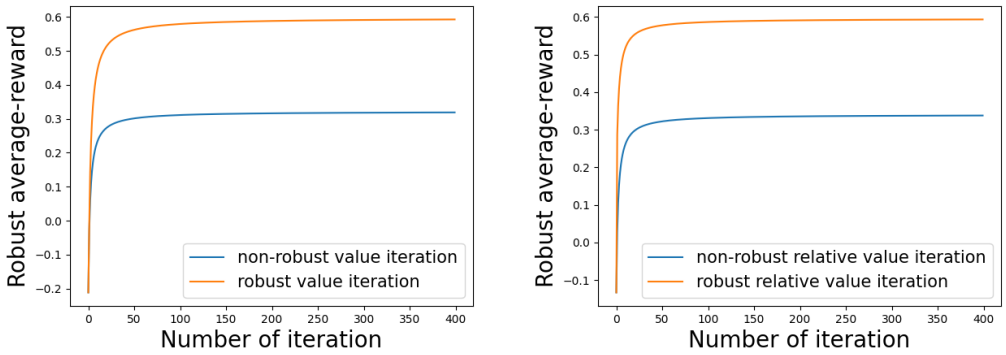


Figure 3: Comparison on contamination model with $R = 0.4$.

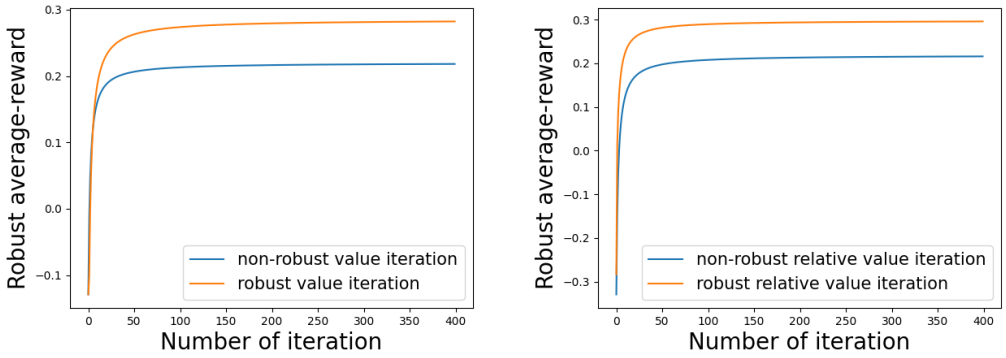


Figure 4: Comparison on total variation model with $R = 0.6$.

robust relative value iteration converge to the optimal robust policies, which validates our theoretical results.

6.2.2 Robustness of Model-Free Approach: 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: $\mathcal{S} = \{\text{low and high}\}$. The robot can either 1) search for empty cans; 2) remain stationary and wait for someone to bring it a can; or 3) go back to its home base to recharge, i.e., $\mathcal{A} = \{\text{wait, search, recharge}\}$. 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 humans. The reward of finding a can is set to be $+5$, the reward of finding nothing and running out of battery is -5, and $r = 0$ for waiting. 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 $\zeta = 0.4$ and implement our algorithms and

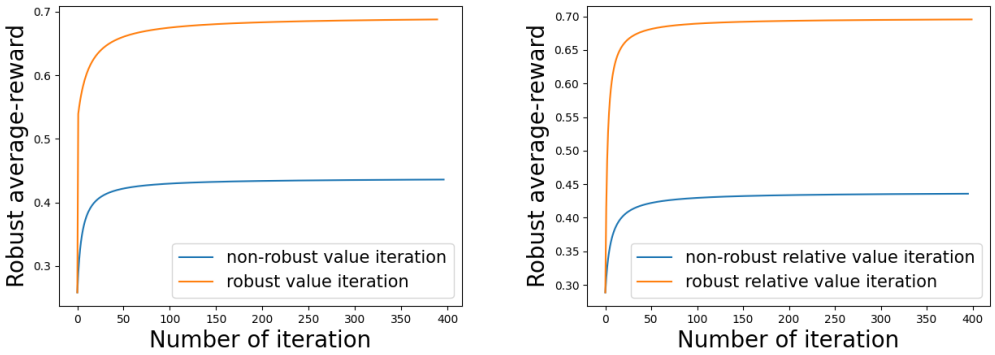


Figure 5: Comparison on KL-divergence model with $R = 0.8$.

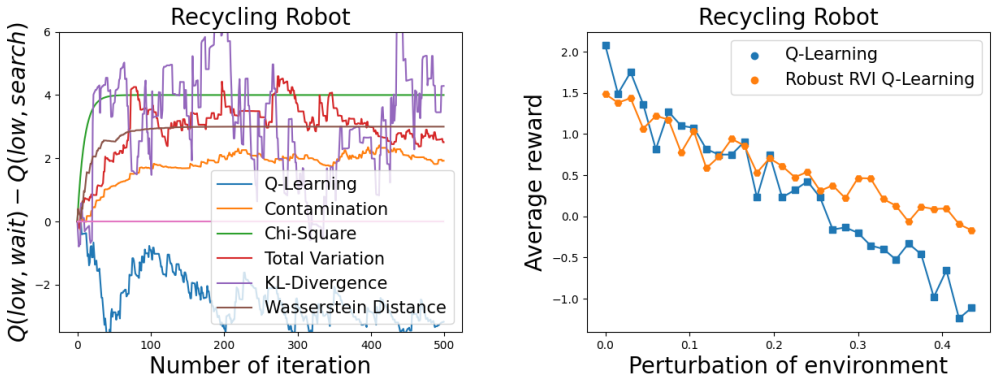
vanilla Q-learning under the nominal environment ($\alpha = \beta = 0.5$) with stepsize 0.01. To show the difference among the policies that the algorithms learned, we plot the difference of Q values at low battery level in Figure 6a. The results showed the average value of the difference between the Q-function among 10 independent experiments. In the low battery level, the robust algorithms find conservative policies that 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 calculate the exact robust average reward functions of the two policies over the testing uncertainty set $(0.5 - x, 0.5 + x)$ using the model-based approach Alg 1, and plot them in Figure 6b. It can be seen that when the perturbation is small, the true worst-case kernels (w.r.t. ζ 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.

6.2.3 Robustness of Model-Free Approach: 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



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

Figure 6: Recycling Robot.

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 $\zeta = 0.4$ and $\alpha_t = 0.01$, and implement our algorithms and vanilla Q-learning under the nominal environment where $D_t \sim \textbf{Uniform}(0, 16)$ 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

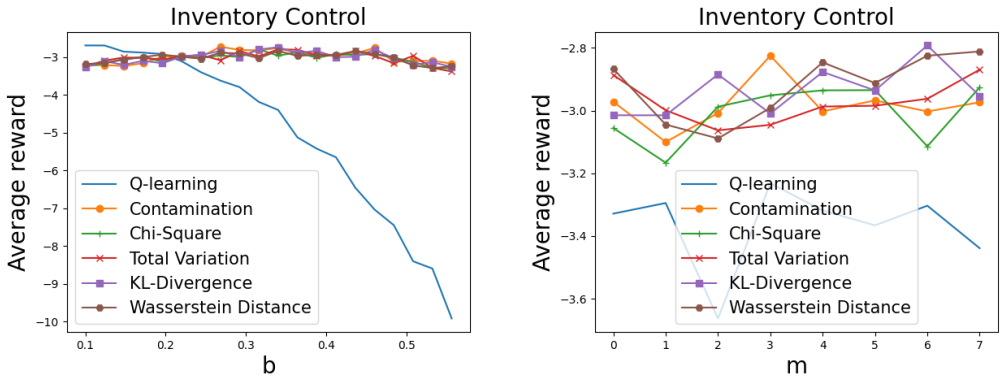
$$U_{(m,b)}(x) = \begin{cases} \frac{1}{|\mathcal{S}|} + b\frac{|\mathcal{S}|-2}{2|\mathcal{S}|}, & \text{if } x \in \{m, m+1\}, \\ \frac{1-b}{|\mathcal{S}|}, & \text{else.} \end{cases}$$

The results are plotted in Figure 7. We first fix $m = 0$ and plot the performance under different values of b in Figure 7a, then we fix $b = 0.25$ and plot the performance under different values of m in Figure 7b.

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.

7. Conclusion

In this paper, we investigated the problem of robust MDPs under the average-reward criterion. We first developed the fundamental characterization of their structures using two approaches: the limit approach and the direct one. We first established *uniform* convergence of the discounted value function to average-reward and showed the common properties



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

Figure 7: Inventory Control.

shared by robust MDPs under both reward criteria. We then designed a direct approach for robust average-reward MDPs, where we derived the robust Bellman equation for robust average-reward MDPs. Based on these results, we further designed two model-based algorithms, robust VI and robust RVI, with convergence and robustness guarantees. We then generalized such approaches to the model-free setting, where we constructed an unbiased estimator of the robust Bellman operator and proposed robust RVI TD and Q-learning algorithms, and further theoretically proved their convergence and optimality.

8. Acknowledgement

This work is supported by the National Science Foundation under Grants CCF-2007783, CCF-2106560, ECCS-2337375 (CAREER), and CCF-2106339.

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.

Appendix A. 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.

A.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.

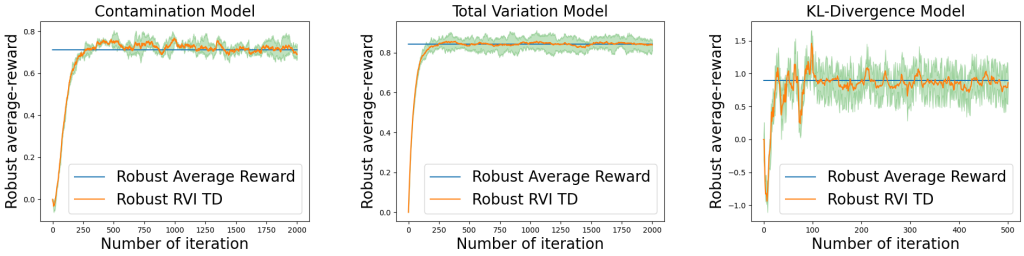


Figure 8: Robust RVI TD Algorithm under Garnet Problem.

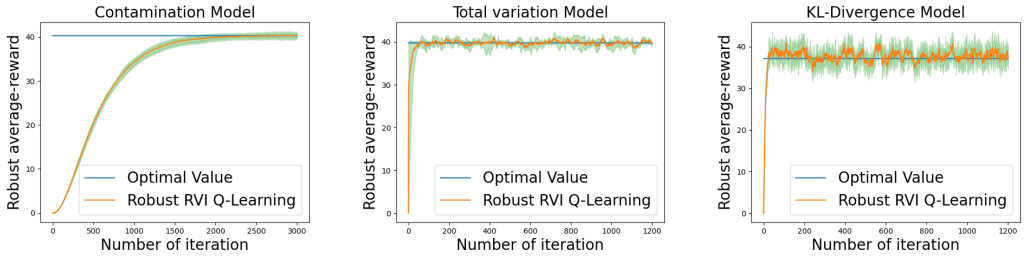


Figure 9: Robust RVI Q-Learning Algorithm under Garnet Problem.

A.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 $\zeta = 0.4$, $\alpha_n = 0.01$ and plot the (optimal) robust average-reward computed using model-based methods 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.

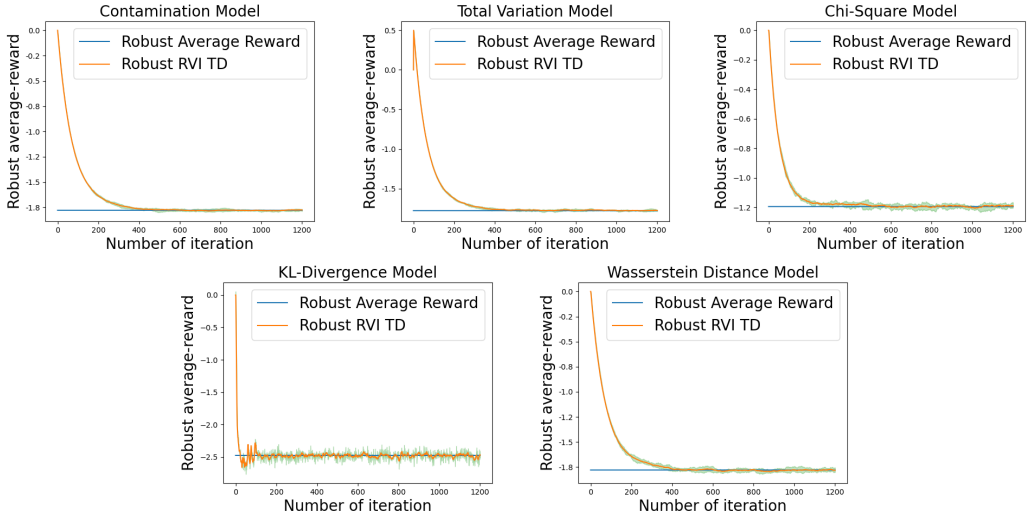


Figure 10: Robust RVI TD Algorithm under Frozen-Lake environment.

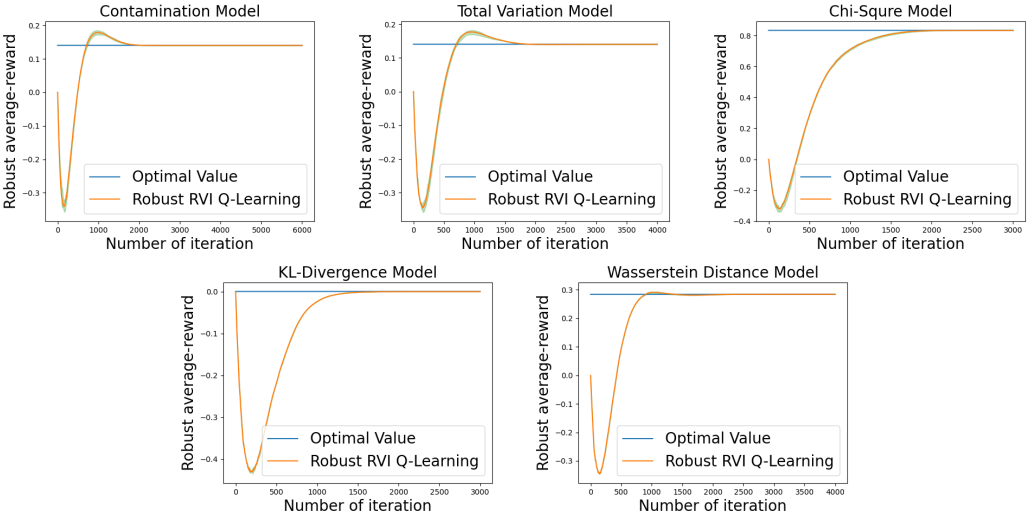


Figure 11: Robust RVI Q-learning Algorithm under Frozen-Lake environment.

A.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, 2022), 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 12, 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.

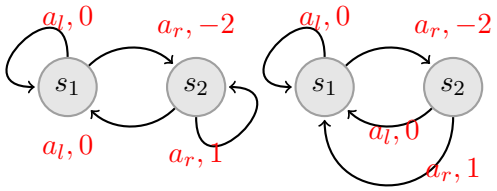
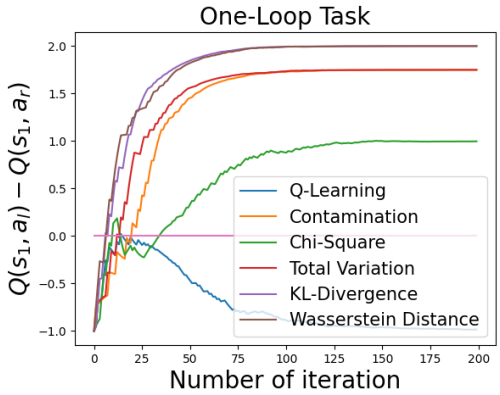


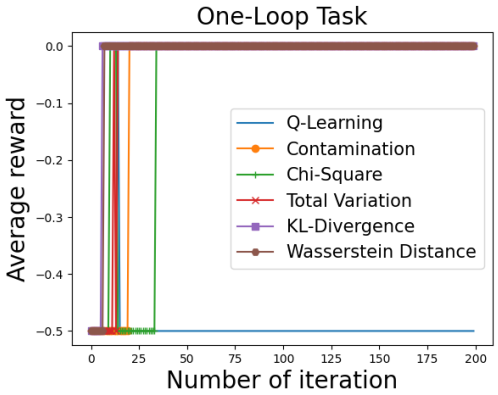
Figure 12: One-Loop Task.

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 13a. 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 12. We plot the average-reward of policies π_t under this perturbed environment. The results are shown in Figure 13b.



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



(b) Average-Reward under Testing MDP

Figure 13: 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.

Appendix B. Review of Robust Discounted MDPs

In this section, we provide a brief review on the existing methods and results for robust discounted MDPs.

B.1 Robust Policy Evaluation

We first consider the robust policy evaluation problem, where we aim to estimate the robust value function $V^{\pi}_{\mathcal{P},\gamma}$ for any policy π. It has been shown that the robust Bellman operator $\mathbf{T}_{\pi}$ is a γ-contraction, and the robust value function $V^{\pi}_{\mathcal{P},\gamma}$ is its unique fixed-point. Hence by recursively applying the robust Bellman operator, we can find the robust discounted value function (Nilim & El Ghaoui, 2004; Iyengar, 2005).

Algorithm 6 Policy evaluation for robust discounted MDPs

Input: π, V_0, T

for $t = 0, 1, ..., T-1$ **do**
  for all $s \in \mathcal{S}$ **do**
    $V_{t+1}(s) \leftarrow \mathbb{E}_{\pi}[r(s, A) + \gamma \sigma_{\mathcal{P}^A_s}(V_t)]$
  end for
end for
return V_T

B.2 Robust Optimal Control

Another important problem in robust MDP is finding the optimal policy that maximizes the robust discounted value function:

$$\pi^* = \arg\max_{\pi} V^{\pi}_{\mathcal{P},\gamma}. \tag{26}$$

A robust value iteration approach is developed in (Nilim & El Ghaoui, 2004; Iyengar, 2005) as follows.

Algorithm 7 Optimal Control for robust discounted MDPs

Input: V_0, T

for $t = 0, 1, ..., T-1$ **do**
  for all $s \in \mathcal{S}$ **do**
    $V_{t+1}(s) \leftarrow \max_a \left\{ r(s, a) + \gamma \sigma_{\mathcal{P}^a_s}(V_t) \right\}$
  end for
end for
$\pi^*(s) \leftarrow \arg\max_a \left\{ r(s, a) + \gamma \sigma_{\mathcal{P}^a_s}(V_T) \right\}, \forall s$
return π^*

Appendix C. Equivalence between Time-Varying and Stationary Models

We first provide an equivalence result between time-varying and stationary transition kernel models under stationary policies, which is an analog result to the one for robust discounted MDPs (Iyengar, 2005; Nilim & El Ghaoui, 2004). This result will be used in our following proofs.

Recall the definitions of the robust discounted value function and worst-case average-reward in (4),(5), the worst-case is taken w.r.t. $\kappa = (\mathsf{P}_0, \mathsf{P}_1 ...) \in \bigotimes_{t\geq 0} \mathcal{P}$, therefore, the transition kernel at each time step could be different. This model is referred to as the time-varying transition kernel model (as in (Iyengar, 2005; Nilim & El Ghaoui, 2004)). Another commonly used setting is that the transition kernels at different time steps are the same, which is referred to as the stationary model (Iyengar, 2005; Nilim & El Ghaoui, 2004). In this paper, we use the following notations to distinguish the two models. By $\mathbb{E}_{\mathsf{P}}[\cdot]$, we denote the expectation when the transition kernels at all time steps are the same, P, i.e., the stationary model. We also denote by $g^{\pi}_{\mathsf{P}}(s) \triangleq \lim_{n\to\infty} \mathbb{E}_{\mathsf{P},\pi}\left[\frac{1}{n}\sum_{t=0}^{n-1} r_t \big| S_0 = s\right]$ and $V^{\pi}_{\mathsf{P},\gamma}(s) \triangleq \mathbb{E}_{\mathsf{P},\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_t \big| S_0 = s\right]$ being the expected average-reward and expected discounted value function under the stationary model P. By $\mathbb{E}_{\kappa}[\cdot]$, we denote the expectation when the transition kernel at time t is P_t, i.e., the time-varying model.

For the discounted setting, it has been shown in (Nilim & El Ghaoui, 2004) that for a stationary policy π, any $\gamma \in [0,1)$, and any $s \in \mathcal{S}$,

$$\begin{aligned} V^{\pi}_{\mathcal{P},\gamma}(s) &= \min_{\kappa \in \bigotimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\pi,\kappa}\left[\sum_{t=0}^{\infty}\gamma^t r_t | S_0 = s\right] \\ &= \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\pi,\mathsf{P}}\left[\sum_{t=0}^{\infty}\gamma^t r_t | S_0 = s\right]. \end{aligned} \tag{27}$$

In the following theorem, we prove an analog of (27) for robust average-reward MDPs that if we consider stationary policies, then the robust average-reward problem with the time-varying model can be equivalently solved by a stationary model.

Specifically, we define the worst-case average-reward for the stationary transition kernel model as follows:

$$\min_{\mathsf{P}\in\mathcal{P}} \lim_{n\to\infty} \mathbb{E}_{\pi,\mathsf{P}}\left[\frac{1}{n}\sum_{t=0}^{n-1} r_t \big| S_0 = s\right]. \tag{28}$$

Recall the worst-case average-reward for the time-varying model in (5). We will show that for any stationary policy, (5) can be equivalently solved by solving (28).

Theorem 13. *Consider an arbitrary stationary policy π. Then, the worst-case average-reward under the time-varying model is the same as the one under the stationary model:*

$$\begin{aligned} g^{\pi}_{\mathcal{P}}(s) &\triangleq \min_{\kappa \in \bigotimes_{t\geq 0}\mathcal{P}} \lim_{n\to\infty} \mathbb{E}_{\kappa,\pi}\left[\frac{1}{n}\sum_{t=0}^{n-1} r_t | S_0 = s\right] \\ &= \min_{\mathsf{P}\in\mathcal{P}} \lim_{n\to\infty} \mathbb{E}_{\mathsf{P},\pi}\left[\frac{1}{n}\sum_{t=0}^{n-1} r_t \big| S_0 = s\right]. \end{aligned} \tag{29}$$

Similar result also holds for the robust relative value function:

$$V_{\mathcal{P}}^{\pi}(s) \triangleq \min_{\kappa \in \bigotimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\kappa,\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})|S_0 = s\right]$$
$$= \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\mathsf{P},\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})|S_0 = s\right]. \tag{30}$$

Proof. From the robust Bellman equation in Theorem 7 [5], we have that

$$V_{\mathcal{P}}^{\pi}(s) + g_{\mathcal{P}}^{\pi} = \sum_a \pi(a|s)\left(r(s,a) + \sigma_{\mathcal{P}_s^a}(V_{\mathcal{P}}^{\pi})\right). \tag{31}$$

Denote by $\arg\min_{p\in\mathcal{P}_s^a}(p)^\top V_{\mathcal{P}}^{\pi} \triangleq p_s^{a6}$, and denote by $\mathsf{P}^{\pi} \triangleq \{p_s^a : s \in \mathcal{S}, a \in \mathcal{A}\}$. It then follows that

$$\begin{aligned}
V_{\mathcal{P}}^{\pi}(s) &= \sum_a \pi(a|s)\left(r(s,a) - g_{\mathcal{P}}^{\pi} + \sigma_{\mathcal{P}_s^a}(V_{\mathcal{P}}^{\pi})\right)\\
&= \sum_a \pi(a|s)(r(s,a) - g_{\mathcal{P}}^{\pi}) + \sum_a \pi(a|s)\mathbb{E}_{\mathsf{P}^{\pi}}[V_{\mathcal{P}}^{\pi}(S_1)|S_0 = s, A_0 = a]\\
&= \sum_a \pi(a|s)(r(s,a) - g_{\mathcal{P}}^{\pi}) + \mathbb{E}_{\mathsf{P}^{\pi},\pi}[V_{\mathcal{P}}^{\pi}(S_1)|S_0 = s]\\
&= \sum_a \pi(a|s)(r(s,a) - g_{\mathcal{P}}^{\pi}) + \mathbb{E}_{\mathsf{P}^{\pi},\pi}\left[\sum_a \pi(a|S_1)(r(S_1,a) - g_{\mathcal{P}}^{\pi})|S_0 = s\right]\\
&\quad + \mathbb{E}_{\mathsf{P}^{\pi},\pi}\left[\sum_a \pi(a|S_1)\sigma_{\mathcal{P}_{S_1}^a}(V_{\mathcal{P}}^{\pi})|S_0 = s\right]\\
&= \sum_a \pi(a|s)(r(s,a) - g_{\mathcal{P}}^{\pi}) + \mathbb{E}_{\mathsf{P}^{\pi},\pi}\left[r_1 - g_{\mathcal{P}}^{\pi}|S_0 = s\right] + \mathbb{E}_{\mathsf{P}^{\pi},\pi}\left[\sigma_{\mathcal{P}_{S_1}^{A_1}}(V_{\mathcal{P}}^{\pi})|S_0 = s\right]\\
&= \sum_a \pi(a|s)(r(s,a) - g_{\mathcal{P}}^{\pi}) + \mathbb{E}_{\mathsf{P}^{\pi},\pi}\left[r_1 - g_{\mathcal{P}}^{\pi}\middle|S_0 = s\right] + \mathbb{E}_{\mathsf{P}^{\pi},\pi}\left[(p_{S_1}^{A_1})^\top V_{\mathcal{P}}^{\pi}|S_0 = s\right]\\
&= \mathbb{E}_{\mathsf{P}^{\pi},\pi}\left[r_0 - g_{\mathcal{P}}^{\pi} + r_1 - g_{\mathcal{P}}^{\pi}|S_0 = s\right] + \mathbb{E}_{\mathsf{P}^{\pi},\pi}[V_{\mathcal{P}}^{\pi}(S_2)|S_0 = s]\\
&\quad \cdots\cdots\\
&= \mathbb{E}_{\mathsf{P}^{\pi},\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})|s\right].
\end{aligned} \tag{32}$$

By the definition, the following always hold:

$$\min_{\kappa\in\bigotimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\kappa,\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})|S_0 = s\right] \leq \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\mathsf{P},\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})|S_0 = s\right]. \tag{33}$$

5. The proof of Theorem 7 is independent of Theorem 13 and does not rely on the results to be shown here.
6. We pick one arbitrarily if there are multiple minimizers.

This hence implies that a stationary transition kernel sequence $\kappa = (\mathsf{P}^\pi, \mathsf{P}^\pi, ...)$ is one of the worst-case transition kernels for $V^\pi_{\mathcal{P}}$. Therefore, (30) can be proved.

Consider the transition kernel P^π. We denote its non-robust average-reward and the non-robust relative value function by $g^\pi_{\mathsf{P}^\pi}$ and $V^\pi_{\mathsf{P}^\pi}$. By the non-robust Bellman equation (Sutton & Barto, 2018), we have that

$$V^\pi_{\mathsf{P}^\pi}(s) = \sum_a \pi(a|s)(r(s,a) - g^\pi_{\mathsf{P}^\pi}) + \mathbb{E}_{\mathsf{P}^\pi,\pi}[V^\pi_{\mathsf{P}^\pi}(S_1)|s]. \tag{34}$$

On the other hand, the robust Bellman equation shows that

$$V^\pi_{\mathcal{P}}(s) = V^\pi_{\mathsf{P}^\pi}(s) = \sum_a \pi(a|s)(r(s,a) - g^\pi_{\mathcal{P}}) + \mathbb{E}_{\mathsf{P}^\pi,\pi}[V^\pi_{\mathsf{P}^\pi}(S_1)|s]. \tag{35}$$

These two equations imply that $g^\pi_{\mathcal{P}} = g^\pi_{\mathsf{P}^\pi}$, and hence the stationary kernel $(\mathsf{P}^\pi, \mathsf{P}^\pi, ...)$ is also a worst-case kernel of robust average-reward in the time-varying setting. This proves (29). □

Appendix D. Limit Approach

D.1 Proof of Theorem 1

In the proof, unless otherwise specified, we denote by $\|v\|$ the l_∞ norm of a vector v, and for a matrix A, we denote by $\|A\|$ its matrix norm induced by l_∞ norm, i.e., $\|A\| = \sup_{x\in\mathbb{R}^d} \frac{\|Ax\|_\infty}{\|x\|_\infty}$.

Lemma 3. *[Theorem 8.2.3 in (Puterman, 1994)] For any* P, γ, π,

$$V^\pi_{\mathsf{P},\gamma} = \frac{1}{1-\gamma} g^\pi_{\mathsf{P}} + h^\pi_{\mathsf{P}} + f^\pi_{\mathsf{P}}(\gamma), \tag{36}$$

where $h^\pi_{\mathsf{P}} = H^\pi_{\mathsf{P}} r_\pi$, *and* $f^\pi_{\mathsf{P}}(\gamma) = \frac{1}{\gamma}\sum_{n=1}^\infty (-1)^n \left(\frac{1-\gamma}{\gamma}\right)^n (H^\pi_{\mathsf{P}})^{n+1} r_\pi$.

Following Proposition 8.4.6 in (Puterman, 1994), we can show the following lemma.

Lemma 4. H^π_{P} *is continuous on* $\Pi \times \mathcal{P}$. *If* Π *and* $\mathcal{P}$ *are compact,* $\|H^\pi_{\mathsf{P}}\|$ *is uniformly bounded on* $\Pi \times \mathcal{P}$, *i.e., there exists a constant* h, *such that* $\|H^\pi_{\mathsf{P}}\| \le h$ *for any* π, P.

For simplicity, denote by

$$\begin{aligned} S^\pi_\infty(\mathsf{P},\gamma) &\triangleq \frac{1}{\gamma}\sum_{n=1}^\infty (-1)^n \left(\frac{1-\gamma}{\gamma}\right)^n (H^\pi_{\mathsf{P}})^{n+1} r_\pi, \\ S^\pi_N(\mathsf{P},\gamma) &\triangleq \frac{1}{\gamma}\sum_{n=1}^N (-1)^n \left(\frac{1-\gamma}{\gamma}\right)^n (H^\pi_{\mathsf{P}})^{n+1} r_\pi. \end{aligned} \tag{37}$$

Clearly $S^\pi_\infty(\mathsf{P},\gamma) = f^\pi_{\mathsf{P}}(\gamma)$ and $\lim_{N\to\infty} S^\pi_N(\mathsf{P},\gamma) = S^\pi_\infty(\mathsf{P},\gamma)$ for any specific π, P.

Lemma 5. *There exists* $\delta \in (0,1)$, *such that*

$$\lim_{N\to\infty} S^\pi_N(\mathsf{P},\gamma) = S^\pi_\infty(\mathsf{P},\gamma) \tag{38}$$

uniformly on $\Pi \times \mathcal{P} \times [\delta, 1]$.

Proof. Note that $\|H_{\mathsf{P}}^{\pi}\| \leq h$, hence there exists δ, s.t.

$$\frac{1-\delta}{\delta}h \leq k < 1 \tag{39}$$

for some constant k. Then for any $\gamma \geq \delta$,

$$\frac{1-\gamma}{\gamma}h \leq \frac{1-\delta}{\delta}h \leq k. \tag{40}$$

Moreover, note that

$$\left\|\frac{1}{\gamma}(-1)^n\left(\frac{1-\gamma}{\gamma}\right)^n (H_{\mathsf{P}}^{\pi})^{n+1}r\right\| \leq \frac{1}{\gamma}\left(\frac{1-\gamma}{\gamma}\right)^n h^{n+1} \leq \frac{hk^n}{\delta} \triangleq M_n, \tag{41}$$

which is because $\|A+B\| \leq \|A\|+\|B\|$ for induced l_∞ norm, $\|Ax\| \leq \|A\|\|x\|$ and $\|r_\pi\|_\infty \leq 1$.

Note that

$$\sum_{n=1}^{\infty} M_n = \frac{h}{\delta}\frac{k}{1-k}, \tag{42}$$

hence by Weierstrass M-test (Rudin, 2022), $S_N^{\pi}(\mathsf{P},\gamma)$ uniformly converges to $S_\infty^{\pi}(\mathsf{P},\gamma)$ on $\Pi \times \mathcal{P} \times [\delta, 1]$. □

Lemma 6. *There exists a uniform constant L, such that*

$$\|S_N^{\pi}(\mathsf{P},\gamma_1) - S_N^{\pi}(\mathsf{P},\gamma_2)\| \leq L|\gamma_1 - \gamma_2|, \tag{43}$$

for any N, π, $\mathsf{P}, \gamma_1, \gamma_2 \in [\delta, 1]$.

Proof. We first show that $\gamma S_N^{\pi}(\mathsf{P},\gamma) = \sum_{n=1}^{N}(-1)^n\left(\frac{1-\gamma}{\gamma}\right)^n (H_{\mathsf{P}}^{\pi})^{n+1}r_\pi \triangleq T_N^{\pi}(\mathsf{P},\gamma)$ is uniformly Lipschitz w.r.t. the l_∞ norm, i.e.,

$$\|T_N^{\pi}(\mathsf{P},\gamma_1) - T_N^{\pi}(\mathsf{P},\gamma_2)\| \leq l|\gamma_1 - \gamma_2|, \tag{44}$$

for any N, π, $\mathsf{P}, \gamma_1, \gamma_2 \in [\delta, 1]$ and some constant l.

Clearly, it can be shown by verifying $\nabla T_N^{\pi}(\mathsf{P},\gamma)$ is uniformly bounded for any π, N, P or γ.

First, it can be shown that

$$\nabla T_N^{\pi}(\mathsf{P},\gamma) = \sum_{n=1}^{N}(-1)^n n\left(\frac{1-\gamma}{\gamma}\right)^{n-1}\frac{-1}{\gamma^2}(H_{\mathsf{P}}^{\pi})^{n+1}r_\pi, \tag{45}$$

and moreover

$$\|\nabla T_N^{\pi}(\mathsf{P},\gamma)\| \leq \sum_{n=1}^{N} n\left(\frac{1-\gamma}{\gamma}\right)^{n-1}\frac{1}{\gamma^2}h^{n+1} \triangleq l_N(\gamma). \tag{46}$$

Note that

$$h\frac{1-\gamma}{\gamma}l_N(\gamma) = \sum_{n=1}^{N} n\left(\frac{1-\gamma}{\gamma}\right)^n \frac{1}{\gamma^2}h^{n+2}, \tag{47}$$

then, we can show that

$$\begin{aligned}
&\left(1-h\frac{1-\gamma}{\gamma}\right) l_N(\gamma) \\
&= \sum_{n=1}^{N} n\left(\frac{1-\gamma}{\gamma}\right)^{n-1} \frac{1}{\gamma^2}h^{n+1} - \sum_{n=1}^{N} n\left(\frac{1-\gamma}{\gamma}\right)^n \frac{1}{\gamma^2}h^{n+2} \\
&= \frac{1}{\gamma^2}h^2 - N\left(\frac{1-\gamma}{\gamma}\right)^N \frac{1}{\gamma^2}h^{N+2} + \sum_{n=2}^{N}\left(\frac{1-\gamma}{\gamma}\right)^{n-1} \frac{1}{\gamma^2}h^{n+1} \\
&\leq \frac{1}{\gamma^2}h^2 + \frac{h^2}{\gamma^2}\frac{1-\gamma}{\gamma}h\frac{1}{1-\frac{1-\gamma}{\gamma}h} \\
&= \frac{h^2}{\gamma^2} + \frac{h^2}{\gamma^2}\frac{1-\gamma}{\gamma}h\frac{1}{1-\frac{1-\gamma}{\gamma}h}.
\end{aligned} \tag{48}$$

Hence, we have that

$$\begin{aligned}
\|\nabla T_N^\pi(\mathsf{P},\gamma)\| \leq l_N(\gamma) &\leq \frac{1}{1-h\frac{1-\gamma}{\gamma}}\left(\frac{h^2}{\gamma^2} + \frac{h^2}{\gamma^2}\frac{1-\gamma}{\gamma}h\frac{1}{1-\frac{1-\gamma}{\gamma}h}\right) \\
&\leq \frac{1}{1-k}\left(\frac{h^2}{\delta^2} + \frac{h^2}{\delta^2}\frac{k}{1-k}\right),
\end{aligned} \tag{49}$$

which implies a uniform bound on $\|\nabla T_N^\pi(\mathsf{P},\gamma)\|$.

Now, we have that

$$\begin{aligned}
&|S_N^\pi(\mathsf{P},\gamma_1) - S_N^\pi(\mathsf{P},\gamma_2)| \\
&\leq \frac{|\gamma_2-\gamma_1|}{\gamma_1\gamma_2}\|T_N^\pi(\mathsf{P},\gamma_1)\| + \frac{\|T_N^\pi(\mathsf{P},\gamma_1) - T_N^\pi(\mathsf{P},\gamma_2)\|}{\gamma_2}.
\end{aligned} \tag{50}$$

To show $\|T_N^\pi(\mathsf{P},\gamma)\|$ is uniformly bounded, we have that

$$\begin{aligned}
\|T_N^\pi(\mathsf{P},\gamma)\| &\leq \sum_{n=1}^{N}\left\|\left(\frac{1-\gamma}{\gamma}\right)^n (H_{\mathsf{P}}^\pi)^{n+1} r\right\| \\
&\leq \sum_{n=1}^{N}\left(\frac{1-\gamma}{\gamma}\right)^n h^{n+1} \\
&\leq \sum_{n=1}^{N} k^n h \\
&\leq h\frac{k}{1-k}.
\end{aligned} \tag{51}$$

Then, it follows that

$$\begin{aligned}
&\|S_N^\pi(\mathsf{P},\gamma_1) - S_N^\pi(\mathsf{P},\gamma_2)\| \\
&= \left\| \frac{\gamma_2-\gamma_1}{\gamma_1\gamma_2} T_N^\pi(\mathsf{P},\gamma_1) + \frac{T_N^\pi(\mathsf{P},\gamma_1) - T_N^\pi(\mathsf{P},\gamma_2)}{\gamma_2} \right\| \\
&\leq \left(\frac{1}{\delta^2} h \frac{k}{1-k} + \frac{1}{\delta} \frac{1}{1-k} \left(\frac{h^2}{\delta^2} + \frac{h^2}{\delta^2} \frac{k}{1-k} \right) \right) |\gamma_1 - \gamma_2| \\
&\triangleq L|\gamma_1 - \gamma_2|, \qquad (52)
\end{aligned}$$

where $L = \left(\frac{1}{\delta^2} h \frac{k}{1-k} + \frac{1}{\delta} \frac{1}{1-k} \left(\frac{h^2}{\delta^2} + \frac{h^2}{\delta^2} \frac{k}{1-k} \right) \right)$ is a universal constant that does not depend on N, P, π or γ. □

Lemma 7. *$S_\infty^\pi(\mathsf{P},\gamma)$ uniformly converges as $\gamma \to 1$ on $\Pi \times \mathcal{P}$. Also, $S_\infty^\pi(\mathsf{P},\gamma)$ is L-Lipschitz for any $\gamma > \delta$: for any π, P and any $\gamma_1, \gamma_2 \in (\delta, 1]$.*

$$\|S_\infty^\pi(\mathsf{P},\gamma_1) - S_\infty^\pi(\mathsf{P},\gamma_2)\| \leq L|\gamma_1 - \gamma_2|. \qquad (53)$$

Proof. From Lemma 5, for any ϵ, there exists N_ϵ, such that for any $n \geq N_\epsilon$, π, P, $\gamma > \delta$,

$$\|S_\infty^\pi(\mathsf{P},\gamma) - S_n^\pi(\mathsf{P},\gamma)\| < \epsilon. \qquad (54)$$

Thus for any $\gamma_1, \gamma_2 \in (\delta, 1]$,

$$\begin{aligned}
&\|S_\infty^\pi(\mathsf{P},\gamma_1) - S_\infty^\pi(\mathsf{P},\gamma_2)\| \\
&\leq \|S_\infty^\pi(\mathsf{P},\gamma_1) - S_n^\pi(\mathsf{P},\gamma_1)\| + \|S_n^\pi(\mathsf{P},\gamma_1) - S_n^\pi(\mathsf{P},\gamma_2)\| + \|S_n^\pi(\mathsf{P},\gamma_2) - S_\infty^\pi(\mathsf{P},\gamma_2)\| \\
&\leq 2\epsilon + \|S_n^\pi(\mathsf{P},\gamma_1) - S_n^\pi(\mathsf{P},\gamma_2)\| \\
&\leq 2\epsilon + L|\gamma_1 - \gamma_2|, \qquad (55)
\end{aligned}$$

where the last step is from Lemma 6.

Thus, for any ϵ, there exists $\omega = \max\{\delta, 1-\epsilon\}$, such that for any $\gamma_1, \gamma_2 > \omega$,

$$\|S_\infty^\pi(\mathsf{P},\gamma_1) - S_\infty^\pi(\mathsf{P},\gamma_2)\| < (2+L)\epsilon, \qquad (56)$$

and hence by Cauchy's criterion we conclude that $S_\infty^\pi(\mathsf{P},\gamma)$ converges uniformly on $\Pi \times \mathcal{P}$.

On the other hand, since (55) holds for any ϵ, it implies that

$$\|S_\infty^\pi(\mathsf{P},\gamma_1) - S_\infty^\pi(\mathsf{P},\gamma_2)\| \leq L|\gamma_1 - \gamma_2|, \qquad (57)$$

which completes the proof. □

We now prove Theorem 1. For any P, π, we have that

$$V_{\mathsf{P},\gamma}^\pi = \frac{1}{1-\gamma} g_{\mathsf{P}}^\pi + h_{\mathsf{P}}^\pi + f_{\mathsf{P}}^\pi(\gamma). \qquad (58)$$

It then follows that

$$(1-\gamma) V_{\mathsf{P},\gamma}^\pi = g_{\mathsf{P}}^\pi + (1-\gamma) h_{\mathsf{P}}^\pi + (1-\gamma) f_{\mathsf{P}}^\pi(\gamma). \qquad (59)$$

Clearly $(1-\gamma)h^\pi_{\mathsf{P}} \to 0$ uniformly on $\Pi \times \mathcal{P}$ because $\|h^\pi_{\mathsf{P}}\| = \|H^\pi_{\mathsf{P}} r_\pi\| \leq h$ is uniformly bounded. Then,

$$\begin{aligned}
&\|(1-\gamma_1)f^\pi_{\mathsf{P}}(\gamma_1) - (1-\gamma_2)f^\pi_{\mathsf{P}}(\gamma_2)\| \\
&\leq \|(1-\gamma_1)f^\pi_{\mathsf{P}}(\gamma_1) - (1-\gamma_1)f^\pi_{\mathsf{P}}(\gamma_2)\| + \|(1-\gamma_1)f^\pi_{\mathsf{P}}(\gamma_2) - (1-\gamma_2)f^\pi_{\mathsf{P}}(\gamma_2)\| \\
&\leq (1-\gamma_1)L|\gamma_1-\gamma_2| + \|f^\pi_{\mathsf{P}}(\gamma_2)\||\gamma_1-\gamma_2|. \qquad (60)
\end{aligned}$$

For any $\pi, \mathsf{P}, \gamma > \delta$,

$$\begin{aligned}
\|f^\pi_{\mathsf{P}}(\gamma)\| &= \left\| \frac{1}{\gamma} \sum_{n=1}^{\infty} (-1)^n \left(\frac{1-\gamma}{\gamma}\right)^n (H^\pi_{\mathsf{P}})^{n+1} r_\pi \right\| \\
&\leq \left| \frac{1}{\gamma} \sum_{n=1}^{\infty} \left(\frac{1-\gamma}{\gamma}\right)^n h^{n+1} \right| \\
&\leq \frac{h}{\delta} \frac{1-\gamma}{\gamma} h \frac{1}{1-\frac{1-\gamma}{\gamma}h} \\
&\leq \frac{h}{\delta} \frac{k}{1-k} \\
&\triangleq c_f. \qquad (61)
\end{aligned}$$

Hence, $(1-\gamma)f^\pi_{\mathsf{P}}(\gamma) \to 0$ uniformly on $\Pi \times \mathcal{P}$ due to the fact that $\|f^\pi_{\mathsf{P}}(\gamma)\|$ is uniformly bounded for any $\pi, \gamma > \delta, \mathsf{P}$.

Then we have that $\lim_{\gamma\to 1}(1-\gamma)V^\pi_{\mathsf{P},\gamma} = g^\pi_{\mathsf{P}}$ uniformly on $\mathcal{P} \times \Pi$. This completes the proof of Theorem 1.

D.2 Proof of Theorem 2

We first show a lemma which allows us to interchange the order of lim and max.

Lemma 8. *If a function $f(x,y)$ converges uniformly to $F(x)$ on $\mathcal{X}$ as $y \to y_0$, then*

$$\max_x \lim_{y\to y_0} f(x,y) = \lim_{y\to y_0} \max_x f(x,y). \qquad (62)$$

Proof. For each $f(x,y)$, denote by $\arg\max_x f(x,y) = x_y$, and hence $f(x_y,y) \geq f(x,y)$ for any x, y. Also denote by $\arg\max_x F(x) = x'$. Now because $f(x,y)$ uniformly converges to $F(x)$, then for any ϵ, there exists δ', such that $\forall |y-y_0| < \delta'$,

$$|f(x,y) - F(x)| \leq \epsilon \qquad (63)$$

for any x. Now consider $|f(x_y,y) - F(x')|$ for $|y-y_0| < \delta'$. If $f(x_y,y) - F(x') > 0$, then

$$|f(x_y,y) - F(x')| = f(x_y,y) - F(x') = f(x_y,y) - F(x_y) + F(x_y) - F(x') \leq \epsilon; \qquad (64)$$

On the other hand if $f(x_y,y) - F(x') < 0$, then

$$|f(x_y,y) - F(x')| = F(x') - f(x_y,y) = F(x') - f(x',y) + f(x',y) - f(x_y,y) \leq \epsilon. \qquad (65)$$

Hence, we showed that for any ϵ, there exists δ', such that $\forall |y - y_0| < \delta'$,

$$|f(x_y, y) - F(x')| = |\max_x f(x, y) - \max_x F(x)| \leq \epsilon, \tag{66}$$

and hence

$$\lim_{y \to y_0} \max_x f(x, y) = \max_x F(x) = \max_x \lim_{y \to y_0} f(x, y), \tag{67}$$

and this completes the proof. □

Then, we show that the robust discounted value function converges uniformly to the robust average-reward as the discounted factor approaches 1.

Theorem 14 (Restatement of Theorem 2)**.** *The robust discounted value function converges uniformly to the robust average-reward on* Π*:*

$$\lim_{\gamma \to 1} (1 - \gamma) V^\pi_{\mathcal{P},\gamma} = g^\pi_{\mathcal{P}}. \tag{68}$$

Proof. Due to Theorem 13, for any stationary policy π, $g^\pi_{\mathcal{P}}(s) = \min_{\mathsf{P} \in \mathcal{P}} g^\pi_{\mathsf{P}}(s)$ under the stationary model. Hence from the uniform convergence in Theorem 1, we first show the following:

$$\begin{aligned} g^\pi_{\mathcal{P}} &= \min_{\mathsf{P} \in \mathcal{P}} g^\pi_{\mathsf{P}} \\ &= \min_{\mathsf{P} \in \mathcal{P}} \lim_{\gamma \to 1} (1 - \gamma) V^\pi_{\mathsf{P},\gamma} \\ &\overset{(a)}{=} \lim_{\gamma \to 1} \min_{\mathsf{P} \in \mathcal{P}} (1 - \gamma) V^\pi_{\mathsf{P},\gamma} \\ &= \lim_{\gamma \to 1} (1 - \gamma) V^\pi_{\mathcal{P},\gamma}, \end{aligned} \tag{69}$$

where (a) is because Lemma 8. Moreover, note that $\lim_{\gamma \to 1} (1 - \gamma) V^\pi_{\mathsf{P},\gamma} = g^\pi_{\mathsf{P}}$ uniformly on $\Pi \times \mathcal{P}$, hence the convergence in (69) is also uniform on Π. Thus, we complete the proof. □

D.3 Proof of Theorem 5

Theorem 15 (Restatement of Theorem 5)**.** *V_T generated by Algorithm 1 converges to the robust average-reward $g^\pi_{\mathcal{P}}$ as $T \to \infty$.*

Proof. From discounted robust Bellman equation (Nilim & El Ghaoui, 2004), it can be shown that

$$(1 - \gamma_t) V^\pi_{\mathcal{P},\gamma_t} = (1 - \gamma_t) \sum_a \pi(a|s) (r(s, a) + \gamma_t \sigma_{\mathcal{P}^a_s}(V^\pi_{\mathcal{P},\gamma_t})). \tag{70}$$

Then we can show that for any $s \in \mathcal{S}$,

$$
\begin{aligned}
&|V_{t+1}(s) - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}(s)| \\
&= |V_{t+1}(s) - (1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}(s) + (1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}(s) - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}(s)| \qquad (71)\\
&\leq |(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}(s) - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}(s)| + |V_{t+1}(s) - (1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}(s)| \\
&= |(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}(s) - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}(s)| \\
&\quad + \left|\sum_a \pi(a|s)\Big((1-\gamma_t)r(s,a) + \gamma_t \sigma_{\mathcal{P}^a_s}(V_t) - ((1-\gamma_t)r(s,a) + \gamma_t \sigma_{\mathcal{P}^a_s}((1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}))\Big)\right| \\
&= |(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}(s) - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}(s)| + \left|\sum_a \pi(a|s)\Big(\gamma_t \sigma_{\mathcal{P}^a_s}(V_t) - \gamma_t \sigma_{\mathcal{P}^a_s}((1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t})\Big)\right| \\
&= |(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}(s) - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}(s)| + \gamma_t\left|\sum_a \pi(a|s)\Big(\sigma_{\mathcal{P}^a_s}(V_t) - \sigma_{\mathcal{P}^a_s}((1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t})\Big)\right|. \qquad (72)
\end{aligned}
$$

If we denote by $\Delta_t \triangleq \|V_t - (1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}\|_\infty$, then

$$
\begin{aligned}
&\Delta_{t+1} \qquad (73)\\
&\leq \|(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t} - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}\|_\infty + \gamma_t \max_s \left\{\sum_a \pi(a|s)\left|\sigma_{\mathcal{P}^a_s}(V_t) - \sigma_{\mathcal{P}^a_s}((1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t})\right|\right\}.
\end{aligned}
$$

It can be easily verified that $\sigma_{\mathcal{P}^a_s}(V)$ is a 1-Lipschitz function, thus the second term in (73) can be further bounded as

$$
\begin{aligned}
&\sum_a \pi(a|s)\left|\sigma_{\mathcal{P}^a_s}(V_t) - \sigma_{\mathcal{P}^a_s}((1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t})\right| \\
&\leq \sum_a \pi(a|s)\|V_t - (1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}\|_\infty \\
&= \|V_t - (1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}\|_\infty, \qquad (74)
\end{aligned}
$$

and hence

$$
\Delta_{t+1} \leq \|(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t} - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}\|_\infty + \gamma_t \Delta_t. \qquad (75)
$$

Recall that

$$
(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t} = (1-\gamma_t)\min_{\mathsf{P}} V^{\pi}_{\mathsf{P},\gamma_t}. \qquad (76)
$$

Let $s^*_t \triangleq \arg\max_s |(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}(s) - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}(s)|$. Then it follows that

$$
\|(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t} - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}\|_\infty = |(1-\gamma_t)V^{\pi}_{\mathcal{P},\gamma_t}(s^*_t) - (1-\gamma_{t+1})V^{\pi}_{\mathcal{P},\gamma_{t+1}}(s^*_t)|. \qquad (77)
$$

Note that from (Nilim & El Ghaoui, 2004; Iyengar, 2005), for any stationary policy π, there exists a stationary model P such that $V^{\pi}_{\mathcal{P},\gamma}(s) = \mathbb{E}_{\mathsf{P},\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_t | S_0 = s\right] \triangleq V^{\pi}_{\mathsf{P},\gamma}$.

Hence in the following, for each γ_t, we denote the worst-case transition kernel of $V^\pi_{\mathcal{P},\gamma_t}$ by P_t.

If $(1-\gamma_t)V^\pi_{\mathcal{P},\gamma_t}(s^*_t) \geq (1-\gamma_{t+1})V^\pi_{\mathcal{P},\gamma_{t+1}}(s^*_t)$, then

$$
\begin{aligned}
&|(1-\gamma_t)V^\pi_{\mathcal{P},\gamma_t}(s^*_t) - (1-\gamma_{t+1})V^\pi_{\mathcal{P},\gamma_{t+1}}(s^*_t)| \\
&= \min_{\mathsf{P}}(1-\gamma_t)V^\pi_{\mathsf{P},\gamma_t}(s^*_t) - \min_{\mathsf{P}}(1-\gamma_{t+1})V^\pi_{\mathsf{P},\gamma_{t+1}}(s^*_t) \\
&= (1-\gamma_t)V^\pi_{\mathsf{P}_t,\gamma_t}(s^*_t) - (1-\gamma_{t+1})V^\pi_{\mathsf{P}_{t+1},\gamma_{t+1}}(s^*_t) \\
&= (1-\gamma_t)V^\pi_{\mathsf{P}_t,\gamma_t}(s^*_t) - (1-\gamma_t)V^\pi_{\mathsf{P}_{t+1},\gamma_t}(s^*_t) + (1-\gamma_t)V^\pi_{\mathsf{P}_{t+1},\gamma_t}(s^*_t) - (1-\gamma_{t+1})V^\pi_{\mathsf{P}_{t+1},\gamma_{t+1}}(s^*_t) \\
&\overset{(a)}{\leq} (1-\gamma_t)V^\pi_{\mathsf{P}_{t+1},\gamma_t}(s^*_t) - (1-\gamma_{t+1})V^\pi_{\mathsf{P}_{t+1},\gamma_{t+1}}(s^*_t) \\
&\leq \|(1-\gamma_t)V^\pi_{\mathsf{P}_{t+1},\gamma_t} - (1-\gamma_{t+1})V^\pi_{\mathsf{P}_{t+1},\gamma_{t+1}}\|_\infty, \qquad (78)
\end{aligned}
$$

where (a) is due to $(1-\gamma_t)V^\pi_{\mathsf{P}_t,\gamma_t}(s^*_t) = \min_{\mathsf{P}}(1-\gamma_t)V^\pi_{\mathsf{P},\gamma_t}(s^*_t) \leq (1-\gamma_t)V^\pi_{\mathsf{P}_{t+1},\gamma_t}(s^*_t)$.

Now, according to Lemma 3,

$$
\begin{aligned}
(1-\gamma_t)V^\pi_{\mathsf{P}_{t+1},\gamma_t} &= g^\pi_{\mathsf{P}_{t+1}} + (1-\gamma_t)h^\pi_{\mathsf{P}_{t+1}} + (1-\gamma_t)f^\pi_{\mathsf{P}_{t+1}}(\gamma_t), \qquad (79) \\
(1-\gamma_{t+1})V^\pi_{\mathsf{P}_{t+1},\gamma_{t+1}} &= g^\pi_{\mathsf{P}_{t+1}} + (1-\gamma_{t+1})h^\pi_{\mathsf{P}_{t+1}} + (1-\gamma_{t+1})f^\pi_{\mathsf{P}_{t+1}}(\gamma_{t+1}). \qquad (80)
\end{aligned}
$$

Hence, for any $\gamma_t > \delta$, (78) can be further bounded as

$$
\begin{aligned}
&\|(1-\gamma_t)V^\pi_{\mathsf{P}_{t+1},\gamma_t} - (1-\gamma_{t+1})V^\pi_{\mathsf{P}_{t+1},\gamma_{t+1}}\|_\infty \\
&= \|(\gamma_{t+1}-\gamma_t)h^\pi_{\mathsf{P}_{t+1}} + (1-\gamma_t)f^\pi_{\mathsf{P}_{t+1}}(\gamma_t) - (1-\gamma_{t+1})f^\pi_{\mathsf{P}_{t+1}}(\gamma_{t+1})\|_\infty \\
&\leq (\gamma_{t+1}-\gamma_t)\|h^\pi_{\mathsf{P}_{t+1}}\|_\infty + \|f^\pi_{\mathsf{P}_{t+1}}(\gamma_t) - f^\pi_{\mathsf{P}_{t+1}}(\gamma_{t+1})\|_\infty + \|\gamma_{t+1}f^\pi_{\mathsf{P}_{t+1}}(\gamma_{t+1}) - \gamma_t f^\pi_{\mathsf{P}_{t+1}}(\gamma_t)\|_\infty \\
&\overset{(a)}{\leq} h(\gamma_{t+1}-\gamma_t) + L(\gamma_{t+1}-\gamma_t) + \|\gamma_{t+1}f^\pi_{\mathsf{P}_{t+1}}(\gamma_{t+1}) - \gamma_t f^\pi_{\mathsf{P}_{t+1}}(\gamma_t)\|_\infty \\
&\leq h(\gamma_{t+1}-\gamma_t) + L(\gamma_{t+1}-\gamma_t) + \|\gamma_{t+1}f^\pi_{\mathsf{P}_{t+1}}(\gamma_{t+1}) - \gamma_{t+1}f^\pi_{\mathsf{P}_{t+1}}(\gamma_t)\|_\infty \\
&\quad + \|\gamma_{t+1}f^\pi_{\mathsf{P}_{t+1}}(\gamma_t) - \gamma_t f^\pi_{\mathsf{P}_{t+1}}(\gamma_t)\|_\infty \\
&\leq h(\gamma_{t+1}-\gamma_t) + L(\gamma_{t+1}-\gamma_t) + \gamma_{t+1}\|f^\pi_{\mathsf{P}_{t+1}}(\gamma_{t+1}) - f^\pi_{\mathsf{P}_{t+1}}(\gamma_t)\|_\infty + \|f^\pi_{\mathsf{P}_{t+1}}(\gamma_t)\|_\infty(\gamma_{t+1}-\gamma_t) \\
&\overset{(b)}{\leq} (h + L + \gamma_{t+1}L + \sup_{\pi,\mathsf{P},\gamma}\|f^\pi_{\mathsf{P}}(\gamma)\|_\infty)(\gamma_{t+1}-\gamma_t) \\
&\leq K(\gamma_{t+1}-\gamma_t), \qquad (81)
\end{aligned}
$$

where (a) is from Lemma 7 for any $\gamma_t > \delta$, c_f is defined in (61) and $K \triangleq h + 2L + c_f$ is a uniform constant; And (b) is from Lemma 7.

Similarly, the inequality also holds for the case when $(1-\gamma_t)V^\pi_{\mathcal{P},\gamma_t}(s^*_t) \leq (1-\gamma_{t+1})V^\pi_{\mathcal{P},\gamma_{t+1}}(s^*_t)$. Thus we have that for any t such that $\gamma_t > \delta$,

$$\Delta_{t+1} \leq K(\gamma_{t+1}-\gamma_t) + \gamma_t\Delta_t, \qquad (82)$$

where K is a uniform constant.

Following Lemma 8 from (Tewari & Bartlett, 2007), we have that $\Delta_t \to 0$. Note that

$$\|V_t - g^\pi_{\mathcal{P}}\|_\infty \leq \|V_t - (1-\gamma_t)V^\pi_{\mathcal{P},\gamma_t}\|_\infty + \|(1-\gamma_t)V^\pi_{\mathcal{P},\gamma_t} - g^\pi_{\mathcal{P}}\|_\infty = \Delta_t + \|(1-\gamma_t)V^\pi_{\mathcal{P},\gamma_t} - g^\pi_{\mathcal{P}}\|_\infty. \qquad (83)$$

Together with Theorem 2, we further have that

$$\lim_{t\to\infty} \|V_t - g^\pi_{\mathcal{P}}\|_\infty = 0, \tag{84}$$

which completes the proof.

□

D.4 Proof of Theorem 6

Note that the optimal robust average-reward is defined as

$$g^*_{\mathcal{P}}(s) \triangleq \max_\pi g^\pi_{\mathcal{P}}(s). \tag{85}$$

We further define

$$V^*_{\mathcal{P},\gamma}(s) \triangleq \max_\pi V^\pi_{\mathcal{P},\gamma}(s). \tag{86}$$

Theorem 16 (Restatement of Theorem 6)**.** *V_T generated by Algorithm 2 converges to the optimal robust average-reward $g^*_{\mathcal{P}}$ as $T \to \infty$.*

Proof. Firstly, from the uniform convergence in Theorem 2, it can be shown that

$$\lim_{t\to\infty} (1-\gamma_t) V^*_{\mathcal{P},\gamma_t} = g^*_{\mathcal{P}}. \tag{87}$$

We then show that for any $s \in \mathcal{S}$,

$$\begin{aligned}
&|V_{t+1}(s) - (1-\gamma_{t+1})V^*_{\mathcal{P},\gamma_{t+1}}(s)| \\
&\le |V_{t+1}(s) - (1-\gamma_t)V^*_{\mathcal{P},\gamma_t}(s)| + |(1-\gamma_t)V^*_{\mathcal{P},\gamma_t}(s) - (1-\gamma_{t+1})V^*_{\mathcal{P},\gamma_{t+1}}(s)| \\
&\overset{(a)}{=} |(1-\gamma_t)V^*_{\mathcal{P},\gamma_t}(s) - (1-\gamma_{t+1})V^*_{\mathcal{P},\gamma_{t+1}}(s)| \\
&\quad + \left| \max_a \left((1-\gamma_t) r(s,a) + \gamma_t \sigma_{\mathcal{P}^a_s}(V_t) \right) - \max_a \left(((1-\gamma_t) r(s,a) + \gamma_t \sigma_{\mathcal{P}^a_s}((1-\gamma_t)V^*_{\mathcal{P},\gamma_t})) \right) \right| \\
&\le |(1-\gamma_t)V^*_{\mathcal{P},\gamma_t}(s) - (1-\gamma_{t+1})V^*_{\mathcal{P},\gamma_{t+1}}(s)| \\
&\quad + \max_a \left| (1-\gamma_t) r(s,a) + \gamma_t \sigma_{\mathcal{P}^a_s}(V_t) - ((1-\gamma_t) r(s,a) + \gamma_t \sigma_{\mathcal{P}^a_s}((1-\gamma_t)V^*_{\mathcal{P},\gamma_t})) \right|,
\end{aligned} \tag{88}$$

where (a) is because the optimal robust Bellman equation, and the last inequality is from the fact that $|\max_x f(x) - \max_x g(x)| \le \max_x |f(x) - g(x)|$.

Hence (88) can be further bounded as

$$\begin{aligned}
&|V_{t+1}(s) - (1-\gamma_{t+1})V^*_{\mathcal{P},\gamma_{t+1}}(s)| \\
&\le |(1-\gamma_t)V^*_{\mathcal{P},\gamma_t}(s) - (1-\gamma_{t+1})V^*_{\mathcal{P},\gamma_{t+1}}(s)| + \gamma_t \max_a \left| \sigma_{\mathcal{P}^a_s}(V_t) - \sigma_{\mathcal{P}^a_s}((1-\gamma_t)V^*_{\mathcal{P},\gamma_t}) \right|.
\end{aligned} \tag{89}$$

If we denote by $\Delta_t \triangleq \|V_t - (1-\gamma_t)V^*_{\mathcal{P},\gamma_t}\|_\infty$, then

$$\Delta_{t+1} \le \|(1-\gamma_t)V^*_{\mathcal{P},\gamma_t} - (1-\gamma_{t+1})V^*_{\mathcal{P},\gamma_{t+1}}\|_\infty + \gamma_t \max_{s.a} \left| \sigma_{\mathcal{P}^a_s}(V_t) - \sigma_{\mathcal{P}^a_s}((1-\gamma_t)V^*_{\mathcal{P},\gamma_t}) \right|. \tag{90}$$

Since the support function $\sigma_{\mathcal{P}_s^a}(V)$ is 1-Lipschitz, then it can be shown that for any s, a,

$$\left|\sigma_{\mathcal{P}_s^a}(V_t) - \sigma_{\mathcal{P}_s^a}((1-\gamma_t)V^*_{\mathcal{P},\gamma_t})\right| \leq \|V_t - (1-\gamma_t)V^*_{\mathcal{P},\gamma_t}\|_\infty. \tag{91}$$

Hence

$$\Delta_{t+1} \leq \|(1-\gamma_t)V^*_{\mathcal{P},\gamma_t} - (1-\gamma_{t+1})V^*_{\mathcal{P},\gamma_{t+1}}\|_\infty + \gamma_t\Delta_t. \tag{92}$$

Similar to (81) in Theorem 5, we can show that

$$\|(1-\gamma_t)V^*_{\mathcal{P},\gamma_t} - (1-\gamma_{t+1})V^*_{\mathcal{P},\gamma_{t+1}}\|_\infty \leq K|\gamma_t - \gamma_{t+1}|, \tag{93}$$

and similar to Lemma 8 from (Tewari & Bartlett, 2007),

$$\lim_{t\to\infty} \Delta_t = 0. \tag{94}$$

Moreover, note that

$$\begin{aligned}
&\|V_t - g^*_{\mathcal{P}}\|_\infty \\
&\leq \|V_t - (1-\gamma_t)V^*_{\mathcal{P},\gamma_t}\|_\infty + \|(1-\gamma_t)V^*_{\mathcal{P},\gamma_t} - g^*_{\mathcal{P}}\|_\infty \\
&= \Delta_t + \|(1-\gamma_t)V^*_{\mathcal{P},\gamma_t} - g^*_{\mathcal{P}}\|_\infty,
\end{aligned} \tag{95}$$

which together with (87) implies that

$$\|V_t - g^*_{\mathcal{P}}\|_\infty \to 0, \tag{96}$$

and hence it completes the proof.

□

Lemma 9. *There exists a deterministic optimal policy, i.e., $\exists \pi^* \in \Pi_D$, s.t. $g^{\pi^*}_{\mathcal{P}} = g^*_{\mathcal{P}} = \max_{\pi\in\Pi} g^\pi_{\mathcal{P}}$.*

D.5 Proof of Lemma 9

Lemma 10. *(Restatement of Lemma 9). There exists a deterministic optimal policy, i.e., $\exists \pi^* \in \Pi_D$, s.t. $g^{\pi^*}_{\mathcal{P}} = g^*_{\mathcal{P}} = \max_{\pi\in\Pi} g^\pi_{\mathcal{P}}$.*

Proof. Assume that there is no deterministic optimal robust policy, i.e., there exists a strictly random policy $\pi_r \in \Pi$, such that for any deterministic policy $\pi \in \Pi_D$,

$$g^{\pi_r}_{\mathcal{P}} > g^\pi_{\mathcal{P}}. \tag{97}$$

According to Theorem 2, we have that

$$\lim_{\gamma\to 1}(1-\gamma)V^{\pi_r}_{\mathcal{P},\gamma} = g^{\pi_r}_{\mathcal{P}}, \tag{98}$$

$$\lim_{\gamma\to 1}(1-\gamma)V^{\pi}_{\mathcal{P},\gamma} = g^{\pi}_{\mathcal{P}}, \forall \pi \in \Pi_D. \tag{99}$$

Since there are only finite number of deterministic policies, there exists $\delta < 1$, such that for any $\gamma > \delta$,

$$V_{\mathcal{P},\gamma}^{\pi_r} > V_{\mathcal{P},\gamma}^{\pi}, \forall \pi \in \Pi_D. \tag{100}$$

This implies that for $\gamma > \delta$, the random policy π_r is better than all the deterministic policies, i.e.,

$$V_{\mathcal{P},\gamma}^{\pi_r} > \max_{\pi \in \Pi_D} V_{\mathcal{P},\gamma}^{\pi}. \tag{101}$$

However, Theorem 3.1 of (Iyengar, 2005) implies that there exists deterministic optimal robust policy, i.e.,

$$\max_{\pi \in \Pi_D} V_{\mathcal{P},\gamma}^{\pi} = \max_{\pi \in \Pi} V_{\mathcal{P},\gamma}^{\pi} \geq V_{\mathcal{P},\gamma}^{\pi_r}, \tag{102}$$

which contradicts to (101). Hence it implies that there exists a deterministic optimal robust policy, and completes the proof. □

D.6 Proof of Theorem 4

Theorem 17 (Restatement of Theorem 4)**.** *There exists $0 < \delta < 1$, such that for any $\gamma > \delta$, a deterministic optimal robust policy for robust discounted value function $V_{\mathcal{P},\gamma}^{*}$ is also an optimal policy for robust average-reward, i.e.,*

$$V_{\mathcal{P},\gamma}^{\pi^*} = V_{\mathcal{P},\gamma}^{*}. \tag{103}$$

Moreover, when $\arg\max_{\pi \in \Pi^D} g_{\mathcal{P}}^{\pi}$ is a singleton, there exists a unique Blackwell optimal policy.

Proof. According to Lemma 9, there exists $\pi^* \in \Pi^D$ such that

$$g_{\mathcal{P}}^{*} = g_{\mathcal{P}}^{\pi^*}. \tag{104}$$

Assume the robust average-reward of all deterministic policies are sorted in a descending order:

$$g_{\mathcal{P}}^{*} = g_{\mathcal{P}}^{\pi_1^*} = g_{\mathcal{P}}^{\pi_2^*} = ... = g_{\mathcal{P}}^{\pi_m^*} > g_{\mathcal{P}}^{\pi_1} \geq ... \geq g_{\mathcal{P}}^{\pi_n} \tag{105}$$

for all $\pi_i^*, \pi_i \in \Pi^D$, and we define $\Pi^* = \{\pi_i^* : i = 1, ..., m\}$. Denote by $d = g_{\mathcal{P}}^{\pi_i^*} - g_{\mathcal{P}}^{\pi_1}$.

From Theorem 2, we know that for any $\pi \in \Pi^D$,

$$\lim_{\gamma \to 1} (1-\gamma) V_{\mathcal{P},\gamma}^{\pi} = g_{\mathcal{P}}^{\pi}. \tag{106}$$

Because the set Π^D is finite, for any $\epsilon < \frac{d}{2}$, there exists $\delta' < 1$, such that for any $\gamma > \delta'$, π_i^* and π_j,

$$|(1-\gamma) V_{\mathcal{P},\gamma}^{\pi_i^*} - g_{\mathcal{P}}^{*}| < \epsilon, \tag{107}$$

$$|(1-\gamma) V_{\mathcal{P},\gamma}^{\pi_j} - g_{\mathcal{P}}^{\pi_j}| < \epsilon. \tag{108}$$

It hence implies that

$$(1-\gamma)V_{\mathcal{P},\gamma}^{\pi_i^*} \geq (d-2\epsilon) + (1-\gamma)V_{\mathcal{P},\gamma}^{\pi_j} > (1-\gamma)V_{\mathcal{P},\gamma}^{\pi_j}, \tag{109}$$

and

$$V_{\mathcal{P},\gamma}^{\pi_i^*} > V_{\mathcal{P},\gamma}^{\pi_j}. \tag{110}$$

Note that from Theorem 3.1 in (Iyengar, 2005), i.e., $\max_{\pi\in\Pi^D} V_{\mathcal{P},\gamma}^{\pi} = V_{\mathcal{P},\gamma}^{*}$, we have that for any γ, there exists a deterministic policy $\pi \in \Pi^D$, such that $V_{\mathcal{P},\gamma}^{*} = V_{\mathcal{P},\gamma}^{\pi}$. Together with (110), it implies that all the possible optimal robust polices of $V_{\mathcal{P},\gamma}^{\pi}$ belong to $\{\pi_1^*, ...\pi_m^*\}$, i.e., the set Π^*. Hence, there exists $\pi_j^* \in \Pi^*$, such that

$$V_{\mathcal{P},\gamma}^{\pi_j^*} = \max_{\pi\in\Pi^D} V_{\mathcal{P},\gamma}^{\pi} = V_{\mathcal{P},\gamma}^{*}. \tag{111}$$

For the second part, when the optimal robust policy of robust average-reward is unique, i.e., $\Pi^* = \{\pi^*\}$. Then from the results above, there exists δ', such that for any $\gamma > \delta'$, $V_{\mathcal{P},\gamma}^{\pi^*} > V_{\mathcal{P},\gamma}^{\pi}$ for any $\pi^* \neq \pi \in \Pi^D$, and hence π^* is the optimal policy for discounted robust MDPs, which is the unique Blackwell optimal policy.

□

Appendix E. Direct Approach

Recall that

$$V_{\mathcal{P}}^{\pi}(s) \triangleq \min_{\kappa\in\bigotimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\kappa,\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})\big|S_0 = s\right], \tag{112}$$

where

$$g_{\mathcal{P}}^{\pi} = \min_{\kappa\in\bigotimes_{t\geq 0}\mathcal{P}} \lim_{n\to\infty} \mathbb{E}_{\kappa,\pi}\left[\frac{1}{n}\sum_{t=0}^{n-1} r_t | S_0 = s\right]. \tag{113}$$

We first show that the robust relative function is always finite.

Lemma 11. *For any π, $V_{\mathcal{P}}^{\pi}$ is finite.*

Proof. According to Theorem 13, $V_{\mathcal{P}}^{\pi} = \min_{\mathsf{P}\in\mathcal{P}} V_{\mathsf{P}}^{\pi} = \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\mathsf{P},\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})\right]$. Note that $V_{\mathcal{P}}^{\pi}$ can be rewritten as

$$\begin{aligned}
V_{\mathcal{P}}^{\pi} &= \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\mathsf{P},\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})\right] \\
&= \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\mathsf{P},\pi}\left[\lim_{n\to\infty}\sum_{t=0}^{n}(r_t - g_{\mathcal{P}}^{\pi})\right] \\
&= \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\mathsf{P},\pi}\left[\lim_{n\to\infty}\sum_{t=0}^{n}(r_t - g_{\mathsf{P}}^{\pi} + g_{\mathsf{P}}^{\pi} - g_{\mathcal{P}}^{\pi})\right] \\
&= \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\mathsf{P},\pi}\left[\lim_{n\to\infty}(R_n - ng_{\mathsf{P}}^{\pi} + ng_{\mathsf{P}}^{\pi} - ng_{\mathcal{P}}^{\pi})\right],
\end{aligned} \tag{114}$$

where $R_n = \sum_{t=0}^{n} r_t$. Note that for any $\mathsf{P} \in \mathcal{P}$ and n, $ng_{\mathsf{P}}^{\pi} \geq ng_{\mathcal{P}}^{\pi}$, hence

$$\lim_{n\to\infty}(R_n - ng_{\mathsf{P}}^{\pi} + ng_{\mathsf{P}}^{\pi} - ng_{\mathcal{P}}^{\pi}) \geq \lim_{n\to\infty}(R_n - ng_{\mathsf{P}}^{\pi}), \tag{115}$$

and thus the lower bound of $V_{\mathcal{P}}^{\pi}$ can be derived as follows,

$$\begin{aligned} V_{\mathcal{P}}^{\pi} &\geq \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\mathsf{P},\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathsf{P}}^{\pi})\right] \\ &= \min_{\mathsf{P}\in\mathcal{P}} V_{\mathsf{P}}^{\pi} \\ &= \min_{\mathsf{P}\in\mathcal{P}} H_{\mathsf{P}}^{\pi} r_{\pi}. \end{aligned} \tag{116}$$

which is finite due to the fact that H_{P}^{π} is continuous on the compact set $\mathcal{P}$.

From Theorem 13, we denote the stationary worst-case transition kernel of $g_{\mathcal{P}}^{\pi}$ by P_g. Then the upper bound of $V_{\mathcal{P}}^{\pi}$ can be bounded by noting that

$$\begin{aligned} V_{\mathcal{P}}^{\pi} &= \min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\mathsf{P},\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathsf{P}_g}^{\pi})\right] \\ &\leq \mathbb{E}_{\mathsf{P}_g,\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathsf{P}_g}^{\pi})\right] \\ &= V_{\mathsf{P}_g}^{\pi}, \end{aligned} \tag{117}$$

which is also finite and P_g denotes the worst-case transition kernel of $g_{\mathcal{P}}^{\pi}$. Hence we show that $V_{\mathcal{P}}^{\pi}$ is finite for any π and hence complete the proof. □

After showing that the robust relative value function is well-defined, we show the following robust Bellman equation for average-reward robust MDPs.

Theorem 18 (Restatement of Theorem 7)**.** *For any s and π, $(V_{\mathcal{P}}^{\pi}, g_{\mathcal{P}}^{\pi})$ is a solution to the following robust Bellman equation:*

$$V(s) + g = \sum_a \pi(a|s)\left(r(s,a) + \sigma_{\mathcal{P}_s^a}(V)\right). \tag{118}$$

And if (g, V) is a solution to the robust Bellman equation

$$V(s) = \sum_a \pi(a|s)(r(s,a) - g + \sigma_{\mathcal{P}_s^a}(V)), \forall s, \tag{119}$$

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

Proof. We first show the first part. From the definition,

$$V_{\mathcal{P}}^{\pi}(s) = \min_{\kappa\in\bigotimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\kappa,\pi}\left[\sum_{t=0}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})\middle| S_0 = s\right], \tag{120}$$

hence

$$
\begin{aligned}
V_{\mathcal{P}}^{\pi}(s) &= \min_{\kappa \in \bigotimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\kappa,\pi}\Big[\sum_{t=0}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})\big|S_0 = s\Big] \\
&= \min_{\kappa \in \bigotimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\kappa,\pi}\Big[(r_0 - g_{\mathcal{P}}^{\pi}) + \sum_{t=1}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})\big|S_0 = s\Big] \\
&= \min_{\kappa \in \bigotimes_{t\geq 0}\mathcal{P}} \left\{\sum_a \pi(a|s) r(s,a) - g_{\mathcal{P}}^{\pi} + \mathbb{E}_{\kappa,\pi}\Big[\sum_{t=1}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})\big|S_0 = s\Big]\right\} \\
&= \sum_a \pi(a|s)\left(r(s,a) - g_{\mathcal{P}}^{\pi}\right) + \min_{\kappa \in \bigotimes_{t\geq 0}\mathcal{P}} \left\{\sum_{a,s'} \pi(a|s)\mathsf{P}_{s,s'}^{a}\mathbb{E}_{\kappa,\pi}\Big[\sum_{t=1}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})|S_1 = s'\Big]\right\} \\
&= \sum_a \pi(a|s)\left(r(s,a) - g_{\mathcal{P}}^{\pi}\right) \\
&\quad + \min_{\mathsf{P}_0 \in \mathcal{P}} \min_{\kappa = (\mathsf{P}_1,\ldots) \in \bigotimes_{t\geq 1}\mathcal{P}} \left\{\sum_{a,s'} \pi(a|s)(\mathsf{P}_0)_{s,s'}^{a}\mathbb{E}_{\kappa,\pi}\Big[\sum_{t=1}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})|S_1 = s'\Big]\right\} \\
&= \sum_a \pi(a|s)\left(r(s,a) - g_{\mathcal{P}}^{\pi}\right) \\
&\quad + \min_{\mathsf{P}_0 \in \mathcal{P}} \left\{\sum_{a,s'} \pi(a|s)(\mathsf{P}_0)_{s,s'}^{a} \min_{\kappa = (\mathsf{P}_1,\ldots) \in \bigotimes_{t\geq 1}\mathcal{P}} \left\{\mathbb{E}_{\kappa,\pi}\Big[\sum_{t=1}^{\infty}(r_t - g_{\mathcal{P}}^{\pi})|S_1 = s'\Big]\right\}\right\} \\
&= \sum_a \pi(a|s)\left(r(s,a) - g_{\mathcal{P}}^{\pi}\right) + \sum_a \pi(a|s) \sum_{s'} \min_{p_{s,s'}^{a} \in \mathcal{P}_s^a} p_{s,s'}^{a} V_{\mathcal{P}}^{\pi}(s') \\
&= \sum_a \pi(a|s)\left(r(s,a) - g_{\mathcal{P}}^{\pi}\right) + \sum_a \pi(a|s)\sigma_{\mathcal{P}_s^a}\left(V_{\mathcal{P}}^{\pi}\right) \\
&= \sum_a \pi(a|s)\left(r(s,a) - g_{\mathcal{P}}^{\pi} + \sigma_{\mathcal{P}_s^a}(V_{\mathcal{P}}^{\pi})\right). \qquad (121)
\end{aligned}
$$

This hence completes the proof.

We then show the second part. 1). The robust Bellman equation in (119) can be rewritten as

$$
g + V(s) - r_{\pi}(s) = \sigma_{\mathcal{P}_s}(V), \forall s \in \mathcal{S}. \qquad (122)
$$

From the definition, it follows that

$$
\sigma_{\mathcal{P}_s}(V) = \sum_a \pi(a|s) \min_{\mathsf{P}_s^a \in \mathcal{P}_s^a} \mathsf{P}_s^a V. \qquad (123)
$$

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

$$
g + V(s) - r_{\pi}(s) - \sum_a \pi(a|s)\mathsf{P}_s^a V \leq 0, \forall s. \qquad (124)
$$

It can be further rewritten in matrix form as:

$$ge \leq r_\pi + (\mathsf{P}^\pi - I)V, \tag{125}$$

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

$$\sum_a \pi(a|s)\mathsf{P}_s^a. \tag{126}$$

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

$$\begin{aligned} \mathsf{P}^\pi ge = ge &\leq \mathsf{P}^\pi r_\pi + \mathsf{P}^\pi(\mathsf{P}^\pi - I)V, \\ ge &\leq (\mathsf{P}^\pi)^2 r_\pi + (\mathsf{P}^\pi)^2(\mathsf{P}^\pi - I)V, \\ &... \\ ge &\leq (\mathsf{P}^\pi)^{n-1} r_\pi + (\mathsf{P}^\pi)^{n-1}(\mathsf{P}^\pi - I)V. \end{aligned} \tag{127}$$

Now, by summing up all these inequalities in (125) and (127), we have that

$$nge \leq \sum_{i=0}^{n-1}(\mathsf{P}^\pi)^i r_\pi + ((\mathsf{P}^\pi)^n - I)V, \tag{128}$$

and hence,

$$ge \leq \frac{\sum_{i=0}^{n-1}(\mathsf{P}^\pi)^i r_\pi}{n} + \frac{((\mathsf{P}^\pi)^n - I)V}{n}. \tag{129}$$

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

$$\begin{aligned} ge &\leq \lim_{n\to\infty} \frac{\sum_{i=0}^{n-1}(\mathsf{P}^\pi)^i r_\pi}{n} + \lim_{n\to\infty} \frac{((\mathsf{P}^\pi)^n - I)V}{n} \\ &= g_{\mathsf{P}}^\pi e, \end{aligned} \tag{130}$$

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

$$ge = r_\pi - V + \mathsf{P}_V^\pi V. \tag{131}$$

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

$$xe = r_\pi - Y + \mathsf{P}_V^\pi Y. \tag{132}$$

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

$$g = g_{\mathsf{P}_V}^\pi, \tag{133}$$

$$V = V_{\mathsf{P}_V}^\pi + ce, \text{ for some } c \in \mathbb{R}. \tag{134}$$

However, note that

$$g^\pi_{\mathsf{P}_V} = g \le g^\pi_{\mathcal{P}} = \min_{\mathsf{P}\in\mathcal{P}} g^\pi_{\mathsf{P}} \le g^\pi_{\mathsf{P}_V}, \tag{135}$$

thus,

$$g^\pi_{\mathsf{P}_V} = g = g^\pi_{\mathcal{P}}. \tag{136}$$

2). From (136),

$$g^\pi_{\mathsf{P}_V} = g^\pi_{\mathcal{P}}\,. \tag{137}$$

It then follows from the definition of Ω^π_g that $\mathsf{P}_V \in \Omega^\pi_g$.

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

$$xe = r_\pi - Y + \mathsf{P}^\pi_V Y, \tag{138}$$

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

Theorem 19. *[Restatement of Theorem 8] If (g, Q) is a solution to the optimal robust Bellman equation*

$$Q(s,a) = r(s,a) - g + \sigma_{\mathcal{P}^a_s}(V_Q), \forall s, a, \tag{139}$$

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

Proof. In this proof, for two vectors $v, w \in \mathbb{R}^n$, $v \ge w$ denotes that $v(s) \ge w(s)$ entry-wise.

Taking the maximum on both sides of (139) w.r.t. a, we have that

$$\max_a Q(s,a) = \max_a \{r(s,a) - g + \sigma_{\mathcal{P}^a_s}(V_Q)\}, \forall s \in \mathcal{S}. \tag{140}$$

This is equivalent to

$$V_Q(s) = \max_a \{r(s,a) - g + \sigma_{\mathcal{P}^a_s}(V_Q)\}, \forall s \in \mathcal{S}, \tag{141}$$

and hence (g, V_Q) is a solution to (141). We hence only need to show the conclusion for any solution (g, V) to (141).

Let $B(g,V)(s) \triangleq \max_a \left\{r(s,a) - g + \sigma_{\mathcal{P}^a_s}(V) - V(s)\right\}$. Since (g, V) is a solution to (14), hence for any $a \in \mathcal{A}$ and any $s \in \mathcal{S}$,

$$r(s,a) - g + \sigma_{\mathcal{P}^a_s}(V) - V(s) \le 0, \tag{142}$$

from which it follows that for any policy π,

$$g(s) \ge r_\pi(s) + \sum_a \pi(a|s)\sigma_{\mathcal{P}^a_s}(V) - V(s) \triangleq r_\pi(s) + \sum_a \pi(a|s)(p^a_s)^\top V - V(s), \tag{143}$$

where $r_\pi(s) \triangleq \sum_a \pi(a|s) r(s,a)$, $p^a_s \triangleq \arg\min_{p\in\mathcal{P}^a_s} p^\top V$, and $\mathsf{P}_V = \{p^a_s : s \in \mathcal{S}, a \in \mathcal{A}\}$. We also denotes the state transition matrix induced by π and P_V by P^π_V.

Using these notations, and rewrite (143), we have that

$$g\mathbf{1} \geq r_\pi + (\mathsf{P}_V^\pi - I)V. \tag{144}$$

Since the inequality in (144) holds entry-wise, all entries of P_V^π are positive, then by multiplying both sides of (144) by P_V^π, we have that

$$g\mathbf{1} = g\mathsf{P}_V^\pi\mathbf{1} \geq \mathsf{P}_V^\pi r_\pi + \mathsf{P}_V^\pi(\mathsf{P}_V^\pi - I)V. \tag{145}$$

Multiplying the both sides of (145) by P_V^π, and repeatedly doing that, we have that

$$g\mathbf{1} \geq (\mathsf{P}_V^\pi)^2 r_\pi + (\mathsf{P}_V^\pi)^2(\mathsf{P}_V^\pi - I)V, \tag{146}$$

$$\vdots \qquad\qquad \vdots \tag{147}$$

$$g\mathbf{1} \geq (\mathsf{P}_V^\pi)^{n-1} r_\pi + (\mathsf{P}_V^\pi)^{n-1}(\mathsf{P}_V^\pi - I)V. \tag{148}$$

Summing up these inequalities from (144) to (148), we have that

$$ng\mathbf{1} \geq (I + \mathsf{P}_V^\pi + ... + (\mathsf{P}_V^\pi)^{n-1})r_\pi + (I + \mathsf{P}_V^\pi + ... + (\mathsf{P}_V^\pi)^{n-1})(\mathsf{P}_V^\pi - I)V, \tag{149}$$

and from which, it follows that

$$\begin{aligned} g\mathbf{1} &\geq \frac{1}{n}(I + \mathsf{P}_V^\pi + ... + (\mathsf{P}_V^\pi)^{n-1})r_\pi + \frac{1}{n}(I + \mathsf{P}_V^\pi + ... + (\mathsf{P}_V^\pi)^{n-1})(\mathsf{P}_V^\pi - I)V \\ &= \frac{1}{n}(I + \mathsf{P}_V^\pi + ... + (\mathsf{P}_V^\pi)^{n-1})r_\pi + \frac{1}{n}((\mathsf{P}_V^\pi)^n - I)V. \end{aligned} \tag{150}$$

It can be easily verified that $\lim_{n\to\infty} \frac{1}{n}((\mathsf{P}_V^\pi)^n - I)V = 0$, and hence it implies that

$$\begin{aligned} g\mathbf{1} &\geq \lim_{n\to\infty} \frac{1}{n}(I + \mathsf{P}_V^\pi + ... + (\mathsf{P}_V^\pi)^{n-1})r_\pi \\ &= \lim_{n\to\infty} \frac{1}{n}\mathbb{E}_{\mathsf{P}_V^\pi,\pi}\left[\sum_{t=0}^{n} r_t\right] \\ &= g^\pi_{\mathsf{P}_V^\pi}\mathbf{1} \\ &\geq g^\pi_{\mathcal{P}}\mathbf{1}. \end{aligned} \tag{151}$$

Since (151) holds for any policy π, it follows that $g \geq g^*_{\mathcal{P}}$. On the other hand, since $B(g, V) = 0$, there exists a policy τ such that

$$g\mathbf{1} = r_\tau + (\mathsf{P}_V^\tau - I)V, \tag{152}$$

where $r_\tau, \mathsf{P}_V^\tau$ are similarly defined as for π. From Theorem 13, there exists a stationary transition kernel $\mathsf{P}^\tau_{\text{ave}}$ such that $g^\tau_{\mathcal{P}} = g^\tau_{\mathsf{P}^\tau_{\text{ave}}}$. We denote the state transition matrix induced by τ and $\mathsf{P}^\tau_{\text{ave}}$ by P^τ. Then because P_V^τ is the worst-case transition of V, it follows that

$$\mathsf{P}_V^\tau V \leq \mathsf{P}^\tau V. \tag{153}$$

Thus

$$g\mathbf{1} \leq r_\tau + (\mathsf{P}^\tau - I)V. \tag{154}$$

Similarly, we have that

$$g\mathbf{1} \leq (\mathsf{P}^\tau)^{j-1} r_\tau + (\mathsf{P}^\tau)^{j-1}(\mathsf{P}^\tau - I)V, \tag{155}$$

for $j = 2, ..., n$. Summing these inequalities together we have that

$$\begin{aligned} ng\mathbf{1} &\leq (I + \mathsf{P}^\tau + ... + (\mathsf{P}^\tau)^{n-1}) r_\tau + (I + \mathsf{P}^\tau + ... + (\mathsf{P}^\tau)^{n-1})(\mathsf{P}^\tau)^{n-1}(\mathsf{P}^\tau - I)V \\ &= (I + \mathsf{P}^\tau + ... + (\mathsf{P}^\tau)^{n-1}) r_\tau + ((\mathsf{P}^\tau)^n - I)V. \end{aligned} \tag{156}$$

Hence

$$g\mathbf{1} \leq \lim_{n\to\infty} \frac{1}{n} \mathbb{E}_{\mathsf{P}^\tau_{\text{ave}}, \tau} \left[\sum_{t=0}^{n} r_t \right] = g^\tau_{\mathsf{P}^\tau_{\text{ave}}} \mathbf{1} = g^\tau_{\mathcal{P}} \mathbf{1} \leq g^*_{\mathcal{P}} \mathbf{1}. \tag{157}$$

Thus $g = g^*_{\mathcal{P}}$. This hence proves (1).

To prove (2), note that for any stationary policy π, we denote by

$$\sigma_{\mathcal{P}^\pi}(V) \triangleq (\sum_a \pi(a|s_1)\sigma_{\mathcal{P}^a_{s_1}}(V), ..., \sum_a \pi(a|s_{|\mathcal{S}|})\sigma_{\mathcal{P}^a_{s_{|\mathcal{S}|}}}(V))$$

being a vector in $\mathbb{R}^{|\mathcal{S}|}$. Then (15) is equivalent to

$$r_{\pi^*} + \sigma_{\mathcal{P}^{\pi^*}}(V) = \max_\pi \left\{ r_\pi + \sigma_{\mathcal{P}^\pi}(V) \right\}. \tag{158}$$

Hence,

$$r_{\pi^*} - g + \sigma_{\mathcal{P}^{\pi^*}}(V) - V = \max_\pi \left\{ r_\pi - g + \sigma_{\mathcal{P}^\pi}(V) - V \right\}. \tag{159}$$

Since (g, V) is a solution to (14), it follows that

$$r_{\pi^*} - g + \sigma_{\mathcal{P}^{\pi^*}}(V) - V = 0. \tag{160}$$

According to the robust Bellman equation (12), $(g^{\pi^*}_{\mathcal{P}}, V^{\pi^*}_{\mathcal{P}})$ is a solution to (160). Thus from Theorem 19, $g^{\pi^*}_{\mathcal{P}} = g^*_{\mathcal{P}}$, and hence π^* is an optimal robust policy.

To prove (3), recall that $V_Q(s) = \max_a Q(s,a)$. It can be also written as

$$V_Q(s) = \sum_a \pi_Q(a|s) Q(s,a). \tag{161}$$

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 (140) can be rewritten as

$$Q(s, \pi_Q(s)) = r(s, \pi_Q(s)) - g + \sigma_{\mathcal{P}^{\pi_Q(s)}_s} \left(\sum_a \pi_Q(a|\cdot) Q(\cdot, a) \right). \tag{162}$$

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

$$W(s) = \sum_a \pi_Q(a|s)(r(s,a) - g + \sigma_{\mathcal{P}^a_s}(W)). \tag{163}$$

Therefore, (W, g) is a solution to the robust Bellman equation for the policy π_Q in Theorem 7. By Theorem 7, we have that

$$g = g_{\mathcal{P}}^{\pi_Q}, \tag{164}$$
$$W = V_{\mathsf{P}}^{\pi_Q} + ce, \tag{165}$$

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. □

Theorem 20 (Restatement of Theorem 9)**.** *(w_T, V_t) in Algorithm 3 converges to a solution of* (14)*.*

Before showing this theorem, we first present the robust aperiodic transform in the next lemma.

Lemma 12. *(Robust Aperiodic Transform) Assume a robust MDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$ satisfies Assumption 1. Construct another uncertainty set as follows:*

$$\tilde{\mathcal{P}}_s^a = \{\tilde{\mathsf{P}}_s^a = (1-\tau)\mathsf{P}_s^a + \tau \boldsymbol{1}_s : \mathsf{P}_s^a \in \mathcal{P}_s^a\},$$

where $\tau \in (0, 1)$. Then the optimal robust policies for both robust MDPs are the same.

Proof. The result can be straightforwardly derived by the following claim: for any π and $\mathsf{P} \in \mathcal{P}$, $g_{\mathsf{P}}^{\pi} = g_{\tilde{\mathsf{P}}}^{\pi}$.

Note that the discounted value function $V_{\mathsf{P},\gamma}^{\pi} = (I - \gamma \mathsf{P}^{\pi})^{-1} r_{\pi}$. Hence we have that

$$\begin{aligned}
V_{\tilde{\mathsf{P}},\gamma}^{\pi} &= (I - \gamma \tilde{\mathsf{P}}^{\pi})^{-1} r_{\pi} \\
&= (I - \gamma((1-\tau)\mathsf{P}^{\pi} + \tau I))^{-1} r_{\pi} \\
&= ((1-\tau\gamma)I - \gamma(1-\tau)\mathsf{P}^{\pi})^{-1} r_{\pi} \\
&= \left((1-\tau\gamma)\left(I - \frac{\gamma(1-\tau)}{1-\tau\gamma}\mathsf{P}^{\pi}\right)\right)^{-1} r_{\pi} \\
&= \frac{1}{1-\tau\gamma}\left(I - \frac{\gamma(1-\tau)}{1-\tau\gamma}\mathsf{P}^{\pi}\right)^{-1} r_{\pi} \\
&= \frac{1}{1-\tau\gamma} V_{\mathsf{P},\frac{\gamma(1-\tau)}{1-\tau\gamma}}^{\pi}.
\end{aligned} \tag{166}$$

Moreover, by noting $1 - \frac{\gamma(1-\tau)}{1-\tau\gamma} = \frac{1-\gamma}{1-\tau\gamma}$, we have

$$\begin{aligned}
(1-\gamma)V_{\tilde{\mathsf{P}},\gamma}^{\pi} &= \frac{1-\gamma}{1-\tau\gamma} V_{\mathsf{P},\frac{\gamma(1-\tau)}{1-\tau\gamma}}^{\pi} \\
&= \frac{1-\gamma}{1-\tau\gamma} V_{\mathsf{P},\frac{\gamma(1-\tau)}{1-\tau\gamma}}^{\pi} \\
&= \left(1 - \frac{\gamma(1-\tau)}{1-\tau\gamma}\right) V_{\mathsf{P},\frac{\gamma(1-\tau)}{1-\tau\gamma}}^{\pi}.
\end{aligned} \tag{167}$$

Now set $\gamma \to 1$ on both sides, we have that

$$g^\pi_{\tilde{\mathsf{P}}} = \lim_{\gamma\to 1}(1-\gamma)V^\pi_{\tilde{\mathsf{P}},\gamma} = \lim_{\gamma\to 1}\left(1-\frac{\gamma(1-\tau)}{1-\tau\gamma}\right)V^\pi_{\mathsf{P},\frac{\gamma(1-\tau)}{1-\tau\gamma}} = g^\pi_{\mathsf{P}}, \tag{168}$$

where the last equation is from the fact that $\gamma \to 1$ implies $\frac{\gamma(1-\tau)}{1-\tau\gamma} \to 1$, and hence proves the claim and the lemma. □

The reason we apply such a robust aperiodic transform is that the modified transition kernel is always aperiodic (since $\tilde{\mathsf{P}}^\pi(s|s), (\tilde{\mathsf{P}}^\pi)^2(s|s) > 0, \forall s$). Hence Assumption 1 is equivalent to the following strong assumption:

Assumption 5. *There exists a positive integer J such that for any $\mathsf{P} = \{p^a_s \in \Delta(\mathcal{S})\} \in \mathcal{P}$ and any stationary deterministic policy π, there exists $\kappa > 0$ and a state $s \in \mathcal{S}$, such that $((\mathsf{P}^\pi)^J)_{x,s} \geq \kappa, \forall x \in \mathcal{S}$.*

This assumption is shown to be equivalent to assuming each transition kernel in the uncertainty set can induce a unichain and aperiodic Markov chain under any determinisit policy (Bertsekas, 2011). Under the robust aperiodic transform, the modified uncertainty set hence satisfy this assumption. Without loss of generality, we prove the convergence under this assumption.

Proof. We first denote the update operator as

$$Lv(s) \triangleq \max_a(r(s,a) + \sigma_{\mathcal{P}^a_s}(v)). \tag{169}$$

Now, consider $sp(Lv - Lu)$. Denote by $\acute{s} \triangleq \arg\max_s(Lv(s) - Lu(s))$ and $\grave{s} \triangleq \arg\min_s(Lv(s) - Lu(s))$. Also denote by $a_v \triangleq \arg\max_a(r(\acute{s},a) + \sigma_{\mathcal{P}^a_{\acute{s}}}(v))$ and $a_u \triangleq \arg\max_a(r(\acute{s},a) + \sigma_{\mathcal{P}^a_{\acute{s}}}(u))$ Then

$$\begin{aligned}
Lv(\acute{s}) - Lu(\acute{s}) &= \max_a(r(\acute{s},a) + \sigma_{\mathcal{P}^a_{\acute{s}}}(v)) - \max_a(r(\acute{s},a) + \sigma_{\mathcal{P}^a_{\acute{s}}}(u)) \\
&\triangleq r(\acute{s},a_v) + \sigma_{\mathcal{P}^{a_v}_{\acute{s}}}(v) - (r(\acute{s},a_u) + \sigma_{\mathcal{P}^{a_u}_{\acute{s}}}(u)) \\
&\leq r(\acute{s},a_v) + \sigma_{\mathcal{P}^{a_v}_{\acute{s}}}(v) - (r(\acute{s},a_v) + \sigma_{\mathcal{P}^{a_v}_{\acute{s}}}(u)) \\
&= \sigma_{\mathcal{P}^{a_v}_{\acute{s}}}(v) - \sigma_{\mathcal{P}^{a_v}_{\acute{s}}}(u) \\
&\triangleq (p^{a_v,v}_{\acute{s}})^\top v - (p^{a_v,u}_{\acute{s}})^\top u,
\end{aligned} \tag{170}$$

where $p^{a_v,v}_{\acute{s}} = \arg\min_{p\in\mathcal{P}^{a_v}_{\acute{s}}} p^\top v$ and $p^{a_v,u}_{\acute{s}} = \arg\min_{p\in\mathcal{P}^{a_v}_{\acute{s}}} p^\top u$. Thus (170) can be further bounded as

$$\begin{aligned}
&Lv(\acute{s}) - Lu(\acute{s}) \\
&\leq (p^{a_v,v}_{\acute{s}})^\top v - (p^{a_v,u}_{\acute{s}})^\top u \\
&\leq (p^{a_v,u}_{\acute{s}})^\top (v-u).
\end{aligned} \tag{171}$$

Similarly,

$$Lv(\grave{s}) - Lu(\grave{s}) \geq (p^{a_u,v}_{\grave{s}})^\top (v-u). \tag{172}$$

Thus

$$sp(Lv - Lu) \leq (p_{\acute{s}}^{a_v,u})^\top (v-u) - (p_{\grave{s}}^{a_u,v})^\top (v-u). \tag{173}$$

Now denote by $v - u \triangleq (x_1, x_2, ..., x_n)$, $p_{\acute{s}}^{a_v,u} = (p_1, ..., p_n)$ and $p_{\grave{s}}^{a_u,v} = (q_1, ..., q_n)$. Further denote by $b_i \triangleq \min\{p_i, q_i\}$ Then

$$\begin{aligned}
&\sum_{i=1}^n p_i x_i - \sum_{i=1}^n q_i x_i \\
&= \sum_{i=1}^n (p_i - b_i) x_i - \sum_{i=1}^n (q_i - b_i) x_i \\
&\leq \sum_{i=1}^n (p_i - b_i) \max\{x_i\} - \sum_{i=1}^n (q_i - b_i) \min\{x_i\} \\
&= \sum_{i=1}^n (p_i - b_i) sp(x) + \left(\sum_{i=1}^n (p_i - b_i) - \sum_{i=1}^n (q_i - b_i) \right) \min\{x_i\} \\
&= \left(1 - \sum_{i=1}^n b_i\right) sp(x).
\end{aligned} \tag{174}$$

Thus we showed that

$$sp(Lv - Lu) \leq \left(1 - \sum_{i=1}^n b_i\right) sp(v-u). \tag{175}$$

Now from Assumption 1, and following Theorem 8.5.3 from (Puterman, 1994), it can be shown that there exists $1 > \lambda > 0$, such that for any a, u, v,

$$\sum_{i=1}^n b_i \geq \lambda. \tag{176}$$

Further, following Theorem 8.5.2 in (Puterman, 1994), it can be shown that L is a J-step contraction operator for some integer J, i.e.,

$$sp(L^J v - L^J u) \leq (1-\lambda) sp(v-u). \tag{177}$$

Then, it can be shown that the relative value iteration converges to a solution of the optimal equation similar to the relative value iteration for non-robust MDPs under the average-reward criterion (Theorem 8.5.7 in (Puterman, 1994), Section 1.6.4 in(Sigaud & Buffet, 2013)), and hence (w_t, V_t) converges to a solution to (14) as $\epsilon \to 0$. □

Appendix F. Robust RVI TD Method for Policy Evaluation

We define the following notation:

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

F.1 Proof of Lemma 1

We construct the following example.

Example 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 3 with $g^\pi_{\mathsf{P}_1}(1) = g^\pi_{\mathsf{P}_1}(2) = g^\pi_{\mathsf{P}_1}(3), g^\pi_{\mathsf{P}_2}(1) = g^\pi_{\mathsf{P}_2}(2) = g^\pi_{\mathsf{P}_2}(3)$.

Under both transition kernels $\mathsf{P}_1, \mathsf{P}_2$, the average-reward are identical: $g^\pi_{\mathsf{P}_1} = g^\pi_{\mathsf{P}_2} = 0.5r_2 + 0.5r_3$. 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

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

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

When $r_3 > r_2$, only $V^\pi_{\mathsf{P}_1}$ is the solution to (12); and when $r_2 > r_3$, only $V^\pi_{\mathsf{P}_2}$ is the solution to (12). Hence, this proves Lemma 1 and implies that not any relative value function w.r.t. a worst-case transition kernel is a solution to (12).

F.2 Proof of Theorem 10

Theorem 21. *(Restatement of Theorem 10) Under Assumptions 3,2,4, $(f(V_n), V_n)$ converges to a (possible sample path dependent) solution to* (12) *a.s..*

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

Lemma 13. *Algorithm 4 remains bounded during the update, i.e.,*

$$\sup_n \|V_n\| < \infty, a.s.. \tag{178}$$

Proof. Denote by

$$h(V) \triangleq r_\pi + \sigma_{\mathcal{P}}(V) - f(V)e - V. \tag{179}$$

Then the update of robust RVI TD can be rewritten as

$$V_{n+1} = V_n + \alpha_n(h(V_n) + M_{n+1}), \tag{180}$$

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_∞:

$$h_\infty(V) \triangleq \lim_{c\to\infty} \frac{h(cV)}{c}. \tag{181}$$

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

$$h_\infty(V) = \lim_{c\to\infty} \frac{r_\pi}{c} + \sigma_{\mathcal{P}}(V) - f(V)e - V = \sigma_{\mathcal{P}}(V) - f(V)e - V. \tag{182}$$

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 4;

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

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

First, note that

$$\begin{aligned}
&\|h(V_1) - h(V_2)\| \\
&= \max_s \left| \sum_a \pi(a|s)(\sigma_{\mathcal{P}_s^a}(V_1) - \sigma_{\mathcal{P}_s^a}(V_2)) - (f(V_1) - f(V_2)) - (V_1(s) - V_2(s)) \right| \\
&\leq \max_s \left\{ \left| \sum_a \pi(a|s)(\sigma_{\mathcal{P}_s^a}(V_1) - \sigma_{\mathcal{P}_s^a}(V_2)) \right| + |(f(V_1) - f(V_2))| + |(V_1(s) - V_2(s))| \right\} \\
&\leq (2 + L_f)\|V_1 - V_2\|,
\end{aligned} \tag{183}$$

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 $\mathrm{Var}[\hat{\mathbf{T}}V_n|\mathcal{F}_n] \leq K(1+\|V_n\|^2)$. As discussed in Section 5.1, we assume the existence of an unbiased estimator $\hat{\mathbf{T}}$ with bounded variance here, and we will construct the estimator in Section 5.3.

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

$$\dot{x}(t) = h_\infty(x(t)) \tag{184}$$

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 (184) satisfies

$$\mathbf{T}_0(W) - f(W)e - W = 0. \tag{185}$$

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

$$\begin{cases} W = \mathbf{T}_0(W) - ge, \\ g = f(W). \end{cases} \tag{186}$$

The equation in (186) is the robust Bellman equation for a zero-reward robust MDP. Hence, from Theorem 7, any solution (g, W) to (186) satisfies

$$g = g_{\mathcal{P}}^\pi, W = V_{\mathsf{P}}^\pi + ce, \tag{187}$$

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 (184) satisfies

$$W = V_{\mathsf{P}}^{\pi} + ce, f(W) = g_{\mathcal{P}}^{\pi}. \tag{188}$$

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

$$g_{\mathsf{P}}^{\pi} = \lim_{T\to\infty} \mathbb{E}_{\mathsf{P}}\left[\sum_{t=0}^{T-1} \frac{r_t}{T}\right] = 0, \tag{189}$$

$$V_{\mathsf{P}}^{\pi} = \mathbb{E}_{\mathsf{P}}\left[\sum_{t=0}^{\infty} (r_t - g_{\mathsf{P}}^{\pi})\right] = 0, \tag{190}$$

thus $g_{\mathcal{P}}^{\pi} = 0$ and $W = ce$ for some $c \in \mathbb{R}$. From (188), 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

$$W = V_{\mathcal{P}}^{\pi} + ce = 0. \tag{191}$$

Thus, the only equilibrium of (184) is 0.

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

$$\mathbf{T}_0 V(s) = \sum_a \pi(a|s)(\sigma_{\mathcal{P}_s^a}(V)). \tag{192}$$

We further introduce two operators:

$$\mathbf{T}_0' V \triangleq \mathbf{T}_0 V - f(V)e, \tag{193}$$

$$\tilde{\mathbf{T}}_0 V \triangleq \mathbf{T}_0 V - g_{\mathcal{P}}^{\pi} e. \tag{194}$$

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:

$$\dot{x} = \mathbf{T}_0' x - x, \tag{195}$$

$$\dot{y} = \tilde{\mathbf{T}}_0 y - y. \tag{196}$$

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 (195) is the same as the ODE in (184).

Since the second equation (196) 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 (196) converges to the set of equilibrium points, i.e.,

$$y(t) \to \left\{W : W = \tilde{\mathbf{T}}_0 W\right\}, a.s.. \tag{197}$$

Similar to the discussion for $\mathbf{T}_0$, our Theorem 7 implies that the set of equilibrium points of (196) is $\{W = ce : c \in \mathbb{R}\}$. Hence, for any solution $y(t)$ to (196), $y(t) \to ce$ for some constant k that may depend on the initial value of $y(t)$.

Now, consider the solution $x(t)$ to (195). According to Lemma 20 (note that $\mathbf{T}_0$ here is a special case of $\mathbf{T}$ in Lemma 20 with $r = 0$), if the solutions $x(t), y(t)$ have the same initial value $x(0) = y(0)$, then

$$x(t) = y(t) + r(t)e, \tag{198}$$

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

$$r(t) = \int_0^t e^{-(t-s)}(g_{\mathcal{P}}^{\pi} - f(y(s)))ds \tag{199}$$

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.,

$$x(t) \to ce - f(ce)e = 0. \tag{200}$$

Hence, any solution $x(t)$ to (195) 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 21.

Proof. In Lemma 13, we have shown that

$$\sup_n \|V_n\| < \infty, a.s.. \tag{201}$$

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

$$\mathbf{T}'V \triangleq \mathbf{T}V - f(V)e, \tag{202}$$

$$\tilde{\mathbf{T}}V \triangleq \mathbf{T}V - g_{\mathcal{P}}^{\pi}e. \tag{203}$$

From Lemma 20, we know that if $x(t), y(t)$ are the solutions to equations

$$\dot{x} = \mathbf{T}'x - x, \tag{204}$$

$$\dot{y} = \tilde{\mathbf{T}}y - y, \tag{205}$$

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

$$x(t) = y(t) + r(t)e, \tag{206}$$

where $r(t)$ satisfies

$$\dot{r}(t) = -r(t) + g_{\mathcal{P}}^{\pi} - f(y(t)), r(0) = 0. \tag{207}$$

Thus, by the variation of constants formula,

$$r(t) = \int_0^t e^{-(t-s)}(g_{\mathcal{P}}^{\pi} - f(y(s)))ds. \tag{208}$$

Note that $\tilde{\mathbf{T}}$ is also non-expansive, hence $y(t)$ converges to some equilibrium of (205) (Theorem 3.1 of (Borkar & Soumyanatha, 1997)). The set of equilibrium points of (205) can be characterized as

$$\begin{aligned}
&\left\{W : \tilde{\mathbf{T}}W = W\right\} \\
&= \{W : W = TW - g_{\mathcal{P}}^{\pi} e\} \\
&= \left\{W : W(s) = \sum_a \pi(a|s)(r(s,a) - g_{\mathcal{P}}^{\pi} + \sigma_{\mathcal{P}_s^a}(W)), \forall s \in \mathcal{S}\right\}.
\end{aligned} \tag{209}$$

From Theorem 7, any equilibrium of (205) can be rewritten as

$$W = V_{\mathsf{P}}^{\pi} + ce, \text{ for some } \mathsf{P} \in \Omega_g^{\pi}, c \in \mathbb{R}. \tag{210}$$

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

$$y(t) \to y^* \triangleq V_{\mathsf{P}^*}^{\pi} + c^* e, \text{ for some } \mathsf{P}^* \in \Omega_g^{\pi}, c^* \in \mathbb{R}. \tag{211}$$

Similar to Lemma 13, it can be showed that $r(t) \to g_{\mathcal{P}}^{\pi} - f(y^*)$ (Lemma B.4 in (Wan et al., 2021), Theorem 3.4 in (Abounadi et al., 2001)). This further implies that

$$x(t) \to y^* + (g_{\mathcal{P}}^{\pi} - f(y^*))e = V_{\mathsf{P}^*}^{\pi} + (c^* + g_{\mathcal{P}}^{\pi} - f(y^*))e, \tag{212}$$

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

$$\begin{aligned}
f(x(t)) &\to f(V_{\mathsf{P}^*}^{\pi} + (c^* + g_{\mathcal{P}}^{\pi} - f(y^*))e) \\
&= f(V_{\mathsf{P}^*}^{\pi}) + c^* + g_{\mathcal{P}}^{\pi} - f(y^*) \\
&= f(V_{\mathsf{P}^*}^{\pi}) + c^* + g_{\mathcal{P}}^{\pi} - f(V_{\mathsf{P}^*}^{\pi} + c^* e) \\
&= f(V_{\mathsf{P}^*}^{\pi}) + c^* + g_{\mathcal{P}}^{\pi} - f(V_{\mathsf{P}^*}^{\pi}) - c^* \\
&= g_{\mathcal{P}}^{\pi}.
\end{aligned} \tag{213}$$

Hence, we show that

$$x(t) \to V_{\mathsf{P}^*}^{\pi} + m^* e, \tag{214}$$

$$f(x(t)) \to g_{\mathcal{P}}^{\pi}. \tag{215}$$

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

$$V_n \to V_{\mathsf{P}^*}^{\pi} + m^* e, \tag{216}$$

$$f(V_n) \to g_{\mathcal{P}}^{\pi}, \tag{217}$$

which completes the proof. □

Appendix G. Robust RVI Q-Learning

G.1 Proof of Theorem 11

Lemma 14. *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 $Var(\hat{\mathbf{H}}Q(s,a)) \leq C(1+\|Q\|^2)$ for some constant C, then under Assumptions 2, 1 and 4, Algorithm 5 remains bounded during the update almost surely, i.e.,*

$$\sup_n \|Q_n\| < \infty, a.s.. \tag{218}$$

Proof. Denote by

$$h(Q) \triangleq r_\pi + \sigma_{\mathcal{P}}(V_Q) - f(Q)e - Q. \tag{219}$$

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

$$Q_{n+1} = Q_n + \alpha_n(h(Q_n) + M_{n+1}), \tag{220}$$

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_∞:

$$h_\infty(Q) \triangleq \lim_{c\to\infty} \frac{h(cQ)}{c}. \tag{221}$$

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

$$h_\infty(Q) = \lim_{c\to\infty} \frac{r_\pi}{c} + \sigma_{\mathcal{P}}(V_Q) - f(Q)e - Q = \sigma_{\mathcal{P}}(V_Q) - f(Q)e - Q. \tag{222}$$

Similar to the proof of Theorem 13, it suffices to verify the following conditions:
(1). h is Lipschitz;
(2). Stepsize α_n satisfies Assumption 4;
(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 Theorem 13. We then verify (1) and (4). Firstly, it can be shown that

$$\begin{aligned}
&|h(Q_1)(s,a) - h(Q_2)(s,a)| \\
&= \left|\sigma_{\mathcal{P}_s^a}(V_{Q_1}) - f(Q_1) - Q_1(s,a) - \sigma_{\mathcal{P}_s^a}(V_{Q_2}) - f(Q_2) - Q_2(s,a)\right| \\
&\leq \left|\sigma_{\mathcal{P}_s^a}(V_{Q_1}) - \sigma_{\mathcal{P}_s^a}(V_{Q_2})\right| + |f(Q_1) - f(Q_2)| + |Q_1(s,a) - Q_2(s,a)| \\
&\leq \|V_{Q_1} - V_{Q_2}\| + L_f\|Q_1 - Q_2\| + \|Q_1 - Q_2\| \\
&\leq (2 + L_f)\|Q_1 - Q_2\|,
\end{aligned} \tag{223}$$

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

$$\dot{X}(t) = h_\infty(X(t)) = \sigma_{\mathcal{P}}(V_X(t)) - f(X(t))e - X(t), \tag{224}$$

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 (224) satisfies that

$$Q(s,a) = \sigma_{\mathcal{P}_s^a}(V_Q) - f(Q)e, \tag{225}$$

which can be viewed as an optimal robust Bellman equation (13) with zero reward. Hence, by Theorem 19, it implies that

$$f(Q) = g_{\mathcal{P}}^* = 0, \tag{226}$$

$$V_Q = V_{\mathsf{P}}^{\pi_Q} + ce \text{ for some } \mathsf{P} \in \Omega_g^{\pi_Q}, c \in \mathbb{R}. \tag{227}$$

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 (225), Q satisfies that

$$Q(s,a) = \sigma_{\mathcal{P}_s^a}(V_Q) = \sigma_{\mathcal{P}_s^a}(ce) = c. \tag{228}$$

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

$$f(Q) = f(ce) = c = 0. \tag{229}$$

Therefore,

$$c = 0, \tag{230}$$

$$Q = 0. \tag{231}$$

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

$$\mathbf{H}_0 Q(s,a) = \sigma_{\mathcal{P}_s^a}(V_Q), \tag{232}$$

and further introduce two operators

$$\mathbf{H}_0' Q(s,a) = \sigma_{\mathcal{P}_s^a}(V_Q) - f(Q), \tag{233}$$

$$\tilde{\mathbf{H}}_0 Q(s,a) = \sigma_{\mathcal{P}_s^a}(V_Q) - g_{\mathcal{P}}^*. \tag{234}$$

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

$$\dot{y} = \tilde{\mathbf{H}}_0 y - y \tag{235}$$

converges to the set of equilibrium points

$$\{W : W(s,a) = \sigma_{\mathcal{P}_s^a}(V_W) - g_{\mathcal{P}}^*\}, a.s.. \tag{236}$$

This again can be viewed as an optimal robust Bellman equation with zero-reward. Hence, any equilibrium W of (235) satisfies

$$\max_a W(s,a) = c, \forall s. \tag{237}$$

This together with (236) further implies that the equilibrium W of (235) satisfies

$$W(s,a) = \sigma_{\mathcal{P}_s^a}(V_W) = \sigma_{\mathcal{P}_s^a}(ce) = c, \tag{238}$$

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

Lemma 25 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 13, 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,

$$x(t) \to 0, \tag{239}$$

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 22 (Restatement of Theorem 11)**.** *The sequence $\{Q_n\}$ generated by Algorithm 5 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,

$$\mathbf{H}Q(s,a) = r(s,a) + \sigma_{\mathcal{P}_s^a}(V_Q), \tag{240}$$

define two operators

$$\mathbf{H}'Q \triangleq \mathbf{H}Q - f(Q)e, \tag{241}$$

$$\tilde{\mathbf{H}}Q \triangleq \mathbf{H}Q - g_{\mathcal{P}}^* e. \tag{242}$$

From Lemma 25, we know that if $x(t), y(t)$ are the solutions to equations

$$\dot{x} = \mathbf{H}'x - x, \tag{243}$$

$$\dot{y} = \tilde{\mathbf{H}}y - y, \tag{244}$$

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

$$x(t) = y(t) + r(t)e, \tag{245}$$

where $r(t)$ satisfies

$$\dot{r}(t) = -r(t) + g_{\mathcal{P}}^* - f(y(t)), r(0) = 0. \tag{246}$$

It can be easily verified that $\tilde{\mathbf{H}}$ is non-expansive. Hence $y(t)$ converges to the set of equilibrium points of of (244) (Theorem 3.1 of (Borkar & Soumyanatha, 1997)), which can be characterized as

$$\begin{aligned}&\left\{W : \tilde{\mathbf{H}}W = W\right\}\\ &= \{W : W = \mathbf{H}W - g^*_{\mathcal{P}} e\}\\ &= \left\{W : W(s,a) = r(s,a) - g^*_{\mathcal{P}} + \sigma_{\mathcal{P}^a_s}(V_W), \forall s, a\right\}. \end{aligned} \tag{247}$$

From Theorem 19, any equilibrium W satisfies

$$V_W = V_{\mathsf{P}}^{\pi_W} + ce, \text{ for some } \mathsf{P} \in \Omega_g^{\pi_W}, c \in \mathbb{R}, \tag{248}$$

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

Similar to (212) to (216), it can be shown that $r(t) \to g^*_{\mathcal{P}} - f(W^*)$. This further implies that

$$x(t) \to W^* + (g^*_{\mathcal{P}} - f(W^*))e \triangleq W^* + m^* e, \tag{249}$$

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 (13). Moreover, since f is continuous (because it is Lipschitz), we have that

$$\begin{aligned} f(x(t)) &\to f(W^* + m^* e)\\ &= f(W^*) + g^*_{\mathcal{P}} - f(W^*)\\ &= g^*_{\mathcal{P}}. \end{aligned} \tag{250}$$

This completes the proof. □

Appendix H. Case Studies for Robust RVI TD

In this section, we provide the proof of the first part of Theorem 12, 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 15. *If*

$$\mathbb{E}[\hat{\sigma}_{\mathcal{P}^a_s} V] = \sigma_{\mathcal{P}^a_s}(V), \forall s, a, \tag{251}$$

and moreover, there exists a constant C, such that

$$Var(\hat{\sigma}_{\mathcal{P}^a_s} V) \le C(1 + \|V\|^2), \forall s, a, \tag{252}$$

then

$$\mathbb{E}[\hat{\mathbf{T}} V(s)] = \mathbf{T} V(s), \forall s, \tag{253}$$

and

$$Var(\hat{\mathbf{T}} V(s)) \le |\mathcal{A}| C(1 + \|V\|^2), \forall s. \tag{254}$$

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

$$\begin{aligned}
\mathbb{E}[\hat{\mathbf{T}}V(s)] &= \mathbb{E}\bigg[\sum_a \pi(a|s)(r(s,a) + \hat{\sigma}_{\mathcal{P}_s^a} V)\bigg] \\
&= \sum_a \pi(a|s)(r(s,a) + \mathbb{E}[\hat{\sigma}_{\mathcal{P}_s^a} V]) \\
&= \sum_a \pi(a|s)(r(s,a) + \sigma_{\mathcal{P}_s^a}(V)) = \mathbf{T}V(s),
\end{aligned} \tag{255}$$

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

$$\begin{aligned}
\mathrm{Var}(\hat{\mathbf{T}}V(s)) &= \mathbb{E}\bigg[\bigg(\sum_a \pi(a|s)(r(s,a) + \hat{\sigma}_{\mathcal{P}_s^a} V) - \mathbb{E}\bigg[\sum_a \pi(a|s)(r(s,a) + \hat{\sigma}_{\mathcal{P}_s^a} V)\bigg]\bigg)^2\bigg] \\
&= \mathbb{E}\bigg[\bigg(\sum_a \pi(a|s)(r(s,a) + \hat{\sigma}_{\mathcal{P}_s^a} V) - \sum_a \pi(a|s)(r(s,a) + \mathbb{E}\big[\hat{\sigma}_{\mathcal{P}_s^a} V\big]\bigg)^2\bigg] \\
&= \mathbb{E}\bigg[\bigg(\sum_a \pi(a|s)(\hat{\sigma}_{\mathcal{P}_s^a} V) - \mathbb{E}\big[\hat{\sigma}_{\mathcal{P}_s^a} V\big]\bigg)^2\bigg] \\
&\overset{(a)}{\leq} \mathbb{E}\bigg[\sum_a \pi(a|s)(\hat{\sigma}_{\mathcal{P}_s^a} V - \mathbb{E}\big[\hat{\sigma}_{\mathcal{P}_s^a} V\big])^2\bigg] \\
&= \sum_a \pi(a|s)\mathbb{E}\bigg[(\hat{\sigma}_{\mathcal{P}_s^a} V - \mathbb{E}\big[\hat{\sigma}_{\mathcal{P}_s^a} V^2\big])^2\bigg] \\
&\leq \sum_a \pi(a|s)\mathrm{Var}(\hat{\sigma}_{\mathcal{P}_s^a} V) \\
&\leq |\mathcal{A}|C(1 + \|V\|^2),
\end{aligned} \tag{256}$$

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

This lemma implies that to prove Theorem 12, it suffices to show that $\hat{\sigma}_{\mathcal{P}_s^a}$ is unbiased and has bounded variance.

H.1 Contamination Uncertainty Set

Theorem 23. $\hat{\mathbf{T}}$ *defined in* (18) *is unbiased and has bounded variance.*

Proof. First, note that

$$\begin{aligned}
V_{n+1}(s) &= V_n(s) + \alpha_n(r(s,a) + ((1-\zeta)V_n(s') + \zeta \min_x V_n(x) - f(V_n) - V_n(s)) \\
&= V_n(s) + \alpha_n(\mathbf{T}V_n(s) - f(V_n) - V_n(s) + M_n(s)),
\end{aligned} \tag{257}$$

where

$$M_n(s) = r(s,a) + (1-\zeta)V_n(s') + \zeta \min_x V_n(x) - \mathbf{T}V_n(s), \tag{258}$$

and

$$\mathbf{T}V_n(s) = \sum_a \pi(a|s)\bigg(r(s,a) + (1-\zeta)\sum_{s'} \mathsf{P}^a_{s,s'} V_n(s') + \zeta \min_x V_n(x)\bigg). \tag{259}$$

Thus,

$$\begin{aligned}
\mathbb{E}[M_n(s)] &= \mathbb{E}\big[r(s,a) + (1-\zeta)V_n(s') + \zeta \min_x V_n(x)\big] \\
&\quad - \sum_a \pi(a|s)\bigg(r(s,a) + (1-\zeta)\sum_{s'} \mathsf{P}^a_{s,s'} V_n(s') + \zeta \min_x V_n(x)\bigg) \\
&= \sum_a \pi(a|s)\bigg(r(s,a) + (1-\zeta)\sum_{s'} \mathsf{P}^a_{s,s'} V_n(s') + \zeta \min_x V_n(x)\bigg) \\
&\quad - \sum_a \pi(a|s)\bigg(r(s,a) + (1-\zeta)\sum_{s'} \mathsf{P}^a_{s,s'} V_n(s') + \zeta \min_x V_n(x))\bigg) \\
&= 0.
\end{aligned} \tag{260}$$

Hence, the operator is unbiased.

We also have that

$$\begin{aligned}
\mathbb{E}[|M_n(s)|^2] &= \mathbb{E}\bigg[\bigg(r(s,a) + (1-\zeta)V_n(s') + \zeta \min_x V_n(x) - \mathbf{T}V_n(s)\bigg)^2\bigg] \\
&\leq 2\mathbb{E}\bigg[\bigg(r(s,a) + (1-\zeta)V_n(s') + \zeta \min_x V_n(x)\bigg)^2\bigg] + 2\mathbb{E}[(\mathbf{T}V_n(s))^2] \\
&\overset{(a)}{\leq} 8 + 8\|V_n\|^2 \\
&\leq 8(1 + \|V_n\|^2),
\end{aligned} \tag{261}$$

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

The proof is completed. □

H.2 Total Variation Uncertainty Set

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

$$\hat{\sigma}_{\mathcal{P}^a_s}(V) = \max_{\mu \geq 0} \big(\hat{\mathsf{P}}^{a,1}_{s,N+1}(V-\mu) - \zeta \mathrm{Span}(V-\mu)\big) + \frac{\Delta_N(V)}{p_N}, \tag{262}$$

where

$$\begin{aligned}
\Delta_N(V) &= \max_{\mu \geq 0} \big(\hat{\mathsf{P}}^{a}_{s,N+1}(V-\mu) - \zeta \mathrm{Span}(V-\mu)\big) \\
&\quad - \frac{1}{2}\max_{\mu \geq 0} \big(\hat{\mathsf{P}}^{a,O}_{s,N+1}(V-\mu) - \zeta \mathrm{Span}(V-\mu)\big) \\
&\quad - \frac{1}{2}\max_{\mu \geq 0} \big(\hat{\mathsf{P}}^{a,E}_{s,N+1}(V-\mu) - \zeta \mathrm{Span}(V-\mu)\big).
\end{aligned} \tag{263}$$

Theorem 24. *The estimated operator $\hat{\sigma}_{\mathcal{P}_s^a}$ defined in* (262) *is unbiased, i.e.,*

$$\mathbb{E}[\hat{\sigma}_{\mathcal{P}_s^a} V] = \sigma_{\mathcal{P}_s^a}(V). \tag{264}$$

Proof. First, denote the dual function (20) by g:

$$g_{s,a}^V(\mu) = \mathsf{P}_s^a(V-\mu) - \zeta \mathrm{Span}(V-\mu), \tag{265}$$

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

$$\mu_{s,a}^V = \arg\max_{\mu \geq 0} \big(\mathsf{P}_s^a(V-\mu) - \zeta \mathrm{Span}(V-\mu)\big). \tag{266}$$

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

$$g_{s,a,N+1}^V(\mu) = \hat{\mathsf{P}}_{s,N+1}^a(V-\mu) - \zeta \mathrm{Span}(V-\mu), \tag{267}$$

$$g_{s,a,N+1,O}^V(\mu) = \hat{\mathsf{P}}_{s,N+1}^{a,O}(V-\mu) - \zeta \mathrm{Span}(V-\mu), \tag{268}$$

$$g_{s,a,N+1,E}^V(\mu) = \hat{\mathsf{P}}_{s,N+1}^{a,E}(V-\mu) - \zeta \mathrm{Span}(V-\mu), \tag{269}$$

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

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}_{\mathcal{P}_s^a} V] &= \mathbb{E}\bigg[\max_{\mu \geq 0} \big(\hat{\mathsf{P}}_{s,N+1}^{a,1}(V-\mu) - \zeta \mathrm{Span}(V-\mu)\big) + \frac{\Delta_N(V)}{p_N}\bigg] \\
&= \mathbb{E}[g_{s,a,0}^V(\mu_{s,a,0}^V)] + \mathbb{E}\bigg[\frac{\Delta_N(V)}{p_N}\bigg] \\
&= \mathbb{E}[g_{s,a,0}^V(\mu_{s,a,0}^V)] + \sum_{n=0}^{\infty} p(N=n)\mathbb{E}\bigg[\frac{\Delta_N(V)}{p_N} | N=n\bigg] \\
&= \mathbb{E}[g_{s,a,0}^V(\mu_{s,a,0}^V)] + \sum_{n=0}^{\infty} \mathbb{E}[\Delta_n(V)] \\
&= \mathbb{E}[g_{s,a,0}^V(\mu_{s,a,0}^V)] \\
&\quad + \sum_{n=0}^{\infty} \mathbb{E}\bigg[g_{s,a,n+1}^V(\mu_{s,a,n+1}^V) - \frac{g_{s,a,n+1,O}^V(\mu_{s,a,n+1,O}^V) + g_{s,a,n+1,E}^V(\mu_{s,a,n+1,E}^V)}{2}\bigg] \\
&= \mathbb{E}[g_{s,a,0}^V(\mu_{s,a,0}^V)] + \sum_{n=0}^{\infty} \mathbb{E}\Big[g_{s,a,n+1}^V(\mu_{s,a,n+1}^V) - g_{s,a,n}^V(\mu_{s,a,n}^V)\Big],
\end{aligned} \tag{270}$$

where the last inequality is from Lemma 21. The last equation can be further rewritten as

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}_{\mathcal{P}_s^a} V] &= \mathbb{E}[g_{s,a,0}^V(\mu_{s,a,0}^V)] + \sum_{n=0}^{\infty} \mathbb{E}\Big[g_{s,a,n+1}^V(\mu_{s,a,n+1}^V) - g_{s,a,n}^V(\mu_{s,a,n}^V)\Big] \\
&= \lim_{n\to\infty} \mathbb{E}\Big[g_{s,a,n}^V(\mu_{s,a,n}^V)\Big].
\end{aligned} \tag{271}$$

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

$$\lim_{n\to\infty} \mathbb{E}\Big[g_{s,a,n}^V(\mu_{s,a,n}^V)\Big] = g_{s,a}^V(\mu_{s,a}^V). \tag{272}$$

For any arbitrary i.i.d. samples $\{X_i\}$ and its corresponding function $g^V_{s,a,n}$, together with Lemma 22, we have that

$$\begin{aligned}
&|g^V_{s,a,n}(\mu^V_{s,a,n}) - g^V_{s,a}(\mu^V_{s,a})| \\
&= |\max_{0\le\mu\le V+\|V\|e} g^V_{s,a}(\mu) - \max_{0\le\mu\le V+\|V\|e} g^V_{s,a,n}(\mu)| \\
&\le \max_{0\le\mu\le V+\|V\|e} |g^V_{s,a}(\mu) - g^V_{s,a,n}(\mu)| \\
&= \max_{0\le\mu\le V+\|V\|e} |\mathsf{P}^a_s(V-\mu) - \zeta\mathrm{Span}(V-\mu) - \hat{\mathsf{P}}^a_{s,n}(V-\mu) + \zeta\mathrm{Span}(V-\mu)| \\
&= \max_{0\le\mu\le V+\|V\|e} |\mathsf{P}^a_s(V-\mu) - \hat{\mathsf{P}}^a_{s,n}(V-\mu)| \\
&\le \max_{0\le\mu\le V+\|V\|e} \|V-\mu\|\|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n}\|_1 \\
&\le 3\|V\|\|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n}\|_1. \qquad (273)
\end{aligned}$$

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

$$\mathbb{E}[|g^V_{s,a,n}(\mu^V_{s,a,n}) - g^V_{s,a}(\mu^V_{s,a})|] \le 3\|V\|\frac{|\mathcal{S}|^2\sqrt{\pi}}{2^{\frac{n+1}{2}}}, \qquad (274)$$

which implies that

$$\lim_{n\to\infty} \mathbb{E}\left[g^V_{s,a,n}(\mu^V_{s,a,n})\right] = g^V_{s,a}(\mu^V_{s,a}), \qquad (275)$$

completing the proof. □

Theorem 25. *The estimated operator $\hat{\sigma}_{\mathcal{P}^a_s}$ defined in* (262) *has bounded variance, i.e., there exists a constant C_0, such that*

$$Var(\hat{\sigma}_{\mathcal{P}^a_s}V) \le (1 + 18(1+2\zeta)^2 + 2C_0)\|V\|^2. \qquad (276)$$

Proof. Similar to Theorem 24, we have that

$$\begin{aligned}
&\mathrm{Var}(\hat{\sigma}_{\mathcal{P}^a_s}V) \\
&= \mathbb{E}[(\hat{\sigma}_{\mathcal{P}^a_s}V)^2] - \sigma_{\mathcal{P}^a_s}(V)^2 \\
&\le \mathbb{E}\left[\left(g^V_{s,a,0}(\mu^V_{s,a,0}) + \frac{\Delta_N(V)}{p_N}\right)^2\right] + (\sigma_{\mathcal{P}^a_s}(V))^2 \\
&\le 2\mathbb{E}\left[\left(g^V_{s,a,0}(\mu^V_{s,a,0})\right)^2\right] + 2\mathbb{E}\left[\left(\frac{\Delta_N(V)}{p_N}\right)^2\right] + (\sigma_{\mathcal{P}^a_s}(V))^2 \\
&\le (1 + 18(1+2\zeta)^2)\|V\|^2 + 2\sum_{i=0}^{\infty}\frac{\mathbb{E}[(\Delta_i(V))^2]}{p_i}, \qquad (277)
\end{aligned}$$

where the last inequality is from Lemma 22. For any $n \geq 1$, we have that

$$\begin{aligned}
&\mathbb{E}[(\Delta_n(V))^2] \\
&= \mathbb{E}\left[\left(g^V_{s,a,n}(\mu^V_{s,a,n}) - \frac{g^V_{s,a,n,E}(\mu^V_{s,a,n,E}) + g^V_{s,a,n,O}(\mu^V_{s,a,n,O})}{2}\right)^2\right] \\
&= \mathbb{E}\left[\left(g^V_{s,a,n}(\mu^V_{s,a,n}) - g^V_{s,a}(\mu^V_{s,a}) + g^V_{s,a}(\mu^V_{s,a}) - \frac{g^V_{s,a,n,E}(\mu^V_{s,a,n,E}) + g^V_{s,a,n,O}(\mu^V_{s,a,n,O})}{2}\right)^2\right] \\
&\leq 2\mathbb{E}[(g^V_{s,a,n}(\mu^V_{s,a,n}) - g^V_{s,a}(\mu^V_{s,a}))^2] \\
&\quad + 2\mathbb{E}\left[\left(g^V_{s,a}(\mu^V_{s,a}) - \frac{g^V_{s,a,n,E}(\mu^V_{s,a,n,E}) + g^V_{s,a,n,O}(\mu^V_{s,a,n,O})}{2}\right)^2\right] \\
&\overset{(a)}{=} 2\mathbb{E}[(g^V_{s,a,n}(\mu^V_{s,a,n}) - g^V_{s,a}(\mu^V_{s,a}))^2] + 2\mathbb{E}[(g^V_{s,a,n-1}(\mu^V_{s,a,n-1}) - g^V_{s,a}(\mu^V_{s,a}))^2] \\
&\leq 18\|V\|^2\mathbb{E}[\|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n}\|_1^2] + 18\|V\|^2\mathbb{E}[\|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n-1}\|_1^2],
\end{aligned} \tag{278}$$

where (a) is due to Lemma 21 and the last inequality follows a similar argument to (273). 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

$$\sum_{i=0}^{\infty} \frac{\mathbb{E}[(\Delta_i(V))^2]}{p_i} \leq C_0\|V\|^2. \tag{279}$$

Thus,

$$\mathrm{Var}(\hat{\sigma}_{\mathcal{P}^a_s} V) \leq (1 + 18(1+2\zeta)^2)\|V\|^2 + 2C_0\|V\|^2 = (1 + 18(1+2\zeta)^2 + 2C_0)\|V\|^2\,. \tag{280}$$

□

H.3 Chi-Square Uncertainty Set

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

$$\begin{aligned}
\hat{\sigma}_{\mathcal{P}^a_s} V = &\max_{\mu \geq 0} \left(\hat{\mathsf{P}}^{a,1}_{s,N+1}(V-\mu) - \sqrt{\zeta \mathrm{Var}_{\hat{\mathsf{P}}^{a,1}_{s,N+1}}(V-\mu)}\right) \\
&+ \frac{\Delta_N(V)}{p_N},
\end{aligned} \tag{281}$$

where

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

Theorem 26. *The estimated operator defined in* (281) *is unbiased, i.e.,*

$$\mathbb{E}[\hat{\sigma}_{\mathcal{P}^a_s} V] = \sigma_{\mathcal{P}^a_s}(V). \tag{282}$$

Proof. Denote the dual function (21) by g:

$$g^V_{s,a}(\mu) = \mathsf{P}^a_s(V-\mu) - \sqrt{\zeta \mathrm{Var}_{\mathsf{P}^a_s}(V-\mu)}, \tag{283}$$

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

$$\mu^V_{s,a} = \arg\max_{\mu \geq 0} \left(\mathsf{P}^a_s(V-\mu) - \sqrt{\zeta \mathrm{Var}_{\mathsf{P}^a_s}(V-\mu)}\right). \tag{284}$$

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

$$g^V_{s,a,N+1}(\mu) = \hat{\mathsf{P}}^a_{s,N+1}(V-\mu) - \sqrt{\zeta \mathrm{Var}_{\hat{\mathsf{P}}^a_{s,N+1}}(V-\mu)}, \tag{285}$$

$$g^V_{s,a,N+1,O}(\mu) = \hat{\mathsf{P}}^{a,O}_{s,N+1}(V-\mu) - \sqrt{\zeta \mathrm{Var}_{\hat{\mathsf{P}}^{a,O}_{s,N+1}}(V-\mu)}, \tag{286}$$

$$g^V_{s,a,N+1,E}(\mu) = \hat{\mathsf{P}}^{a,E}_{s,N+1}(V-\mu) - \sqrt{\zeta \mathrm{Var}_{\hat{\mathsf{P}}^{a,E}_{s,N+1}}(V-\mu)}, \tag{287}$$

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

$$\begin{aligned}
&\mathbb{E}[\hat{\sigma}_{\mathcal{P}^a_s} V] \\
&= \mathbb{E}[g^V_{s,a,0}(\mu^V_{s,a,0})] + \mathbb{E}\left[\frac{\Delta_N(V)}{p_N}\right] \\
&= \mathbb{E}[g^V_{s,a,0}(\mu^V_{s,a,0})] + \sum_{n=0}^{\infty} p(N=n)\mathbb{E}\left[\frac{\Delta_N(V)}{p_N} | N=n\right] \\
&= \mathbb{E}[g^V_{s,a,0}(\mu^V_{s,a,0})] + \sum_{n=0}^{\infty} \mathbb{E}[\Delta_n] \\
&= \mathbb{E}[g^V_{s,a,0}(\mu^V_{s,a,0})] \\
&\quad + \sum_{n=0}^{\infty} \mathbb{E}\left[g^V_{s,a,n+1}(\mu^V_{s,a,n+1}) - \frac{g^V_{s,a,n+1,O}(\mu^V_{s,a,n+1,O}) + g^V_{s,a,n+1,E}(\mu^V_{s,a,n+1,E})}{2}\right] \\
&= \mathbb{E}[g^V_{s,a,0}(\mu^V_{s,a,0})] + \sum_{n=0}^{\infty} \mathbb{E}\left[g^V_{s,a,n+1}(\mu^V_{s,a,n+1}) - g^V_{s,a,n}(\mu^V_{s,a,n})\right],
\end{aligned} \tag{288}$$

where the last inequality is from Lemma 21. The last equation can be further rewritten as

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}_{\mathcal{P}^a_s} V] &= \mathbb{E}[g^V_{s,a,0}(\mu^V_{s,a,0})] + \sum_{n=0}^{\infty} \mathbb{E}\left[g^V_{s,a,n+1}(\mu^V_{s,a,n+1}) - g^V_{s,a,n}(\mu^V_{s,a,n})\right] \\
&= \lim_{n\to\infty} \mathbb{E}\left[g^V_{s,a,n}(\mu^V_{s,a,n})\right].
\end{aligned} \tag{289}$$

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

$$\lim_{n\to\infty} \mathbb{E}\left[g^V_{s,a,n}(\mu^V_{s,a,n})\right] = g^V_{s,a}(\mu^V_{s,a}). \tag{290}$$

For any arbitrary i.i.d. samples $\{X_i\}$ and its corresponding function $g^V_{s,a,n}$, together with Lemma 23, we have that

$$
\begin{aligned}
&|g^V_{s,a,n}(\mu^V_{s,a,n}) - g^V_{s,a}(\mu^V_{s,a})| \\
&= |\max_{0\le\mu\le V+\|V\|e} g^V_{s,a}(\mu) - \max_{0\le\mu\le V+\|V\|e} g^V_{s,a,n}(\mu)| \\
&\le \max_{0\le\mu\le V+\|V\|e} |g^V_{s,a}(\mu) - g^V_{s,a,n}(\mu)| \\
&= \max_{0\le\mu\le V+\|V\|e} \left|\mathsf{P}^a_s(V-\mu) - \hat{\mathsf{P}}^a_{s,n}(V-\mu) - \left(\sqrt{\zeta \mathrm{Var}_{\mathsf{P}^a_s}(V-\mu)} - \sqrt{\zeta \mathrm{Var}_{\hat{\mathsf{P}}^a_{s,n}}(V-\mu)}\right)\right| \\
&\le \max_{0\le\mu\le V+\|V\|e} |\mathsf{P}^a_s(V-\mu) - \hat{\mathsf{P}}^a_{s,n}(V-\mu)| \\
&\quad + \max_{0\le\mu\le V+\|V\|e} \left|\left(\sqrt{\zeta \mathrm{Var}_{\mathsf{P}^a_s}(V-\mu)} - \sqrt{\zeta \mathrm{Var}_{\hat{\mathsf{P}}^a_{s,n}}(V-\mu)}\right)\right| \\
&\overset{(a)}{\le} \max_{0\le\mu\le V+\|V\|e} \|V-\mu\| \|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n}\|_1 \\
&\quad + \max_{0\le\mu\le V+\|V\|e} \sqrt{|\zeta \mathrm{Var}_{\mathsf{P}^a_s}(V-\mu) - \zeta \mathrm{Var}_{\hat{\mathsf{P}}^a_{s,n}}(V-\mu)|},
\end{aligned} \tag{291}
$$

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

$$
\begin{aligned}
|\mathrm{Var}_p[X] - \mathrm{Var}_q[X]| &= |\mathbb{E}_p[X^2] - \mathbb{E}_p[X]^2 - \mathbb{E}_q[X^2] + \mathbb{E}_q[X]^2| \\
&\le |\mathbb{E}_p[X^2] - \mathbb{E}_q[X^2]| + |(\mathbb{E}_p[X] + \mathbb{E}_q[X])(\mathbb{E}_p[X] - \mathbb{E}_q[X])| \\
&\le \sup|X^2| \|p-q\|_1 + 2(\sup|X|)^2 \|p-q\|_1.
\end{aligned} \tag{292}
$$

Hence,

$$
\sqrt{|\zeta \mathrm{Var}_{\mathsf{P}^a_s}(V-\mu) - \zeta \mathrm{Var}_{\hat{\mathsf{P}}^a_{s,n}}(V-\mu)|} \le \sqrt{3\zeta \|V-\mu\|^2 \|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n}\|_1}. \tag{293}
$$

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

$$
\mathbb{E}[|g^V_{s,a,n}(\mu^V_{s,a,n}) - g^V_{s,a}(\mu^V_{s,a})|] \le 3\|V\| \left(\frac{|\mathcal{S}|^2\sqrt{\pi}}{2^{\frac{n+1}{2}}} + \sqrt{\frac{3\zeta|\mathcal{S}|^2\sqrt{\pi}}{2^{\frac{n+1}{2}}}}\right), \tag{294}
$$

which implies that

$$
\lim_{n\to\infty} \mathbb{E}\left[g^V_{s,a,n}(\mu^V_{s,a,n})\right] = g^V_{s,a}(\mu^V_{s,a}), \tag{295}
$$

which completes the proof. □

Theorem 27. *The estimated operator $\hat{\sigma}_{\mathcal{P}^a_s}$ defined in* (281) *has bounded variance, i.e., there exists a constant C_0, such that*

$$
Var(\hat{\sigma}_{\mathcal{P}^a_s} V) \le (1 + 18(1+\sqrt{2\zeta})^2 + 2C_0)\|V\|^2. \tag{296}
$$

Proof. We have that

$$\begin{aligned}
&\mathrm{Var}(\hat{\sigma}_{\mathcal{P}_s^a} V) \\
&= \mathbb{E}[(\hat{\sigma}_{\mathcal{P}_s^a} V)^2] - \sigma_{\mathcal{P}_s^a}(V)^2 \\
&\leq \mathbb{E}\left[\left(g^V_{s,a,0}(\mu^V_{s,a,0}) + \frac{\Delta_N(V)}{p_N}\right)^2\right] + (\sigma_{\mathcal{P}_s^a}(V))^2 \\
&\leq 2\mathbb{E}\left[\left(g^V_{s,a,0}(\mu^V_{s,a,0})\right)^2\right] + 2\mathbb{E}\left[\left(\frac{\Delta_N(V)}{p_N}\right)^2\right] + (\sigma_{\mathcal{P}_s^a}(V))^2 \\
&\leq (1 + 18(1+\sqrt{2\zeta})^2)\|V\|^2 + 2\sum_{i=0}^{\infty} \frac{\mathbb{E}[(\Delta_i(V))^2]}{p_i},
\end{aligned} \tag{297}$$

where the last inequality is from Lemma 23. For any $n \geq 1$, we have that

$$\begin{aligned}
&\mathbb{E}[(\Delta_n(V))^2] \\
&= \mathbb{E}\left[\left(g^V_{s,a,n}(\mu^V_{s,a,n}) - \frac{g^V_{s,a,n,E}(\mu^V_{s,a,n,E}) + g^V_{s,a,n,O}(\mu^V_{s,a,n,O})}{2}\right)^2\right] \\
&= \mathbb{E}\left[\left(g^V_{s,a,n}(\mu^V_{s,a,n}) - g^V_{s,a}(\mu^V_{s,a}) + g^V_{s,a}(\mu^V_{s,a}) - \frac{g^V_{s,a,n,E}(\mu^V_{s,a,n,E}) + g^V_{s,a,n,O}(\mu^V_{s,a,n,O})}{2}\right)^2\right] \\
&\leq 2\mathbb{E}[(g^V_{s,a,n}(\mu^V_{s,a,n}) - g^V_{s,a}(\mu^V_{s,a}))^2] \\
&\quad + 2\mathbb{E}\left[\left(g^V_{s,a}(\mu^V_{s,a}) - \frac{g^V_{s,a,n,E}(\mu^V_{s,a,n,E}) + g^V_{s,a,n,O}(\mu^V_{s,a,n,O})}{2}\right)^2\right] \\
&\overset{(a)}{=} 2\mathbb{E}[(g^V_{s,a,n}(\mu^V_{s,a,n}) - g^V_{s,a}(\mu^V_{s,a}))^2] + 2\mathbb{E}[(g^V_{s,a,n-1}(\mu^V_{s,a,n-1}) - g^V_{s,a}(\mu^V_{s,a}))^2] \\
&\leq 18(1+\sqrt{3\zeta})^2\|V\|^2\mathbb{E}[\|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n}\|_1^2 + \|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n}\|_1] \\
&\quad + 18(1+\sqrt{3\zeta})^2\|V\|^2\mathbb{E}[\|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n-1}\|_1^2 + \|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n-1}\|_1],
\end{aligned} \tag{298}$$

where (a) is due to Lemma 21 and the last inequality follows a similar argument to (291). 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

$$\sum_{i=0}^{\infty} \frac{\mathbb{E}[(\Delta_i(V))^2]}{p_i} \leq C_0\|V\|^2. \tag{299}$$

Thus,

$$\mathrm{Var}(\hat{\sigma}_{\mathcal{P}_s^a} V) \leq (1 + 18(1+\sqrt{2\zeta})^2)\|V\|^2 + 2C_0\|V\|^2 = (1 + 18(1+\sqrt{2\zeta})^2 + 2C_0)\|V\|^2. \tag{300}$$

□

H.4 KL-Divergence Uncertainty Sets

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

$$\hat{\sigma}_{\mathcal{P}_s^a} V \triangleq -\min_{\alpha \geq 0}\left(\zeta\alpha + \alpha \log\left(e^{\frac{-V(s_1')}{\alpha}}\right)\right) + \frac{\Delta_N(V)}{p_N},$$

where

$$\begin{aligned}\Delta_N(V) = &-\min_{\alpha\geq 0}\left(\zeta\alpha+\alpha\log\left(\mathbb{E}_{\hat{\mathsf{P}}^a_{s,N+1}}\left[e^{\frac{-V}{\alpha}}\right]\right)\right)\\ &+\frac{1}{2}\min_{\alpha\geq 0}\left(\zeta\alpha+\alpha\log\left(\mathbb{E}_{\hat{\mathsf{P}}^{a,O}_{s,N+1}}\left[e^{\frac{-V}{\alpha}}\right]\right)\right)\\ &+\frac{1}{2}\min_{\alpha\geq 0}\left(\zeta\alpha+\alpha\log\left(\mathbb{E}_{\hat{\mathsf{P}}^{a,E}_{s,N+1}}\left[e^{\frac{-V}{\alpha}}\right]\right)\right).\end{aligned} \tag{301}$$

Theorem 28. *(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 $Var(\hat{\sigma}_{\mathcal{P}^a_s}V) \leq C_0(1+\|V\|^2)$.*

H.5 Wasserstein Distance Uncertainty Sets

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

Definition 2. *For any function $f : \mathcal{Z} \to \mathbb{R}$, $\lambda \geq 0$ and $x \in \mathcal{Z}$, define the regularization operator*

$$\Phi(\lambda, x) \triangleq \inf_{y\in\mathcal{Z}}(\lambda d(x,y)^l + f(y)). \tag{302}$$

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

$$\kappa_q \triangleq \inf\left(\lambda\geq 0 : \sum_{x\in\mathcal{Z}} q(x)\Phi(\lambda,x) > -\infty\right). \tag{303}$$

Lemma 16. *(Gao & Kleywegt, 2023) Consider the distributional robust optimization of a function f:*

$$\inf_{W_l(q,p)\leq\zeta} \mathbb{E}_{x\sim q}[f(x)], \tag{304}$$

and define its dual problem as

$$\sup_{\lambda\geq 0}(-\lambda\zeta^l + \sum_{x\in\mathcal{Z}} p(x)\inf_{y\in\mathcal{Z}}(f(y)+\lambda d(x,y)^l)). \tag{305}$$

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

$$\inf_{W_l(q,p)\leq\zeta} \mathbb{E}_{x\sim q}[f(x)] = \sup_{\lambda\geq 0}(-\lambda\zeta^l + \sum_{x\in\mathcal{Z}} p(x)\inf_{y\in\mathcal{Z}}(f(y)+\lambda d(x,y)^l)). \tag{306}$$

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

Lemma 17. *(Restatement of (25)) It holds that*

$$\sigma_{\mathcal{P}^a_s}(V) = \sup_{\lambda\geq 0}\left(-\lambda\zeta^l + \sum_x \mathsf{P}^a_{s,x}\inf_y(V(y)+\lambda d(x,y)^l)\right). \tag{307}$$

Proof. In our case, the regularization operator is

$$\Phi(\lambda, x) = \inf_{s\in\mathcal{S}}(\lambda d(s,x)^l + V(s)). \tag{308}$$

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

$$\sum_{x\in\mathcal{S}} \mathsf{P}_s^a(x)\Phi(\lambda,x) = \sum_{x\in\mathcal{S}} \mathsf{P}_s^a(x) \inf_{s\in\mathcal{S}}(\lambda d(s,x)^l + V(s)) \geq -\|V\| > -\infty. \tag{309}$$

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

$$\hat{\sigma}_{\mathcal{P}_s^a} V \triangleq \sup_{\lambda\geq 0}\bigg(-\lambda\zeta^l + \inf_y (V(y) + \lambda d(s_1', y)^l)\bigg) + \frac{\Delta_N(V)}{p_N} + r(s,a), \tag{310}$$

where

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

Theorem 29. *The estimated operator defined in* (310) *is unbiased, i.e.,*

$$\mathbb{E}[\hat{\sigma}_{\mathcal{P}_s^a} V] = \sigma_{\mathcal{P}_s^a}(V). \tag{311}$$

Proof. Denote the dual function (25) by g:

$$g_{s,a}^V(\lambda) = -\lambda\zeta^l + \mathbb{E}_{S\sim\mathsf{P}_s^a}[\inf_{x\in\mathcal{S}}(V(x) + \lambda d(S,x)^l)], \tag{312}$$

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

$$\lambda_{s,a}^V = \arg\max_{\lambda\geq 0}\bigg(-\lambda\zeta^l + \mathbb{E}_{S\sim\mathsf{P}_s^a}[\inf_{x\in\mathcal{S}}(V(x) + \lambda d(S,x)^l)]\bigg). \tag{313}$$

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

$$\begin{aligned}
&\mathbb{E}[\hat{\sigma}_{\mathcal{P}_s^a} V] \\
&= \mathbb{E}[g^V_{s,a,0}(\lambda^V_{s,a,0})] + \mathbb{E}\left[\frac{\Delta_N(V)}{p_N}\right] \\
&= \mathbb{E}[g^V_{s,a,0}(\lambda^V_{s,a,0})] + \sum_{n=0}^{\infty} p(N=n)\mathbb{E}\left[\frac{\Delta_N(V)}{p_N}|N=n\right] \\
&= \mathbb{E}[g^V_{s,a,0}(\lambda^V_{s,a,0})] + \sum_{n=0}^{\infty} \mathbb{E}[\Delta_n] \\
&= \mathbb{E}[g^V_{s,a,0}(\lambda^V_{s,a,0})] \\
&\quad + \sum_{n=0}^{\infty} \mathbb{E}\left[g^V_{s,a,n+1}(\lambda^V_{s,a,n+1}) - \frac{g^V_{s,a,n+1,O}(\lambda^V_{s,a,n+1,O}) + g^V_{s,a,n+1,E}(\lambda^V_{s,a,n+1,E})}{2}\right] \\
&= \mathbb{E}[g^V_{s,a,0}(\lambda^V_{s,a,0})] + \sum_{n=0}^{\infty} \mathbb{E}\left[g^V_{s,a,n+1}(\lambda^V_{s,a,n+1}) - g^V_{s,a,n}(\lambda^V_{s,a,n})\right], \qquad (314)
\end{aligned}$$

where the last inequality is from Lemma 21. The last equation can be further rewritten as

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}_{\mathcal{P}_s^a} V] &= \mathbb{E}[g^V_{s,a,0}(\lambda^V_{s,a,0})] + \sum_{n=0}^{\infty} \mathbb{E}\left[g^V_{s,a,n+1}(\lambda^V_{s,a,n+1}) - g^V_{s,a,n}(\lambda^V_{s,a,n})\right] \\
&= \lim_{n\to\infty} \mathbb{E}\left[g^V_{s,a,n}(\lambda^V_{s,a,n})\right]. \qquad (315)
\end{aligned}$$

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

$$\lim_{n\to\infty} \mathbb{E}\left[g^V_{s,a,n}(\lambda^V_{s,a,n})\right] = g^V_{s,a}(\lambda^V_{s,a}). \qquad (316)$$

For any arbitrary i.i.d. samples $\{X_i\}$ and its corresponding function $g^V_{s,a,n}$, together with Lemma 24, we have that

$$\begin{aligned}
&|g^V_{s,a,n}(\lambda^V_{s,a,n}) - g^V_{s,a}(\lambda^V_{s,a})| \\
&= |\max_{0\le\lambda\le\frac{2\|V\|}{\zeta^l}} g^V_{s,a}(\lambda) - \max_{0\le\lambda\le\frac{2\|V\|}{\zeta^l}} g^V_{s,a,n}(\lambda)| \\
&\le \max_{0\le\lambda\le\frac{2\|V\|}{\zeta^l}} |g^V_{s,a}(\lambda) - g^V_{s,a,n}(\lambda)| \\
&= \max_{0\le\lambda\le\frac{2\|V\|}{\zeta^l}} \left|\mathbb{E}_{S\sim\mathsf{P}_s^a}[\inf_{x\in\mathcal{S}}(V(x)+\lambda d(S,x)^l)] - \mathbb{E}_{S\sim\hat{\mathsf{P}}^a_{s,n}}[\inf_{x\in\mathcal{S}}(V(x)+\lambda d(S,x)^l)]\right| \\
&\le \max_{0\le\lambda\le\frac{2\|V\|}{\zeta^l}} \|\mathsf{P}_s^a - \hat{\mathsf{P}}^a_{s,n}\|_1 \sup_{x,S\in\mathcal{S}}(|V(x)+\lambda d(S,x)^l|) \\
&\le \left(1+\frac{2D^l}{\zeta^l}\right)\|V\|\|\mathsf{P}_s^a - \hat{\mathsf{P}}^a_{s,n}\|_1, \qquad (317)
\end{aligned}$$

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

$$\mathbb{E}[|g^V_{s,a,n}(\lambda^V_{s,a,n}) - g^V_{s,a}(\lambda^V_{s,a})|] \leq \left(1 + \frac{2D^l}{\zeta^l}\right)\left(\frac{|\mathcal{S}|^2\sqrt{\pi}}{2^{\frac{n+1}{2}}}\right)\|V\|, \tag{318}$$

which implies that

$$\lim_{n\to\infty} \mathbb{E}\left[g^V_{s,a,n}(\lambda^V_{s,a,n})\right] = g^V_{s,a}(\lambda^V_{s,a}). \tag{319}$$

This completes the proof. □

Theorem 30. *The estimated operator $\hat{\sigma}_{\mathcal{P}^a_s}$ defined in* (310) *has bounded variance, i.e., there exists a constant C_0, such that*

$$Var(\hat{\sigma}_{\mathcal{P}^a_s} V) \leq (3 + 2C_0)\|V\|^2. \tag{320}$$

Proof. We first have that

$$\begin{aligned}
&\mathrm{Var}(\hat{\sigma}_{\mathcal{P}^a_s} V) \\
&= \mathbb{E}[(\hat{\sigma}_{\mathcal{P}^a_s} V)^2] - \sigma_{\mathcal{P}^a_s}(V)^2 \\
&\leq \mathbb{E}\left[\left(g^V_{s,a,0}(\lambda^V_{s,a,0}) + \frac{\Delta_N(V)}{p_N}\right)^2\right] + (\sigma_{\mathcal{P}^a_s}(V))^2 \\
&\leq 2\mathbb{E}\left[\left(g^V_{s,a,0}(\lambda^V_{s,a,0})\right)^2\right] + 2\mathbb{E}\left[\left(\frac{\Delta_N(V)}{p_N}\right)^2\right] + (\sigma_{\mathcal{P}^a_s}(V))^2 \\
&\leq 3\|V\|^2 + 2\sum_{i=0}^{\infty} \frac{\mathbb{E}[(\Delta_i(V))^2]}{p_i},
\end{aligned} \tag{321}$$

where the last inequality is from Lemma 23. For any $n \geq 1$, we have that

$$\begin{aligned}
&\mathbb{E}[(\Delta_n(V))^2] \\
&= \mathbb{E}\left[\left(g^V_{s,a,n}(\lambda^V_{s,a,n}) - \frac{g^V_{s,a,n,E}(\lambda^V_{s,a,n,E}) + g^V_{s,a,n,O}(\lambda^V_{s,a,n,O})}{2}\right)^2\right] \\
&= \mathbb{E}\left[\left(g^V_{s,a,n}(\lambda^V_{s,a,n}) - g^V_{s,a}(\lambda^V_{s,a}) + g^V_{s,a}(\lambda^V_{s,a}) - \frac{g^V_{s,a,n,E}(\lambda^V_{s,a,n,E}) + g^V_{s,a,n,O}(\lambda^V_{s,a,n,O})}{2}\right)^2\right] \\
&\leq 2\mathbb{E}[(g^V_{s,a,n}(\lambda^V_{s,a,n}) - g^V_{s,a}(\lambda^V_{s,a}))^2] \\
&\quad + 2\mathbb{E}\left[\left(g^V_{s,a}(\lambda^V_{s,a}) - \frac{g^V_{s,a,n,E}(\lambda^V_{s,a,n,E}) + g^V_{s,a,n,O}(\lambda^V_{s,a,n,O})}{2}\right)^2\right] \\
&\overset{(a)}{=} 2\mathbb{E}[(g^V_{s,a,n}(\lambda^V_{s,a,n}) - g^V_{s,a}(\lambda^V_{s,a}))^2] + 2\mathbb{E}[(g^V_{s,a,n-1}(\lambda^V_{s,a,n-1}) - g^V_{s,a}(\lambda^V_{s,a}))^2] \\
&\leq 2\left(1 + \frac{2D^l}{\zeta^l}\right)^2 \|V\|^2 \mathbb{E}[\|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n}\|_1^2] + 2\left(1 + \frac{2D^l}{\zeta^l}\right)^2 \|V\|^2 \mathbb{E}[\|\mathsf{P}^a_s - \hat{\mathsf{P}}^a_{s,n-1}\|_1^2],
\end{aligned} \tag{322}$$

where (a) is due to Lemma 21 and the last inequality follows a similar argument to (318). 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

$$\sum_{i=0}^{\infty} \frac{\mathbb{E}[(\Delta_i(V))^2]}{p_i} \leq C_0 \|V\|^2. \tag{323}$$

Thus, we have that

$$\mathrm{Var}(\hat{\sigma}_{\mathcal{P}_s^a} V) \leq 3\|V\|^2 + 2C_0\|V\|^2 = (3 + 2C_0)\|V\|^2. \tag{324}$$

□

Appendix I. Case Studies for Robust RVI Q-Learning

In this section, we provide the proof of the second part of Theorem 12, 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 Section H.

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

Lemma 18. *It holds that*

$$\|V_Q\| \leq \|Q\|. \tag{325}$$

Proof. From the definition of V_Q, we have that

$$\|V_Q\| = \max_s |V_Q(s)| = \max_s |\max_a Q(s,a)| \triangleq |Q(s^*, a^*)|. \tag{326}$$

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

$$\|V_Q\| \leq \|Q\|. \tag{327}$$

□

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

Lemma 19. *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

$$\mathbb{E}[\hat{\mathbf{H}} Q(s,a)] = \mathbb{E}[r(s,a) + \hat{\sigma}_{\mathcal{P}_s^a} V_Q(s)] = r(s,a) + \sigma_{\mathcal{P}_s^a}(V_Q) = \mathbf{H} Q(s,a). \tag{328}$$

For boundedness, note that

$$\begin{aligned}
\mathrm{Var}(\hat{\mathbf{H}} Q(s,a)) &= \mathbb{E}\Big[\big(\hat{\mathbf{H}} Q(s,a) - \mathbf{H} Q(s,a)\big)^2\Big] \\
&= \mathbb{E}\big[\big(\hat{\sigma}_{\mathcal{P}_s^a} V_Q(s) - \sigma_{\mathcal{P}_s^a}(V_Q)\big)^2\big] \\
&\leq C(1 + \|V_Q\|^2) \\
&\leq C(1 + \|Q\|^2),
\end{aligned} \tag{329}$$

where the last inequality is from Lemma 18. □

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 Section H. We thus omit the proofs for this part.

Appendix J. Technical Lemmas

Lemma 20. *For a robust Bellman operator* $\mathbf{T}$*, define*

$$\mathbf{T}'V \triangleq \mathbf{T}V - f(V)e, \tag{330}$$

$$\tilde{\mathbf{T}}V \triangleq \mathbf{T}V - g_{\mathcal{P}}^{\pi}e. \tag{331}$$

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

$$\dot{x} = \mathbf{T}'x - x, \tag{332}$$

$$\dot{y} = \tilde{\mathbf{T}}y - y, \tag{333}$$

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

$$x(t) = y(t) + r(t)e, \tag{334}$$

where $r(t)$ *satisfies*

$$\dot{r}(t) = -r(t) + g_{\mathcal{P}}^{\pi} - f(y(t)). \tag{335}$$

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

$$x(t) = x_0e^{-t} + \int_0^t e^{-(t-s)}\tilde{\mathbf{T}}(x(s))ds + \left(\int_0^t e^{-(t-s)}(g_{\mathcal{P}}^{\pi} - f(x(s)))ds\right)e, \tag{336}$$

$$y(t) = x_0e^{-t} + \int_0^t e^{-(t-s)}\tilde{\mathbf{T}}(y(s))ds. \tag{337}$$

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

$$\max_i(x_i(t) - y_i(t)) \leq \int_0^t e^{-(t-s)} \max_i(\tilde{\mathbf{T}}_i(x(s)) - \tilde{\mathbf{T}}_i(y(s)))ds + \int_0^t e^{-(t-s)}(g_{\mathcal{P}}^{\pi} - f(x(s)))ds,$$

$$\min_i(x_i(t) - y_i(t)) \geq \int_0^t e^{-(t-s)} \min_i(\tilde{\mathbf{T}}_i(x(s)) - \tilde{\mathbf{T}}_i(y(s)))ds + \int_0^t e^{-(t-s)}(g_{\mathcal{P}}^{\pi} - f(x(s)))ds.$$

This hence implies that

$$\begin{aligned}\mathrm{Span}(x(t) - y(t)) &\leq \int_0^t e^{-(t-s)}\mathrm{Span}(\tilde{\mathbf{T}}(x(s)) - \tilde{\mathbf{T}}(y(s)))ds \\ &\leq \int_0^t e^{-(t-s)}\mathrm{Span}(x(s) - y(s))ds,\end{aligned} \tag{338}$$

where the last inequality is because $\tilde{\mathbf{T}}$ is non-expansive w.r.t. the span semi-norm.

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

$$\begin{aligned}\dot{r}(t)e &= \dot{x}(t) - \dot{y}(t) \\ &= \tilde{\mathbf{T}}x(t) + (g_{\mathcal{P}}^{\pi} - f(x(t)))e - x(t) - \tilde{\mathbf{T}}y(t) + y(t) \\ &= (-r(t) + g_{\mathcal{P}}^{\pi} - f(y(t)))e, \end{aligned} \tag{339}$$

where the last equation is because

$$\tilde{\mathbf{T}}x(t) = \tilde{\mathbf{T}}(y(t) + r(t)e) = \tilde{\mathbf{T}}(y(t)) + r(t)e, \tag{340}$$

$$f(x(t)) = f(y(t) + r(t)e) = f(y(t)) + r(t). \tag{341}$$

This completes the proof. □

Lemma 21. *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,*

$$\mathbb{E}[g(\hat{q}_{n+1,O})] = \mathbb{E}[g(\hat{q}_{n+1,E})] = \mathbb{E}[g(\hat{q}_n)]. \tag{342}$$

Proof. Note that

$$\hat{q}_{n+1,O}(s) = \frac{\sum_{i=1}^{2^n} \mathbf{1}_{X_{2i-1}=s}}{2^n}, \tag{343}$$

hence,

$$\begin{aligned}\mathbb{E}[g(\hat{q}_{n+1,O})] &= \sum_{p=(p_1,...,p_{|\mathcal{S}|})\in\Delta(|\mathcal{S}|)} g(p)\mathbb{P}(\hat{q}_{n+1,O} = p) \\ &= \sum_{p=(p_1,...,p_{|\mathcal{S}|})\in\Delta(|\mathcal{S}|)} \mathbb{P}\left(\frac{\sum_{i=1}^{2^n} \mathbf{1}_{X_{2i-1}=s_1}}{2^n} = p_1, ..., \frac{\sum_{i=1}^{2^n} \mathbf{1}_{X_{2i-1}=s_{|\mathcal{S}|}}}{2^n} = p_{|\mathcal{S}|}\middle| q\right) g(p), \end{aligned} \tag{344}$$

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

$$\begin{aligned}\mathbb{E}[g(\hat{q}_n)] &= \sum_{p=(p_1,...,p_{|\mathcal{S}|})\in\Delta(|\mathcal{S}|)} g(p)\mathbb{P}(\hat{q}_n = p) \\ &= \sum_{p=(p_1,...,p_{|\mathcal{S}|})\in\Delta(|\mathcal{S}|)} \mathbb{P}\left(\frac{\sum_{i=1}^{2^n} \mathbf{1}_{X_i=s_1}}{2^n} = p_1, ..., \frac{\sum_{i=1}^{2^n} \mathbf{1}_{X_i=s_{|\mathcal{S}|}}}{2^n} = p_{|\mathcal{S}|}\middle| q\right) g(p). \end{aligned} \tag{345}$$

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

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

Thus,

$$\mathbb{E}[g(\hat{q}_{n+1,O})] = \mathbb{E}[g(\hat{q}_n)]. \tag{346}$$

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

Lemma 22. *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,*

$$\mu^V_{s,a}, \mu^V_{s,a,N+1}, \mu^V_{s,a,N+1,E}, \mu^V_{s,a,N+1,O} \leq V + \|V\|e, \tag{347}$$

$$\|\mu^V_{s,a}\|, \|\mu^V_{s,a,N+1}\|, \|\mu^V_{s,a,N+1,E}\|, \|\mu^V_{s,a,N+1,O}\| \leq 2\|V\|, \tag{348}$$

$$|g^V_{s,a}(\mu^V_{s,a})|, |g^V_{s,a,N+1}(\mu^V_{s,a,N+1})| \leq 3(1+2\zeta)\|V\|,$$
$$|g^V_{s,a,N+1,E}(\mu^V_{s,a,N+1,E})|, |g^V_{s,a,N+1,O}(\mu^V_{s,a,N+1,O})| \leq 3(1+2\zeta)\|V\|. \tag{349}$$

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,

$$\begin{aligned}\mu^W_{s,a} &= \arg\max_{\mu\geq 0} \big(\mathsf{P}^a_s(W-\mu) - \zeta\mathrm{Span}(W-\mu)\big)\\ &= \arg\max_{\mu\geq 0} \big(-w + \mathsf{P}^a_s(V-\mu) - \zeta\mathrm{Span}(V-\mu)\big),\end{aligned} \tag{350}$$

which is because $\mathrm{Span}(V+ke) = \mathrm{Span}(V)$ 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 $W \geq 0$, hence $\mu^W_{s,a}$ is bounded: $\mu^W_{s,a} \leq W$, this further implies that

$$\|\mu^V_{s,a}\| = \|\mu^W_{s,a}\| \leq \|W\| \leq 2\|V\|. \tag{351}$$

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,

$$\begin{aligned}g^V_{s,a}(\mu^V_{s,a}) &= \mathsf{P}^a_s(V-\mu^V_{s,a}) - \zeta\mathrm{Span}(V-\mu^V_{s,a})\\ &\leq \|V\| + \|\mu^V_{s,a}\| + \zeta|\max_i(V(i)-\mu^V_{s,a}(i))| + \zeta|\min_i(V(i)-\mu^V_{s,a}(i))|\\ &\leq 3\|V\| + 2\zeta(\|V\| + \|\mu^V_{s,a}\|)\\ &\leq 3(1+2\zeta)\|V\|.\end{aligned} \tag{352}$$

On the other hand,

$$g^V_{s,a}(\mu^V_{s,a}) \geq g^V_{s,a}(0) = \mathsf{P}^a_s V - \zeta\mathrm{Span}(V) = \mathsf{P}^a_s V - \zeta\max_i V(i) + \zeta\min_i V(i). \tag{353}$$

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

$$\begin{aligned}&\mathsf{P}^a_s V - \zeta\max_i V(i) + \zeta\min_i V(i)\\ &= \sum_x \mathsf{P}^a_{s,x} V(x) - \zeta V(M) + \zeta V(m)\\ &\geq -\|V\| - 2\zeta\|V\|,\end{aligned} \tag{354}$$

where the last inequality is from $\|V\| \geq V(i) \geq -\|V\|$ for any entry i. Thus, combining (353) and (354) implies that

$$-(1+2\zeta)\|V\| \leq g^V_{s,a}(\mu^V_{s,a}) \leq 3(1+2\zeta)\|V\|. \tag{355}$$

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 23. *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,*

$$\mu^V_{s,a}, \mu^V_{s,a,N+1}, \mu^V_{s,a,N+1,E}, \mu^V_{s,a,N+1,O} \leq V + \|V\|e, \tag{356}$$

$$\|\mu^V_{s,a}\|, \|\mu^V_{s,a,N+1}\|, \|\mu^V_{s,a,N+1,E}\|, \|\mu^V_{s,a,N+1,O}\| \leq 2\|V\|, \tag{357}$$

$$|g^V_{s,a}(\mu^V_{s,a})|, |g^V_{s,a,N+1}(\mu^V_{s,a,N+1})| \leq 3(1+\sqrt{2\zeta})\|V\|, \tag{358}$$

$$|g^V_{s,a,N+1,O}(\mu^V_{s,a,N+1,O})|, |g^V_{s,a,N+1,E}(\mu^V_{s,a,N+1,E})| \leq 3(1+\sqrt{2\zeta})\|V\|. \tag{359}$$

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,

$$\begin{aligned}\mu^W_{s,a} &= \arg\max_{W \geq \mu \geq 0} \big(\mathsf{P}^a_s(W-\mu) - \sqrt{\zeta \mathrm{Var}_{\mathsf{P}^a_s}(W-\mu)}\big) \\ &= \arg\max_{W \geq \mu \geq 0} \big(-w + \mathsf{P}^a_s(V-\mu) - \sqrt{\zeta \mathrm{Var}_{\mathsf{P}^a_s}(V-\mu)}\big),\end{aligned} \tag{360}$$

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

$$\|\mu^V_{s,a}\| = \|\mu^W_{s,a}\| \leq \|W\| \leq 2\|V\|. \tag{361}$$

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,

$$\begin{aligned}|g^V_{s,a}(\mu^V_{s,a})| &= |\mathsf{P}^a_s(V-\mu^V_{s,a}) - \sqrt{\zeta \mathrm{Var}_{\mathsf{P}^a_s}(V-\mu^V_{s,a})}| \\ &\leq \|V\| + \|\mu^V_{s,a}\| + \sqrt{2\zeta\|V-\mu^V_{s,a}\|^2} \\ &\leq 3\|V\| + \sqrt{2\zeta}(\|V\| + \|\mu^V_{s,a}\|) \\ &\leq 3(1+\sqrt{2\zeta})\|V\|.\end{aligned} \tag{362}$$

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 24. *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,*

$$\lambda^V_{s,a}, \lambda^V_{s,a,n}, \lambda^V_{s,a,n,O}, \lambda^V_{s,a,n,E} \leq \frac{2\|V\|}{\zeta^l}, \tag{363}$$

$$|g^V_{s,a}(\lambda^V_{s,a})|, |g^V_{s,a,n}(\lambda^V_{s,a,n})|, |g^V_{s,a,n,O}(\lambda^V_{s,a,n,O})|, |g^V_{s,a,n,E}(\lambda^V_{s,a,n,E})| \leq \|V\|. \tag{364}$$

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,

$$g^V_{s,a}(\lambda) = -\lambda\zeta^l + \mathbb{E}_{S\sim \mathsf{P}^a_s}[\lambda d(S, s^S_\lambda)^l + V(s^S_\lambda)]. \tag{365}$$

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

$$-\lambda^V_{s,a}\zeta^l + \mathbb{E}_{S\sim \mathsf{P}^a_s}[\lambda^V_{s,a} d(S, s^S_{\lambda^V_{s,a}})^l + V(s^S_{\lambda^V_{s,a}})] \geq g^V_{s,a}(0) = \mathbb{E}_{S\sim \mathsf{P}^a_s}[V(s^S_0)] = \min_x V(x), \tag{366}$$

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,

$$\begin{aligned}
&\mathbb{E}_{S\sim \mathsf{P}^a_s}[\lambda^V_{s,a} d(S, s^S_{\lambda^V_{s,a}})^l + V(s^S_{\lambda^V_{s,a}})] \\
&= \sum_x \mathsf{P}^a_{s,x}\big(\lambda^V_{s,a} d(x, s^x_{\lambda^V_{s,a}})^l + V(s^x_{\lambda^V_{s,a}})\big) \\
&\overset{(a)}{\leq} \sum_x \mathsf{P}^a_{s,x}\big(\lambda^V_{s,a} d(x,x)^l + V(x)\big) \\
&= \mathbb{E}_{\mathsf{P}^a_s}[V(S)],
\end{aligned} \tag{367}$$

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 (366) and (367), then we further have that

$$\min_x V(x) \leq -\lambda^V_{s,a}\zeta^l + \mathbb{E}_{S\sim \mathsf{P}^a_s}[\lambda^V_{s,a} d(S, s^S_{\lambda^V_{s,a}})^l + V(s^S_{\lambda^V_{s,a}})] \leq -\lambda^V_{s,a}\zeta^l + \mathbb{E}_{\mathsf{P}^a_s}[V(S)]. \tag{368}$$

This implies that

$$\lambda^V_{s,a} \leq \frac{\mathbb{E}_{\mathsf{P}^a_s}[V(S)] - \min_x V(x)}{\zeta^l} \leq \frac{2\|V\|}{\zeta^l}, \tag{369}$$

and hence $\lambda^V_{s,a}$ is bounded.

On the other hand, note that $g^V_{s,a}(\lambda^V_{s,a}) = \sigma_{\mathcal{P}^a_s}[V(S)]$, hence,

$$\big|g^V_{s,a}(\lambda^V_{s,a})\big| \leq \|V\|. \tag{370}$$

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

$$\mathbf{H}'Q \triangleq \mathbf{H}Q - f(Q)e, \tag{371}$$

$$\tilde{\mathbf{H}}Q \triangleq \mathbf{H}Q - g^*_{\mathcal{P}}e. \tag{372}$$

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

$$\dot{x} = \mathbf{H}'x - x, \tag{373}$$

$$\dot{y} = \tilde{\mathbf{H}}y - y, \tag{374}$$

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 20 if we show that $\tilde{\mathbf{H}}$ is non-expansion w.r.t. the span semi-norm.

It can be shown that

$$\text{Span}(\tilde{\mathbf{H}}(Q_1) - \tilde{\mathbf{H}}(Q_2)) \leq \text{Span}(V_{Q_1} - V_{Q_2}). \tag{375}$$

Let

$$s = \arg\max_i \{\max_a Q_1(i,a) - \max_a Q_2(i,a)\}, \tag{376}$$

$$t = \arg\min_i \{\max_a Q_1(i,a) - \max_a Q_2(i,a)\}. \tag{377}$$

Then,

$$\begin{aligned}
\text{Span}(V_{Q_1} - V_{Q_2}) &= (\max_a Q_1(s,a) - \max_a Q_2(s,a)) - (\max_a Q_1(t,a) - \max_a Q_2(t,a)) \\
&\leq Q_1(s,a_s) - Q_2(s,a_s) - (Q_1(t,a_t) - Q_2(t,a_t)) \\
&\leq \max_{x,b}(Q_1(x,b) - Q_2(x,b)) - \min_{x,b}(Q_1(x,b) - Q_2(x,b)) \\
&= \text{Span}(Q_1 - Q_2).
\end{aligned} \tag{378}$$

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

References

Abdullah, M. A., Ren, H., Ammar, H. B., Milenkovic, V., Luo, R., Zhang, M., & Wang, J. (2019). Wasserstein robust reinforcement learning. *arXiv preprint arXiv:1907.13196*, *abs/1907.13196*.

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

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

Atia, G. K., Beckus, A., Alkhouri, I., & Velasquez, A. (2021). Steady-state planning in expected reward multichain mdps. *Journal of Artificial Intelligence Research*, *72*, 1029–1082.

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

Bagnell, J. A., Ng, A. Y., & Schneider, J. G. (2001). Solving uncertain Markov decision processes.. *1*.

Bertsekas, D. P. (2011). *Dynamic Programming and Optimal Control 3rd edition*, Vol. II.

Blackwell, D. (1962). Discrete Dynamic Programming. *The Annals of Mathematical Statistics*, *33*(2), 719 – 726.

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

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

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

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

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

Chatterjee, K., Goharshady, E. K., Karrabi, M., Novotny, P., & Zikelic, D. (2023). Solving long-run average reward robust mdps via stochastic games. *arXiv preprint arXiv:2312.13912*, *abs/2312.13912*.

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

Derman, C. (1970). *Finite state Markovian decision processes*. Academic Press, Inc.

Gao, R., & Kleywegt, A. (2023). Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, *48*(2), 603–655.

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

Goyal, V., & Grand-Clement, J. (2018). Robust Markov decision process: Beyond rectangularity. *arXiv preprint arXiv:1811.00215*, *abs/1811.00215*.

Grand-Clément, J., & Petrik, M. (2023). Reducing blackwell and average optimality to discounted mdps via the blackwell discount factor. *arXiv preprint arXiv:2302.00036*, *abs/2302.00036*.

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

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

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

Hordijk, A., & Yushkevich, A. A. (2002). Blackwell optimality. In *Handbook of Markov decision processes*, pp. 231–267. Springer.

Hou, L., Pang, L., Hong, X., Lan, Y., Ma, Z., & Yin, D. (2020). Robust reinforcement learning with Wasserstein constraint. *arXiv preprint arXiv:2006.00945*, *abs/2006.00945*.

Hu, Z., & Hong, L. J. (2013). Kullback-Leibler divergence constrained distributionally robust optimization. *Optimization Online*, *1*, 1695–1724.

Huang, S., Papernot, N., Goodfellow, I., Duan, Y., & Abbeel, P. (2017). Adversarial attacks on neural network policies. In *Proc. International Conference on Learning Representations (ICLR)*.

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

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

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

Kazemi, M., Perez, M., Somenzi, F., Soudjani, S., Trivedi, A., & Velasquez, A. (2022). 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.

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

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

Kos, J., & Song, D. (2017). Delving into adversarial attacks on deep policies. In *Proc. International Conference on Learning Representations (ICLR)*.

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

Liang, Z., Ma, X., Blanchet, J., Zhang, J., & Zhou, Z. (2023). Single-trajectory distributionally robust reinforcement learning. *arXiv preprint arXiv:2301.11721*, *abs/2301.11721*.

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

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

Lin, Y.-C., Hong, Z.-W., Liao, Y.-H., Shih, M.-L., Liu, M.-Y., & Sun, M. (2017). Tactics of adversarial attack on deep reinforcement learning agents. In *Proc. International Joint Conferences on Artificial Intelligence (IJCAI)*, pp. 3756–3762.

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

Mandlekar, A., Zhu, Y., Garg, A., Fei-Fei, L., & Savarese, S. (2017). Adversarially robust policy learning: Active construction of physically-plausible perturbations. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3932–3939. IEEE.

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

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

Pattanaik, A., Tang, Z., Liu, S., Bommannan, G., & Chowdhary, G. (2018). Robust deep reinforcement learning with adversarial attacks. In *Proc. International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2040–2042.

Pinto, L., Davidson, J., Sukthankar, R., & Gupta, A. (2017). Robust adversarial reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, pp. 2817–2826. PMLR.

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

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

Rajeswaran, A., Ghotra, S., Ravindran, B., & Levine, S. (2017). Epopt: Learning robust neural network policies using model ensembles. In *Proc. International Conference on Learning Representations (ICLR)*.

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

Rudin, W. (2022). *Functional Analysis* (2nd edition). McGraw-Hill Science &Engineering &Math.

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

Si, N., Zhang, F., Zhou, Z., & Blanchet, J. (2020). Distributionally robust policy evaluation and learning in offline contextual bandits. In *Proc. International Conference on Machine Learning (ICML)*, pp. 8884–8894. PMLR.

Sigaud, O., & Buffet, O. (2013). *Markov decision processes in artificial intelligence*. John Wiley & Sons.

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

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

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

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

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

Vinitsky, E., Du, Y., Parvate, K., Jang, K., Abbeel, P., & Bayen, A. (2020). Robust reinforcement learning using adversarial populations. *arXiv preprint arXiv:2008.01825*, *2008.01825*.

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

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

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

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

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

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

Wei, C.-Y., Jahromi, M. J., Luo, H., Sharma, H., & Jain, R. (2020). Model-free reinforcement learning in infinite-horizon average-reward Markov decision processes. In *International conference on machine learning*, pp. 10170–10180. PMLR.

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

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

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

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

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

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

Zhang, S., Zhang, Z., & Maguluri, S. T. (2021b). Finite sample analysis of average-reward TD learning and Q-learning. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 34, pp. 1230–1242.

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

Zhou, Z., Bai, Q., Zhou, Z., Qiu, L., Blanchet, J., & Glynn, P. (2021). 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.