Robust Average-Reward Markov Decision Processes

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

[1] University at Buffalo, The State University of New York
[2] University of Colorado Boulder
[3] University of Central Florida
[4] Air Force Research Laboratory
ywang294@buffalo.com, alvaro.velasquez@colorado.edu, george.atia@ucf.edu, ashley.prater-bennette@us.af.mil,
szou3@buffalo.edu

Abstract

In robust Markov decision processes (MDPs), the uncertainty in the transition kernel is addressed by finding a policy that optimizes the worst-case performance over an uncertainty set of MDPs. While much of the literature has focused on discounted MDPs, robust average-reward MDPs remain largely unexplored. In this paper, we focus on robust average-reward MDPs, where the goal is to find a policy that optimizes the worst-case average reward over an uncertainty set. We first take an approach that approximates average-reward MDPs using discounted MDPs. We prove that the robust discounted value function converges to the robust average-reward as the discount factor goes to 1, and moreover when it is large, any optimal policy of the robust discounted MDP is also an optimal policy of the robust average-reward. We further design a robust dynamic programming approach, and theoretically characterize its convergence to the optimum. Then, we investigate robust average-reward MDPs directly without using discounted MDPs as an intermediate step. We derive the robust Bellman equation for robust average-reward MDPs, prove that the optimal policy can be derived from its solution, and further design a robust relative value iteration algorithm that provably finds its solution, or equivalently, the optimal robust policy.

Introduction

A Markov decision process (MDP) is an effective mathematical tool for sequential decision-making 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, in practice there could exist a mismatch between the assumed MDP model and the underlying environment due to various factors, such as non-stationarity of the environment, modeling error, exogenous perturbation, partial observability, and adversarial attacks. The ensuing model mismatch could result in solution policies with poor performance.

This challenge spurred noteworthy efforts on developing and analyzing a framework of robust MDPs e.g., (Bagnell, Ng, and Schneider 2001; Nilim and El Ghaoui 2004; Iyengar 2005). Rather than adopting a fixed MDP model, in the robust MDP setting, one seeks to optimize the worst-case performance over an uncertainty set of possible MDP models. The solution to the robust MDP problem provides performance guarantee for all uncertain MDP models, and is thus robust to the model mismatch.

Robust MDP problems falling under different reward optimality criteria are fundamentally different. In robust discounted MDPs, the goal is to find a policy that maximizes the discounted cumulative reward in the worst case. In this setting, as the agent interacts with the environment, the reward received diminishes exponentially over time. Much of the prior work in the robust setting has focused on the discounted reward formulation. The model-based method, e.g., (Iyengar 2005; Nilim and El Ghaoui 2004; Bagnell, Ng, and Schneider 2001; Satia and Lave Jr 1973; Wiesemann, Kuhn, and Rustem 2013; Tamar, Mannor, and Xu 2014; Lim and Autef 2019; Xu and Mannor 2010; Yu and Xu 2015; Lim, Xu, and Mannor 2013), where information about the uncertainty set is assumed to be known to the learner, unveiled several fundamental characterizations of robust discounted MDPs. This was further extended to the more practical model-free setting in which only samples from a simulator (the centroid of the uncertainty set) are available to the learner. For example, the value-based method (Roy, Xu, and Pokutta 2017; Badrinath and Kalathil 2021; Wang and Zou 2021; Tessler, Efroni, and Mannor 2019; Zhou et al. 2021; Yang, Zhang, and Zhang 2021; Panaganti and Kalathil 2021; Goyal and Grand-Clement 2018; Kaufman and Schaefer 2013; Ho, Petrik, and Wiesemann 2018, 2021; Si et al. 2020) optimizes the worst-case performance using the robust value function as an intermediate step; on the other hand, the model-free policy-based method (Russel, Benosman, and Van Baar 2020; Derman, Geist, and Mannor 2021; Eysenbach and Levine 2021; Wang and Zou 2022) directly optimizes the policy and is thus scalable to large/continuous state and action spaces.

Although discounted MDPs induce an elegant Bellman operator that is a contraction, and have been studied extensively, the policy obtained usually has poor long-term performance when a system operates for an extended period of time. When the discount factor is very 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, Bagnell, and Peters 2013). Therefore, it is also important to optimize the long-term average performance of a system.

However, robust MDPs under the average-reward criterion are largely understudied. Compared to the discounted setting, the average-reward setting depends on the limiting behavior of the underlying stochastic process, and hence 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 dependence of 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. Heretofore, only a handful of studies have considered average-reward MDPs in the robust setting. The first work by (Tewari and 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, Xu, and Mannor 2013) proposed an algorithm for robust average-reward MDPs under the ℓ_1 uncertainty set. However, obtaining fundamental characterizations of the problem and convergence guarantee remains elusive.

Challenges and Contributions

In this paper, we derive characterizations of robust average-reward MDPs with general uncertainty sets, and develop model-based approaches with provable theoretical guarantee. Our approach is fundamentally different from previous work on robust discounted MDPs, 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* setting and for a specific 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 and 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 and Yushkevich 2002) for the non-robust setting, however, our proof is fundamentally different. Technically, the proof in (Blackwell 1962; Hordijk and 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 while increasing the discount factor at a certain rate. 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 design a robust relative value iteration method without using the discounted MDPs as an intermediate step. We further pursue a direct approach that solves the robust average-reward MDPs without using the limit method, i.e., without using discounted MDPs as an intermediate step. We derive a robust Bellman equation for robust average-reward MDPs, and show that the pair of robust relative value function and robust average-reward is a solution to the robust Bellman equation under the average-reward setting. We further prove that if we can find any solution to the robust Bellman equation, then the optimal policy can be derived by a greedy approach. The problem hence can be equivalently solved by solving the robust Bellman equation. We then design a robust value iteration method which provably converges to the solution of the robust Bellman equation, i.e., solve the optimal policy for the robust average-reward MDP problem.

Related Work

Robust discounted MDPs. Model-based methods for robust discounted MDPs were studied in, e.g., (Iyengar 2005; Nilim and El Ghaoui 2004; Bagnell, Ng, and Schneider 2001; Satia and Lave Jr 1973; Wiesemann, Kuhn, and Rustem 2013; Lim and Autef 2019; Xu and Mannor 2010; Lim, Xu, and Mannor 2013), 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 centroid MDP of the uncertainty set are available in an online fashion (Roy, Xu, and Pokutta 2017; Badrinath and Kalathil 2021; Wang and Zou 2021, 2022; Tessler, Efroni, and Mannor 2019) and an offline fashion (Zhou et al. 2021; Yang, Zhang, and Zhang 2021; Panaganti and Kalathil 2021; Goyal and Grand-Clement 2018; Ho, Petrik, and Wiesemann 2021). 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; Huang 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. However, 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 and Bartlett 2007), where the authors showed the existence of a Blackwell optimal policy, i.e., there exists some $\delta \in [0,1)$, such that the optimal robust policy exists and remains unchanged for any discount factor $\gamma \in [\delta, 1)$. However, this result depends on the structure of the uncertainty set. For general uncertainty sets, the existence of a Blackwell optimal policy may not be guaranteed. More recently, (Lim, Xu, and Mannor 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.

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 state space $\mathcal{S}$, an action space $\mathcal{A}$, a transition kernel $\mathsf{P} = \{p^a_s \in \Delta(\mathcal{S}), a \in \mathcal{A}, s \in \mathcal{S}\}^1$, where p^a_s is the distribution of the next state over $\mathcal{S}$ upon taking action a in state s (with $p^a_{s,s'}$ 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 $p^{a_t}_{s_t}$, 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 $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ 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: $V^\pi_{\mathsf{P},\gamma}(s) \triangleq \mathbb{E}_{\pi,\mathsf{P}}\left[\sum_{t=0}^{\infty} \gamma^t r_t | S_0 = s\right]$.

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 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

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

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

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

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):

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

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 and 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

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

where $\kappa = (\mathsf{P}_0, \mathsf{P}_1 ...) \in \bigotimes_{t\geq 0} \mathcal{P}$.

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

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

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

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

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 and El Ghaoui 2004; Iyengar 2005; Puterman 1994):

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

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 and El Ghaoui 2004; Iyengar 2005). 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:

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

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.

Limit Approach for Robust Average-Reward MDPs

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.,

$$\lim_{\gamma\to 1}(1-\gamma)V^\pi_{\mathsf{P},\gamma} = g^\pi_{\mathsf{P}}. \quad (8)$$

We take a similar idea, and show that the same result holds in the robust case: $\lim_{\gamma\to1}(1-\gamma)V^\pi_{\mathcal{P},\gamma} = g^\pi_{\mathcal{P}}$ under a mild assumption. Based on this result, we further design algorithms (Algorithms 1 and 2) that apply a sequence of robust discounted Bellman operators while increasing the discount factor at a certain rate. We then theoretically prove that our algorithms converge to the optimal solutions.

In the following, we first show that the convergence $\lim_{\gamma\to1}(1-\gamma)V^\pi_{\mathsf{P},\gamma} = g^\pi_{\mathsf{P}}$ 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 are considered, e.g., unichain and communicating (Wei et al. 2020; Zhang and Ross 2021; Chen, Jain, and Luo 2022; Wan, Naik, and Sutton 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 and Bartlett 2007), total-variation (Rahimian, Bayraksan, and De-Mello 2022) and KL-divergence (Hu and Hong 2013). The unichain assumption is also widely used in studies of average-reward MDPs, e.g., (Puterman 1994; Wan, Naik, and Sutton 2021; Zhang and Ross 2021; Lan 2020; Zhang, Zhang, and Maguluri 2021). Also it is worth noting that under the unichain assumption, the robust 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 $\|H^\pi_{\mathsf{P}}\|, \forall\pi\in\Pi, \mathsf{P}\in\mathcal{P}$. Assumption 1 is one sufficient condition.*

In (Puterman 1994), the convergence $\lim_{\gamma\to1}(1-\gamma)V^\pi_{\mathsf{P},\gamma} = g^\pi_{\mathsf{P}}$ 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 exchanged 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 $\gamma\to1$, i.e.,*

$$\lim_{\gamma\to1}(1-\gamma)V^\pi_{\mathsf{P},\gamma} = g^\pi_{\mathsf{P}}, \textit{ uniformly.} \quad (9)$$

With uniform convergence in Theorem 1, the order of the limit $\gamma\to1$ and $\min_{\mathsf{P}}$ can be interchanged, then the following convergence of the robust discounted value function can be established.

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

$$\lim_{\gamma\to1}(1-\gamma)V^\pi_{\mathcal{P},\gamma} = g^\pi_{\mathcal{P}} \textit{ uniformly.} \quad (10)$$

We note that a similar convergence result is shown in (Tewari and Bartlett 2007), but only for a special uncertainty set of finite interval. Our Theorem 2 holds for general compact uncertainty sets. Moreover, it is worth highlighting that our proof technique is fundamentally different from the one in (Tewari and Bartlett 2007). Specifically, under the finite interval uncertainty set, 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 previous two convergence results play a fundamental role in limit method for robust average-reward MDPs, and are of key importance to motivate the design of the following two algorithms, the basic idea of which is to apply a sequence of robust discounted Bellman operators on an arbitrary initialization while increasing the discount factor.

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 and Bartlett 2007) for a specific finite interval uncertainty set. We present the a robust value iteration (robust

Algorithm 1: Robust VI: Policy Evaluation

```
Input: $\pi, V_0(s) = 0, \forall s, T$
for $t = 0, 1, ..., T-1$ do
    $\gamma_t \leftarrow \frac{t+1}{t+2}$
    for all $s \in \mathcal{S}$ do
        $V_{t+1}(s) \leftarrow \mathbb{E}_\pi[(1-\gamma_t)r(s,A) + \gamma_t \sigma_{\mathcal{P}_s^A}(V_t)]$
    end for
end for
return $V_T$
```

VI) algorithm for evaluating the robust average-reward with general uncertainty sets in Algorithm 1. At each time step t, the discount factor γ_t is set to $\frac{t+1}{t+2}$, 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 3 *Algorithm 1 converges to robust average reward.*

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.

Similar to Algorithm 1, at each time step, the discount factor γ_t is set to be closer to 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, hence can find the optimal robust policy.

Theorem 4 *The output V_T in Algorithm 2 converges to the optimal robust average-reward $g^*_{\mathcal{P}}$: $V_T \to g^*_{\mathcal{P}}$ as $T \to \infty$.*

As discussed in (Blackwell 1962; Hordijk and 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. Towards this issue, 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 all 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.

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 $V_T, \pi_T$
```

Following a similar idea, we justify that the obtained policy from Algorithm 2 is not only optimal in the robust average-reward setting, but also sensitive to early rewards.

Denote by Π^*_D the set of all the deterministic optimal policies for robust average-reward, i.e. $\Pi^*_D = \{\pi \in \Pi_D : g^\pi_{\mathcal{P}} = g^*_{\mathcal{P}}\}$.

Theorem 5 (Blackwell optimality) *There exists $0 < \delta < 1$, such that for any $\gamma > \delta$, the deterministic optimal robust policy for robust discounted value function $V^*_{\mathcal{P},\gamma}$ belongs to Π^*_D. Moreover, when Π^*_D is a singleton, there exists a unique Blackwell optimal policy.*

This result implies that using the limit method in this section to find 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: $f_{\pi,\nu}(\gamma) \triangleq V^\pi_{\mathsf{P},\gamma} - V^\mu_{\mathsf{P},\gamma}$ 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 and Bartlett 2007) for the finite interval uncertainty set. Specifically, in (Tewari and Bartlett 2007), it was shown that the worst-case transition kernels for any π, γ are from a finite set $\mathcal{M}$, hence $f_{\pi,\nu}(\gamma) \triangleq \min_{\mathsf{P} \in \mathcal{M}} V^\pi_{\mathsf{P},\gamma} - \min_{\mathsf{P} \in \mathcal{M}} V^\mu_{\mathsf{P},\gamma}$ 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 and Bartlett 2007) inapplicable to our problem.

Direct Approach for Robust Average-Reward MDPs

The limit approach in Section is based on the uniform convergence of the 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.

For average-reward MDPs, the relative value iteration (RVI) approach (Puterman 1994) is commonly used since it is numerically stable and has convergence guarantee. In the following, we generalize the RVI algorithm to the robust setting, and design the robust RVI algorithm in Algorithm 3.

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*

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

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

The following theorem presents a robust Bellman equation for robust average-reward MDPs.

Theorem 6 *For any s and π, $(V^{\pi}_{\mathcal{P}}, g^{\pi}_{\mathcal{P}})$ is a solution to the following robust Bellman equation:*

$$V(s) + g = \sum_a \pi(a|s)\left(r(s,a) + \sigma_{\mathcal{P}^a_s}(V)\right). \quad (12)$$

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 and 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. In sharp contrast, in the average-reward setting, the robust Bellman 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.

We first derive the following equivalent optimality condition for robust average-reward MDPs.

Theorem 7 *For any (g, V) that is a solution to*

$$\max_a \left\{r(s,a) - g + \sigma_{\mathcal{P}^a_s}(V) - V(s)\right\} = 0, \forall s, \quad (13)$$

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

$$\pi^*(s) = \arg\max_a \left\{r(s,a) + \sigma_{\mathcal{P}^a_s}(V)\right\} \quad (14)$$

for any $s \in \mathcal{S}$, then π^* is an optimal robust policy.

Theorem 7 suggests that as long as we find a solution (g, V) to (13), which though may not be unique, then g is the optimal robust average-reward $g^*_{\mathcal{P}}$, and the greedy policy π^* is the optimal policy to our robust average-reward MDP problem in (7).

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 (13), and further the optimal policy could be obtained by (14). Here **1** denotes the all-ones vector, and sp denotes the span semi-norm: $sp(w) = \max_s w(s) - \min_s w(s)$. Different from Algorithm 2, in Algorithm 3, we do not need to apply the robust discounted Bellman operator. The method directly solves the robust optimal control problem for average-reward robust MDPs.

Algorithm 3: Robust RVI

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

To study the convergence of the robust RVI algorithm, we first make an additional assumption as follows.

Assumption 2 *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 unichain and aperiodic (Bertsekas 2011). It can be also replaced using some weaker ones, e.g., Proposition 4.3.2 of (Bertsekas 2011), or be removed by designing a variant of RVI, e.g., Proposition 4.3.4 of (Bertsekas 2011). In the following theorem, we show that our Algorithm 3 converges to a solution of (13), hence according to Theorem 7 if we set π according to (14), then π is the optimal robust policy.

Theorem 8 *(w_t, V_t) converges to a solution (w, V) to* (13) *as $\epsilon \to 0$.*

Remark 2 *In this section, we mainly present the robust RVI algorithm for the robust optimal control problem, and its 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 8 can also be similarly derived.*

Examples and Numerical Results

In this section, 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 Algorithms 1,2,3, for different uncertainty sets, the only difference lies in how the support function $\sigma_{\mathcal{P}^a_s}(V)$ 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 methods v.s. non-robust (relative) value iteration method on different uncertainty sets. Our experiments are based on the Garnet problem $\mathcal{G}(20, 40)$ (Archibald, McKinnon, and Thomas 1995). More specifically, there are 20 states and 30 actions; the nominal transition kernel

$\mathsf{P} = \{p_s^a \in \Delta(\mathcal{S})\}$ is randomly generated according to the uniform distribution, and the reward functions $r(s,a) \sim \mathcal{N}(0, \sigma_{s,a})$, where $\sigma_{s,a} \sim \text{Uniform}[0,1]$. In our experiments, the uncertainty sets are designed to be centered at the nominal transition kernel. 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. For any (s,a) the uncertainty set $\mathcal{P}_s^a$ is defined as $\mathcal{P}_s^a = \{q : q = (1-R)p_s^a + Rp', p' \in \Delta(\mathcal{S})\}$, where p_s^a is the nominal transition kernel. It can be viewed as an adversarial model, where at each time-step, the environment transits according to the nominal transition kernel p with probability $1-R$, and according to an arbitrary kernel p' with probability R. Note that $\sigma_{\mathcal{P}_s^a}(V) = (1-R)(p_s^a)^\top V + R \min_s V(s)$. Our experimental results under the contamination model are shown in Fig 1.

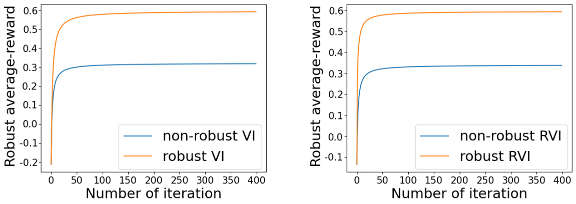

Figure 1: Comparison on contamination model, $R = 0.4$.

Total variation. The total variation distance is another commonly used distance metric to measure the difference between two distributions. For two distributions p and q, it is defined as $D_{TV}(p,q) = \frac{1}{2}\|p-q\|_1$. Consider an uncertainty set defined via total variation: $\mathcal{P}_s^a = \{q : D_{TV}(q\|p_s^a) \leq R\}$. Then, its support function can be efficiently solved as follows (Iyengar 2005): $\sigma_{\mathcal{P}_s^a}(V) = p^\top V - R\min_{\mu \geq 0}\{\max_s(V(s) - \mu(s)) - \min_s(V(s) - \mu(s))\}$.

Our experimental results under the total variation model are shown in Fig 2.

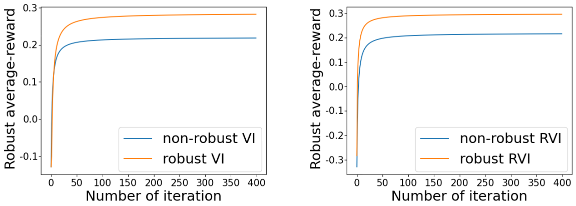

Figure 2: Comparison on total variation model, $R = 0.6$.

Kullback-Lerbler (KL) divergence. The Kullback–Leibler divergence is widely used to measure the distance between two probability distributions. For distributions p, q, it is defined as $D_{KL}(q\|p) = \sum_s q(s)\log\frac{q(s)}{p(s)}$. Consider an uncertainty set defined via KL divergence: $\mathcal{P}_s^a = \{q : D_{KL}(q\|p_s^a) \leq R\}$. Then, its support function can be efficiently solved using the duality result in (Hu and Hong 2013): $\sigma_{\mathcal{P}_s^a}(V) = -\min_{\alpha \geq 0}\left\{R\alpha + \alpha\log\left(p^\top e^{\frac{-V}{\alpha}}\right)\right\}$. Our experimental results under the KL-divergence model are shown in Fig 3.

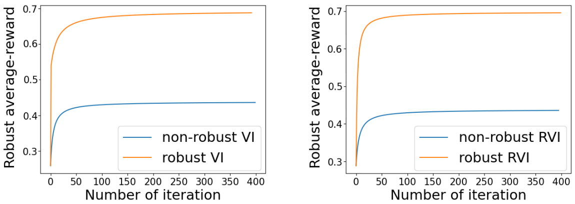

Figure 3: Comparison on KL-divergence model, $R = 0.8$.

It can be seen that our robust methods can obtain policies that achieve higher worst-case reward. Also, both our limit-based robust value iteration and our direct method of robust relative value iteration converge to the optimal robust policies, which validates our theoretical results.

Conclusion

In this paper, we investigated the problem of robust MDPs under the average-reward setting. We established *uniform* convergence of the discounted value function to average-reward, which further implies the uniform convergence of the *robust* discounted value function to *robust* average-reward. Based on this insight, we designed a robust dynamic programming approach using the robust discounted MDPs as an approximation (the limit method). We theoretically proved their convergence and optimality and proved a robust version of the Blackwell optimality (Blackwell 1962). We then designed a direct approach for robust average-reward MDPs, where we derived the robust Bellman equation for robust average-reward MDPs. We further designed a robust RVI method, which was proven to converge to the optimal robust solution. Technically, our proof techniques are fundamentally different from existing studies on average-reward robust MDPs, e.g., those in (Blackwell 1962; Tewari and Bartlett 2007).

Acknowledgments

This work was supported by the National Science Foundation under Grants CCF-2106560, CCF-2007783, CCF-2106339, and CCF-1552497.

References

Abdullah, M. A.; Ren, H.; Ammar, H. B.; Milenkovic, V.; Luo, R.; Zhang, M.; and Wang, J. 2019. Wasserstein robust reinforcement learning. *arXiv preprint arXiv:1907.13196*.

Archibald, T.; McKinnon, K.; and 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.; and Velasquez, A. 2021. Steady-State Planning in Expected Reward Multichain MDPs. *Journal of Artificial Intelligence Research*, 72: 1029–1082.

Badrinath, K. P.; and Kalathil, D. 2021. Robust Reinforcement Learning using Least Squares Policy Iteration with Provable Performance Guarantees. In *Proc. International Conference on Machine Learning (ICML)*, 511–520. PMLR.

Bagnell, J. A.; Ng, A. Y.; and Schneider, J. G. 2001. Solving uncertain Markov decision processes. *Robotics Commons*.

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

Blackwell, D. 1962. Discrete dynamic programming. *The Annals of Mathematical Statistics*, 719–726.

Chen, L.; Jain, R.; and Luo, H. 2022. Learning Infinite-Horizon Average-Reward Markov Decision Processes with Constraints. *arXiv preprint arXiv:2202.00150*.

Derman, C. 1970. *Finite state Markovian decision processes*. Academic Press, Inc.

Derman, E.; Geist, M.; and Mannor, S. 2021. Twice regularized MDPs and the equivalence between robustness and regularization. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*.

Eysenbach, B.; and Levine, S. 2021. Maximum entropy RL (provably) solves some robust RL problems. *arXiv preprint arXiv:2103.06257*.

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

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

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

Hordijk, A.; and Yushkevich, A. A. 2002. Blackwell optimality. In *Handbook of Markov decision processes*, 231–267. Springer.

Hou, L.; Pang, L.; Hong, X.; Lan, Y.; Ma, Z.; and Yin, D. 2020. Robust Reinforcement Learning with Wasserstein Constraint. *arXiv preprint arXiv:2006.00945*.

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

Huang, S.; Papernot, N.; Goodfellow, I.; Duan, Y.; and 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.; and Schaefer, A. J. 2013. Robust modified policy iteration. *INFORMS Journal on Computing*, 25(3): 396–410.

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

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

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

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

Mandlekar, A.; Zhu, Y.; Garg, A.; Fei-Fei, L.; and 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)*, 3932–3939. IEEE.

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

Panaganti, K.; and Kalathil, D. 2021. Sample Complexity of Robust Reinforcement Learning with a Generative Model. *arXiv preprint arXiv:2112.01506*.

Pattanaik, A.; Tang, Z.; Liu, S.; Bommannan, G.; and Chowdhary, G. 2018. Robust Deep Reinforcement Learning with Adversarial Attacks. In *Proc. International Conference on Autonomous Agents and MultiAgent Systems*, 2040–2042.

Pinto, L.; Davidson, J.; Sukthankar, R.; and Gupta, A. 2017. Robust adversarial reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, 2817–2826. PMLR.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc.

Rahimian, H.; Bayraksan, G.; and 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.

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

Russel, R. H.; Benosman, M.; and Van Baar, J. 2020. Robust Constrained-MDPs: Soft-Constrained Robust Policy Optimization under Model Uncertainty. *arXiv preprint arXiv:2010.04870*.

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

Si, N.; Zhang, F.; Zhou, Z.; and Blanchet, J. 2020. Distributionally robust policy evaluation and learning in offline contextual bandits. In *Proc. International Conference on Machine Learning (ICML)*, 8884–8894. PMLR.

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

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

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

Vinitsky, E.; Du, Y.; Parvate, K.; Jang, K.; Abbeel, P.; and Bayen, A. 2020. Robust Reinforcement Learning using Adversarial Populations. *arXiv preprint arXiv:2008.01825*.

Wan, Y.; Naik, A.; and Sutton, R. S. 2021. Learning and planning in average-reward markov decision processes. In *International Conference on Machine Learning*, 10653–10662. PMLR.

Wang, Y.; and Zou, S. 2021. Online Robust Reinforcement Learning with Model Uncertainty. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, Y.; and Zou, S. 2022. Policy Gradient Method For Robust Reinforcement Learning. In *Proc. International Conference on Machine Learning (ICML)*, volume 162, 23484–23526. PMLR.

Wei, C.-Y.; Jahromi, M. J.; Luo, H.; Sharma, H.; and Jain, R. 2020. Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *International conference on machine learning*, 10170–10180. PMLR.

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

Xu, H.; and Mannor, S. 2010. Distributionally Robust Markov Decision Processes. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2505–2513.

Yang, W.; Zhang, L.; and Zhang, Z. 2021. Towards Theoretical Understandings of Robust Markov Decision Processes: Sample Complexity and Asymptotics. *arXiv preprint arXiv:2105.03863*.

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

Zhang, S.; Zhang, Z.; and Maguluri, S. T. 2021. Finite Sample Analysis of Average-Reward TD Learning and Q-Learning. *Advances in Neural Information Processing Systems*, 34: 1230–1242.

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

Zhou, Z.; Bai, Q.; Zhou, Z.; Qiu, L.; Blanchet, J.; and Glynn, P. 2021. Finite-Sample Regret Bound for Distributionally Robust Offline Tabular Reinforcement Learning. In *Proc. International Conference on Artifical Intelligence and Statistics (AISTATS)*, 3331–3339. PMLR.