# Finite-Sample Analysis of Policy Evaluation for Robust Average Reward Reinforcement Learning

**Yang Xu**
Purdue University
West Lafayette, IN 47907, USA
xu1720@purdue.edu

**Washim Uddin Mondal**
Indian Institute of Technology Kanpur
Kanpur, UP, India 208016
wmondal@iitk.ac.in

**Vaneet Aggarwal**
Purdue University
West Lafayette, IN 47907, USA
vaneet@purdue.edu

## Abstract

We present the first finite-sample analysis of policy evaluation in robust average-reward Markov Decision Processes (MDPs). Prior work in this setting have established only asymptotic convergence guarantees, leaving open the question of sample complexity. In this work, we address this gap by showing that the robust Bellman operator is a contraction under a carefully constructed semi-norm, and developing a stochastic approximation framework with controlled bias. Our approach builds upon Multi-Level Monte Carlo (MLMC) techniques to estimate the robust Bellman operator efficiently. To overcome the infinite expected sample complexity inherent in standard MLMC, we introduce a truncation mechanism based on a geometric distribution, ensuring a finite expected sample complexity while maintaining a small bias that decays exponentially with the truncation level. Our method achieves the order-optimal sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2})$ for robust policy evaluation and robust average reward estimation, marking a significant advancement in robust reinforcement learning theory.

## 1 Introduction

Reinforcement learning (RL) has achieved notable success in domains such as robotics [33], finance [17], healthcare [44], transportation [1], and large language models [30] by enabling agents to learn optimal decision-making strategies through interaction with an environment. However, in many real-world applications, direct interaction is impractical due to safety concerns, high costs, or limited data collection budgets [32, 19]. This challenge is particularly evident in scenarios where agents are trained in simulated environments before being deployed in the real world, such as in robotic control and autonomous driving. The mismatch between simulated and real environments, known as the simulation-to-reality gap, often leads to performance degradation when the learned policy encounters unmodeled uncertainties. Robust reinforcement learning (robust RL) addresses this challenge by formulating the learning problem as an optimization over an uncertainty set of transition probabilities, ensuring reliable performance under worst-case conditions. In this work, we focus on the problem of evaluating the robust value function and robust average reward for a given policy using only data sampled from a simulator (nominal model), aiming to enhance generalization and mitigate the impact of transition uncertainty in real-world deployment.

Reinforcement learning problems under infinite time horizons are typically studied under two primary reward formulations: the discounted-reward setting, where future rewards are exponentially

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

discounted, and the average-reward setting, which focuses on optimizing long-term performance. While the discounted-reward formulation is widely used, it may lead to myopic policies that underperform in applications requiring sustained long-term efficiency, such as queueing systems, inventory management, and network control. In contrast, the average-reward setting is more suitable for environments where decisions impact long-term operational efficiency. Despite its advantages, robust reinforcement learning under the average-reward criterion remains largely unexplored. Existing works on robust average-reward RL primarily provide asymptotic guarantees [37, 39, 38], lacking algorithms with finite-time performance bounds. This gap highlights the need for principled approaches that ensure robustness against model uncertainties while maintaining strong long-term performance guarantees.

Solving the robust average-reward reinforcement learning problem is significantly more challenging than its non-robust counterpart, with the primary difficulty arising in policy evaluation. Specifically, the goal is to compute the worst-case value function and worst-case average reward over an entire uncertainty set of transition models while having access only to samples from a nominal transition model. In this paper, we investigate three types of uncertainty sets: Contamination uncertainty sets, total variation (TV) distance uncertainty sets, and Wasserstein distance uncertainty sets. Unlike the standard average-reward setting, where value functions and average rewards can be estimated directly from observed trajectories [41, 2, 12, 13, 14], the robust setting introduces an additional layer of complexity due to the need to optimize against adversarial transitions. Consequently, conventional approaches based on direct estimation such as [41, 2, 12, 13, 14] immediately fail, as they do not account for the worst-case nature of the problem. Overcoming this challenge requires new algorithmic techniques that can infer the worst-case dynamics using only limited samples from the nominal model.

### 1.1 Challenges and Contributions

A common approach to policy evaluation in robust RL is to solve the corresponding robust Bellman operator. However, robust average-reward RL presents additional difficulties compared to the robust discounted-reward setting. In the discounted case, the presence of a discount factor induces a contraction property in the robust Bellman operator [40, 46], facilitating stable iterative updates. In contrast, the average-reward Bellman operator lacks a contraction property with respect to any norm even in the non-robust setting [45], making standard fixed-point analysis inapplicable. Due to this fundamental limitation, existing works on robust average-reward RL such as [39] rely on asymptotic techniques, primarily leveraging ordinary differential equation (ODE) analysis to examine the behavior of temporal difference (TD) learning. These methods exploit the asymptotic stability of the corresponding ODE [6] to establish almost sure convergence but fail to provide finite-sample performance guarantees. Addressing this limitation requires novel analytical tools and algorithmic techniques capable of providing explicit finite-sample bounds for robust policy evaluation and optimization.

In this work, we first establish and exploit a key structural property of the robust average-reward Bellman operator with uncertainty set $\mathcal{P}$ under the ergodicity of the nominal model: it is a contraction under some semi-norm, denoted as $\|\cdot\|_{\mathcal{P}}$, where the detailed construction is specified in Theorem 4.2 and (15). Constructing $\|\cdot\|_{\mathcal{P}}$ is not straightforward, because ergodicity alone only guarantees that the chain mixes over multiple steps and fails to produce a single-step contraction for familiar measures such as the span semi-norm. To overcome this, we group together all the worst-case transition dynamics under uncertainty into one compact family of linear mappings, and observe that their "worst-case gain" over any number of steps stays strictly below 1. From this we build an extremal norm, which by construction shrinks every non-constant component by the same fixed factor in a single step. Finally, we add a small "quotient" correction that exactly annihilates constant shifts, producing a semi-norm that vanishes only on constant functions but still inherits the one-step shrinkage. The above construction yields a uniform, strict contraction for the robust Bellman operator.

This fundamental result above enables the use of stochastic approximation techniques similar to [45] to analyze and bound the error in policy evaluation, overcoming the lack of a standard contraction property that has hindered prior finite-sample analyses. Building on this insight, we develop a novel stochastic approximation framework tailored to the robust average-reward setting. Our approach simultaneously estimates both the robust value function and the robust average reward, leading to

an efficient iterative procedure for solving the robust Bellman equation. A critical challenge in this framework under TV and Wasserstein distance uncertainty sets is accurately estimating the worst-case transition effects, which requires computing the support function of the uncertainty set. While previous works [4, 5, 39] have leveraged Multi-Level Monte Carlo (MLMC) for this task, their MLMC-based estimators suffer from infinite expected sample complexity due to the unbounded nature of the required geometric sampling, leading to only asymptotic convergence. To address this, we introduce a truncation mechanism based on a truncated geometric distribution, ensuring that the sample complexity remains finite while maintaining an exponentially decaying bias. With these techniques, we derive the first finite-sample complexity guarantee for policy evaluation in robust average-reward RL, achieving an optimal $\tilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity bound. The main contributions of this paper are summarized as follows:

- We prove that under the ergodicity assumption of the nominal model, the robust average-reward Bellman operator is a contraction with respect to a suitably constructed semi-norm (Theorem 4.2). This key result enables the application of stochastic approximation techniques for policy evaluation.
- We prove the convergence of stochastic approximation under the semi-norm contraction and under i.i.d. with noise with non-zero bias (Theorem B.1) as an intermediate result.
- We develop an efficient method for computing estimates for the robust Bellman operator under TV distance and Wasserstein distance uncertainty sets. By modifying MLMC with a truncated geometric sampling scheme, we ensure finite expected sample complexity while keeping variance controlled and bias decaying exponentially with truncation level (Theorem 5.1-5.4).
- We propose a novel temporal difference learning method that iteratively updates the robust value function and the robust average reward, facilitating efficient policy evaluation in robust average-reward RL. We establish the first non-asymptotic sample complexity result for policy evaluation in robust average-reward RL, proving an order-optimal $\tilde{\mathcal{O}}(\epsilon^{-2})$ complexity for policy evaluation (Theorem 6.1), along with a $\tilde{\mathcal{O}}(\epsilon^{-2})$ complexity for robust average-reward estimation (Theorem 6.2).

## 2 Related Work

The theoretical guarantees of robust average-reward reinforcement learning have been studied by the following works. [37] takes a model-based perspective, approximating robust average-reward MDPs with discounted MDPs and proving uniform convergence of the robust discounted value function as the discount factor approaches one, employing dynamic programming and Blackwell optimality arguments to characterize optimal policies. [39] proposes a model-free approach by developing robust relative value iteration (RVI) TD and Q-learning algorithms, proving their almost sure convergence using stochastic approximation, martingale theory, and Multi-Level Monte Carlo estimators to handle non-linearity in the robust Bellman operator. While these studies provide fundamental insights into robust average-reward RL, they do not establish explicit convergence rate guarantees due to the lack of contraction properties in the robust Bellman operator. In addition, [31, 35] study the policy optimization of average-reward robust MDPs assuming direct queries of the sub-gradient information.

Policy evaluation in robust discounted-reward reinforcement learning with finite sample guarantees has been extensively studied, with the key recent works [40, 46, 25, 24, 23] focusing on solving the robust Bellman equation by finding its fixed-point solution. This approach is made feasible by the contraction property of the robust Bellman operator under the sup-norm, which arises due to the presence of a discount factor $\gamma < 1$. However, this fundamental approach does not directly extend to the robust average-reward setting, where the absence of a discount factor removes the contraction property under any norm. As a result, existing robust discounted methods cannot be applied in the robust average-reward RL setting.

Recently, a growing body of concurrent work has established finite-sample guarantees for robust average-reward reinforcement learning. Model-based approaches include [29, 10], and [28] develops a model-free value-iteration method under contamination and $\ell_p$-ball uncertainty sets. While these results significantly advance the area, the specific problem of policy evaluation in robust average-reward MDPs has not yet been addressed in terms of sample complexity. Our work targets this gap.

## 3 Formulation

### 3.1 Robust average-reward MDPs.

For a robust MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$ while $|\mathcal{S}| = S$ and $|\mathcal{A}| = A$, the transition kernel is assumed to be in some uncertainty set $\mathcal{P}$. At each time step, the environment transits to the next state according to an arbitrary transition kernel $\mathsf{P} \in \mathcal{P}$. In this paper, we focus on the $(s,a)$-rectangular compact uncertainty set [27, 22], i.e., $\mathcal{P} = \bigotimes_{s,a} \mathcal{P}_s^a$, where $\mathcal{P}_s^a \subseteq \Delta(\mathcal{S})$, and $\Delta$ denotes the probability simplex. Popular uncertainty sets include those defined by the contamination model [21, 40], total variation [26], and Wasserstein distance [15].

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

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

where $\kappa = (\mathsf{P}_0, \mathsf{P}_1 ...) \in \bigotimes_{n\geq 0} \mathcal{P}$. It was shown in [37] that the worst case under the time-varying model is equivalent to the one under the stationary model:

$$g_{\mathcal{P}}^{\pi}(s) = \min_{\mathsf{P}\in\mathcal{P}} \lim_{T\to\infty} \mathbb{E}_{\pi,\mathsf{P}}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_t | S_0 = s\right]. \tag{2}$$

Therefore, we limit our focus to the stationary model. We refer to the minimizers of (2) as the worst-case transition kernels for the policy $\pi$, and denote the set of all possible worst-case transition kernels by $\Omega_g^{\pi}$, i.e., $\Omega_g^{\pi} \triangleq \{\mathsf{P} \in \mathcal{P} : g_{\mathsf{P}}^{\pi} = g_{\mathcal{P}}^{\pi}\}$, where $g_{\mathsf{P}}^{\pi}$ denotes the average reward of policy $\pi$ under the single transition $\mathsf{P} \in \mathcal{P}$:

$$g_{\mathsf{P}}^{\pi}(s) \triangleq \lim_{T\to\infty} \mathbb{E}_{\pi,\mathsf{P}}\left[\frac{1}{T}\sum_{n=0}^{T-1} r_t | S_0 = s\right]. \tag{3}$$

We focus on the model-free setting, where only samples from the nominal MDP denoted as $\tilde{\mathsf{P}}$ (the centroid of the uncertainty set) are available. We investigate the problem of robust policy evaluation and robust average reward estimation, which means for a given policy $\pi$, we aim to estimate the robust value function and the robust average reward. Throughout this paper, we make the following standard assumption regarding the structure of the induced Markov chain.

**Assumption 3.1.** The Markov chain induced by $\pi$ is irreducible and aperiodic for the nominal model $\tilde{\mathsf{P}}$.

In contrast to many current works on robust average-reward RL [37, 39, 38, 31, 28], Assumption 3.1 requires only that the center of the uncertainty set be irreducible and aperiodic. We note that when the radius of uncertainty sets is small enough, Assumption 3.1 can ensure that $\mathsf{P}^{\pi}$ is irreducible and aperiodic for all $\mathsf{P} \in \mathcal{P}$. This ensures that, under any transition model within the uncertainty set, the policy $\pi$ induces a single recurrent communicating class. A well-known result in average-reward MDPs states that under Assumption 3.1, the average reward is independent of the starting state, i.e., for any $\mathsf{P} \in \mathcal{P}$ and all $s, s' \in \mathcal{S}$, we have $g_{\mathsf{P}}^{\pi}(s) = g_{\mathsf{P}}^{\pi}(s')$. Thus, we can drop the dependence on the initial state and simply write $g_{\mathsf{P}}^{\pi}$ as the robust average reward. We now formally define the robust value function $V_{\mathsf{P}_V}^{\pi}$ by connecting it with the following robust Bellman equation:

**Theorem 3.2** (Robust Bellman Equation, Theorem 3.1 in [39])**.** *If $(g, V)$ is a solution to the robust Bellman equation*

$$V(s) = \sum_a \pi(a|s)\big(r(s,a) - g + \sigma_{\mathcal{P}_s^a}(V)\big), \quad \forall s \in \mathcal{S}, \tag{4}$$

*where $\sigma_{\mathcal{P}_s^a}(V) = \min_{p\in\mathcal{P}_s^a} p^{\top} V$ is denoted as the support function, then the scalar $g$ corresponds to the robust average reward, i.e., $g = g_{\mathcal{P}}^{\pi}$, and the worst-case transition kernel $\mathsf{P}_V$ belongs to the set of minimizing transition kernels, i.e., $\mathsf{P}_V \in \Omega_g^{\pi}$, where $\Omega_g^{\pi} \triangleq \{\mathsf{P} \in \mathcal{P} : g_{\mathsf{P}}^{\pi} = g_{\mathcal{P}}^{\pi}\}$. Furthermore, the function $V$ is unique up to an additive constant, where if $V$ is a solution to the Bellman equation,*

*then we have* $V = V^{\pi}_{\mathsf{P}_V} + c\mathbf{e}$*, where* $c \in \mathbb{R}$ *and* $\mathbf{e}$ *is the all-ones vector in* $\mathbb{R}^S$*, and* $V^{\pi}_{\mathsf{P}_V}$ *is defined as the relative value function of the policy* $\pi$ *under the single transition* $\mathsf{P}_V$ *as follows:*

$$V^{\pi}_{\mathsf{P}_V}(s) \triangleq \mathbb{E}_{\pi,\mathsf{P}_V}\left[\sum_{t=0}^{\infty}(r_t - g^{\pi}_{\mathsf{P}_V})|S_0 = s\right]. \tag{5}$$

Theorem 3.2 implies that the robust Bellman equation (4) identifies both the worst-case average reward $g$ and a corresponding value function $V$ that is determined only up to an additive constant. In particular, $\sigma_{\mathcal{P}^a_s}(V)$ represents the worst-case transition effect over the uncertainty set $\mathcal{P}^a_s$. Unlike the robust discounted case, where the contraction property of the Bellman operator under the sup-norm enables straightforward fixed-point iteration, the robust average-reward Bellman equation does not induce contraction under any norm, making direct iterative methods inapplicable. Throughout the paper, we denote $\mathbf{e}$ as the all-ones vector in $\mathbb{R}^S$. We now characterize the explicit forms of $\sigma_{\mathcal{P}^a_s}(V)$ for different compact uncertainty sets as follows:

**Contamination Uncertainty Set** The contamination uncertainty models outliers or rare faults [7]. Specifically, the $\delta$-contamination uncertainty set is $\mathcal{P}^a_s = \{(1-\delta)\tilde{\mathsf{P}}^a_s + \delta q : q \in \Delta(\mathcal{S})\}$, where $0 < \delta < 1$ is the radius. Under this uncertainty set, the support function can be computed as

$$\sigma_{\mathcal{P}^a_s}(V) = (1-\delta)(\tilde{\mathsf{P}}^a_s)^{\top}V + \delta \min_s V(s), \tag{6}$$

and this is linear in the nominal transition kernel $\tilde{\mathsf{P}}^a_s$.

**Total Variation Uncertainty Set.** The total variation (TV) distance uncertainty set models categorical misspecification or discretization error [18], and is characterized as $\mathcal{P}^a_s = \{q \in \Delta(|\mathcal{S}|) : \frac{1}{2}\|q - \tilde{\mathsf{P}}^a_s\|_1 \le \delta\}$, define $\|\cdot\|_{\mathrm{sp}}$ as the span semi-norm and the support function can be computed using its dual function [22]:

$$\sigma_{\mathcal{P}^a_s}(V) = \max_{\mu \ge \mathbf{0}} \big((\tilde{\mathsf{P}}^a_s)^{\top}(V-\mu) - \delta\|V-\mu\|_{\mathrm{sp}}\big). \tag{7}$$

**Wasserstein Distance Uncertainty Sets.** The Wasserstein distance uncertainty Models smooth model drift when states have a geometry [11]. 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 $\|d\|_{\mu,l} = \big(\mathbb{E}_{(X,Y)\sim\mu}\big[d(X,Y)^l\big]\big)^{1/l}$. The Wasserstein distance uncertainty set is then defined as

$$\mathcal{P}^a_s = \left\{q \in \Delta(\mathcal{S}) : W_l(\tilde{\mathsf{P}}^a_s, q) \le \delta\right\}. \tag{8}$$

The support function w.r.t. the Wasserstein distance set, can be calculated as follows [15]:

$$\sigma_{\mathcal{P}^a_s}(V) = \sup_{\lambda\ge 0}\left(-\lambda\delta^l + \mathbb{E}_{\tilde{\mathsf{P}}^a_s}\big[\inf_y\big(V(y) + \lambda d(S,y)^l\big)\big]\right). \tag{9}$$

### 3.2 Robust Bellman Operator

Motivated by Theorem 3.2, we define the robust Bellman operator, which forms the basis for our policy evaluation procedure.

**Definition 3.3** (Robust Bellman Operator, [39])**.** The robust Bellman operator $\mathbf{T}_g$ is defined as:

$$\mathbf{T}_g(V)(s) = \sum_a \pi(a|s)\big[r(s,a) - g + \sigma_{\mathcal{P}^a_s}(V)\big], \quad \forall s \in \mathcal{S}. \tag{10}$$

The operator $\mathbf{T}_g$ transforms a value function $V$ by incorporating the worst-case transition effect. A key challenge in solving the robust Bellman equation is that $\mathbf{T}_g$ does not satisfy contraction under standard norms, preventing the use of conventional fixed-point iteration. To cope with this problem, we establish that $\mathbf{T}_g$ is a contraction under some constructed semi-norm. This allows us to further develop provably efficient stochastic approximation algorithms.

## 4 Semi-Norm Contraction of Robust Bellman Operators

Under Assumption 3.1, we are able to establish the semi-norm contraction property. For motivation, we first establish the semi-norm contraction property of the non-robust average-reward Bellman operator for a policy $\pi$ under transition $\mathsf{P}$ defined as follows:

$$\mathbf{T}_g^{\mathsf{P}}(V)(s) = \sum_a \pi(a|s)\big[r(s,a) - g + \sum_{s'} \mathsf{P}(s'|s,a)V(s')\big], \quad \forall s \in \mathcal{S}. \tag{11}$$

**Lemma 4.1.** *Let $\mathcal{S}$ be a finite state space, and let $\pi$ be a stationary policy. If the Markov chain induced by $\pi$ under the transition $\mathsf{P}$ is irreducible and aperiodic, there exists a semi-norm $\|\cdot\|_{\mathsf{P}}$ with kernel $\{c\mathbf{e} : c \in \mathbb{R}\}$ and a constant $\beta \in (0,1)$ such that for all $V_1, V_2 \in \mathbb{R}^S$ and any $g \in \mathbb{R}$,*

$$\big\|\mathbf{T}_g^{\mathsf{P}}(V_1) - \mathbf{T}_g^{\mathsf{P}}(V_2)\big\|_{\mathsf{P}} \le \beta\|V_1 - V_2\|_{\mathsf{P}}. \tag{12}$$

**Proof Sketch** Under ergodicity, the one-step transition matrix (denoted as $\mathsf{P}^\pi$) has a unique stationary distribution $d^\pi$, define the stationary projector $E = \mathbf{e}^\top d^\pi$, then the fluctuation matrix (defined as $Q^\pi = \mathsf{P}^\pi - E$) has all eigenvalues strictly inside the unit circle. Standard finite-dimensional theory (via the discrete Lyapunov equation [20]) would produce a norm $\|\cdot\|_Q$ on $\mathbb{R}^S$ such that there is a constant $\alpha \in (0,1)$ such that for any $x \in \mathbb{R}^S$, $\|Q^\pi x\|_Q \le \alpha\|x\|_Q$. We then build the semi-norm as follows:

$$\|x\|_{\mathsf{P}} = \|Q^\pi x\|_Q + \epsilon \inf_{c\in\mathbb{R}} \|x - c\mathbf{e}\|_Q, \quad 0 < \epsilon < 1 - \alpha, \tag{13}$$

so that its kernel is exactly the constant vectors (the second term vanishes only on shifts of $\mathbf{e}$) and the first term enforces a one-step shrinkage by $\beta = \alpha + \epsilon < 1$. A short calculation then shows $\|\mathsf{P}^\pi x\|_{\mathsf{P}} \le \beta\|x\|_{\mathsf{P}}$, yielding the desired contraction, which leads to the overall result.

The concrete proof of Lemma 4.1 including the detailed construction of the semi-norm $\|\cdot\|_{\mathsf{P}}$ is in Appendix A.1, where the properties of irreducible and aperiodic finite state Markov chain are utilized. Thus, we show the (non-robust) average-reward Bellman operator $\mathbf{T}_g^{\mathsf{P}}$ is a strict contraction under $\|\cdot\|_{\mathsf{P}}$. Based on the above motivations, we now formally establish the contraction property of the robust average-reward Bellman operator by leveraging Lemma 4.1 and the compactness of the uncertainty sets.

**Theorem 4.2.** *Under Assumption 3.1, if $\mathcal{P}$ is compact, with certain restrictions on the radius of the uncertainty sets, there exists a semi-norm $\|\cdot\|_{\mathcal{P}}$ with kernel $\{c\mathbf{e} : c \in \mathbb{R}\}$ such that the robust Bellman operator $\mathbf{T}_g$ is a contraction. Specifically, there exist $\gamma \in (0,1)$ such that*

$$\|\mathbf{T}_g(V_1) - \mathbf{T}_g(V_2)\|_{\mathcal{P}} \;\le\; \gamma\,\|V_1 - V_2\|_{\mathcal{P}},\ \forall\, V_1, V_2 \in \mathbb{R}^S,\ g \in \mathbb{R}. \tag{14}$$

**Proof Sketch** For any $\mathsf{P} \in \mathcal{P}$, the one-step transition matrix $\mathsf{P}^\pi$ has a unique stationary projector $E_{\mathsf{P}}$ due to ergodicity. Since $\mathcal{P}$ is compact, the family of fluctuation matrices $\{Q_{\mathsf{P}}^\pi = \mathsf{P}^\pi - E_{\mathsf{P}} : \mathsf{P} \in \mathcal{P}\}$ has joint spectral radius strictly less than 1. By Lemma F.1in [3], one is able to construct an "extremal norm" (denoted as $\|\cdot\|_{\text{ext}}$) under which every $Q_{\mathsf{P}}^\pi$ contracts by a uniform factor $\alpha \in (0,1)$. Mimicking the non-robust case in Lemma 4.1, we similarly define

$$\|x\|_{\mathcal{P}} = \sup_{\mathsf{P}\in\mathcal{P}} \|Q_{\mathsf{P}}^\pi x\|_{\text{ext}} + \epsilon \inf_{c\in\mathbb{R}} \|x - c\mathbf{e}\|_{\text{ext}}, \quad 0 < \epsilon < 1 - \alpha. \tag{15}$$

The supremum term zeros out if $x \in \{c\mathbf{e} : c \in \mathbb{R}\}$ , and it inherits the uniform one-step shrinkage by $\alpha$. Adding the small quotient term fixes the kernel without spoiling $\gamma = \alpha + \epsilon < 1$, so one shows at once

$$\|\mathbf{T}_g^{\mathsf{P}}(V_1) - \mathbf{T}_g^{\mathsf{P}}(V_2)\|_{\mathcal{P}} \le \gamma\|V_1 - V_2\|_{\mathcal{P}} \quad \text{for all } \mathsf{P} \in \mathcal{P} \tag{16}$$

The above leads to the desired results.

The concrete proof of Theorem 4.2 along with the detailed construction of the semi-norm $\|\cdot\|_{\mathcal{P}}$ and the specific radius restrictions on various uncertainty sets are in Appendix A.2. Since all the uncertainty sets listed in Section 3.1 are closed and bounded in a real vector space, these uncertainty sets are all compact and satisfy the contraction property in Theorem 4.2. We also note that the contraction factor $\gamma$ relates to the joint spectral gap of the family $\{Q_{\mathsf{P}}^\pi : \mathsf{P} \in \mathcal{P}\}$.

## 5 Efficient Estimators for Uncertainty Sets

To utilize the contraction property in Section 4 to obtain convergence rate results, our idea is perform the following iterative stochastic approximation:

$$V_{t+1}(s) \leftarrow V_t(s) + \eta_t \left( \hat{\mathbf{T}}_g(V_t)(s) - V_t(s) \right), \quad \forall s \in \mathcal{S} \tag{17}$$

where the learning rate $\eta_t$ would be specified in Section 6. The detailed analysis and complexities of the general stochastic approximation in the form of (17) is provided in Appendix B. Theorem B.1 implies that if $\hat{\mathbf{T}}_g(V)$, being an estimator of $\mathbf{T}_g(V)$, could be constructed with bounded variance and small bias, $V_t$ converges to a solution of the Bellman equation in (4). However, the challenge of constructing our desired $\hat{\mathbf{T}}_g(V)$ lies in the construction of the support function estimator $\hat{\sigma}_{\mathcal{P}_s^a}(V)$.

In this section, we aim to construct an estimator $\hat{\sigma}_{\mathcal{P}_s^a}(V)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$ in various uncertainty sets. Recall that the support function $\sigma_{\mathcal{P}_s^a}(V)$ represents the worst-case transition effect over the uncertainty set $\mathcal{P}_s^a$ as defined in the robust Bellman equation in Theorem 3.2. The explicit forms of $\sigma_{\mathcal{P}_s^a}(V)$ for different uncertainty sets were characterized in (6)-(9). Our goal in this section is to construct efficient estimators $\hat{\sigma}_{\mathcal{P}_s^a}(V)$ that approximates $\sigma_{\mathcal{P}_s^a}(V)$ while maintaining controlled variance and finite sample complexity.

**Linear Contamination Uncertainty Set** Recall that the $\delta$-contamination uncertainty set is $\mathcal{P}_s^a = \{(1-\delta)\tilde{\mathsf{P}}_s^a + \delta q : q \in \Delta(\mathcal{S})\}$, where $0 < \delta < 1$ is the radius. Since the support function can be computed by (6) and the expression is linear in the nominal transition kernel $\tilde{\mathsf{P}}_s^a$. A direct approach is to use the transition to the subsequent state to construct our estimator:

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

where $s'$ is a subsequent state sample after $(s,a)$. Hence, the sample complexity of (18) is just one. Lemma F.3 from [39] states that $\hat{\sigma}_{\mathcal{P}_s^a}(V)$ obtained by (18) is unbiased and has bounded variance as follows:

$$\mathbb{E}\left[\hat{\sigma}_{\mathcal{P}_s^a}(V)\right] = \sigma_{\mathcal{P}_s^a}(V), \quad \text{and} \quad \mathrm{Var}(\hat{\sigma}_{\mathcal{P}_s^a}(V)) \leq \|V\|^2 \tag{19}$$

**Nonlinear Contamination Sets** Regarding TV and Wasserstein distance uncertainty sets, they have a nonlinear relationship between the nominal distribution $\tilde{\mathsf{P}}_s^a$ and the support function $\sigma_{\mathcal{P}_s^a}(V)$. Previous works such as [4, 5, 39] have proposed a Multi-Level Monte-Carlo (MLMC) method for obtaining an unbiased estimator of $\sigma_{\mathcal{P}_s^a}(V)$ with bounded variance. However, their approaches require drawing $2^{N+1}$ samples where $N$ is sampled from a geometric distribution $\mathrm{Geom}(\Psi)$ with parameter $\Psi \in (0, 0.5)$. This operation would need infinite samples in expectation for obtaining each single estimator as $\mathbb{E}[2^{N+1}] = \sum_{N=0}^{\infty} 2^{N+1}\Psi(1-\Psi)^N = \sum_{N=0}^{\infty} 2\Psi(2-2\Psi)^N \to \infty$. To handle the above problem, we aim to provide an estimator $\hat{\sigma}_{\mathcal{P}_s^a}(V)$ with finite sample complexity and small enough bias. We construct a truncated-MLMC estimator under geometric sampling with parameter $\Psi = 0.5$ as shown in Algorithm 1.

In particular, if $n < N_{\max}$, then $\{N' = n\} = \{N = n\}$ with probability $(\frac{1}{2})^{n+1}$, while $\{N' = N_{\max}\}$ has probability $\sum_{m=N_{\max}}^{\infty}(1/2)^{m+1} = 2^{-N_{\max}}$. After obtaining $N'$, Algorithm 1 then collects a set of $2^{N'+1}$ i.i.d. samples from the nominal transition model to construct empirical estimators for different transition distributions. The core of the approach lies in computing the support function estimates for TV and Wasserstein uncertainty sets using a correction term $\Delta_{N'}(V)$, which accounts for the bias introduced by truncation. This correction ensures that the final estimator maintains a low bias while achieving a finite sample complexity. This truncation technique has been widely used in prior work across different settings such as [36, 14, 43]. We now present several crucial properties of Algorithm 1.

**Theorem 5.1** (Finite Sample Complexity)**.** *Under Algorithm 1, denote $M = 2^{N'+1}$ as the random number of samples (where $N' = \min\{N, N_{\max}\}$). Then*

$$\mathbb{E}[M] = N_{\max} + 2 = \mathcal{O}(N_{\max}). \tag{20}$$

The proof of Theorem 5.1 is in Appendix C.1, which demonstrates that setting the geometric sampling parameter to $\Psi = 0.5$ ensures that the expected number of samples follows a linear growth

**Algorithm 1** Truncated MLMC Estimator for TV and Wasserstein Uncertainty Sets

**Input**: $s \in \mathcal{S}$, $a \in \mathcal{A}$, Max level $N_{\max}$, Value function $V$

Sample $N \sim \text{Geom}(0.5)$

$N' \leftarrow \min\{N, N_{\max}\}$

Collect $2^{N'+1}$ i.i.d. samples of $\{s'_i\}_{i=1}^{2^{N'+1}}$ with $s'_i \sim \tilde{\mathsf{P}}_s^a$ for each $i$

$\hat{\mathsf{P}}^{a,E}_{s,N'+1} \leftarrow \frac{1}{2^{N'}} \sum_{i=1}^{2^{N'}} \mathbb{1}_{\{s'_{2i}\}}$

$\hat{\mathsf{P}}^{a,O}_{s,N'+1} \leftarrow \frac{1}{2^{N'}} \sum_{i=1}^{2^{N'}} \mathbb{1}_{\{s'_{2i-1}\}}$

$\hat{\mathsf{P}}^{a}_{s,N'+1} \leftarrow \frac{1}{2^{N'+1}} \sum_{i=1}^{2^{N'+1}} \mathbb{1}_{\{s'_i\}}$

$\hat{\mathsf{P}}^{a,1}_{s,N'+1} \leftarrow \mathbb{1}_{\{s'_1\}}$

**if** TV **then** Obtain $\sigma_{\hat{\mathsf{P}}^{a,1}_{s,N'+1}}(V), \sigma_{\hat{\mathsf{P}}^{a}_{s,N'+1}}(V), \sigma_{\hat{\mathsf{P}}^{a,E}_{s,N'+1}}(V), \sigma_{\hat{\mathsf{P}}^{a,O}_{s,N'+1}}(V)$ from (7)

**else if** Wasserstein **then** Obtain $\sigma_{\hat{\mathsf{P}}^{a,1}_{s,N'+1}}(V), \sigma_{\hat{\mathsf{P}}^{a}_{s,N'+1}}(V), \sigma_{\hat{\mathsf{P}}^{a,E}_{s,N'+1}}(V), \sigma_{\hat{\mathsf{P}}^{a,O}_{s,N'+1}}(V)$ from (9)

**end if**

$\Delta_{N'}(V) \leftarrow \sigma_{\hat{\mathsf{P}}^{a}_{s,N'+1}}(V) - \frac{1}{2}\Big[\sigma_{\hat{\mathsf{P}}^{a,E}_{s,N'+1}}(V) + \sigma_{\hat{\mathsf{P}}^{a,O}_{s,N'+1}}(V)\Big]$

$\hat{\sigma}_{\mathcal{P}_s^a}(V) \leftarrow \sigma_{\hat{\mathsf{P}}^{a,1}_{s,N'+1}}(V) + \frac{\Delta_{N'}(V)}{\mathbb{P}(N'=n)}$, where $p'(n) = \mathbb{P}(N' = n)$ **return** $\hat{\sigma}_{\mathcal{P}_s^a}(V)$

pattern rather than an exponential one. This choice precisely cancels out the effect of the exponential sampling inherent in the truncated MLMC estimator, preventing infinite expected sample complexity. This result shows that the expected number of queries grows only linearly with $N_{\max}$, ensuring that the sampling cost remains manageable even for large truncation levels. The key factor enabling this behavior is setting the geometric distribution parameter to $0.5$, which balances the probability mass across different truncation levels, preventing an exponential increase in sample complexity.

**Theorem 5.2** (Exponentially Decaying Bias)**.** *Let $\hat{\sigma}_{\mathcal{P}_s^a}(V)$ be the estimator of $\sigma_{\mathcal{P}_s^a}(V)$ obtained from Algorithm 1 then under TV uncertainty set, we have:*

$$\left|\mathbb{E}\left[\hat{\sigma}_{\mathcal{P}_s^a}(V) - \sigma_{\mathcal{P}_s^a}(V)\right]\right| \leq 6(1 + \frac{1}{\delta})2^{-\frac{N_{\max}}{2}} \|V\|_{\text{sp}} \tag{21}$$

*where $\delta$ denotes the radius of TV distance. Under Wasserstein uncertainty set, we have:*

$$\left|\mathbb{E}\left[\hat{\sigma}_{\mathcal{P}_s^a}(V) - \sigma_{\mathcal{P}_s^a}(V)\right]\right| \leq 6 \cdot 2^{-\frac{N_{\max}}{2}} \|V\|_{\text{sp}} \tag{22}$$

Theorem 5.2 establishes that the bias of the truncated MLMC estimator decays exponentially with $N_{\max}$, ensuring that truncation does not significantly affect accuracy. This result follows from observing that the deviation introduced by truncation can be expressed as a sum of differences between support function estimates at different level, and each of which is controlled by the $\ell_1$-distance between transition distributions. Thus, we can use binomial concentration property to ensure the exponentially decaying bias.

The proof of Theorem 5.2 is in Appendix C.2. One important lemma used in the proof is the following Lemma 5.3, where we show the Lipschitz property for both TV and Wasserstein distance uncertainty sets.

**Lemma 5.3.** *For any $p, q \in \Delta(\mathcal{S})$, let $\mathcal{P}_{TV}$ and $\mathcal{Q}_{TV}$ denote the TV distance uncertainty set with radius $\delta$ centering at $p$ and $q$ respectively, and let $\mathcal{P}_W$ and $\mathcal{Q}_W$ denote the Wasserstein distance uncertainty set with radius $\delta$ centering at $p$ and $q$ respectively. Then for any value function $V$, we have:*

$$|\sigma_{\mathcal{P}_{TV}}(V) - \sigma_{\mathcal{Q}_{TV}}(V)| \leq (1 + \frac{1}{\delta})\|V\|_{\text{sp}}\|p - q\|_1 \text{ and } |\sigma_{\mathcal{P}_W}(V) - \sigma_{\mathcal{Q}_W}(V)| \leq \|V\|_{\text{sp}}\|p - q\|_1 \tag{23}$$

We refer the proof of Theorem 5.2 to Appendix C.3.

**Theorem 5.4** (Linear Variance)**.** *Let $\hat{\sigma}_{\mathcal{P}_s^a}(V)$ be the estimator of $\sigma_{\mathcal{P}_s^a}(V)$ obtained from Algorithm 1 then under TV distance uncertainty set, we have:*

$$\text{Var}(\hat{\sigma}_{\mathcal{P}_s^a}(V)) \leq 3\|V\|_{\text{sp}}^2 + 144(1 + \frac{1}{\delta})^2\|V\|_{\text{sp}}^2 N_{\max} \tag{24}$$

*and under Wasserstein distance uncertainty set, we have:*

$$\mathrm{Var}(\hat{\sigma}_{\mathcal{P}_s^a}(V)) \leq 3\|V\|_{\mathrm{sp}}^2 + 144\|V\|_{\mathrm{sp}}^2 N_{\max} \tag{25}$$

Theorem 5.4 establishes that the variance of the truncated MLMC estimator grows linearly with $N_{\max}$, ensuring that the estimator remains stable even as the truncation level increases. The proof of Theorem 5.4 is in Appendix C.4, which follows from bounding the second moment of the estimator by analyzing the variance decomposition across different MLMC levels. Specifically, by expressing the estimator in terms of successive refinements of the transition model, we show that the variance accumulates additively across levels due to the binomial concentration property.

## 6 Robust Average-Reward TD Learning

Equipped with the methods of constructing $\hat{\sigma}_{\mathcal{P}_s^a}(V)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we now present the formal algorithm for robust policy evaluation and robust average reward for a given policy $\pi$ in Algorithm 2. Algorithm 2 presents a robust temporal difference (TD) learning method for policy evaluation in robust average-reward MDPs. This algorithm builds upon the truncated MLMC estimator (Algorithm 1) and the biased stochastic approximation framework in Section B, ensuring both efficient sample complexity and finite-time convergence guarantees.

The algorithm is divided into two main phases. The first phase (Lines 1-7) estimates the robust value function. The noisy Bellman operator is computed using the estimator $\hat{\sigma}_{\mathcal{P}_s^a}(V_t)$ obtained depending on the uncertainty set type. Then the iterative update follows a stochastic approximation scheme with stepsize $\eta_t$, ensuring convergence while maintaining stability. Finally, the value function is centered at an anchor state $s_0$ to remove the ambiguity due to its additive invariance. The second phase (Lines 8-14) estimates the robust average reward by utilizing $V_T$ from the output of the first phase. The expected Bellman residual $\delta_t(s)$ is computed across all states and averaging it to obtain $\bar{\delta}_t$. A separate stochastic approximation update with stepsize $\beta_t$ is then applied to refine $g_t$, ensuring convergence to the robust worst-case average reward. By combining these two phases, Algorithm 2 provides an efficient and provably convergent method for robust policy evaluation under average-reward criteria, marking a significant advancement over prior methods that only provided asymptotic guarantees.

**Algorithm 2** Robust Average-Reward TD

**Input**: Policy $\pi$, Initial values $V_0$, $g_0 = 0$, Stepsizes $\eta_t$, $\beta_t$, Max level $N_{\max}$, Anchor state $s_0 \in \mathcal{S}$

1: **for** $t = 0, 1, \ldots, T-1$ **do**
2:  **for** each $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**
3:  **if** Contamination **then** Sample $\hat{\sigma}_{\mathcal{P}_s^a}(V_t)$ according to (18)
4:  **else if** TV or Wasserstein **then** Sample $\hat{\sigma}_{\mathcal{P}_s^a}(V_t)$ according to Algorithm 1
5:  **end if**
6:  **end for**
7:  $\hat{\mathbf{T}}_{g_0}(V_t)(s) \leftarrow \sum_a \pi(a|s)\big[r(s,a) - g_0 + \hat{\sigma}_{\mathcal{P}_s^a}(V_t)\big], \quad \forall s \in \mathcal{S}$
8:  $V_{t+1}(s) \leftarrow V_t(s) + \eta_t \left(\hat{\mathbf{T}}_{g_0}(V_t)(s) - V_t(s)\right), \quad \forall s \in \mathcal{S}$
9:  $V_{t+1}(s) = V_{t+1}(s) - V_{t+1}(s_0), \quad \forall s \in \mathcal{S}$
10: **end for**
11: **for** $t = 0, 1, \ldots, T-1$ **do**
12:  **for** each $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**
13:  **if** Contamination **then** Sample $\hat{\sigma}_{\mathcal{P}_s^a}(V_t)$ according to (18)
14:  **else if** TV or Wasserstein **then** Sample $\hat{\sigma}_{\mathcal{P}_s^a}(V_t)$ according to Algorithm 1
15:  **end if**
16:  **end for**
17:  $\hat{\bar{\delta}}_t(s) \leftarrow \sum_a \pi(a|s)\big[r(s,a) + \hat{\sigma}_{\mathcal{P}_s^a}(V_T)\big] - V_T(s), \quad \forall s \in \mathcal{S}$
18:  $\bar{\delta}_t \leftarrow \frac{1}{S}\sum_s \hat{\delta}_t(s)$
19:  $g_{t+1} \leftarrow g_t + \beta_t(\bar{\delta}_t - g_t)$
20: **end for** **return** $V_T$, $g_T$

To derive the sample complexity of robust policy evaluation, we utilize the semi-norm contraction property of the Bellman operator in Theorem 4.2, and fit Algorithm 2 into the general biased stochastic approximation result in Theorem B.1 while incorporating the bias analysis characterized in Section 5. Since each phase of Algorithm 2 contains a loop of length $T$ with all the states and actions updated together, the total samples needed for the entire algorithm in expectation is $2SAT\mathbb{E}[N_{\max}]$, where $\mathbb{E}[N_{\max}]$ is one for contamination uncertainty sets and is $\mathcal{O}(N_{\max})$ from Theorem 5.1 for TV and Wasserstein distance uncertainty sets.

**Theorem 6.1.** *If $V_t$ is generated by Algorithm 2 and satisfying Assumption 3.1, then if the stepsize $\eta_t \coloneqq \mathcal{O}(\frac{1}{t})$, we require a sample complexity of $\mathcal{O}\left(\frac{SAt_{\text{mix}}^2}{\epsilon^2(1-\gamma)^2}\right)$ for contamination uncertainty set and a sample complexity of $\tilde{\mathcal{O}}\left(\frac{SAt_{\text{mix}}^2}{\epsilon^2(1-\gamma)^2}\right)$ for TV and Wasserstein distance uncertainty set to ensure an $\epsilon$ convergence of $V_T$. Moreover, these results are order-optimal in terms of $\epsilon$.*

**Theorem 6.2.** *If $g_t$ is generated by Algorithm 2 and satisfying Assumption 3.1, then if the stepsize $\beta_t \coloneqq \mathcal{O}(\frac{1}{t})$, we require a sample complexity of $\tilde{\mathcal{O}}\left(\frac{SAt_{\text{mix}}^2}{\epsilon^2(1-\gamma)^2}\right)$ for all contamination, TV, and Wasserstein distance uncertainty set to ensure an $\epsilon$ convergence of $g_T$.*

The formal version of Theorems 6.1 and 6.2 along with the proofs are in Appendix D. Theorem 6.1 provides the order-optimal sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2})$ for Algorithm 2 to achieve an $\epsilon$-accurate estimate of $V_T$. Although Theorem 6.1 claims order-optimal in terms of $\epsilon$, we do not claim tightness in $S$, $A$ and $\gamma$, and treat sharpening these dependencies as open. The proof of Theorem 6.2 extends the analysis of Theorem 6.1 to robust average reward estimation. The key difficulty lies in controlling the propagation of error from value function estimates to reward estimation. By again leveraging the contraction property and appropriately tuning stepsizes, we establish an $\tilde{\mathcal{O}}(\epsilon^{-2})$ complexity bound for robust average reward estimation.

## 7 Conclusion

This paper provides the first finite-sample analysis for policy evaluation in robust average-reward MDPs, bridging a gap where only asymptotic guarantees existed. By introducing a biased stochastic approximation framework and leveraging the properties of various uncertainty sets, we establish finite-time convergence under biased noise. Our algorithm achieves an order-optimal sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2})$ for policy evaluation, despite the added complexity of robustness.

A crucial step in our analysis is proving that the robust Bellman operator is contractive under our constructed semi-norm $\|\cdot\|_{\mathcal{P}}$, ensuring the validity of stochastic approximation updates. We further develop a truncated Multi-Level Monte Carlo estimator that efficiently computes worst-case value functions under total variation and Wasserstein uncertainty, while keeping bias and variance controlled. One limitation of this work is that the results require ergodicity to hold in the setting, as stated in Assumption 3.1. Additionally, scaling the algorithm and results in the paper via function approximations remains an important open problem.

## Acknowledgments

We would like to thank Zaiwei Chen of Purdue University for assistance with identifying relevant literature and for constructive feedback. We also thank Zijun Chen and Nian Si of The Hong Kong University of Science and Technology for helpful discussions regarding Appendix A. Finally, we thank the anonymous reviewers for insightful comments that substantially improved the paper.

## References

[1] Abubakr O Al-Abbasi, Arnob Ghosh, and Vaneet Aggarwal. Deeppool: Distributed model-free algorithm for ride-sharing using deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 20(12):4714–4727, 2019.

[2] Qinbo Bai, Washim Uddin Mondal, and Vaneet Aggarwal. Regret analysis of policy gradient algorithm for infinite horizon average reward markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10980–10988, 2024.

[3] Marc A Berger and Yang Wang. Bounded semigroups of matrices. *Linear Algebra and its Applications*, 166:21–27, 1992.

[4] Jose H Blanchet and Peter W Glynn. Unbiased Monte Carlo for optimization and functions of expectations via multi-level randomization. In *2015 Winter Simulation Conference (WSC)*, pages 3656–3667. IEEE, 2015.

[5] Jose H Blanchet, Peter W Glynn, and Yanan Pei. Unbiased multilevel Monte Carlo: Stochastic optimization, steady-state simulation, quantiles, and other applications. *arXiv preprint arXiv:1904.09929*, 2019.

[6] Vivek S Borkar and Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 100. Springer, 2008.

[7] Mengjie Chen, Chao Gao, and Zhao Ren. A general decision theory for huber's $\epsilon$-contamination model. *Electronic Journal of Statistics*, 10:3752–3774, 2016.

[8] Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. *Advances in Neural Information Processing Systems*, 33:8223–8234, 2020.

[9] Zaiwei Chen, Sheng Zhang, Zhe Zhang, Shaan Ul Haque, and Siva Theja Maguluri. A non-asymptotic theory of seminorm lyapunov stability: From deterministic to stochastic iterative algorithms. *arXiv preprint arXiv:2502.14208*, 2025.

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

[11] Julien Grand Clement and Christian Kroer. First-order methods for wasserstein distributionally robust mdp. In *International Conference on Machine Learning*, pages 2010–2019. PMLR, 2021.

[12] Swetha Ganesh and Vaneet Aggarwal. Regret analysis of average-reward unichain mdps via an actor-critic approach. *Advances in Neural Information Processing Systems*, 2025.

[13] Swetha Ganesh, Washim Uddin Mondal, and Vaneet Aggarwal. Order-optimal regret with novel policy gradient approaches in infinite-horizon average reward mdps. In *International Conference on Artificial Intelligence and Statistics*, pages 3421–3429. PMLR, 2025.

[14] Swetha Ganesh, Washim Uddin Mondal, and Vaneet Aggarwal. A sharper global convergence analysis for average reward reinforcement learning via an actor-critic approach. In *Forty-second International Conference on Machine Learning*, 2025.

[15] Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

[16] Stéphane Gaubert and Zheng Qu. Dobrushin's ergodicity coefficient for markov operators on cones. *Integral Equations and Operator Theory*, 81(1):127–150, 2015.

[17] Ben Hambly, Renyuan Xu, and Huining Yang. Recent advances in reinforcement learning in finance. *Mathematical Finance*, 33(3):437–503, 2023.

[18] Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Partial policy iteration for l1-robust markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46, 2021.

[19] Sebastian Höfer, Kostas Bekris, Ankur Handa, Juan Camilo Gamboa, Melissa Mozifian, Florian Golemo, Chris Atkeson, Dieter Fox, Ken Goldberg, John Leonard, et al. Sim2real in robotics and automation: Applications and challenges. *IEEE transactions on automation science and engineering*, 18(2):398–400, 2021.

[20] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[21] Peter J Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pages 1753–1758, 1965.

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

[23] Yufei Kuang, Miao Lu, Jie Wang, Qi Zhou, Bin Li, and Houqiang Li. Learning robust policy against disturbance in transition dynamics via state-conservative policy optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7247–7254, 2022.

[24] Navdeep Kumar, Esther Derman, Matthieu Geist, Kfir Y Levy, and Shie Mannor. Policy gradient for rectangular robust markov decision processes. *Advances in Neural Information Processing Systems*, 36:59477–59501, 2023.

[25] Yan Li, Guanghui Lan, and Tuo Zhao. First-order policy optimization for robust markov decision process. *arXiv preprint arXiv:2209.10579*, 2022.

[26] Shiau Hong Lim, Huan Xu, and Shie Mannor. Reinforcement learning in robust markov decision processes. *Advances in Neural Information Processing Systems*, 26, 2013.

[27] Arnab Nilim and Laurent El Ghaoui. Robustness in markov decision problems with uncertain transition matrices. *Advances in neural information processing systems*, 16, 2003.

[28] Zachary Roch, Chi Zhang, George Atia, and Yue Wang. A finite-sample analysis of distributionally robust average-reward reinforcement learning. *arXiv preprint arXiv:2505.12462*, 2025.

[29] Zachary Andrew Roch, George K Atia, and Yue Wang. A reduction framework for distributionally robust reinforcement learning under average reward. In *Forty-second International Conference on Machine Learning*, 2025.

[30] Saksham Sahai Srivastava and Vaneet Aggarwal. A technical survey of reinforcement learning techniques for large language models. *arXiv preprint arXiv:2507.04136*, 2025.

[31] Zhongchang Sun, Sihong He, Fei Miao, and Shaofeng Zou. Policy optimization for robust average reward mdps. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[32] Niko Sünderhauf, Oliver Brock, Walter Scheirer, Raia Hadsell, Dieter Fox, Jürgen Leitner, Ben Upcroft, Pieter Abbeel, Wolfram Burgard, Michael Milford, et al. The limits and potentials of deep learning for robotics. *The International journal of robotics research*, 37(4-5):405–420, 2018.

[33] Chen Tang, Ben Abbatematteo, Jiaheng Hu, Rohan Chandra, Roberto Martín-Martín, and Peter Stone. Deep reinforcement learning for robotics: A survey of real-world successes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28694–28698, 2025.

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

[35] Qiuhao Wang, Yuqi Zha, Chin Pang Ho, and Marek Petrik. Provable policy gradient for robust average-reward mdps beyond rectangularity. In *Forty-second International Conference on Machine Learning*, 2025.

[36] Yudan Wang, Shaofeng Zou, and Yue Wang. Model-free robust reinforcement learning with sample complexity analysis. In *Uncertainty in Artificial Intelligence*, pages 3470–3513. PMLR, 2024.

[37] Yue Wang, Alvaro Velasquez, George Atia, Ashley Prater-Bennette, and Shaofeng Zou. Robust average-reward markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15215–15223, 2023.

[38] Yue Wang, Alvaro Velasquez, George Atia, Ashley Prater-Bennette, and Shaofeng Zou. Robust average-reward reinforcement learning. *Journal of Artificial Intelligence Research*, 80:719–803, 2024.

[39] Yue Wang, Alvaro Velasquez, George K Atia, Ashley Prater-Bennette, and Shaofeng Zou. Model-free robust average-reward reinforcement learning. In *International Conference on Machine Learning*, pages 36431–36469. PMLR, 2023.

[40] Yue Wang and Shaofeng Zou. Policy gradient method for robust reinforcement learning. In *International conference on machine learning*, pages 23484–23526. PMLR, 2022.

[41] Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain. Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *International conference on machine learning*, pages 10170–10180. PMLR, 2020.

[42] Fabian Wirth. The generalized spectral radius and extremal norms. *Linear Algebra and its Applications*, 342(1-3):17–40, 2002.

[43] Yang Xu and Vaneet Aggarwal. Accelerating quantum reinforcement learning with a quantum natural policy gradient based approach. In *Forty-second International Conference on Machine Learning*, 2025.

[44] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

[45] Sheng Zhang, Zhe Zhang, and Siva Theja Maguluri. Finite sample analysis of average-reward td learning and $q$-learning. *Advances in Neural Information Processing Systems*, 34:1230–1242, 2021.

[46] Ruida Zhou, Tao Liu, Min Cheng, Dileep Kalathil, PR Kumar, and Chao Tian. Natural actor-critic for robust reinforcement learning with function approximation. *Advances in neural information processing systems*, 36, 2024.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction accurately reflects the novelty and contributions of this paper. All details can be found either in the rest of the main text or in the appendix.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The discussion of limitations can be found in the conclusion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions are clearly stated in the main text, the theorem statements and some proof sketches are also included in the main text. The formal theorem statements and their complete proofs are in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: This is a theoretical paper and does not include experiments. The minor numerical examples added during the rebuttal will be released.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: This is a theoretical paper and does not include experiments. The minor numerical examples added during the rebuttal will be released.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: This is a theoretical paper and does not include experiments. The minor numerical examples added during the rebuttal will be released.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: This is a theoretical paper and does not include experiments. The minor numerical examples added during the rebuttal will be released.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: This is a theoretical paper and does not include experiments. The minor numerical examples added during the rebuttal will be released. The minor numerical examples added during the rebuttal will be released.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper is theoretical and conform, in every respect, the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This is a theoretical work and the algorithm could be applied in different applications. There is nothing specific that can be highlighted.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This is a theoretical work and the algorithm could be applied in different applications. There is nothing specific that can be highlighted.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.the paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

# A Semi-Norm Contraction Property of the Bellman Operator

## A.1 Proof of Lemma 4.1

For any $V_1, V_2 \in \mathbb{R}^S$ and define $\Delta = V_1 - V_2$. Denote $\mathsf{P}^\pi$ as the transition matrix under policy $\pi$ and the unique stationary distribution $d^\pi$, and denote $E$ as the matrix with all rows being identical to $d^\pi$. We further define $Q^\pi = \mathsf{P}^\pi - E$. Thus, we would have,

$$\mathbf{T}_g^{\mathsf{P}}(V_1)(s) - \mathbf{T}_g^{\mathsf{P}}(V_2)(s) = \sum_{s' \in \mathcal{S}} \mathsf{P}^\pi(s'|s)\big[V_1(s') - V_2(s')\big] = \mathsf{P}^\pi\,\Delta(s). \tag{26}$$

which implies

$$\mathbf{T}_g^{\mathsf{P}}(V_1) - \mathbf{T}_g^{\mathsf{P}}(V_2) = \mathsf{P}^\pi\Delta = Q^\pi\Delta + E\Delta \tag{27}$$

We now discuss the detailed construction of the semi-norm $\|\cdot\|_{\mathsf{P}}$. Since $\mathsf{P}^\pi$ is ergodic, according to the Perron–Frobenius theorem, $\mathsf{P}^\pi$ has an eigenvalue $\lambda_1 = 1$ of algebraic multiplicity exactly one, with corresponding right eigenvector $\mathbf{e}$. Moreover, all other eigenvalues $\lambda_2 \geq \ldots \geq \lambda_S$ of $\mathsf{P}^\pi$ satisfies $|\lambda_i| < 1$ for all $i \in \{2, \ldots, S\}$.

**Lemma A.1.** *All eigenvalues of $Q^\pi$ lies strictly inside the unit circle.*

*Proof.* Since $E = \mathbf{e}d^{\pi\top}$, $E$ is a rank-one projector onto the span of $\mathbf{e}$. Hence the spectrum of $E$ is $\{1, 0, \ldots, 0\}$. In addition, we can show $\mathsf{P}^\pi$ and $E$ commute by

$$\mathsf{P}^\pi E = \mathsf{P}^\pi(\mathbf{e}(d^\pi)^\top) = (\mathbf{e}(d^\pi)^\top) = E \tag{28}$$

$$E\mathsf{P}^\pi = \mathbf{e}((d^\pi)^\top\mathsf{P}^\pi) = \mathbf{e}(d^\pi)^\top = E \tag{29}$$

Thus, by the Schur's theorem, $\mathsf{P}$ and $E$ are simultaneously upper triangularizable. In a common triangular basis, the diagonals of $\mathsf{P}$ and $E$ list their eigenvalues in descending orders, which are $\{\lambda_1, \lambda_2, \ldots, \lambda_S\}$ and $\{1, 0, \ldots, 0\}$ respectively. Thus, in that same basis, $Q^\pi = \mathsf{P}^\pi - E$ is also triangular, with diagonal entries being $\{\lambda_1 - 1, \lambda_2 - 0, \ldots, \lambda_S - 0\}$. Since $\lambda_1 = 1$, we have the spectrum of $Q^\pi$ is exactly $\{\lambda_2, \ldots, \lambda_S, 0\}$. Since we already have $|\lambda_i| < 1$ for all $i \in \{2, \ldots, S\}$, we conclude the proof. □

Define $\rho(\cdot)$ to be the spectral radius of a matrix, then Lemma A.1 implies that $\rho(Q^\pi) < 1$. Hence by equivalence of norms in $\mathbb{R}^{|\mathcal{S}|}$ it is possible to construct a vector norm $\|\cdot\|_Q$ so that the induced operator norm of $Q^\pi$ is less than 1, specifically

$$0 \leq \rho(Q^\pi) \leq \big\|Q^\pi\big\|_Q \leq \alpha < 1. \tag{30}$$

A concrete construction example is to leverage the discrete-Lyapunov equation [20] of solving $M$ on the space of symmetric matrices for any $\rho(Q^\pi) < \alpha < 1$ as follows:

$$Q^{\pi\top}MQ^\pi - \alpha^2 M = -I \tag{31}$$

Define $B \coloneqq \alpha^{-1}Q^\pi$, then $\rho(B) = \alpha^{-1}\rho(Q^\pi) < 1$. We can express $M$ in the form of Neumann series as

$$\begin{aligned} M &= \alpha^{-2}I + \alpha^{-4}(Q^\pi)^\top Q^\pi + \alpha^{-6}\left((Q^\pi)^\top\right)^2(Q^\pi)^2 + \ldots \\ &= \sum_{k=0}^{\infty} \alpha^{-2(k+1)}\left((Q^\pi)^\top\right)^k (Q^\pi)^k \\ &= \alpha^{-2}\sum_{k=0}^{\infty}\left(B^\top\right)^k B^k. \end{aligned} \tag{32}$$

We now show that $M$ is bounded. Write $B = SJS^{-1}$, where $J = \mathrm{diag}\big(J_{m_1}(\lambda_1), \ldots, J_{m_r}(\lambda_r)\big)$ is the Jordan normal form. By the Jordan block power formula [20],

$$J_m(\lambda) = \lambda I_m + N_m, \tag{33}$$

with $N_m$ the nilpotent matrix having ones on the first superdiagonal and $N_m^m = 0$. Then $B^k = SJ^kS^{-1}$ and $J^k = \operatorname{diag}\big(J_{m_1}(\lambda_1)^k, \ldots, J_{m_r}(\lambda_r)^k\big)$. For each block and each integer $k \geq m$, by the binomial theorem we have $N_m^m = 0$ and

$$J_m(\lambda)^k = (\lambda I_m + N_m)^k = \sum_{j=0}^{m-1} \binom{k}{j} \lambda^{\,k-j}\, N_m^j. \tag{34}$$

For $k \geq m$, use $\binom{k}{j} \leq \frac{k^j}{j!}$ and factor $|\lambda|^{\,k}$:

$$\|J_m(\lambda)^k\|_2 \leq |\lambda|^{\,k} \sum_{j=0}^{m-1} \frac{k^j}{j!} |\lambda|^{-j} \|N_m^j\|_2 \leq c_{m,\lambda} k^{m-1} |\lambda|^k, \tag{35}$$

where $c_{m,\lambda} \coloneqq \sum_{j=0}^{m-1} \frac{|\lambda|^{-j}}{j!} \|N_m^j\|_2$. Thus, let $s = \max_i m_i$ be the size of the largest Jordan block of $B$. Since $J^k$ is block diagonal,

$$\|J^k\|_2 \leq \sum_{i=1}^{r} \|J_{m_i}(\lambda_i)^k\|_2 \leq \Big(\sum_{i=1}^{r} c_{m_i,\lambda_i}\Big) k^{s-1} \Big(\max_i |\lambda_i|\Big)^k = C_J k^{s-1} \rho(B)^k \tag{36}$$

for all $k \geq s$, where $C_J := \sum_{i=1}^{r} c_{m_i,\lambda_i}$ and $\rho(B) = \max_i |\lambda_i|$.

Since similarity does not change eigenvalues but may scale norms by the condition number, we can derive that

$$\|B^k\|_2 = \|SJ^kS^{-1}\|_2 \leq \|S\|_2 \|S^{-1}\|_2 \|J^k\|_2 \leq \kappa(S)\, C_J k^{s-1} \rho(B)^k \qquad (k \geq s),$$

where $\kappa(S) := \|S\|_2 \|S^{-1}\|_2$. By choosing the appropriate constant, the same bound holds for all $k \geq 0$:

$$\exists\, C_B > 0 \text{ such that } \quad \|B^k\|_2 \;\leq\; C_B\, k^{s-1}\, \rho(B)^{\,k} \quad (k \geq 0).$$

In spectral norm, this implies

$$\|(B^\top)^k B^k\|_2 = \|(B^k)^\top B^k\|_2 = \sigma_{\max}(B^k)^2 = \|B^k\|_2^2 \leq C_B^2 k^{2(s-1)} \rho(B)^{2k}. \tag{37}$$

Thus, the scalar series $\sum_{k=0}^{\infty} k^{2(s-1)}\, \rho(B)^{\,2k}$ is in the form of polynomial times geometric with ratio less than 1, which converges, and the partial sum expression in (32) converges absolutely as a geometric-type series.

Also, since each term in (32) is positive semi-definite, and the first term $\alpha^{-2} I$ being positive definite, we can conclude that $M$ being the summation is well-defined and is positive definite. Thus, using the positive definite $M$ defined in (31), we can define our desired norm $\|\cdot\|_Q$ as

$$\|x\|_Q \coloneqq \sqrt{x^\top M x} \tag{38}$$

which implies

$$\|Q^\pi\|_Q = \sup_{x \neq \mathbf{0}} \frac{\|Q^\pi x\|_Q}{\|x\|_Q} = \sup_{x \neq \mathbf{0}} \frac{\sqrt{(Q^\pi x)^\top M Q^\pi x}}{\sqrt{x^\top M x}} \overset{(a)}{\leq} \alpha < 1 \tag{39}$$

Where $(a)$ is because for any $x \neq \mathbf{0}$, from (31) we have

$$(Qx)^{\pi\top} M Q^\pi x - \alpha^2 x^\top M x = -x^\top x \quad \Rightarrow \quad (Qx)^{\pi\top} M Q^\pi x = \alpha^2 x^\top M x - \|x\|_2^2 \tag{40}$$

Since $\|x\|_2^2$ is always non-negative dividing both sides of the second equation of (40) by $x^\top M x$ and further taking the square root on both sides yields the inequality of $(a)$.

Based on the above construction of the norm $\|\cdot\|_Q$, define the operator $\|\cdot\|_{\mathsf{P}}$ as

$$\|x\|_{\mathsf{P}} \coloneqq \big\|Q^\pi x\big\|_Q + \epsilon \inf_{c \in \mathbb{R}} \big\|x - c\mathbf{e}\big\|_Q \tag{41}$$

where $0 < \epsilon < 1 - \alpha$.

**Lemma A.2.** *The operator $\|\cdot\|_{\mathsf{P}}$ is a valid semi-norm with kernel being exactly $\{c\mathbf{e} : c \in \mathbb{R}\}$. Furthermore, for all $x \in \mathbb{R}^S$, we have $\left\|\mathsf{P}^\pi x\right\|_{\mathsf{P}} \le (\alpha+\epsilon)\left\|x\right\|_{\mathsf{P}}$.*

*Proof.* Regarding positive homogeneity and nonnegativity, for any scalar $\lambda$ and $x \in \mathbb{R}^S$,

$$\|\lambda x\|_{\mathsf{P}} = \left\|Q^\pi(\lambda x)\right\|_Q + \epsilon \inf_c \left\|\lambda x - c\mathbf{e}\right\|_Q = |\lambda|\left\|Q^\pi x\right\|_Q + \epsilon\,|\lambda| \inf_c \left\|x - c\mathbf{e}\right\|_Q = |\lambda|\,\|x\|_{\mathsf{P}},$$

and clearly $\|x\|_{\mathsf{P}} \ge 0$, with equality only when both $\|Q^\pi x\|_Q = 0$ and $\inf_c \|x - ce\|_Q = 0$. Regarding triangle inequality, for any $x, y \in \mathbb{R}^S$,

$$\begin{aligned}
\|x+y\|_{\mathsf{P}} &= \left\|Q^\pi(x+y)\right\|_Q + \epsilon \inf_c \left\|x+y-c\mathbf{e}\right\|_Q \\
&\le \left\|Q^\pi x\right\|_Q + \left\|Q^\pi y\right\|_Q + \epsilon \inf_{a,b} \left\|x - a\mathbf{e} + y - b\mathbf{e}\right\|_Q \\
&\le \left\|Q^\pi x\right\|_Q + \left\|Q^\pi y\right\|_Q + \epsilon \inf_a \left\|x - a\mathbf{e}\right\|_Q + \epsilon \inf_b \left\|y - b\mathbf{e}\right\|_Q \\
&= \|x\|_{\mathsf{P}} + \|y\|_{\mathsf{P}}.
\end{aligned}$$

Regarding the kernel, if $x = k\mathbf{e}$ for some $k \in \mathbb{R}$, then we have

$$\begin{aligned}
\|x\|_{\mathsf{P}} &= \|kQ^\pi\mathbf{e}\|_Q + \epsilon \inf_c \|k\mathbf{e} - c\mathbf{e}\|_Q \\
&= \|k(\mathsf{P}^\pi - E)\mathbf{e}\|_Q + \epsilon\|k\mathbf{e} - k\mathbf{e}\|_Q \\
&= \|k\mathbf{e} - k\mathbf{e}\|_Q + \epsilon\|0\|_Q = 0
\end{aligned} \tag{42}$$

On the other hand, if $x \notin \{c\mathbf{e} : c \in \mathbb{R}\}$, we know that

$$\|x\|_{\mathsf{P}} \ge \epsilon \inf_c \|x - c\mathbf{e}\|_Q > 0 \tag{43}$$

Thus, the kernel of $\|\cdot\|_{\mathsf{P}}$ is exactly $\{c\mathbf{e} : c \in \mathbb{R}\}$. We now show that, for any $x \in \mathbb{R}^S$,

$$\begin{aligned}
\left\|\mathsf{P}^\pi x\right\|_{\mathsf{P}} &= \left\|Q^\pi(\mathsf{P}^\pi x)\right\|_Q + \epsilon \inf_c \left\|\mathsf{P}^\pi x - c\mathbf{e}\right\|_Q \\
&= \left\|Q^\pi Q^\pi x + Q^\pi E x\right\|_Q + \epsilon \inf_c \left\|Q^\pi x - c\mathbf{e}\right\|_Q \\
&\le \alpha\|Q^\pi x\|_Q + \epsilon\|Q^\pi x\|_Q \\
&= (\alpha+\epsilon)\|Q^\pi x\|_Q \\
&\le (\alpha+\epsilon)\|x\|_{\mathsf{P}}.
\end{aligned} \tag{44}$$

□

Let $\beta = \alpha + \epsilon$, by (30) and (41), we have $\alpha \in (0,1)$ and $\epsilon \in (0, 1-\alpha)$. Thus, $\beta \in (0,1)$ and combining $\beta$ with the semi-norm $\|\cdot\|_{\mathsf{P}}$ confirms Lemma 4.1.

### A.2 Proof of Theorem 4.2

We override the terms $\alpha, \lambda$ and $\epsilon$ from the previous section. For any $V_1, V_2$ and $s \in \mathcal{S}$,

$$\begin{aligned}
\mathbf{T}_g(V_1)(s) - \mathbf{T}_g(V_2)(s) &= \sum_{a\in\mathcal{A}} \pi(a|s)[\sigma_{p_s^a}(V_1) - \sigma_{p_s^a}(V_2)] \\
&= \sum_{a\in\mathcal{A}} \pi(a|s)[\min_{p\in\mathcal{P}_s^a} \sum_{s'\in\mathcal{S}} p(s')V_1(s') - \min_{p\in\mathcal{P}_s^a} \sum_{s'\in\mathcal{S}} p(s')V_2(s')] \\
&\le \sum_{a\in\mathcal{A}} \pi(a|s) \max_{p\in\mathcal{P}_s^a} \left[\sum_{s'\in\mathcal{S}} p(s')V_1(s') - \sum_{s'\in\mathcal{S}} p(s')V_2(s')\right] \\
&\le \sum_a \pi(a|s) \sum_{s'} \tilde{p}_{(V_1,V_2)}(s'|s,a)\big[V_1(s') - V_2(s')\big]
\end{aligned} \tag{45}$$

where $\tilde{p}_{(V_1,V_2)}(\cdot|s,a) = \arg\max_{p\in\mathcal{P}_s^a}[\sum_{s'\in\mathcal{S}} p(s')V_1(s') - \sum_{s'\in\mathcal{S}} p(s')V_2(s')]$ and each $\tilde{p}_{(V_1,V_2)} \in \mathcal{P}$ for all $V_1, V_2$. We now discuss the construction of the desired semi-norm $\|\cdot\|_{\mathcal{P}}$.

### A.2.1 Joint Spectral Radius of $Q^\pi_\mathcal{P}$

For any $\mathsf{P} \in \mathcal{P}$, denote $\mathsf{P}^\pi$ as the transition matrix under policy $\pi$ and the unique stationary distribution $d^\pi_\mathsf{P}$, and denote $E_\mathsf{P}$ as the matrix with all rows being identical to $d^\pi_\mathsf{P}$ (we will provide the conditions for all $\mathsf{P}^\pi$ having a unique stationary distribution later). We further define the following:

$$Q^\pi_\mathsf{P} = \mathsf{P}^\pi - E_\mathsf{P} \quad \text{and} \quad Q^\pi_\mathcal{P} = \{Q^\pi_\mathsf{P} : \mathsf{P} \in \mathcal{P}\}. \tag{46}$$

To obtain the desired one-step contraction result under Assumption 3.1 along with proper radius restrictions, we need to show the conditions of the radius under the different uncertainty sets such that the joint spectral radius $\hat{\rho}(Q^\pi_\mathcal{P})$ defined in Lemma F.1 satisfies $\hat{\rho}(Q^\pi_\mathcal{P}) < 1$, which is necessary to establish the desired one-step contraction. We first provide an upper bound of the joint spectral radius as follows:

**Lemma A.3.** *Define the Dobrushin's coefficient of an $n$ dimensional Markov matrix $P$ as*

$$\tau(P) \coloneqq 1 - \min_{i<j} \sum_{s=1}^{n} \min(P_{is}, P_{js}), \tag{47}$$

*then the joint spectral radius of the family $Q^\pi_\mathcal{P}$ is upper bounded by the following:*

$$\hat{\rho}(Q^\pi_\mathcal{P}) \leq \inf_{m \geq 1} \left( \sup_{\mathsf{P}_i \in \mathcal{P}} \tau(\mathsf{P}^\pi_1 \cdot \ldots \cdot \mathsf{P}^\pi_m) \right)^{\frac{1}{m}} \tag{48}$$

*Proof.* We start by first connecting $\hat{\rho}(Q^\pi_\mathcal{P})$ to the joint spectral radius of the family $\{\mathsf{P}^\pi : \mathsf{P} \in \mathcal{P}\}$. Define $\mathbb{H} \coloneqq \{x \in \mathbb{R}^S : \mathbf{e}^\top x = 0\}$ to be the zero-sum subspace where the space spanned by $\mathbf{e}$ is removed. Furthermore, choose an orthonormal basis $U = [u_0 \; U_\mathbb{H}] \in \mathbb{R}^{S \times S}$ with

$$u_0 = \frac{1}{\sqrt{S}}\mathbf{e}, \quad U_\mathbb{H}^\top U_\mathbb{H} = I_{S-1}, \quad U_\mathbb{H} U_\mathbb{H}^\top = \Pi \coloneqq I_S - \frac{1}{S}\mathbf{e}\mathbf{e}^\top, \quad \mathbf{e}^\top U_\mathbb{H} = 0$$

where $\Pi$ is the orthogonal projector onto $\mathbb{H}$. Since $U$ is orthogonal, $U^\top = U^{-1}$. With the above notations, for any $Q^\pi_\mathsf{P} \in Q^\pi_\mathcal{P}$, we can construct a similar matrix $\tilde{Q}^\pi_\mathsf{P}$ as

$$\tilde{Q}^\pi_\mathsf{P} \coloneqq U^{-1} Q^\pi_\mathsf{P} U = \begin{bmatrix} 0 & \alpha_\mathsf{P}^\top \\ 0 & B_\mathsf{P} \end{bmatrix}, \quad \text{where} \;\; B_\mathsf{P} \coloneqq U_\mathbb{H}^\top \mathsf{P}^\pi U_\mathbb{H} \in \mathbb{R}^{(S-1)\times(S-1)}. \tag{49}$$

Equivalently, define $T_\mathsf{P} \coloneqq \Pi \mathsf{P}^\pi\big|_\mathbb{H}$, which operates entirely on $\mathbb{H}$. Then $B_\mathsf{P}$ is the matrix of $T_\mathsf{P}$ in the basis of $U_\mathbb{H}$. Since $E_\mathsf{P} = \mathbf{e}(d^\pi_\mathsf{P})^\top$ and $U_\mathbb{H}^\top \mathbf{e} = 0$, we have $U_\mathbb{H}^\top E_\mathsf{P} U_\mathbb{H} = 0$. Hence, the lower-right block in $U_\mathbb{H}^\top Q^\pi_\mathsf{P} U_\mathbb{H}$ is just $B_\mathsf{P}$. Consequently, for any sequence $\mathsf{P}^\pi_1, \ldots, \mathsf{P}^\pi_k$,

$$U^{-1}\big(Q^\pi_{\mathsf{P}_k} \cdots Q^\pi_{\mathsf{P}_1}\big)U = \begin{bmatrix} 0 & * \\ 0 & B_{\mathsf{P}_k} \cdots B_{\mathsf{P}_1} \end{bmatrix}. \tag{50}$$

Hence, by block upper-triangularity [20], the spectral radius of $Q^\pi_{\mathsf{P}_k} \cdots Q^\pi_{\mathsf{P}_1}$ on $\mathbb{R}^S$ equals the spectral radius of $B_{\mathsf{P}_k} \cdots B_{\mathsf{P}_1}$ on $\mathbb{H}$:

$$\begin{aligned} \rho\big(Q^\pi_{\mathsf{P}_k} \cdots Q^\pi_{\mathsf{P}_1}\big) &= \rho\big(B_{\mathsf{P}_k} \cdots B_{\mathsf{P}_1}\big) \\ &= \rho\big(U_\mathbb{H}^\top \mathsf{P}^\pi_k U_\mathbb{H} \cdots U_\mathbb{H}^\top \mathsf{P}^\pi_1 U_\mathbb{H}\big) \\ &= \rho\big(U_\mathbb{H}^\top \mathsf{P}^\pi_k \Pi \mathsf{P}^\pi_{k-1} \Pi \cdots \mathsf{P}^\pi_2 \Pi \mathsf{P}^\pi_1 U_\mathbb{H}\big) \\ &= \rho(T_{\mathsf{P}_k} \ldots T_{\mathsf{P}_1}). \end{aligned} \tag{51}$$

Thus, the spectral radius of the family $Q^\pi_{\mathsf{P}_k} \cdots Q^\pi_{\mathsf{P}_1}$ on $\mathbb{R}^S$ equals the spectral radius of $\mathsf{P}^\pi_k \cdots \mathsf{P}^\pi_1$ on $\mathbb{H}$:

$$\rho\big(Q^\pi_{\mathsf{P}_k} \cdots Q^\pi_{\mathsf{P}_1}\big) = \rho(T_{\mathsf{P}_k} \ldots T_{\mathsf{P}_1}) = \rho(\Pi \mathsf{P}^\pi_k \cdots \mathsf{P}^\pi_1\big|_\mathbb{H}). \tag{52}$$

Given (52), we now study the joint spectral radius of $\mathcal{P}_\mathbb{H} = \{\Pi \mathsf{P}^\pi : \mathsf{P} \in \mathcal{P}\}$ on $\mathbb{H}$.

From Lemma F.1, we have that for an arbitrary norm $\|\cdot\|_{\mathbb{H}\to\mathbb{H}}$ on $\mathbb{H}$,

$$\hat{\rho}(\mathcal{P}_\mathbb{H}) = \lim_{k\to\infty} \sup_{\mathsf{P}_i \in \mathcal{P}} \|\Pi \mathsf{P}^\pi_k \ldots \mathsf{P}^\pi_1\|^{\frac{1}{k}}_{\mathbb{H}\to\mathbb{H}} = \lim_{k\to\infty} \sup_{\mathsf{P}_i \in \mathcal{P}} \|T_{\mathsf{P}_k} \ldots T_{\mathsf{P}_1}\|^{\frac{1}{k}}_{\mathbb{H}\to\mathbb{H}}. \tag{53}$$

Divide $k$ into $m$ partitions of blocks with length $q$ and a residue $r$ as $k = qm + r$, where $q, m, r \in \mathbb{N}$, $0 < q \leq k$ and $0 \leq r < q$. Furthermore, let $M_m \coloneqq \sup_{\mathsf{P}_i \in \mathcal{P}} \|T_{\mathsf{P}_m} \dots T_{\mathsf{P}_1}\|_{\mathbb{H}\to\mathbb{H}}$ and $K_{<m} \coloneqq \max_{0\leq r<m} \sup_{\mathsf{P}_i\in\mathcal{P}} \|T_{\mathsf{P}_r} \dots T_{\mathsf{P}_1}\|_{\mathbb{H}\to\mathbb{H}}$, then by the submultiplicity of operator norm, we have that on $\mathbb{H}$,

$$\sup_{\mathsf{P}_i\in\mathcal{P}} \|T_{\mathsf{P}_k} \dots T_{\mathsf{P}_1}\|_{\mathbb{H}\to\mathbb{H}} \leq M_m^q K_{<m}, \tag{54}$$

taking power of $\frac{1}{k}$ and let $k \to \infty$ implies

$$\lim_{k\to\infty} \sup_{\mathsf{P}_i\in\mathcal{P}} \|T_{\mathsf{P}_k} \dots T_{\mathsf{P}_1}\|_{\mathbb{H}\to\mathbb{H}}^{\frac{1}{k}} \leq \lim_{k\to\infty} M_m^{\frac{q}{k}} K_{<m}^{\frac{1}{k}}. \tag{55}$$

Since $q = \lfloor \frac{k}{m} \rfloor$, we have $\lim_{k\to\infty} \frac{q}{k} = \frac{1}{m}$ and $\lim_{k\to\infty} \frac{1}{k} = 0$, which suggests that for any positive integer $m$ we have,

$$\lim_{k\to\infty} \sup_{\mathsf{P}_i\in\mathcal{P}} \|T_{\mathsf{P}_k} \dots T_{\mathsf{P}_1}\|_{\mathbb{H}\to\mathbb{H}}^{\frac{1}{k}} \leq M_m^{\frac{1}{m}} = \left( \sup_{\mathsf{P}_i\in\mathcal{P}} \|T_{\mathsf{P}_m} \dots T_{\mathsf{P}_1}\|_{\mathbb{H}\to\mathbb{H}} \right)^{\frac{1}{m}}, \tag{56}$$

which implies for any norm $\|\cdot\|_{\mathbb{H}\to\mathbb{H}}$ on $\mathbb{H}$, we have

$$\lim_{k\to\infty} \sup_{\mathsf{P}_i\in\mathcal{P}} \|T_{\mathsf{P}_k} \dots T_{\mathsf{P}_1}\|_{\mathbb{H}\to\mathbb{H}}^{\frac{1}{k}} \leq \inf_{m\geq 1} \left( \sup_{\mathsf{P}_i\in\mathcal{P}} \|T_{\mathsf{P}_m} \dots T_{\mathsf{P}_1}\|_{\mathbb{H}\to\mathbb{H}} \right)^{\frac{1}{m}}. \tag{57}$$

From [16], the Dobrushin's coefficient is a valid norm (the induced matrix span norm) on the zero-sum subspace $\mathbb{H}$, which yields (48). □

Lemma A.3 provides a quantitative method to relate the joint spectral radius of the family $Q_{\mathcal{P}}^{\pi}$ and the Dobrushin's coefficient of the family $\mathcal{P}$. Under Assumption 3.1, we next discuss the radius restrictions of contamination, TV and Wasserstein distance uncertainty sets such that $\hat{\rho}(Q_{\mathcal{P}}^{\pi}) < 1$ is satisfied.

### A.2.2 Discussions on Radius Restrictions

We provide the following Lemma A.4-A.6, which quantifies the radius restrictions regarding all three uncertainty sets of interests for obtaining the desired results.

**Contamination Uncertainty** Regarding contamination uncertainty, where the uncertainty set is characterized as

$$\mathcal{P} := \left\{ \mathsf{P} : \forall (s,a), \mathsf{P}(\cdot|s,a) \;=\; (1-\delta)\tilde{\mathsf{P}}(\cdot|s,a) + \delta\, q(\cdot|s,a), q(\cdot|s,a) \in \Delta(\mathcal{S}) \right\}, \quad 0 \leq \delta < 1,$$

For a fixed policy $\pi$, the induced state–transition matrix $\mathsf{P}^{\pi}$ is expressed as

$$\mathsf{P}^{\pi}(s,s') \coloneqq \sum_a \pi(a|s)\mathsf{P}(s'|s,a) = (1-\delta)\sum_a \pi(a|s)\tilde{\mathsf{P}}(s'|s,a) + \delta \sum_a \pi(a|s) q(s'|s,a) \tag{58}$$

Define the induced uncertainty set $\mathcal{P}^{\pi} \coloneqq \{\mathsf{P}^{\pi} : \mathsf{P} \in \mathcal{P}\}$ and define $\tilde{\mathsf{P}}^{\pi}(s,s') \coloneqq \sum_a \pi(a|s)\tilde{\mathsf{P}}(s'|s,a)$. Then (58) can be expressed as

$$\mathcal{P}^{\pi} \;=\; \left\{ (1-\delta)\tilde{\mathsf{P}}^{\pi} \;+\; \delta q^{\pi} \;:\; q^{\pi} \text{ row–stochastic} \right\}. \tag{59}$$

**Lemma A.4.** *Under the contamination uncertainty set, if the centroid $\tilde{\mathsf{P}}^{\pi}$ is irreducible and aperiodic, then the joint spectral radius of $Q_{\mathcal{P}}^{\pi}$ defined in* (46) *is strictly less than* 1. *Furthermore, $\mathsf{P}^{\pi}$ is irreducible and aperiodic for all $\mathsf{P} \in \mathcal{P}$.*

*Proof.* Since $\tilde{\mathsf{P}}^{\pi}$ is irreducible and aperiodic. Then there exists an $m \in \mathbb{N}$ such that all entries in $(\tilde{\mathsf{P}}^{\pi})^m$ are strictly positive. For any $\mathsf{P}^{\pi} \in \mathcal{P}^{\pi}$ we can write $\mathsf{P}^{\pi} = (1-\delta)\tilde{\mathsf{P}}^{\pi} + \delta q^{\pi}$ with $q^{\pi}$ being row–stochastic, so by multinomial expansion,

$$(\mathsf{P}^{\pi})^m = \left((1-\delta)\tilde{\mathsf{P}}^{\pi} + \delta q^{\pi}\right)^m \geq (1-\delta)^m (\tilde{\mathsf{P}}^{\pi})^m \quad \text{(entrywise)}. \tag{60}$$

Hence $(\mathsf{P}^\pi)^m$ is strictly positive for all $\mathsf{P} \in \mathcal{P}$, which implies every $\mathsf{P}^\pi \in \mathcal{P}^\pi$ is primitive with the same exponent $m$, which further implies $\mathsf{P}^\pi$ is irreducible and aperiodic for all $\mathsf{P} \in \mathcal{P}$.

To bound the joint spectral radius, for the same integer $m$, define the $m$–step overlap constant of the centroid as

$$a_0 \coloneqq \min_{i<j} \sum_{s\in\mathcal{S}} \min\big\{(\tilde{\mathsf{P}}^\pi)_{is}, (\tilde{\mathsf{P}}^\pi)_{js}\big\} \ \in (0,1].$$

For any length–$m$ product $\mathsf{P}_m^\pi \cdots \mathsf{P}_1^\pi$ with $\mathsf{P}_t^\pi \in \mathcal{P}^\pi$, the same entrywise bound (60) gives $\mathsf{P}_m^\pi \cdots \mathsf{P}_1^\pi \geq (1-\delta)^m (\tilde{\mathsf{P}}^\pi)^m$, whence for all $i \neq j$, we have

$$\sum_s \min\{(\mathsf{P}_m^\pi \cdots \mathsf{P}_1^\pi)_{is}, (\mathsf{P}_m^\pi \cdots \mathsf{P}_1^\pi)_{js}\} \geq (1-\delta)^m a_0. \tag{61}$$

By the definition of the Dobrushin's coefficient in (47), the above yields

$$\tau(\mathsf{P}_m^\pi \cdots \mathsf{P}_1^\pi) \ \leq \ 1 - (1-\delta)^m a_0 \ < \ 1. \tag{62}$$

By Lemma A.3,

$$\hat{\rho}(Q_{\mathcal{P}}^\pi) \leq \inf_{t\geq 1} \left( \sup_{\mathsf{P}_i \in \mathcal{P}} \tau(\mathsf{P}_1^\pi \cdot \ldots \cdot \mathsf{P}_t^\pi) \right)^{\frac{1}{t}} \ \leq \ \big(1 - (1-\delta)^m a_0\big)^{1/m} \ < \ 1.$$

□

Therefore, if the center $\tilde{\mathsf{P}}^\pi$ is primitive and $0 \leq \delta < 1$, without having any additional restrictions on the radius, we have that all induced kernels in $\mathcal{P}_\pi$ are irreducible and aperiodic. Furthermore, the joint spectral radius of $Q_{\mathcal{P}}^\pi$ satisfies $\hat{\rho}(Q_{\mathcal{P}}^\pi) < 1$.

**Total Variation (TV) Distance Uncertainty** Regarding TV uncertainty, where the uncertainty set is characterized as

$$\mathcal{P} \coloneqq \Big\{ \mathsf{P} : \ \forall (s,a), \ \mathrm{TV}\big(P(\cdot|s,a), \tilde{\mathsf{P}}(\cdot|s,a)\big) \ \leq \ \delta \Big\}, \qquad \delta \geq 0,$$

where $\mathrm{TV}(p,q) \coloneqq \frac{1}{2}\|p-q\|_1$. For a fixed policy $\pi$, the induced state–transition matrix $\mathsf{P}^\pi$ is expressed as

$$\mathsf{P}^\pi(s,s') \coloneqq \sum_a \pi(a|s) \mathsf{P}(s'|s,a) \tag{63}$$

Then for each state $s$,

$$\begin{aligned} \mathrm{TV}\big(\mathsf{P}^\pi(s,\cdot), \tilde{\mathsf{P}}^\pi(s,\cdot)\big) &= \mathrm{TV}\Big(\sum_a \pi(a|s)\mathsf{P}(\cdot|s,a), \sum_a \pi(a|s)\tilde{\mathsf{P}}(\cdot|s,a)\Big) \\ &\leq \sum_a \pi(a|s) \mathrm{TV}\big(\mathsf{P}(\cdot|s,a), \tilde{\mathsf{P}}(\cdot|s,a)\big) \\ &\leq \delta, \end{aligned} \tag{64}$$

by convexity of $\mathrm{TV}(\cdot,\cdot)$ in each argument. Hence

$$\mathcal{P}^\pi \subseteq \Big\{ M \text{ row–stochastic} : \ \forall s, \ \mathrm{TV}\big(M(s,\cdot), \tilde{\mathsf{P}}^\pi(s,\cdot)\big) \leq \delta \Big\}. \tag{65}$$

**Lemma A.5.** *Under the TV distance uncertainty set, if the centroid $\tilde{\mathsf{P}}^\pi$ is irreducible and aperiodic, then there exists $m \in \mathbb{N}$ such that $(\tilde{\mathsf{P}}^\pi)^m$ is strictly positive. Define $b_0 = \min_{i,s}((\tilde{\mathsf{P}}^\pi)^m)_{is} > 0$, then if the radius satisfies $\delta < \frac{b_0}{m}$, the joint spectral radius of $Q_{\mathcal{P}}^\pi$ defined in* (46) *is strictly less than* 1. *Furthermore, $\mathsf{P}^\pi$ is irreducible and aperiodic for all $\mathsf{P} \in \mathcal{P}$.*

*Proof.* Define the $m$–step constant $a_0$ as

$$a_0 = \min_{i<j} \sum_{s\in\mathcal{S}} \min\big\{(\tilde{\mathsf{P}}^\pi)^m)_{is}, ((\tilde{\mathsf{P}}^\pi)^m)_{js}\big\} \ \in (0,1],$$

then $a_0 \geq S\, b_0$ where $S = |\mathcal{S}|$.

Regarding the joint spectral radius, for any length-$m$ product $\mathsf{P}_m^\pi \cdots \mathsf{P}_1^\pi$ with $\mathsf{P}_t^\pi \in \mathcal{P}^\pi$. By a telescoping expansion and nonexpansiveness of TV under right–multiplication by a Markov kernel,

$$\mathrm{TV}\big((\mathsf{P}_m^\pi \cdots \mathsf{P}_1^\pi)(i,\cdot), (\tilde{\mathsf{P}}^\pi)^m(i,\cdot)\big) \;\leq m\delta \qquad \text{for all rows } i.$$

Then, for all $i \neq j$,

$$\begin{aligned}\sum_s \min\{(\mathsf{P}_m^\pi \cdots \mathsf{P}_1^\pi)_{is}, (\mathsf{P}_m^\pi \cdots \mathsf{P}_1^\pi)_{js}\} &\geq 1 - \mathrm{TV}\big((\mathsf{P}_m^\pi \cdots \mathsf{P}_1^\pi)(i,\cdot), (\mathsf{P}_m^\pi \cdots \mathsf{P}_1^\pi)(j,\cdot)\big) \\ &\geq a_0 - 2m\delta, \end{aligned} \tag{66}$$

by the triangle inequality in TV. Hence

$$\tau((\mathsf{P}_m^\pi \cdots \mathsf{P}_1^\pi)) = 1 - \min_{i<j} \sum_s \min\{(\mathsf{P}_m^\pi \cdots \mathsf{P}_1^\pi)_{is}, (\mathsf{P}_m^\pi \cdots \mathsf{P}_1^\pi)_{js}\} \;\leq\; 1 - (a_0 - 2m\delta). \tag{67}$$

By setting $\delta < \frac{a_0}{2m}$, we have $\sup_{\mathsf{P}_i \in \mathcal{P}} \tau((\mathsf{P}_m^\pi \cdots \mathsf{P}_1^\pi)) < 1$, and by Lemma A.3,

$$\hat{\rho}\big(Q_{\mathcal{P}}^\pi\big) \;\leq\; \Big(\sup_{\mathsf{P}_i \in \mathcal{P}} \tau((\mathsf{P}_m^\pi \cdots \mathsf{P}_1^\pi))\Big)^{1/m} \;\leq\; \big(1 - (a_0 - 2m\delta)\big)^{1/m} \;<\; 1. \tag{68}$$

Similarly, the same perturbation bound yields

$$\min_s (\mathsf{P}^\pi)_{is}^m \;\geq\; \min_s ((\tilde{\mathsf{P}}^\pi))^m)_{is} - m\delta \geq\; b_0 - m\delta,$$

so by setting $\delta < \frac{b_0}{m}$, we have that $(\mathsf{P}^\pi)^m$ is strictly positive for every $\mathsf{P} \in \mathcal{P}$; hence all induced kernels are irreducible and aperiodic. Since $a_0 \geq S b_0$, we have $\frac{a_0}{2m} \geq \frac{b_0}{m}$ for $S \geq 2$. Therefore the condition that $\delta < \frac{b_0}{m}$ satisfies both requirements. □

**Wasserstein Distance Uncertainty** Regarding Wasserstein uncertainty with $p \geq 1$, let $(\mathcal{S}, d)$ be the finite metric space. The uncertainty set can be characterized as

$$\mathcal{P} \coloneqq \Big\{\mathsf{P} : \forall (s,a), W_p\big(P(\cdot|s,a), \tilde{\mathsf{P}}(\cdot|s,a); d\big) \leq \delta\Big\}.$$

For a fixed policy $\pi$, the induced state–transition matrix $\mathsf{P}^\pi$ is expressed as

$$\mathsf{P}^\pi(s,s') \coloneqq \sum_a \pi(a|s)\mathsf{P}(s'|s,a) \tag{69}$$

For each state $s$, by joint convexity of $W_p^p(\cdot,\cdot;d)$,

$$\begin{aligned} W_p^p\Big(\mathsf{P}^\pi(s,\cdot), \tilde{\mathsf{P}}^\pi(s,\cdot); d\Big) &= W_p^p\Big(\sum_a \pi(a|s)\mathsf{P}(\cdot|s,a), \sum_a \pi(a|s)\tilde{\mathsf{P}}(\cdot|s,a); d\Big) \\ &\leq \sum_a \pi(a|s)\, W_p^p\big(\mathsf{P}(\cdot|s,a), \tilde{\mathsf{P}}(\cdot|s,a); d\big) \leq \delta^p, \end{aligned}$$

hence $W_p\big(\mathsf{P}^\pi(s,\cdot), \tilde{\mathsf{P}}^\pi(s,\cdot); d\big) \leq \delta$ for all $s$, i.e.

$$\mathcal{P}^\pi \;\subseteq \big\{\, M \text{ row–stochastic} :\; \forall s,\; W_p(M(s,\cdot), \tilde{\mathsf{P}}^\pi(s,\cdot); d) \leq \delta\big\}. \tag{70}$$

We now draw connection between (70) and the TV version in (65). Since the state space is finite, denote $\delta_{\min} := \min_{x \neq y} d(x,y) > 0$. Then, for any distributions $u, v$, we have

$$W_1(u,v;d) \;\geq\; \delta_{\min}\, \mathrm{TV}(u,v) \qquad \text{and} \qquad W_p(u,v;d) \;\geq\; W_1(u,v;d),$$

which implies that

$$\mathrm{TV}(u,v) \;\leq\; \frac{W_1(u,v;d)}{\delta_{\min}} \;\leq \frac{W_p(u,v;d)}{\delta_{\min}}. \tag{71}$$

Therefore we can reduce (70) into a TV distance uncertainty set characterized as follows:

$$\mathcal{P}^\pi \subseteq \Big\{\, M \text{ row–stochastic} :\; \forall s,\; \mathrm{TV}\big(M(s,\cdot), \tilde{\mathsf{P}}^\pi(s,\cdot)\big) \leq \frac{\delta}{\delta_{\min}}\Big\}. \tag{72}$$

**Lemma A.6.** *Under the Wasserstein distance uncertainty set, if the centroid $\tilde{\mathsf{P}}^\pi$ is irreducible and aperiodic, then there exists $m \in \mathbb{N}$ such that $(\tilde{\mathsf{P}}^\pi)^m$ is strictly positive. Define $b_0 = \min_{i,s}((\tilde{\mathsf{P}}^\pi)^m)_{is} > 0$ and $\delta_{\min} := \min_{x \neq y} d(x,y) > 0$, then if the radius satisfies $\delta < \frac{\delta_{\min} b_0}{m}$, the joint spectral radius of $Q^\pi_{\mathcal{P}}$ defined in* (46) *is strictly less than* 1*. Furthermore, $\mathsf{P}^\pi$ is irreducible and aperiodic for all $\mathsf{P} \in \mathcal{P}$.*

*Proof.* This is a direct corollary of Lemma A.5 under the condition of (72). □

*Remarks.* (i) If $d$ is normalized so $\delta_{\min} = 1$, the thresholds simplify accordingly. (ii) One can also argue via $W_p^p \geq \delta_{\min}^p$ TV, which gives the alternative (more conservative when $\varepsilon$ is small) choice $r = \delta^p / \delta_{\min}^p$; the linear reduction $r = \delta / \delta_{\min}$ above is sharper and suffices for the bounds.

### A.2.3 Extremal Norm Construction

Under the radius conditions of Lemma A.4-A.6, we have that :

$$r^* \coloneqq \hat{\rho}(Q^\pi_{\mathcal{P}}) < 1 \tag{73}$$

We follow similar process for constructing our desired semi-norm $\|\cdot\|_{\mathcal{P}}$ as in Appendix A.1 by first constructing a norm such that all $Q^\pi_{\mathsf{P}}$ are strictly less then one under that norm. We choose $\alpha \in (r^*, 1)$ and we follow the approach in [42] by constructing an extremal norm $\|\cdot\|_{\text{ext}}$ as follows:

$$\|x\|_{\text{ext}} \coloneqq \sup_{k \geq 0} \sup_{Q_1, \ldots, Q_k \in Q^\pi_{\mathcal{P}}} \alpha^{-k} \|Q_k Q_{k-1} \ldots Q_1 x\|_2 \quad \text{where} \quad Q^\pi_{\mathcal{P}} = \{Q^\pi_{\mathsf{P}} : \mathsf{P} \in \mathcal{P}\} \tag{74}$$

Note that we follow the convention that $\|Q_k Q_{k-1} \ldots Q_1 x\|_2 = \|x\|_2$ when $k = 0$.

**Lemma A.7.** *Under Assumption 3.1 and the radius conditions of Lemma A.4-A.6, the operator $\|\cdot\|_{\text{ext}}$ is a valid norm with $\|Q^\pi_{\mathsf{P}}\|_{\text{ext}} < 1$ for all $\mathsf{P} \in \mathcal{P}$*

*Proof.* We first prove that $\|\cdot\|_{\text{ext}}$ is bounded. Following Lemma F.1 and choosing $\lambda \in (r^*, \alpha)$, then there exist a positive constant $C < \infty$ such that

$$\|Q_k Q_{k-1} \ldots Q_1\|_2 \leq C \lambda^k \tag{75}$$

Hence for each $k$ and for all $x \in \mathbb{R}^S$,

$$\alpha^{-k} \|Q_k Q_{k-1} \ldots Q_1 x\|_2 \leq \alpha^{-k} C \lambda^k \|x\|_2 = C \left(\frac{\lambda}{\alpha}\right)^{-k} \|x\|_2 \longrightarrow 0 \quad \text{as} \quad k \to \infty \tag{76}$$

Thus the double supremum in (74) is over a bounded and vanishing sequence, so $\|\cdot\|_{\text{ext}}$ bounded.

To check that $\|\cdot\|_{\text{ext}}$ is a valid norm, note that if $x = \mathbf{0}$, $\|x\|_{\text{ext}}$ is directly $0$. On the other hand, if $\|x\|_{\text{ext}} = 0$, we have

$$\sup_{k=0} \sup_{Q_1, \ldots, Q_k \in Q^\pi_{\mathcal{P}}} \alpha^{-k} \|Q_k Q_{k-1} \ldots Q_1 x\|_2 = \|x\|_2 = 0 \Rightarrow x = \mathbf{0} \tag{77}$$

Regarding homogeneity, observe that for any $c \in \mathbb{R}$ and $x \in \mathbb{R}^S$,

$$\|cx\|_{\text{ext}} = \sup_{k \geq 0} \sup_{Q_1, \ldots, Q_k \in Q^\pi_{\mathcal{P}}} \alpha^{-k} \|Q_k Q_{k-1} \ldots Q_1 (cx)\|_2 = |c| \|x\|_{\text{ext}} \tag{78}$$

Regarding triangle inequality, using $\|Q_k \ldots Q_1 (x+y)\|_2 \leq \|Q_k \ldots Q_1 x\|_2 + \|Q_k \ldots Q_1 y\|_2$ for any $x, y \in \mathbb{R}^S$, we obtain,

$$\|x+y\|_{\text{ext}} = \sup_{k \geq 0} \sup_{Q_1, \ldots, Q_k \in Q^\pi_{\mathcal{P}}} \alpha^{-k} \|Q_k Q_{k-1} \ldots Q_1 (x+y)\|_2 \leq \|x\|_{\text{ext}} + \|y\|_{\text{ext}} \tag{79}$$

For any $\mathsf{P} \in \mathcal{P}$, we have

$$\begin{aligned}
\|Q_{\mathsf{P}}^{\pi} x\|_{\text{ext}} &= \sup_{k \geq 0} \sup_{Q_1, \ldots, Q_k \in Q_{\mathcal{P}}^{\pi}} \alpha^{-k} \|Q_k Q_{k-1} \ldots Q_1 (Q_{\mathsf{P}}^{\pi} x)\|_2 \\
&\leq \sup_{k \geq 1} \sup_{Q_1, \ldots, Q_k \in Q_{\mathcal{P}}^{\pi}} \alpha^{-(k-1)} \|Q_k Q_{k-1} \ldots Q_1 x\|_2 \\
&= \alpha \sup_{k \geq 1} \sup_{Q_1, \ldots, Q_k \in Q_{\mathcal{P}}^{\pi}} \alpha^{-k} \|Q_k Q_{k-1} \ldots Q_1 x\|_2 \\
&\leq \alpha \sup_{k \geq 0} \sup_{Q_1, \ldots, Q_k \in Q_{\mathcal{P}}^{\pi}} \alpha^{-k} \|Q_k Q_{k-1} \ldots Q_1 x\|_2 \\
&= \alpha \|x\|_{\text{ext}} \qquad (80)
\end{aligned}$$

Since $\mathsf{P}$ is arbitrary, (80) implies that for any $\mathsf{P} \in \mathcal{P}$,

$$\|Q_{\mathsf{P}}^{\pi}\|_{\text{ext}} = \sup_{x \neq \mathbf{0}} \frac{\|Q_{\mathsf{P}}^{\pi} x\|_{\text{ext}}}{\|x\|_{\text{ext}}} \leq \alpha < 1 \qquad (81)$$

□

### A.2.4 Semi-Norm Contraction for Robust Bellman Operator

We now follow the same method as (41) to construct the semi-norm $\|\cdot\|_{\mathcal{P}}$. Define the operator $\|\cdot\|_{\mathcal{P}}$ as

$$\|x\|_{\mathcal{P}} \coloneqq \sup_{\mathsf{P} \in \mathcal{P}} \left\|Q_{\mathsf{P}}^{\pi} x\right\|_{\text{ext}} + \epsilon \inf_{c \in \mathbb{R}} \left\|x - c\mathbf{e}\right\|_{\text{ext}} \qquad (82)$$

where $0 < \epsilon < 1 - \alpha$.

**Lemma A.8.** *The operator $\|\cdot\|_{\mathcal{P}}$ is a valid semi-norm with kernel being exactly $\{c\mathbf{e} : c \in \mathbb{R}\}$ under Assumption 3.1 and the radius conditions of Lemma A.4-A.6. Furthermore, for all $x \in \mathbb{R}^S$, we have $\left\|\mathsf{P}^{\pi} x\right\|_{\mathcal{P}} \leq (\alpha + \epsilon)\left\|x\right\|_{\mathcal{P}}$ for all $\mathsf{P} \in \mathcal{P}$.*

*Proof.* Regarding positive homogeneity and nonnegativity, for any scalar $\lambda$ and $x \in \mathbb{R}^S$,

$$\|\lambda x\|_{\mathcal{P}} = \sup_{\mathsf{P} \in \mathcal{P}} \left\|Q_{\mathsf{P}}^{\pi}(\lambda x)\right\|_{\text{ext}} + \epsilon \inf_{c \in \mathbb{R}} \left\|\lambda x - c\mathbf{e}\right\|_{\text{ext}} = |\lambda| \sup_{\mathsf{P} \in \mathcal{P}} \left\|Q_{\mathsf{P}}^{\pi} x\right\|_{\text{ext}} + \epsilon |\lambda| \inf_{c \in \mathbb{R}} \left\|x - c\mathbf{e}\right\|_{\text{ext}} = |\lambda| \|x\|_{\text{ext}}$$

and $\|x\|_{\text{ext}} \geq 0$. Regarding triangle inequality, for any $x, y \in \mathbb{R}^S$, note that for any $\mathsf{P} \in \mathcal{P}$,

$$\|Q_{\mathsf{P}}^{\pi}(x + y)\|_{\text{ext}} \leq \|Q_{\mathsf{P}}^{\pi} x\|_{\text{ext}} + \|Q_{\mathsf{P}}^{\pi} y\|_{\text{ext}} \qquad (83)$$

Taking supremum over $\mathsf{P}$ on both sides yields

$$\sup_{\mathsf{P} \in \mathcal{P}} \|Q_{\mathsf{P}}^{\pi}(x + y)\|_{\text{ext}} \leq \sup_{\mathsf{P} \in \mathcal{P}} \|Q_{\mathsf{P}}^{\pi} x\|_{\text{ext}} + \sup_{\mathsf{P} \in \mathcal{P}} \|Q_{\mathsf{P}}^{\pi} y\|_{\text{ext}} \qquad (84)$$

Thus, we have

$$\begin{aligned}
\|x + y\|_{\mathcal{P}} &= \sup_{\mathsf{P} \in \mathcal{P}} \left\|Q_{\mathsf{P}}^{\pi}(x + y)\right\|_{\text{ext}} + \epsilon \inf_{c} \left\|x + y - c\mathbf{e}\right\|_{\text{ext}} \\
&\leq \sup_{\mathsf{P} \in \mathcal{P}} \|Q_{\mathsf{P}}^{\pi} x\|_{\text{ext}} + \sup_{\mathsf{P} \in \mathcal{P}} \|Q_{\mathsf{P}}^{\pi} y\|_{\text{ext}} + \epsilon \inf_{a,b} \left\|x - a\mathbf{e} + y - b\mathbf{e}\right\|_{\text{ext}} \\
&\leq \sup_{\mathsf{P} \in \mathcal{P}} \|Q_{\mathsf{P}}^{\pi} x\|_{\text{ext}} + \sup_{\mathsf{P} \in \mathcal{P}} \|Q_{\mathsf{P}}^{\pi} y\|_{\text{ext}} + \epsilon \inf_{a} \left\|x - a\mathbf{e}\right\|_{\text{ext}} + \epsilon \inf_{b} \left\|y - b\mathbf{e}\right\|_{\text{ext}} \\
&= \|x\|_{\mathcal{P}} + \|y\|_{\mathcal{P}}.
\end{aligned}$$

Regarding the kernel, if $x = k\mathbf{e}$ for some $k \in \mathbb{R}$, then similar to (42), we have

$$\|x\|_{\mathcal{P}} = \sup_{\mathsf{P} \in \mathcal{P}} \left\|Q_{\mathsf{P}}^{\pi}(k\mathbf{e})\right\|_{\text{ext}} + \epsilon \inf_{c} \|k\mathbf{e} - c\mathbf{e}\|_{\text{ext}} = 0 + 0 = 0 \qquad (85)$$

On the other hand, if $x \notin \{c\mathbf{e} : c \in \mathbb{R}\}$, we know that

$$\|x\|_{\mathcal{P}} \geq \epsilon \inf_{c} \|x - c\mathbf{e}\|_{\text{ext}} > 0 \qquad (86)$$

Thus, the kernel of $\|\cdot\|_{\mathcal{P}}$ is exactly $\{c\mathbf{e} : c \in \mathbb{R}\}$. We now show that, for any $x \in \mathbb{R}^S$ and $\mathsf{P} \in \mathcal{P}$,

$$\begin{aligned}
\left\|\mathsf{P}^\pi x\right\|_{\mathcal{P}} &= \sup_{Q \in\, Q^\pi_{\mathcal{P}}} \left\|Q\mathsf{P}^\pi x\right\|_{\text{ext}} + \epsilon \inf_c \left\|\mathsf{P}^\pi x - c\mathbf{e}\right\|_{\text{ext}} \\
&= \sup_{Q \in\, Q^\pi_{\mathcal{P}}} \left\|QQ^\pi_{\mathsf{P}} x + QE_{\mathsf{P}} x\right\|_{\text{ext}} + \epsilon \inf_c \left\|Q^\pi_{\mathsf{P}} x + E_{\mathsf{P}} x - c\mathbf{e}\right\|_{\text{ext}} \\
&= \sup_{Q \in\, Q^\pi_{\mathcal{P}}} \left\|QQ^\pi_{\mathsf{P}} x\|_{\text{ext}} + \epsilon \inf_c \left\|Q^\pi_{\mathsf{P}} x - c\mathbf{e}\right\|_{\text{ext}} \right. \\
&\leq \alpha \sup_{Q \in\, Q^\pi_{\mathcal{P}}} \left\|Qx\|_{\text{ext}} + \epsilon \|Q^\pi x\|_{\text{ext}} \right. \\
&= (\alpha + \epsilon)\|Q^\pi x\|_{\text{ext}} \\
&\leq (\alpha + \epsilon)\|x\|_{\mathcal{P}}. \qquad (87)
\end{aligned}$$

□

Since $\alpha \in (0, 1)$ and $\epsilon \in (0, 1 - \alpha)$. Thus, let $\gamma = \alpha + \epsilon$ then $\gamma \in (0, 1)$. Substituting the above result back to (45), we obtain

$$\|\mathbf{T}_g(V_1) - \mathbf{T}_g(V_2)\|_{\mathcal{P}} \leq \|\pi(a|s)\tilde{p}_{(V_1,V_2)}\big[V_1(s') - V_2(s')\big]\|_{\mathcal{P}} \leq \gamma \|V_1 - V_2\|_{\mathcal{P}} \qquad (88)$$

# B Biased Stochastic Approximation Convergence Rate

In Section 4, we established that the robust Bellman operator is a contraction under the semi-norm $\|\cdot\|_{\mathcal{P}}$, ensuring that policy evaluation can be analyzed within a well-posed stochastic approximation framework. However, conventional stochastic approximation methods typically assume unbiased noise, where variance diminishes over time without introducing systematic drift. In contrast, the noise in robust policy evaluation under TV and Wasserstein distance uncertainty sets exhibits a small but persistent bias, arising from the estimators of the support functions $\hat{\sigma}_{\mathcal{P}^a_s}(V)$ (discussed in Section 5). This bias, if not properly addressed, can lead to uncontrolled error accumulation, affecting the reliability of policy evaluation. To address this challenge, this section introduces a novel analysis of biased stochastic approximation, leveraging properties of dual norms to ensure that the bias remains controlled and does not significantly impact the convergence rate. Our results extend prior work on unbiased settings and provide the first explicit finite-time guarantees, which are further used to establish the sample complexity of policy evaluation in robust average-reward RL. Specifically, we analyze the iteration complexity for solving the fixed equivalence class equation $H(x^*) - x^* \in \overline{E}$ where $\overline{E} \coloneqq \{c\mathbf{e} : c \in \mathbb{R}\}$ with $\mathbf{e}$ being the all-ones vector. The stochastic approximation iteration being used is as follows:

$$x^{t+1} = x^t + \eta_t\big[\widehat{H}(x^t) - x^t\big], \quad \text{where} \quad \widehat{H}(x^t) = H(x^t) + w^t. \qquad (89)$$

with $\eta_t > 0$ being the step-size sequence. We assume that there exist $\gamma \in (0, 1)$ such that

$$\|H(x) - H(y)\|_{\mathcal{P}} \leq \gamma \|x - y\|_{\mathcal{P}}, \quad \forall x, y \qquad (90)$$

We also assume that the noise terms $\omega^t$ are i.i.d. and have bounded bias and variance

$$\mathbb{E}[\,\|w^t\|^2_{\mathcal{P}}|\mathcal{F}^t] \leq A + B\,\|x^t - x^*\|^2_{\mathcal{P}} \quad \text{and} \quad \left\|\mathbb{E}[\,w^t|\mathcal{F}^t]\right\|_{\mathcal{P}} \leq \varepsilon_{\text{bias}} \qquad (91)$$

**Theorem B.1.** *If $x^t$ is generated by* (89) *with all assumptions in* (90) *and* (91) *satisfied, then if the stepsize $\eta_t \coloneqq \mathcal{O}(\frac{1}{t})$,*

$$\mathbb{E}\Big[\|x^T - x^*\|^2_{\mathcal{P}}\Big] \leq \mathcal{O}\left(\frac{1}{T^2}\right)\|x^0 - x^*\|^2_{\mathcal{P}} + \mathcal{O}\left(\frac{A}{(1-\gamma)^2 T}\right) + \mathcal{O}\left(\frac{x_{\sup}\varepsilon_{\textit{bias}} \log T}{1-\gamma}\right) \qquad (92)$$

*where $x_{\sup} \coloneqq \sup_x \|x\|_{\mathcal{P}}$ is the upper bound of the $\|\cdot\|_{\mathcal{P}}$ semi-norm for all $x^t$.*

Theorem B.1 adapts the analysis of [45] and extends it to a biased i.i.d. noise setting. To manage the bias terms, we leverage properties of dual norms to bound the inner product between the error term and the gradient, ensuring that the bias influence remains logarithmic in $T$ rather than growing

unbounded, while also carefully structuring the stepsize decay to mitigate long-term accumulation. This results in an extra $\varepsilon_{\text{bias}}$ term with logarithmic dependence of the total iteration $T$.

We perform analysis of the biased-noise extension to the semi-norm stochastic approximation (SA) problem by constructing a smooth convex semi-Lyapunov function for forming the negative drift [45, 9] and using properties in dual norms for managing the bias.

### B.1 Proof of Theorem B.1

#### B.1.1 Setup and Notation.

In this section, we override the notation of the semi-norm $\|\cdot\|_{\mathcal{P}}$ by re-writing it as the norm $\|\cdot\|_{\mathcal{N}}$ (defined in (96)) to the equivalence class of constant vectors. For any norm $\|\cdot\|_c$ and equivalence class $\overline{E}$, define the indicator function $\delta_{\overline{E}}$ as

$$\delta_{\bar{E}}(x) := \begin{cases} 0 & x \in \bar{E}, \\ \infty & \text{otherwise.} \end{cases} \tag{93}$$

then by [45], the semi-norm induced by norm $\|\cdot\|_c$ and equivalence class $\overline{E}$ is the infimal convolution of $\|\cdot\|_c$ and the indicator function $\delta_{\overline{E}}$ can be defined as follows

$$\|x\|_{c,\overline{E}} := (\|\cdot\|_c *_{\inf} \delta_{\overline{E}})(x) = \inf_y(\|x-y\|_c + \delta_{\overline{E}}(y)) = \inf_{e\in\overline{E}} \|x-e\|_c \quad \forall x, \tag{94}$$

where $*_{\inf}$ denotes the infimal convolution operator. Throughout the remaining section, we let $\overline{E} := \{c\mathbf{e} : c \in \mathbb{R}\}$ with $\mathbf{e}$ being the all-ones vector. Since $\|\cdot\|_{\mathcal{P}}$ constructed in (82) is a semi-norm with kernel being $\overline{E}$, we can construct a norm $\|\cdot\|_{\mathcal{N}}$ such that

$$\|x\|_{\mathcal{N},\overline{E}} := (\|\cdot\|_{\mathcal{N}} *_{\inf} \delta_{\overline{E}})(x) = \|x\|_{\mathcal{P}} \tag{95}$$

We construct $\|\cdot\|_{\mathcal{N}}$ as follows:

$$\|x\|_{\mathcal{N}} := \sup_{\mathsf{P}\in\mathcal{P}} \left\|Q^{\pi}_{\mathsf{P}} x\right\|_{\text{ext}} + \epsilon \inf_{c\in\mathbb{R}} \left\|x - c\mathbf{e}\right\|_{\text{ext}} + \epsilon\|x\|_{\text{ext}} \tag{96}$$

where $Q^{\pi}_{\mathsf{P}}$ and $\epsilon$ are defined in (82).

**Lemma B.2.** *The operator $\|\cdot\|_{\mathcal{N}}$ defined in* (96) *is a norm satisfying* (95)*.*

*Proof.* We first verify that $\|\cdot\|_{\mathcal{N}}$ is a norm. Regarding positivity, since all terms in (96) are non-negative, $\|x\|_{\mathcal{N}} \geq 0$ for all $x \in \mathbb{R}^S$ and $\|\mathbf{0}\|_{\mathcal{N}} = 0$. If $x \neq \mathbf{0}$, since $\|\cdot\|_{\text{ext}}$ is a valid norm and $\epsilon > 0$, we have

$$\|x\|_{\mathcal{N}} \geq \epsilon\|x\|_{\text{ext}} > 0.$$

Regarding homogeneity, For any $\lambda \in \mathbb{R}$, we have

$$\begin{aligned} \|\lambda x\|_{\mathcal{N}} &= \sup_{\mathsf{P}\in\mathcal{P}} \left\|Q^{\pi}_{\mathsf{P}}(\lambda x)\right\|_{\text{ext}} + \epsilon \inf_c \left\|(\lambda x) - c\mathbf{e}\right\|_{\text{ext}} + \epsilon\|\lambda x\|_{\text{ext}} \\ &= |\lambda| \sup_{\mathsf{P}\in\mathcal{P}} \left\|Q^{\pi}_{\mathsf{P}} x\right\|_{\text{ext}} + \epsilon|\lambda| \inf_c \left\|x - c\mathbf{e}\right\|_{\text{ext}} + \epsilon|\lambda|\|x\|_{\text{ext}} \\ &= |\lambda|\|x\|_{\mathcal{N}} \end{aligned}$$

Regarding triangle inequality, for any $x, y \in \mathbb{R}^S$, we have

$$\begin{aligned} \|x+y\|_{\mathcal{N}} &= \sup_{\mathsf{P}\in\mathcal{P}} \left\|Q^{\pi}_{\mathsf{P}}(x+y)\right\|_{\text{ext}} + \epsilon \inf_c \left\|(x+y) - c\mathbf{e}\right\|_{\text{ext}} + \epsilon\|x+y\|_{\text{ext}} \\ &\leq \sup_{\mathsf{P}\in\mathcal{P}} \|Q^{\pi}_{\mathsf{P}} x\|_{\text{ext}} + \sup_{\mathsf{P}\in\mathcal{P}} \|Q^{\pi}_{\mathsf{P}} y\|_{\text{ext}} + \epsilon \inf_{a,b} \left\|x - a\mathbf{e} + y - b\mathbf{e}\right\|_{\text{ext}} + \epsilon(\|x\|_{\text{ext}} + \|y\|_{\text{ext}}) \\ &\leq \sup_{\mathsf{P}\in\mathcal{P}} \|Q^{\pi}_{\mathsf{P}} x\|_{\text{ext}} + \sup_{\mathsf{P}\in\mathcal{P}} \|Q^{\pi}_{\mathsf{P}} y\|_{\text{ext}} + \epsilon \inf_a \left\|x - a\mathbf{e}\right\|_{\text{ext}} + \epsilon \inf_b \left\|y - b\mathbf{e}\right\|_{\text{ext}} + \epsilon\|x\|_{\text{ext}} + \|y\|_{\text{ext}} \\ &= \|x\|_{\mathcal{N}} + \|y\|_{\mathcal{N}}. \end{aligned}$$

We now show that since $Q^\pi_{\mathsf{P}}\mathbf{e} = 0$ for all $\mathsf{P} \in \mathcal{P}$, by the definition of infimal convolution, we have that for all $x \in \mathbb{R}^S$,

$$\begin{aligned}
(\|\cdot\|_{\mathcal{N}} *_{\inf} \delta_{\overline{E}})(x) &= \inf_{k\in\mathbb{R}} \|x - k\mathbf{e}\|_{\mathcal{N}} \\
&= \inf_{k\in\mathbb{R}} \left( \sup_{\mathsf{P}\in\mathcal{P}} \left\|Q^\pi_{\mathsf{P}} x - kQ^\pi_{\mathsf{P}}\mathbf{e}\right\|_{\text{ext}} + \epsilon \inf_{c\in\mathbb{R}} \left\|x - c\mathbf{e} - k\mathbf{e}\right\|_{\text{ext}} + \epsilon\|x - k\mathbf{e}\|_{\text{ext}} \right) \\
&= \inf_{k\in\mathbb{R}} \left( \sup_{\mathsf{P}\in\mathcal{P}} \left\|Q^\pi_{\mathsf{P}} x\right\|_{\text{ext}} + \epsilon \inf_{c\in\mathbb{R}} \left\|x - c\mathbf{e}\right\|_{\text{ext}} + \epsilon\|x - k\mathbf{e}\|_{\text{ext}} \right) \\
&= \sup_{\mathsf{P}\in\mathcal{P}} \left\|Q^\pi_{\mathsf{P}} x\right\|_{\text{ext}} + \epsilon \inf_{c\in\mathbb{R}} \left\|x - c\mathbf{e}\right\|_{\text{ext}} + \inf_{k\in\mathbb{R}} \left(\epsilon\|x - k\mathbf{e}\|_{\text{ext}}\right) \\
&= \|x\|_{\mathcal{P}}
\end{aligned}$$

□

We thus restate our problem of analyzing the iteration complexity for solving the fixed equivalence class equation $H(x^*) - x^* \in \overline{E}$, with the operator $H : \mathbb{R}^n \to \mathbb{R}^n$ satisfying the contraction property as follows:

$$\|H(x) - H(y)\|_{\mathcal{N},\overline{E}} \le \gamma \|x - y\|_{\mathcal{N},\overline{E}}, \quad \gamma \in (0,1), \quad \forall x, y \tag{97}$$

The stochastic approximation iteration being used is as follows

$$x^{t+1} = x^t + \eta_t \big[\widehat{H}(x^t) - x^t\big], \quad \text{where} \quad \widehat{H}(x^t) = H(x^t) + w^t. \tag{98}$$

We assume:

- $\mathbb{E}[\,\|w^t\|^2_{\mathcal{N},\overline{E}} | \mathcal{F}^t] \le A + B\|x^t - x^*\|^2_{\mathcal{N},\overline{E}}$ (In the robust average-reward TD case, $B = 0$).
- $\left\|\mathbb{E}[w^t|\mathcal{F}^t]\right\|_{\mathcal{N},\overline{E}} \le \varepsilon_{\text{bias}}$.
- $\eta_t > 0$ is a chosen stepsize sequence (decreasing or constant).

Note that beside the bias in the noise, the above formulation and assumptions are identical to the unbiased setups in Section B of [45]. Thus, we emphasize mostly on managing the bias.

### B.1.2 Semi-Lyapunov $M_{\overline{E}}(\cdot)$ and Smoothness.

By [45, Proposition 1–2], using the Moreau envelope function $M(x)$ in Definition 2.2 of [8], we define

$$M_{\overline{E}}(x) = \big(M *_{\inf} \delta_{\overline{E}}\big)(x),$$

so that there exist $c_l, c_u > 0$ with

$$c_l M_{\overline{E}}(x) \le \frac{1}{2}\|x\|^2_{\mathcal{N},\overline{E}} \le c_u M_{\overline{E}}(x), \tag{99}$$

and $M_{\overline{E}}$ is $L$-smooth *w.r.t. another semi-norm* $\|\cdot\|_{s,\overline{E}}$. Concretely, $L$-smoothness means:

$$M_{\overline{E}}(y) \le M_{\overline{E}}(x) + \langle \nabla M_{\overline{E}}(x),\, y - x\rangle + \tfrac{L}{2}\,\|\,y - x\|^2_{s,\overline{E}}, \quad \forall\, x, y. \tag{100}$$

Moreover, the gradient of $M_{\overline{E}}$ satisfies $\langle \nabla M_{\overline{E}}(x),\ c\,\mathbf{e}\rangle = 0$ for all $x$, and the dual norm denoted as $\|\cdot\|_{*,s,\overline{E}}$ is also L-smooth:

$$\|\nabla M_{\overline{E}}(x) - \nabla M_{\overline{E}}(y)\|_{*,s,\overline{E}} \le L\|\,y - x\|_{s,\overline{E}}, \quad \forall\, x, y. \tag{101}$$

Note that since $\|\cdot\|_{s,\overline{E}}$ and $\|\cdot\|_{\mathcal{N},\overline{E}}$ are semi-norms on a finite-dimensional space with the same kernel, there exist $\rho_1, \rho_2 > 0$ such that

$$\rho_1\,\|z\|_{\mathcal{N},\overline{E}} \le \|z\|_{s,\overline{E}} \le \rho_2\,\|z\|_{\mathcal{N},\overline{E}},\ \forall\, z. \tag{102}$$

Likewise, their dual norms (denoted $\|\cdot\|_{*,s,\overline{E}}$ and $\|\cdot\|_{*,\mathcal{N},\overline{E}}$) satisfy the following:

$$\frac{1}{\rho_2}\|z\|_{*,s,\overline{E}} \le \|z\|_{*,\mathcal{N},\overline{E}} \le \frac{1}{\rho_1}\|z\|_{*,s,\overline{E}},\ \forall\, z. \tag{103}$$

### B.1.3 Formal Statement of Theorem B.1

By $L$-smoothness w.r.t. $\|\cdot\|_{s,\overline{E}}$ in (100), for each $t$,

$$M_{\overline{E}}(x^{t+1}-x^*) \le M_{\overline{E}}(x^t-x^*) + \big\langle \nabla M_{\overline{E}}(x^t-x^*), x^{t+1}-x^t\big\rangle + \tfrac{L}{2}\|x^{t+1}-x^t\|^2_{s,\overline{E}}. \quad (104)$$

where $x^{t+1}-x^t = \eta_t[\widehat{H}(x^t)-x^t] = \eta_t[H(x^t)+w^t-x^t]$. Taking expectation of the second term of the RHS of (104) conditioned on the filtration $\mathcal{F}^t$ we obtain,

$$\begin{aligned}\mathbb{E}[\langle \nabla M_{\overline{E}}(x^t-x^*), x^{t+1}-x^t\rangle|\mathcal{F}^t] &= \eta_t\mathbb{E}[\langle \nabla M_{\overline{E}}(x^t-x^*), H(x^t)-x^t+\omega^t\rangle|\mathcal{F}^t]\\ &= \eta_t\langle \nabla M_{\overline{E}}(x^t-x^*), H(x^t)-x^t\rangle + \eta_t\mathbb{E}[\langle \nabla M_{\overline{E}}(x^t-x^*), \omega^t\rangle|\mathcal{F}^t]\\ &= \eta_t\langle \nabla M_{\overline{E}}(x^t-x^*), H(x^t)-x^t\rangle + \eta_t\langle \nabla M_{\overline{E}}(x^t-x^*), \mathbb{E}[\omega^t|\mathcal{F}^t]\rangle. \quad (105)\end{aligned}$$

To analyze the additional bias term $\langle \nabla M_{\overline{E}}(x^t-x^*), \mathbb{E}[\omega^t|\mathcal{F}^t]\rangle$, we use the fact that for any (semi-)norm $\|\cdot\|$ with dual (semi-)norm $\|\cdot\|_*$ (defined by $\|u\|_* = \sup\{\langle u,v\rangle : \|v\|\le 1\}$), we have the general inequality

$$\big\langle u,\, v\big\rangle \le \|u\|_*\,\|v\|, \quad \forall\, u, v. \quad (106)$$

In the biased noise setting, $u = \nabla M_{\overline{E}}(x^t-x^*)$ and $v = \mathbb{E}[w^t|\mathcal{F}^t]$, with $\|\cdot\| = \|\cdot\|_{\mathcal{N},\overline{E}}$. So

$$\big\langle \nabla M_{\overline{E}}(x^t-x^*), \mathbb{E}[w^t|\mathcal{F}^t]\big\rangle \le \big\|\nabla M_{\overline{E}}(x^t-x^*)\big\|_{*,\mathcal{N},\overline{E}}\cdot\big\|\mathbb{E}[w^t|\mathcal{F}^t]\big\|_{\mathcal{N},\overline{E}}. \quad (107)$$

Since $\|\mathbb{E}[w^t|\mathcal{F}^t]\|_{\mathcal{N},\overline{E}} \le \varepsilon_{\text{bias}}$, it remains to bound $\|\nabla M_{\overline{E}}(x^t-x^*)\|_{*,\mathcal{N},\overline{E}}$. By setting $y=0$ in (101), we get

$$\|\nabla M_{\overline{E}}(x) - \nabla M_{\overline{E}}(0)\|_{*,s,\overline{E}} \le L\|x\|_{s,\overline{E}}, \quad \forall\, x. \quad (108)$$

Thus,

$$\|\nabla M_{\overline{E}}(x)\|_{*,s,\overline{E}} \le \|\nabla M_{\overline{E}}(0)\|_{*,s,\overline{E}} + L\|x\|_{s,\overline{E}}, \quad \forall\, x. \quad (109)$$

By (103), we know that there exists $\frac{1}{\rho_2} \le \alpha \le \frac{1}{\rho_1}$ such that

$$\|\nabla M_{\overline{E}}(x)\|_{*,\mathcal{N},\overline{E}} \le \alpha\|\nabla M_{\overline{E}}(x)\|_{*,s,\overline{E}} \quad (110)$$

Thus, combining (109) and (110) would give:

$$\|\nabla M_{\overline{E}}(x)\|_{*,\mathcal{N},\overline{E}} \le \alpha\big(\|\nabla M_{\overline{E}}(0)\|_{*,s,\overline{E}} + L\|x\|_{s,\overline{E}}\big), \quad \forall\, x. \quad (111)$$

By (102), we know that $\|x\|_{s,\overline{E}} \le \|x\|_{\mathcal{N},\overline{E}}$, thus we have:

$$\|\nabla M_{\overline{E}}(x)\|_{*,\mathcal{N},\overline{E}} \le \alpha\big(\|\nabla M_{\overline{E}}(0)\|_{*,s,\overline{E}} + L\rho_2\|x\|_{\mathcal{N},\overline{E}}\big), \quad \forall\, x. \quad (112)$$

Hence, combining the above with (107), there exist some

$$G = \mathcal{O}\big(\frac{1}{\rho_1}\max\{L\rho_2, \|\nabla M_{\overline{E}}(0)\|_{*,s,\overline{E}}\}\big) \quad (113)$$

such that

$$\mathbb{E}\Big[\langle \nabla M_{\overline{E}}(x^t-x^*),\, w^t\rangle|\mathcal{F}^t\Big] = \big\langle \nabla M_{\overline{E}}(x^t-x^*), \mathbb{E}[w^t|\mathcal{F}^t]\big\rangle \le G\big(1+\|x^t-x^*\|_{\mathcal{N},\overline{E}}\big)\varepsilon_{\text{bias}}. \quad (114)$$

Combining (114) with (105) we obtain

$$\begin{aligned}\mathbb{E}[\langle \nabla M_{\overline{E}}(x^t-x^*), x^{t+1}-x^t\rangle|\mathcal{F}^t] \le \eta_t\langle \nabla M_{\overline{E}}(x^t-x^*), H(x^t)-x^t\rangle\\ + \eta_t G\varepsilon_{\text{bias}}\Big(1+\|x^t-x^*\|_{\mathcal{N},\overline{E}}\Big) \quad (115)\end{aligned}$$

To bound the first term in the RHS of (115), note that

$$\begin{aligned}\langle \nabla M_{\bar{E}}(x^t - x^*), H(x^t) - x^t \rangle &= \langle \nabla M_{\bar{E}}(x^t - x^*), H(x^t) - x^* + x^* - x^t \rangle \\ &\overset{(a)}{\leq} M_{\bar{E}}(H(x^t) - x^*) - M_{\bar{E}}(x^t - x^*) \\ &\overset{(b)}{\leq} \frac{1}{2c_l}\|H(x^t) - H(x^*)\|^2_{c,\bar{E}} - M_{\bar{E}}(x^t - x^*) \\ &\overset{(c)}{\leq} \frac{\gamma^2}{2c_l}\|x^t - x^*\|^2_{c,\bar{E}} - M_{\bar{E}}(x^t - x^*) \\ &\leq \left(\frac{\gamma^2 c_u}{c_l} - 1\right) M_{\bar{E}}(x^t - x^*) \\ &\leq -(1 - \gamma\sqrt{c_u/c_l}) M_{\bar{E}}(x^t - x^*), \end{aligned} \tag{116}$$

where $(a)$ follows from the convexity of $M_{\bar{E}}$, $(b)$ follows from $x^*$ belonging to a fixed equivalence class with respect to $H$ and $(c)$ follows from the contraction property of $H$. Combining (116). (115) and Lemma F.2 with (104), we arrive as follows:

$$\begin{aligned}\mathbb{E}\Big[M_{\overline{E}}(x^{t+1} - x^*)|\mathcal{F}_t\Big] &\leq (1 - 2\alpha_2\eta_t + \alpha_3\eta_t^2) M_{\overline{E}}(x^t - x^*) + \alpha_4\eta_t^2 \\ &\quad + \eta_t G\varepsilon_{\text{bias}}\Big(1 + \|x^t - x^*\|_{\mathcal{N},\overline{E}}\Big) \end{aligned} \tag{117}$$

Where $\alpha_2 \coloneqq (1 - \gamma\sqrt{c_u/c_l})$, $\alpha_3 \coloneqq (8 + 2B)c_u\rho_2 L$ and $\alpha_4 \coloneqq A\rho_2 L$. We now present the formal version of Theorem B.1 as follows:

**Theorem B.3** (Formal version of Theorem B.1)**.** *let $\alpha_2$, $\alpha_3$ and $\alpha_4$ be defined in* (117)*, if $x^t$ is generated by* (98) *with all assumptions in B.1.1 satisfied, then if the stepsize $\eta_t \coloneqq \frac{1}{\alpha_2(t+K)}$ while $K \coloneqq \max\{\alpha_3/\alpha_2, 3\}$,*

$$\mathbb{E}\Big[\|x^T - x^*\|^2_{\mathcal{N},\overline{E}}\Big] \leq \frac{K^2 c_u}{(T+K)^2 c_l}\|x^0 - x^*\|^2_{\mathcal{N},\overline{E}} + \frac{8\alpha_4 c_u}{(T+K)\alpha_2^2} + \frac{2c_u C_1 C_2 \varepsilon_{bias}}{\alpha_2} \tag{118}$$

*where $C_1 = G(1+2x_{sup})$, $C_2 = \frac{1}{K} + \log\left(\frac{T-1+K}{K}\right)$, $G$ is defined in* (113) *and $x_{sup} \coloneqq \sup \|x\|_{\mathcal{N},\overline{E}}$ is the upper bound of the $\|\cdot\|_{\mathcal{P}}$ semi-norm for all $x^t$.*

*Proof.* This choice $\eta_t$ satisfies $\alpha_3\eta_t^2 \leq \alpha_2\eta_t$. Thus, by (117) we have

$$\mathbb{E}\big[M_{\overline{E}}(x^{t+1} - x^*)|\mathcal{F}_t\Big] \leq (1 - \alpha_2\eta_t) M_{\overline{E}}(x^t - x^*) + \alpha_4\eta_t^2 + \eta_t C_1\varepsilon_{\text{bias}} \tag{119}$$

we define $\Gamma_t \coloneqq \Pi_{i=o}^{t-1}(1 - \alpha_2\eta_t)$ and further obtain the $T$-step recursion relationship as follows:

$$\begin{aligned}\mathbb{E}\Big[M_{\overline{E}}(x^T - x^*)\Big] &\leq \Gamma_T M_{\overline{E}}(x^0 - x^*) + \Gamma_T \sum_{t=0}^{T-1}(\frac{1}{\Gamma_{t+1}})[\alpha_4\eta_t^2 + \eta_t C_1\varepsilon_{\text{bias}}] \\ &= \Gamma_T M_{\overline{E}}(x^0 - x^*) + \Gamma_T \sum_{t=0}^{T-1}(\frac{1}{\Gamma_{t+1}})[\alpha_4\eta_t^2] + \Gamma_T \sum_{t=0}^{T-1}(\frac{1}{\Gamma_{t+1}})[\eta_t C_1\varepsilon_{\text{bias}}] \\ &= \underbrace{\Gamma_T M_{\overline{E}}(x^0 - x^*) + \frac{\alpha_4\Gamma_T}{\alpha_2}\sum_{t=0}^{T-1}(\frac{1}{\Gamma_{t+1}})[\alpha_2\eta_t^2]}_{R_1} + \underbrace{\Gamma_T \sum_{t=0}^{T-1}(\frac{1}{\Gamma_{t+1}})[\eta_t C_1\varepsilon_{\text{bias}}]}_{R_2} \end{aligned} \tag{120}$$

where the term $R_1$ is identical to the unbiased case in Theorem 3 of [45] which leads to

$$R_1 \leq \frac{K^2}{(T+K)^2} M_{\overline{E}}(x^0 - x^*) + \frac{4\alpha_2}{(T+K)\alpha_2^2} \tag{121}$$

also, $R_2$ can be bounded by a logarithmic dependence of $T$

$$R_2 \leq \sum_{t=0}^{T-1}[\eta_t C_1\varepsilon_{\text{bias}}] = C_1\varepsilon_{\text{bias}}\sum_{t=0}^{T-1}\frac{1}{\alpha_2(t+K)} \leq \frac{C_1 C_2\varepsilon_{\text{bias}}}{\alpha_2} \tag{122}$$

Combining (121) and (122) with (120) would obtain the following:

$$\mathbb{E}\Big[M_{\overline{E}}(x^T - x^*)\Big] \leq \frac{K^2}{(T+K)^2} M_{\overline{E}}(x^0 - x^*) + \frac{4\alpha_2}{(T+K)\alpha_2^2} + \frac{C_1 C_2 \varepsilon_{\text{bias}}}{\alpha_2} \tag{123}$$

Combining (123) with (99) yields (118). □

# C Uncertainty Set Support Function Estimators

## C.1 Proof of Theorem 5.1

We have

$$\begin{aligned}
\mathbb{E}[M] &= \sum_{n=0}^{N_{\max}-1} 2^{n+1}\,\mathbb{P}(N'=n) + 2^{N_{\max}+1}\mathbb{P}(N'=N_{\max}) \\
&= \sum_{n=0}^{N_{\max}-1} 2^{n+1}\,\mathbb{P}(N=n) + 2^{N_{\max}+1}\mathbb{P}(N\geq N_{\max}) \\
&= \sum_{n=0}^{N_{\max}-1} \Big(\frac{2^{n+1}}{2^{n+1}}\Big) + 2^{N_{\max}+1}\mathbb{P}(N\geq N_{\max}) \\
&= N_{\max} + 2^{N_{\max}+1}\mathbb{P}(N\geq N_{\max}) \\
&= N_{\max} + \frac{2^{N_{\max}+1}}{2^{N_{\max}}} \\
&= N_{\max} + 2 = \mathcal{O}(N_{\max}).
\end{aligned} \tag{124}$$

## C.2 Proof of Theorem 5.2

denote $\hat{\sigma}^*_{\mathcal{P}_s^a}(V)$ as the untruncated MLMC estimator obtained by running Algorithm 1 when setting $N_{\max}$ to infinity. From [39], under both TV uncertainty sets and Wasserstein uncertainty sets, we have $\hat{\sigma}^*_{\mathcal{P}_s^a}(V)$ as an unbiased estimator of $\sigma_{\mathcal{P}_s^a}(V)$. Thus,

$$\begin{aligned}
&\mathbb{E}\left[\hat{\sigma}_{\mathcal{P}_s^a}(V) - \sigma_{\mathcal{P}_s^a}(V)\right] = \mathbb{E}\left[\hat{\sigma}_{\mathcal{P}_s^a}(V)\right] - \mathbb{E}\left[\hat{\sigma}^*_{\mathcal{P}_s^a}(V)\right] \\
&= \mathbb{E}\left[\sigma_{\hat{\mathsf{P}}^{a,1}_{s,N'+1}}(V) + \frac{\Delta_{N'}(V)}{\mathbb{P}(N'=n)}\right] - \mathbb{E}\left[\sigma_{\hat{\mathsf{P}}^{a,1}_{s,N+1}}(V) + \frac{\Delta_N(V)}{\mathbb{P}(N=n)}\right] \\
&= \mathbb{E}\left[\frac{\Delta_{N'}(V)}{\mathbb{P}(N'=n)}\right] - \mathbb{E}\left[\frac{\Delta_N(V)}{\mathbb{P}(N=n)}\right] \\
&= \sum_{n=0}^{N_{\max}} \Delta_n(V) - \sum_{n=0}^{\infty} \Delta_n(V) \\
&= \sum_{n=N_{\max}+1}^{\infty} \Delta_n(V)
\end{aligned} \tag{125}$$

For each $\Delta_n(V)$, the expectation of absolute value can be bounded as

$$\begin{aligned}
&\mathbb{E}\left[|\Delta_n(V)|\right] = \mathbb{E}\left[\left|\sigma_{\hat{\mathsf{P}}^a_{s,n+1}}(V) - \sigma_{\mathsf{P}^a_s}(V)\right|\right] \\
&+ \frac{1}{2}\mathbb{E}\left[\left|\sigma_{\hat{\mathsf{P}}^{a,E}_{s,n+1}}(V) - \sigma_{\mathsf{P}^a_s}(V)\right|\right] + \frac{1}{2}\mathbb{E}\left[\left|\sigma_{\hat{\mathsf{P}}^{a,O}_{s,n+1}}(V) - \sigma_{\mathsf{P}^a_s}(V)\right|\right]
\end{aligned} \tag{126}$$

By the binomial concentration and the Lipschitz property of the support function as in Lemma 5.3, we know for TV distance uncertainty, we have

$$\mathbb{E}\left[|\Delta_n(V)|\right] \leq 6(1+\frac{1}{\delta})2^{-\frac{n}{2}}\|V\|_{\text{sp}} \tag{127}$$

and for Wasserstein disance uncertainty, we have

$$\mathbb{E}\left[|\Delta_n(V)|\right] \leq 6\cdot 2^{-\frac{n}{2}}\|V\|_{\text{sp}} \tag{128}$$

Thus, for TV distance uncertainty, we have

$$\left|\mathbb{E}\left[\hat{\sigma}_{\mathcal{P}_s^a}(V) - \sigma_{\mathcal{P}_s^a}(V)\right]\right| \leq \sum_{n=N_{\max}+1}^{\infty} \mathbb{E}\left[|\Delta_n(V)|\right] \leq 6(1+\frac{1}{\delta})2^{-\frac{N_{\max}}{2}}\|V\|_{\text{sp}} \tag{129}$$

and for Wasserstein distance uncertainty, we have

$$\left|\mathbb{E}\left[\hat{\sigma}_{\mathcal{P}_s^a}(V) - \sigma_{\mathcal{P}_s^a}(V)\right]\right| \leq \sum_{n=N_{\max}+1}^{\infty} \mathbb{E}\left[|\Delta_n(V)|\right] \leq 6 \cdot 2^{-\frac{N_{\max}}{2}}\|V\|_{\text{sp}} \tag{130}$$

### C.3 Proof of Lemma 5.3

For TV uncertainty sets, for a fixed $V$, for any $p \in \Delta(\mathcal{S})$, define $f_p(\mu) \coloneqq p(V-\mu) - \delta\|V-\mu\|_{\text{sp}}$ and $\mu_p^* \coloneqq \arg\max_{\mu \geq \mathbf{0}} f_p(\mu)$. Thus, we have

$$\sigma_{\mathcal{P}_{TV}}(V) - \sigma_{\mathcal{Q}_{TV}}(V) = f_p(\mu_p^*) - f_q(\mu_q^*) \tag{131}$$

since, $\mu_p^*$ and $\mu_q^*$ are maximizers of $f_p$ and $f_q$ respectively, we further have

$$f_p(\mu_q^*) - f_q(\mu_q^*) \leq f_p(\mu_p^*) - f_q(\mu_q^*) \leq f_p(\mu_p^*) - f_q(\mu_p^*) \tag{132}$$

Combing (131) and (132) we thus have:

$$\begin{aligned} |\sigma_{\mathcal{P}_{TV}}(V) - \sigma_{\mathcal{Q}_{TV}}(V)| &\leq \max\{|f_p(\mu_p^*) - f_q(\mu_p^*)|, |f_p(\mu_q^*) - f_q(\mu_q^*)|\} \\ &= \max\{|(p-q)(V-\mu_p^*)|, |(p-q)(V-\mu_q^*)|\} \end{aligned} \tag{133}$$

Note that $\sigma_{\mathcal{P}_{TV}}(V)$ can also be expressed as $\sigma_{\mathcal{P}_{TV}}(V) = p\mathbf{x}^* - \delta\|\mathbf{x}^*\|_{\text{sp}}$ where $\mathbf{x}^* \coloneqq \arg\max_{\mathbf{x} \leq V}(p\mathbf{x} - \delta\|\mathbf{x}\|_{\text{sp}})$. Let $M \coloneqq \max_s \mathbf{x}^*(s)$ and $m \coloneqq \min_s \mathbf{x}^*(s)$, then $\|\mathbf{x}\|_{\text{sp}} = M - m$. Denote $\mathbf{e}$ as the all-ones vector, then $\mathbf{x} = \min_s V(s) \cdot \mathbf{e}$ is a feasible solution. Thus,

$$p\mathbf{x}^* - \delta(M-m) \geq p(\min_s V(s) \cdot \mathbf{e}) - \delta\|\min_s V(s) \cdot \mathbf{e}\|_{\text{sp}} = \min_s V(s) \tag{134}$$

Since $p$ is a probability vector, $p\mathbf{x}^* \leq M$, using the fact that $\delta > 0$, we then obtain

$$M - \delta(M-m) \geq \min_s V(s) \Rightarrow M - m \leq \frac{M - \min_s V(s)}{\delta} \tag{135}$$

Since $\mathbf{x}^*$ is a feasible solution, we have

$$M \leq \max_s V(s) \Rightarrow M - \min_s V(s) \leq \max_s V(s) - \min_s V(s) = \|V\|_{\text{sp}} \tag{136}$$

Combining (135) and (136) we obtain

$$M - m \leq \frac{\|V\|_{\text{sp}}}{\delta} \Rightarrow m \geq M - \frac{\|V\|_{\text{sp}}}{\delta} \geq \min_s V(s) - \frac{\|V\|_{\text{sp}}}{\delta} \tag{137}$$

Where the last inequality is from $M \geq \min_s V(s)$, which is a direct result of (135) and the term $\delta(M-m)$ being positive. We finally arrive with

$$\mathbf{x}^*(j) \in [m, M] \subseteq \left[\min_s V(s) - \frac{\|V\|_{\text{sp}}}{\delta}, \max_s V(s)\right] \quad \forall j \in \mathcal{S} \tag{138}$$

Thus, $\|\mathbf{x}^*\|_{\text{sp}} \leq (1+\frac{1}{\delta})\|V\|_{\text{sp}}$, which leads to

$$\|V - \mu_p^*\|_{\text{sp}} \leq (1+\frac{1}{\delta})\|V\|_{\text{sp}}, \quad \|V - \mu_q^*\|_{\text{sp}} \leq (1+\frac{1}{\delta})\|V\|_{\text{sp}} \tag{139}$$

Combining (139) with (133) we obtain the first part of (23).

For Wasserstein uncertainty sets, note that for any $p \in \Delta(\mathcal{S})$ and value function $V$,

$$\sigma_{\mathcal{P}_W}(V) = \sup_{\lambda \geq 0} \left( \overbrace{-\lambda\delta^l + \mathbb{E}_p\big[\underbrace{\inf_{y \in \mathcal{S}} \left(V(y) + \lambda d(S,y)^l\right)}_{\phi(s,\lambda)}\big]}^{g(\lambda,p)} \right). \tag{140}$$

Note that

$$\inf_{y\in\mathcal{S}} V(y) \le \phi(s,\lambda) \le V(s) + \lambda d(S,s)^l = V(s) \tag{141}$$

where the first inequality is because $\lambda d(S,y)^l \ge 0$ for any $d$ and $l$. We can then bound $\phi$ by the span of $V$ as

$$|\phi(s,\lambda)| \le \|V\|_{\mathrm{sp}} \quad \forall \lambda \ge 0 \tag{142}$$

We then further have that for any $p, q \in \Delta(\mathcal{S})$ and $\lambda \ge 0$,

$$|g(\lambda,p) - g(\lambda,q)| \le \sum_{s\in\mathcal{S}} |p(s)-q(s)||\phi(s,\lambda)| \le \|p-q\|_1 \|V\|_{\mathrm{sp}} \tag{143}$$

using (143) and the fact that $|f(\lambda) - g(\lambda)| \le \epsilon \Rightarrow |\sup_\lambda f(\lambda) - \sup_\lambda g(\lambda)| \le \epsilon$, we obtain the second part of (23).

### C.4 Proof of Theorem 5.4

For all $p \in \Delta(\mathcal{S})$, we have $\sigma_p(V) \le \|V\|_{\mathrm{sp}}$, leading to

$$\mathrm{Var}(\hat{\sigma}_{\mathcal{P}_s^a}(V)) \le \mathbb{E}\left[\left(\hat{\sigma}_{\mathcal{P}_s^a}(V)\right)^2\right] + \|V\|_{\mathrm{sp}}^2 \tag{144}$$

To bound the second moment, note that

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{\sigma}_{\mathcal{P}_s^a}(V)\right)^2\right] &= \mathbb{E}\left[\left(\sigma_{\hat{\mathsf{P}}^{a,1}_{s,N'+1}}(V) + \frac{\Delta_{N'}(V)}{\mathbb{P}(N'=n)}\right)^2\right] \\
&\le \mathbb{E}\left[\left(\|V\|_{\mathrm{sp}} + \frac{\Delta_{N'}(V)}{\mathbb{P}(N'=n)}\right)^2\right] \\
&\le 2\|V\|_{\mathrm{sp}}^2 + 2\mathbb{E}\left[\left(\frac{\Delta_{N'}(V)}{\mathbb{P}(N'=n)}\right)^2\right] \\
&\le 2\|V\|_{\mathrm{sp}}^2 + 2\sum_{n=0}^{N_{\max}} \left(\frac{\mathbb{E}[|\Delta_n(V)|]}{\mathbb{P}(N'=n)}\right)^2 \mathbb{P}(N'=n) \\
&= 2\|V\|_{\mathrm{sp}}^2 + 2\sum_{n=0}^{N_{\max}} \frac{\mathbb{E}[|\Delta_n(V)|]^2}{\mathbb{P}(N'=n)}
\end{aligned} \tag{145}$$

Under TV distance uncertainty set, by (127), we further have

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{\sigma}_{\mathcal{P}_s^a}(V)\right)^2\right] &\le 2\|V\|_{\mathrm{sp}}^2 + 2S\sum_{n=0}^{N_{\max}} \frac{36(1+\frac{1}{\delta})^2 2^{-n}\|V\|_{\mathrm{sp}}^2}{2^{-(n+1)}} \\
&= 2\|V\|_{\mathrm{sp}}^2 + 144(1+\frac{1}{\delta})^2 S\|V\|_{\mathrm{sp}}^2 N_{\max}
\end{aligned} \tag{146}$$

Under Wasserstein distance uncertainty set, by (128), we further have

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{\sigma}_{\mathcal{P}_s^a}(V)\right)^2\right] &\le 2\|V\|_{\mathrm{sp}}^2 + 2S\sum_{n=0}^{N_{\max}} \frac{362^{-n}\|V\|_{\mathrm{sp}}^2}{2^{-(n+1)}} \\
&= 2\|V\|_{\mathrm{sp}}^2 + 144S\|V\|_{\mathrm{sp}}^2 N_{\max}
\end{aligned} \tag{147}$$

## D Convergence for Robust TD

### D.1 Formal Statement of Theorem 6.1

The first half of Algorithm 2 (line 1 - line 7) can be treated as a special instance of the SA updates in (98) with the bias and variance of the i.i.d. noise term specified in Section 5. To facilitate deriving

the bounds of the noise terms, we first analyze the bounds in terms of the $l_\infty$ norm, and then translate the bounds in terms of the $\|\cdot\|_{\mathcal{P}}$ semi-norm to obtain the final results.

We start with analyzing the bias and variance of $\hat{\mathbf{T}}_{g_0}(V_t)$ for each $t$. Recall the definition of $\hat{\mathbf{T}}_{g_0}(V_t)$ is as follows:

$$\hat{\mathbf{T}}_{g_0}(V_t)(s) = \sum_a \pi(a|s)\big[r(s,a) - g_0 + \hat{\sigma}_{\mathcal{P}_s^a}(V_t)\big] \quad \forall s \in \mathcal{S}$$

Thus, we have for all $s \in \mathcal{S}$,

$$\left|\mathbb{E}\left[\hat{\mathbf{T}}_{g_0}(V_t)(s)\right] - \mathbf{T}_{g_0}(V_t)(s)\right| \le \sum_a \pi(a|s)\left|\mathbb{E}[\hat{\sigma}_{\mathcal{P}_s^a}(V_t)] - \sigma_{\mathcal{P}_s^a}(V_t)\right| = \left|\mathbb{E}[\hat{\sigma}_{\mathcal{P}_s^a}(V_t)] - \sigma_{\mathcal{P}_s^a}(V_t)\right| \tag{148}$$

Which further implies the bias of $\hat{\mathbf{T}}_{g_0}(V_t)$ is bounded by the bias of $\hat{\sigma}_{\mathcal{P}_s^a}(V_t)$ as follows:

$$\left\|\mathbb{E}\left[\hat{\mathbf{T}}_{g_0}(V_t)\right] - \mathbf{T}_{g_0}(V_t)\right\|_\infty \le \left|\mathbb{E}[\hat{\sigma}_{\mathcal{P}_s^a}(V_t)] - \sigma_{\mathcal{P}_s^a}(V_t)\right| \tag{149}$$

Regarding the variance, note that

$$\begin{aligned}
&\mathbb{E}\left[(\hat{\mathbf{T}}_{g_0}(V_t)(s) - \mathbf{T}_{g_0}(V_t)(s))^2\right] \\
&= \left(\mathbb{E}\left[\hat{\mathbf{T}}_{g_0}(V_t)(s)\right] - \mathbf{T}_{g_0}(V_t)(s)\right)^2 + \mathrm{Var}\left(\hat{\mathbf{T}}_{g_0}(V_t)(s)\right) \\
&\le \left|\mathbb{E}[\hat{\sigma}_{\mathcal{P}_s^a}(V_t)] - \sigma_{\mathcal{P}_s^a}(V_t)\right|^2 + \mathrm{Var}\left(\sum_a \pi(a|s)\hat{\sigma}_{\mathcal{P}_s^a}(V_t)\right) \\
&= \left|\mathbb{E}[\hat{\sigma}_{\mathcal{P}_s^a}(V_t)] - \sigma_{\mathcal{P}_s^a}(V_t)\right|^2 + \sum_a \pi(a|s)^2 \mathrm{Var}\left(\hat{\sigma}_{\mathcal{P}_s^a}(V_t)\right)
\end{aligned} \tag{150}$$

To create an upper bound of $\|V\|_{\mathrm{sp}}$ for all possible $V$, define the mixing time of any $p \in \mathcal{P}$ to be

$$t_{\mathrm{mix}}^p := \arg\min_{t \ge 1}\left\{\max_{\mu_0} \left\|(\mu_0 p_\pi^t)^\top - \nu^\top\right\|_1 \le \frac{1}{2}\right\} \tag{151}$$

where $p_\pi$ is the finite state Markov chain induced by $\pi$, $\mu_0$ is any initial probability distribution on $\mathcal{S}$ and $\nu$ is its invariant distribution. By Assumption 3.1, and Lemma F.4,and for any value function $V$, we have

$$t_{\mathrm{mix}}^p < +\infty \quad \text{and} \quad \|V\|_{\mathrm{sp}} \le 4t_{\mathrm{mix}}^p \le 4t_{\mathrm{mix}} \tag{152}$$

where we define $t_{\mathrm{mix}} := \sup_{p\in\mathcal{P}} t_{\mathrm{mix}}^p$, then $t_{\mathrm{mix}}$ is also finite due to the compactness of $\mathcal{P}$. We now derive the bounds of biases and variances for the three types of uncertainty sets. Regarding contamination uncertainty sets, according to Lemma F.3, $\hat{\sigma}_{\mathcal{P}_s^a}(V)$ is unbiased and has variance bounded by $\|V\|^2$. Thu, define $t_{\mathrm{mix}}$ according to Lemma F.4 and combining the above result with Lemma F.4, we obtain that $\hat{\mathbf{T}}_{g_0}(V_t)$ is also unbiased and the variance satisfies

$$\mathbb{E}\left[\left\|\hat{\mathbf{T}}_{g_0}(V_t) - \mathbf{T}_{g_0}(V_t)\right\|_\infty^2\right] \le \|V_t\|^2 \le 16t_{\mathrm{mix}}^2 \tag{153}$$

Regarding TV distance uncertainty sets, using the property of the bias and variance of $\hat{\sigma}_{\mathcal{P}_s^a}(V)$ in Theorem 5.2 and Theorem 5.4 while combining them with Lemma F.4, we have

$$\left\|\mathbb{E}\left[\hat{\mathbf{T}}_{g_0}(V_t)\right] - \mathbf{T}_{g_0}(V_t)\right\|_\infty \le 6(1+\frac{1}{\delta})\sqrt{S}2^{-\frac{N_{\max}}{2}}\|V\|_{\mathrm{sp}} = 24(1+\frac{1}{\delta})\sqrt{S}2^{-\frac{N_{\max}}{2}}t_{\mathrm{mix}} \tag{154}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\left\|\hat{\mathbf{T}}_{g_0}(V_t) - \mathbf{T}_{g_0}(V_t)\right\|_\infty^2\right] \\
&\le \left(24(1+\frac{1}{\delta})\sqrt{S}2^{-\frac{N_{\max}}{2}}t_{\mathrm{mix}}\right)^2 + 3\|V\|_{\mathrm{sp}}^2 + 144(1+\frac{1}{\delta})^2 S\|V\|_{\mathrm{sp}}^2 N_{\max} \\
&\le \left(24(1+\frac{1}{\delta})\sqrt{S}2^{-\frac{N_{\max}}{2}}t_{\mathrm{mix}}\right)^2 + 48t_{\mathrm{mix}}^2 + 2304(1+\frac{1}{\delta})^2 S t_{\mathrm{mix}}^2 N_{\max}
\end{aligned} \tag{155}$$

Similarly, for Wasserstein distance uncertainty sets, we have

$$\left\|\mathbb{E}\left[\hat{\mathbf{T}}_{g_0}(V_t)\right] - \mathbf{T}_{g_0}(V_t)\right\|_\infty \leq 6\sqrt{S}2^{-\frac{N_{\max}}{2}}\|V\|_{\text{sp}} = 24\sqrt{S}2^{-\frac{N_{\max}}{2}}t_{\text{mix}} \tag{156}$$

and

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\mathbf{T}}_{g_0}(V_t) - \mathbf{T}_{g_0}(V_t)\right\|_\infty^2\right] &\leq \left(24\sqrt{S}2^{-\frac{N_{\max}}{2}}t_{\text{mix}}\right)^2 + 3\|V\|_{\text{sp}}^2 + 144\|V\|_{\text{sp}}^2 N_{\max} \\ &\leq \left(24\sqrt{S}2^{-\frac{N_{\max}}{2}}t_{\text{mix}}\right)^2 + 48t_{\text{mix}}^2 + 2304St_{\text{mix}}^2 N_{\max} \end{aligned} \tag{157}$$

In order to translate the above bounds from the $l_\infty$ norm into the $\|\cdot\|_{\mathcal{P}}$ norm, recall that in line 7 of Algorithm 2, we chose an anchor state $s_0$ set $V_t(s_0) = 0$ for all $t$ to avoid ambiguity. We thus can draw the following relationship:

**Lemma D.1.** *Let $x \in \mathbb{R}^S$ satisfy $x_i = 0$ for some fixed index $i$. Then*

$$\|x\|_\infty \leq \|x\|_{\text{sp}} = \max_{1\leq j\leq n} x_j - \min_{1\leq j\leq n} x_j \leq 2\|x\|_\infty.$$

*Moreover, since all semi-norms with the same kernel spaces are equivalent, there are constants $c_{\mathcal{P}}, C_{\mathcal{P}} > 0$ so that*

$$c_{\mathcal{P}}\|x\|_{\text{sp}} \leq \|x\|_{\mathcal{P}} \leq C_{\mathcal{P}}\|x\|_{\text{sp}} \quad \forall\, x \in \mathbb{R}^n,$$

*then*

$$c_{\mathcal{P}}\|x\|_\infty \leq c_{\mathcal{P}}\|x\|_{\text{sp}} \leq \|x\|_{\mathcal{P}} \leq C_{\mathcal{P}}\|x\|_{\text{sp}} \leq 2C_{\mathcal{P}}\|x\|_\infty. \tag{158}$$

*Proof.* Since $x_i = 0$, for every $j$ we have $-\|x\|_\infty \leq x_j \leq \|x\|_\infty$. Hence

$$\max_j x_j \;\leq\; \|x\|_\infty, \quad \min_j x_j \;\geq\; -\|x\|_\infty,$$

and so

$$\|x\|_\infty = \max\{\max_j x_j - 0, 0 - \min_j x_j\} \leq \|x\|_{\text{sp}} = \max_j x_j - \min_j x_j \leq \|x\|_\infty - (-\|x\|_\infty) = 2\,\|x\|_\infty.$$

Since $\|\cdot\|_{\text{sp}}$ and $\|\cdot\|_{\mathcal{P}}$ both have the same kernel of $\{c\mathbf{e} : c \in \mathbb{R}\}$, by the equivalence of semi-norms, it follows that there exists $c_{\mathcal{P}}$ and $C_{\mathcal{P}}$ such that

$$c_{\mathcal{P}}\|x\|_\infty \leq c_{\mathcal{P}}\|x\|_{\text{sp}} \leq \|x\|_{\mathcal{P}} \leq C_{\mathcal{P}}\|x\|_{\text{sp}} \leq 2C_{\mathcal{P}}\|x\|_\infty$$

as claimed. □

With the relationship established above, line 1 - line 7 of Algorithm 2 can be formally treated as a special instance of the SA updates in (98) with $B = 0$. We now provide the bias and variance of the i.i.d. noise for the different uncertainty sets discussed using Lemma D.1 and the estimation bounds in (153)-(157). For contamination uncertainty sets, we have

$$\varepsilon_{\text{bias}}^{\text{Cont}} = 0 \quad \text{and} \quad A^{\text{Cont}} = 32C_{\mathcal{P}}^2 t_{\text{mix}}^2 \tag{159}$$

for TV distance uncertainty sets, we have

$$\varepsilon_{\text{bias}}^{\text{TV}} = 48C_{\mathcal{P}}(1 + \frac{1}{\delta})\sqrt{S}2^{-\frac{N_{\max}}{2}}t_{\text{mix}} = \mathcal{O}\left(\sqrt{S}2^{-\frac{N_{\max}}{2}}t_{\text{mix}}\right) \tag{160}$$

and

$$\begin{aligned} A^{\text{TV}} &= 2C_{\mathcal{P}}^2\left(24(1+\frac{1}{\delta})\sqrt{S}2^{-\frac{N_{\max}}{2}}t_{\text{mix}}\right)^2 + 96C_{\mathcal{P}}^2 t_{\text{mix}}^2 + 4608C_{\mathcal{P}}^2(1+\frac{1}{\delta})^2 St_{\text{mix}}^2 N_{\max} \\ &= \mathcal{O}\left(t_{\text{mix}}^2 S N_{\max}\right) \end{aligned} \tag{161}$$

and for Wasserstein distance uncertainty sets, we have

$$\varepsilon_{\text{bias}}^{\text{Wass}} = 48C_{\mathcal{P}}\sqrt{S}2^{-\frac{N_{\max}}{2}}t_{\text{mix}} = \mathcal{O}\left(\sqrt{S}2^{-\frac{N_{\max}}{2}}t_{\text{mix}}\right) \tag{162}$$

and

$$\begin{aligned} A^{\text{Wass}} &= 2C_{\mathcal{P}}^2\left(24\sqrt{S}2^{-\frac{N_{\max}}{2}}t_{\text{mix}}\right)^2 + 96C_{\mathcal{P}}^2 t_{\text{mix}}^2 + 4608C_{\mathcal{P}}^2 St_{\text{mix}}^2 N_{\max} \\ &= \mathcal{O}\left(t_{\text{mix}}^2 S N_{\max}\right) \end{aligned} \tag{163}$$

**Theorem D.2** (Formal version of Theorem 6.1)**.** *Let $\alpha_2 \coloneqq (1-\gamma\sqrt{c_u/c_l})$, $\alpha_3 \coloneqq 8c_u\rho_2 L$ and $\alpha_4 \coloneqq \rho_2 L$, if $V_t$ is generated by Algorithm 2. Define $V^*$ to be the anchored robust value function $V^* = V^{\pi}_{\mathsf{P}_V} + c\mathbf{e}$ for some $c$ such that $V^*(s_0) = 0$, then under Assumption 3.1 and the radius conditions of Lemma A.4-A.6, if the stepsize $\eta_t \coloneqq \frac{1}{\alpha_2(t+K)}$ while $K \coloneqq \max\{\alpha_3/\alpha_2, 3\}$, then for contamination uncertainty sets,*

$$\mathbb{E}\Big[\|V_T - V^*\|_\infty^2\Big] \leq \frac{4K^2 c_u C_{\mathcal{P}}^2}{(T+K)^2 c_l c_{\mathcal{P}}^2}\|V_0 - V^*\|_\infty^2 + \frac{8A^{\mathrm{Cont}}\alpha_4 c_u}{(T+K)\alpha_2^2 c_{\mathcal{P}}^2} = \mathcal{O}\left(\frac{1}{T^2} + \frac{t_{\mathrm{mix}}^2}{T(1-\gamma)^2}\right) \tag{164}$$

*for TV distance uncertainty sets,*

$$\mathbb{E}\Big[\|V_T - V^*\|_\infty^2\Big] \leq \frac{4K^2 c_u C_{\mathcal{P}}^2}{(T+K)^2 c_l c_{\mathcal{P}}^2}\|V_0 - V^*\|_\infty^2 + \frac{8A^{\mathrm{TV}}\alpha_4 c_u}{(T+K)\alpha_2^2 c_{\mathcal{P}}^2} + \frac{c_u C_3 C_2 \varepsilon_{\mathrm{bias}}^{\mathrm{TV}}}{\alpha_2 c_{\mathcal{P}}^2} \tag{165}$$

$$= \mathcal{O}\left(\frac{1}{T^2} + \frac{t_{\mathrm{mix}}^2 N_{\max}}{T(1-\gamma)^2} + \frac{t_{\mathrm{mix}}^2 2^{-\frac{N_{\max}}{2}}\log T}{(1-\gamma)^2}\right) \tag{166}$$

*for Wasserstein distance uncertainty sets,*

$$\mathbb{E}\Big[\|V_T - V^*\|_\infty^2\Big] \leq \frac{4K^2 c_u C_{\mathcal{P}}^2}{(T+K)^2 c_l c_{\mathcal{P}}^2}\|V_0 - V^*\|_\infty^2 + \frac{8A^{\mathrm{Wass}}\alpha_4 c_u}{(T+K)\alpha_2^2 c_{\mathcal{P}}^2} + \frac{c_u C_3 C_2 \varepsilon_{\mathrm{bias}}^{\mathrm{Wass}}}{\alpha_2 c_{\mathcal{P}}^2} \tag{167}$$

$$= \mathcal{O}\left(\frac{1}{T^2} + \frac{t_{\mathrm{mix}}^2 N_{\max}}{T(1-\gamma)^2} + \frac{t_{\mathrm{mix}}^2 2^{-\frac{N_{\max}}{2}}\log T}{(1-\gamma)^2}\right) \tag{168}$$

*where the $\varepsilon$ and $A$ terms are defined in* (159)-(163)*, $C_2 = \frac{1}{K} + \log\left(\frac{T-1+K}{K}\right)$, $C_3 = G(1+8C_{\mathcal{P}}t_{\mathrm{mix}})$, $\gamma$ is defined in* (14)*, $c_u, c_l$ are defined in* (99)*, $\rho_2$ is defined in* (102)*, $G$ is defined in* (113)*, and $C_{\mathcal{P}}, c_{\mathcal{P}}$ are defined in Lemma D.1.*

*Proof.* By Lemma D.1 and (152), we have that for any value function $V$ its $\|\cdot\|_{\mathcal{P}}$ norm is bounded as follows:

$$\|V\|_{\mathcal{P}} \leq 4C_{\mathcal{P}} t_{\mathrm{mix}} \tag{169}$$

Substituting the terms of (159)-(163), (169), and Theorem 4.2 to Theorem B.3, we would have for contamination uncertainty sets,

$$\mathbb{E}\Big[\|V_T - V^*\|_{\mathcal{P}}^2\Big] \leq \frac{K^2 c_u}{(T+K)^2 c_l}\|V_0 - V^*\|_{\mathcal{P}}^2 + \frac{8A^{\mathrm{Cont}}\alpha_4 c_u}{(T+K)\alpha_2^2} = \mathcal{O}\left(\frac{1}{T^2} + \frac{t_{\mathrm{mix}}^2}{T(1-\gamma)^2}\right) \tag{170}$$

for TV distance uncertainty sets,

$$\mathbb{E}\Big[\|V_T - V^*\|_{\mathcal{P}}^2\Big] \leq \frac{K^2 c_u}{(T+K)^2 c_l}\|V_0 - V^*\|_{\mathcal{P}}^2 + \frac{8A^{\mathrm{TV}}\alpha_4 c_u}{(T+K)\alpha_2^2} + \frac{c_u C_3 C_2 \varepsilon_{\mathrm{bias}}^{\mathrm{TV}}}{\alpha_2} \tag{171}$$

$$= \mathcal{O}\left(\frac{1}{T^2} + \frac{S t_{\mathrm{mix}}^2 N_{\max}}{T(1-\gamma)^2} + \frac{S t_{\mathrm{mix}}^2 2^{-\frac{N_{\max}}{2}}\log T}{(1-\gamma)^2}\right) \tag{172}$$

for Wasserstein distance uncertainty sets,

$$\mathbb{E}\Big[\|V_T - V^*\|_{\mathcal{P}}^2\Big] \leq \frac{K^2 c_u}{(T+K)^2 c_l}\|V_0 - V^*\|_{\mathcal{P}}^2 + \frac{4A^{\mathrm{Wass}}\alpha_4 c_u}{(T+K)\alpha_2^2} + \frac{c_u C_3 C_2 \varepsilon_{\mathrm{bias}}^{\mathrm{Wass}}}{\alpha_2} \tag{173}$$

$$= \mathcal{O}\left(\frac{1}{T^2} + \frac{S t_{\mathrm{mix}}^2 N_{\max}}{T(1-\gamma)^2} + \frac{S t_{\mathrm{mix}}^2 2^{-\frac{N_{\max}}{2}}\log T}{(1-\gamma)^2}\right) \tag{174}$$

where the $\varepsilon$ and $A$ terms are defined in (159)-(163), $C_2 = \frac{1}{K} + \log\left(\frac{T-1+K}{K}\right)$, $C_3 = G(1+8C_{\mathcal{P}}t_{\mathrm{mix}})$, $\gamma$ is defined in (14), $c_u, c_l$ are defined in (99), $\rho_2$ is defined in (102), $G$ is defined in (113), and $C_{\mathcal{P}}$ is defined in Lemma D.1. We now translate the result back to the standard $l_\infty$ norm by applying Lemma D.1 again to the above, we obtain the desired results. □

### D.2 Proof of Theorem 6.1

We use the result from Theorem D.2, to set $\mathbb{E}\left[\|V_T - V^*\|_\infty^2\right] \leq \epsilon^2$. For contamination uncertainty set we set $T = \mathcal{O}\left(\frac{t_{\text{mix}}^2}{\epsilon^2(1-\gamma)^2}\right)$, resulting in $\mathcal{O}\left(\frac{SAt_{\text{mix}}^2}{\epsilon^2(1-\gamma)^2}\right)$ sample complexity. For TV and Wasserstein uncertainty set, we set $N_{\max} = \mathcal{O}\left(\log \frac{\sqrt{S}t_{\text{mix}}}{\epsilon(1-\gamma)}\right)$ and $T = \mathcal{O}\left(\frac{t_{\text{mix}}^2}{\epsilon^2(1-\gamma)^2} \log \frac{\sqrt{S}t_{\text{mix}}}{\epsilon(1-\gamma)}\right)$, combining with Theorem 5.1, this would result in $\mathcal{O}\left(\frac{SAt_{\text{mix}}^2}{\epsilon^2(1-\gamma)^2} \log^2 \frac{\sqrt{S}t_{\text{mix}}}{\epsilon(1-\gamma)}\right)$ sample complexity.

To show order-optimality, we provide the standard mean estimation as a hard example. Consider the TD learning of the MDP with only two states $\mathcal{S} = \{s_1, s_2\}$, and $Pr(s \to s_1) = p$, $Pr(s \to s_2) = 1 - p$ for each $s \in S$ with $p \in (0, 1)$. Thus, this MDP is indifferent from the actions chosen and we further define $r(s_1) = 1,\ r(s_2) = 0$. Thus, estimating the relative value functions is equivalent of estimating $p$. By the Cramér–Rao or direct variance argument for Bernoulli($p$) estimation, we have that to achieve $|\hat{p_N} - p|^2 \leq \epsilon^2$ requires $N \geq\ 1/2\epsilon^2 = \Omega(\epsilon^{-2})$.

### D.3 Formal Statement of Theorem 6.2

To analyze the second part (line 8 - line 14) of Algorithm 2 and provide the provide the complexity for $g_t$, we first define the noiseless function $\bar{\delta}(V)$ as

$$\bar{\delta}(V) \coloneqq \frac{1}{S} \sum_s \left( \sum_a \pi(a|s) \big[ r(s,a) + \sigma_{\mathcal{P}_s^a}(V) \big] - V(s) \right) \tag{175}$$

Thus, we have

$$\bar{\delta}_t = \bar{\delta}(V_T) + \nu_t \tag{176}$$

where $\nu_t$ is the noise term with bias equal to the bias $\hat{\sigma}_{\mathcal{P}_s^a}(V_T)$

$$\mathbb{E}[|\nu_t|] = \frac{1}{S} \sum_s \sum_a \left( \pi(a|s) \mathbb{E}\big[\, \big| \sigma_{\mathcal{P}_s^a}(V_T) - \hat{\sigma}_{\mathcal{P}_s^a}(V_T) \big| \,\big] \right) = \left| \mathbb{E}\left[ \hat{\sigma}_{\mathcal{P}_s^a}(V_T) - \sigma_{\mathcal{P}_s^a}(V_T) \right] \right| \tag{177}$$

By the Bellman equation in Theorem 3.2, we have $g_{\mathcal{P}}^\pi = \bar{\delta}(V^*)$, which implies

$$\begin{aligned}
\left| \bar{\delta}(V_T) - g_{\mathcal{P}}^\pi \right| &= \left| \bar{\delta}(V_T) - \bar{\delta}(V^*) \right| \\
&\leq \frac{1}{S} \sum_s \left( \sum_a \pi(a|s) \left| \sigma_{\mathcal{P}_s^a}(V_T) - \sigma_{\mathcal{P}_s^a}(V^*) \right| + |V_T(s) - V^*(s)| \right) \\
&\leq \frac{1}{S} \sum_s \left( \sum_a \pi(a|s) \left| \sigma_{\mathcal{P}_s^a}(V_T) - \sigma_{\mathcal{P}_s^a}(V^*) \right| + |V_T(s) - V^*(s)| \right) \\
&\leq \frac{1}{S} \sum_s 2 \|V_T - V^*\|_{\text{sp}} \\
&= 2 \|V_T - V^*\|_{\text{sp}} \\
&\leq 4 \|V_T - V^*\|_\infty
\end{aligned} \tag{178}$$

Where the last inequality is by Lemma D.1. Thus, the following recursion can be formed

$$\begin{aligned}
|g_{t+1} - g_{\mathcal{P}}^\pi| &= \left| g_t + \beta_t(\bar{\delta}_t - g_t) - g_{\mathcal{P}}^\pi \right| \\
&= \left| g_t - g_{\mathcal{P}}^\pi + \beta_t(\bar{\delta}_t - g_{\mathcal{P}}^\pi + g_{\mathcal{P}}^\pi - g_t) \right| \\
&= \left| g_t - g_{\mathcal{P}}^\pi + \beta_t(\bar{\delta}(V^T) - g_{\mathcal{P}}^\pi + \nu_t + g_{\mathcal{P}}^\pi - g_t) \right| \\
&\leq (1 - \beta_t) \left| g_t - g_{\mathcal{P}}^\pi \right| + \beta_t(\left| \bar{\delta}(V^T) - g_{\mathcal{P}}^\pi \right| + |\nu_t|) \\
&\leq (1 - \beta_t) \left| g_t - g_{\mathcal{P}}^\pi \right| + \beta_t(4\|V_T - V^*\|_\infty + |\nu_t|)
\end{aligned} \tag{179}$$

Thus, taking expectation conditioned on the filtration $\mathcal{F}^t$ yields

$$\begin{aligned}
\mathbb{E}\left[ |g_{t+1} - g_{\mathcal{P}}^\pi| \right] &\leq (1 - \beta_t) \left| g_t - g_{\mathcal{P}}^\pi \right| + \beta_t \left( 4\mathbb{E}\left[ \|V_T - V^*\|_\infty \right] + \mathbb{E}[|\nu_t|] \right) \\
&\leq (1 - \beta_t) \left| g_t - g_{\mathcal{P}}^\pi \right| + \beta_t \left( 4\mathbb{E}\left[ \|V_T - V^*\|_\infty \right] + \left| \mathbb{E}\left[ \hat{\sigma}_{\mathcal{P}_s^a}(V_T) - \sigma_{\mathcal{P}_s^a}(V_T) \right] \right| \right)
\end{aligned} \tag{180}$$

By letting $\zeta_t \coloneqq \Pi_{i=0}^{t-1}(1-\beta_t)$, we obtain the $T$-step recursion as follows:

$$\begin{aligned}
&\mathbb{E}\left[|g_T - g_{\mathcal{P}}^{\pi}|\right] \leq \zeta_T \left|g_0 - g_{\mathcal{P}}^{\pi}\right| \\
&\quad + \zeta_T \sum_{t=0}^{T-1} (\frac{1}{\zeta_{t+1}}) \beta_t \left(4\mathbb{E}\left[\|V_T - V^*\|_\infty\right] + \left|\mathbb{E}\left[\hat{\sigma}_{\mathcal{P}_s^a}(V_T) - \sigma_{\mathcal{P}_s^a}(V_T)\right]\right|\right) \\
&\quad = \zeta_T \left|g_0 - g_{\mathcal{P}}^{\pi}\right| + \sum_{t=0}^{T-1} (\frac{\zeta_T}{\zeta_{t+1}}) \beta_t \left(4\mathbb{E}\left[\|V_T - V^*\|_\infty\right] + \left|\mathbb{E}\left[\hat{\sigma}_{\mathcal{P}_s^a}(V_T) - \sigma_{\mathcal{P}_s^a}(V_T)\right]\right|\right) \\
&\quad \leq \zeta_T \left|g_0 - g_{\mathcal{P}}^{\pi}\right| + \sum_{t=0}^{T-1} \beta_t \big(4\mathbb{E}\left[\|V_T - V^*\|_\infty\right] + \left|\mathbb{E}\left[\hat{\sigma}_{\mathcal{P}_s^a}(V_T) - \sigma_{\mathcal{P}_s^a}(V_T)\right]\right|\big) \\
&\quad = \zeta_T \left|g_0 - g_{\mathcal{P}}^{\pi}\right| + \big(4\mathbb{E}\left[\|V_T - V^*\|_\infty\right] + \left|\mathbb{E}\left[\hat{\sigma}_{\mathcal{P}_s^a}(V_T) - \sigma_{\mathcal{P}_s^a}(V_T)\right]\right|\big) \sum_{t=0}^{T-1} \beta_t
\end{aligned} \tag{181}$$

By setting $\beta_t \coloneqq \frac{1}{t+1}$, we have $\zeta_T = \frac{1}{T+1} \leq \frac{1}{T}$ and $\sum_{t=0}^{T-1} \beta_t \leq 2\log T$, (181) implies

$$\mathbb{E}\left[|g_T - g_{\mathcal{P}}^{\pi}|\right] \leq \frac{1}{T}\left|g_0 - g_{\mathcal{P}}^{\pi}\right| + \big(8\mathbb{E}\left[\|V_T - V^*\|_\infty\right] + 2\left|\mathbb{E}\left[\hat{\sigma}_{\mathcal{P}_s^a}(V_T) - \sigma_{\mathcal{P}_s^a}(V_T)\right]\right|\big)\log T \tag{182}$$

**Theorem D.3** (Formal version of Theorem 6.2)**.** *Following all notations and assumptions in Theorem D.2, then for contamination uncertainty sets,*

$$\begin{aligned}
\mathbb{E}\Big[|g_T - g_{\mathcal{P}}^{\pi}|\Big] &\leq \frac{1}{T}\left|g_0 - g_{\mathcal{P}}^{\pi}\right| + \frac{16KC_{\mathcal{P}}\sqrt{c_u}\log T}{(T+K)c_{\mathcal{P}}\sqrt{c_l}}\|V_0 - V^*\|_\infty \\
&\quad + \frac{16\sqrt{2A^{\mathrm{Cont}}\alpha_4 c_u}\log T}{\alpha_2 c_{\mathcal{P}}\sqrt{T+K}} = \mathcal{O}\left(\frac{1}{T} + \frac{\log T}{T} + \frac{t_{\mathrm{mix}}\log T}{\sqrt{T}(1-\gamma)}\right).
\end{aligned} \tag{183}$$

*For TV distance uncertainty sets,*

$$\begin{aligned}
&\mathbb{E}\Big[|g_T - g_{\mathcal{P}}^{\pi}|\Big] \leq \frac{1}{T}\left|g_0 - g_{\mathcal{P}}^{\pi}\right| + \frac{16KC_{\mathcal{P}}\sqrt{c_u}\log T}{(T+K)c_{\mathcal{P}}\sqrt{c_l}}\|V_0 - V^*\|_\infty \\
&\quad + \frac{16\sqrt{2A^{\mathrm{TV}}\alpha_4 c_u}\log T}{\alpha_2 c_{\mathcal{P}}\sqrt{T+K}} + \frac{8\sqrt{c_u C_3 C_2 \varepsilon_{\mathrm{bias}}^{\mathrm{TV}}}\log T}{c_{\mathcal{P}}\sqrt{\alpha_2}} + 48(1+\frac{1}{\delta})\sqrt{S}2^{-\frac{N_{\max}}{2}} t_{\mathrm{mix}}\log T \\
&\quad = \mathcal{O}\left(\frac{1}{T} + \frac{\log T}{T} + \frac{t_{\mathrm{mix}}\sqrt{SN_{\max}}\log T}{\sqrt{T}(1-\gamma)} + \frac{\sqrt{S}t_{\mathrm{mix}}2^{-\frac{N_{\max}}{4}}\log^{\frac{3}{2}} T}{\sqrt{1-\gamma}} + \sqrt{S}2^{-\frac{N_{\max}}{2}} t_{\mathrm{mix}}\log T\right).
\end{aligned} \tag{184}$$

*For Wasserstein distance uncertainty sets,*

$$\mathbb{E}\Big[|g_T - g_{\mathcal{P}}^{\pi}|\Big] \leq \frac{1}{T}|g_0 - g_{\mathcal{P}}^{\pi}| + \frac{16KC_{\mathcal{P}}\sqrt{c_u}\log T}{(T+K)c_{\mathcal{P}}\sqrt{c_l}}\|V_0 - V^*\|_\infty$$
$$+ \frac{16\sqrt{2A^{\text{Wass}}\alpha_4 c_u}\log T}{\alpha_2 c_{\mathcal{P}}\sqrt{T+K}} + \frac{8\sqrt{c_u C_3 C_2 \varepsilon_{\text{bias}}^{\text{Wass}}}\log T}{c_{\mathcal{P}}\sqrt{\alpha_2}} + 48\sqrt{S}2^{-\frac{N_{\max}}{2}}t_{\text{mix}}\log T$$
$$= \mathcal{O}\left(\frac{1}{T} + \frac{\log T}{T} + \frac{t_{\text{mix}}\sqrt{SN_{\max}}\log T}{\sqrt{T}(1-\gamma)} + \frac{\sqrt{S}t_{\text{mix}}2^{-\frac{N_{\max}}{4}}\log^{\frac{3}{2}} T}{\sqrt{1-\gamma}} + \sqrt{S}2^{-\frac{N_{\max}}{2}}t_{\text{mix}}\log T\right). \tag{185}$$

*where all the above variables are defined the same as in Theorem D.2.*

*Proof.* By Theorem D.2, taking square root on both side and utilizing the concavity of square root function, we have for contamination uncertainty sets,

$$\mathbb{E}\Big[\|V_T - V^*\|_\infty\Big] \leq \frac{2KC_{\mathcal{P}}\sqrt{c_u}}{(T+K)c_{\mathcal{P}}\sqrt{c_l}}\|V_0 - V^*\|_\infty + \frac{2\sqrt{2A^{\text{Cont}}\alpha_4 c_u}}{\alpha_2 c_{\mathcal{P}}\sqrt{T+K}} \tag{186}$$

for TV distance uncertainty sets,

$$\mathbb{E}\Big[\|V_T - V^*\|_\infty\Big] \leq \frac{2KC_{\mathcal{P}}\sqrt{c_u}}{(T+K)c_{\mathcal{P}}\sqrt{c_l}}\|V_0 - V^*\|_\infty + \frac{2\sqrt{2A^{\text{TV}}\alpha_4 c_u}}{\alpha_2 c_{\mathcal{P}}\sqrt{T+K}} + \sqrt{\frac{c_u C_3 C_2 \varepsilon_{\text{bias}}^{\text{TV}}}{\alpha_2 c_{\mathcal{P}}^2}} \tag{187}$$

for Wasserstein distance uncertainty sets,

$$\mathbb{E}\Big[\|V_T - V^*\|_\infty\Big] \leq \frac{2KC_{\mathcal{P}}\sqrt{c_u}}{(T+K)c_{\mathcal{P}}\sqrt{c_l}}\|V_0 - V^*\|_\infty + \frac{2\sqrt{2A^{\text{Wass}}\alpha_4 c_u}}{\alpha_2 c_{\mathcal{P}}\sqrt{T+K}} + \sqrt{\frac{c_u C_3 C_2 \varepsilon_{\text{bias}}^{\text{Wass}}}{\alpha_2 c_{\mathcal{P}}^2}} \tag{188}$$

Regarding the bound for the absolute bias of $\hat{\sigma}_{\mathcal{P}_s^a}$, from Lemma F.3, we have for contamination uncertainty,

$$\left|\mathbb{E}\left[\hat{\sigma}_{\mathcal{P}_s^a}(V) - \sigma_{\mathcal{P}_s^a}(V)\right]\right| = 0 \tag{189}$$

In addition, combining (129)-(130) with Lemma F.4, we have for for TV distance uncertainty,

$$\left|\mathbb{E}\left[\hat{\sigma}_{\mathcal{P}_s^a}(V) - \sigma_{\mathcal{P}_s^a}(V)\right]\right| \leq 24\sqrt{S}(1+\frac{1}{\delta})2^{-\frac{N_{\max}}{2}}t_{\text{mix}} \tag{190}$$

and for Wasserstein distance uncertainty, we have

$$\left|\mathbb{E}\left[\hat{\sigma}_{\mathcal{P}_s^a}(V) - \sigma_{\mathcal{P}_s^a}(V)\right]\right| \leq 24\sqrt{S}2^{-\frac{N_{\max}}{2}}t_{\text{mix}} \tag{191}$$

Combining (186)-(191) with (182) gives the desired result. □

### D.4 Proof of Theorem 6.2

We use the result from Theorem D.3, to set $\mathbb{E}\Big[|g_T - g_{\mathcal{P}}^{\pi}|\Big] \leq \epsilon$. For contamination uncertainty sets we set $T = \mathcal{O}\left(\frac{t_{\text{mix}}^2}{\epsilon^2(1-\gamma)^2}\log\frac{t_{\text{mix}}}{\epsilon(1-\gamma)}\right)$, resulting in $\mathcal{O}\left(\frac{SAt_{\text{mix}}^2}{\epsilon^2(1-\gamma)^2}\log\frac{t_{\text{mix}}}{\epsilon(1-\gamma)}\right)$ sample complexity. For TV and Wasserstein uncertainty set, we set $N_{\max} = \mathcal{O}\left(\log\frac{\sqrt{S}t_{\text{mix}}}{\epsilon(1-\gamma)}\right)$ and $T = \mathcal{O}\left(\frac{t_{\text{mix}}^2}{\epsilon^2(1-\gamma)^2}\log^3\frac{\sqrt{S}t_{\text{mix}}}{\epsilon(1-\gamma)}\right)$, combining with Theorem 5.1, this would result in $\mathcal{O}\left(\frac{SAt_{\text{mix}}^2}{\epsilon^2(1-\gamma)^2}\log^4\frac{\sqrt{S}t_{\text{mix}}}{\epsilon(1-\gamma)}\right)$ sample complexity.

# E Numerical Validations for Semi-Norm Contractions

In this section, we provide numerical examples that directly verify the one-step strict contraction across the settings studied. These results empirically support the key structural claims used by our analysis.

## E.1 Evaluations of Lemma 4.1

Lemma A.2 is the technical backbone for Lemma 4.1, as Lemma A.2 constructs the fixed-kernel semi-norm and provides the one-step contraction for a given $P$. Therefore, we perform numerical evaluations on Lemma A.2 to demonstrate the one-step contraction property for Lemma 4.1.

For a kernel $P$ with stationary distribution $d$, we follow the steps in Appendix A.1 and construct $\|\cdot\|_{\mathsf{P}}$ in (41) with

$$\alpha = \min\{0.99, \tfrac{1+\rho(Q)}{2}\}, \qquad \epsilon = 0.25(1-\alpha), \tag{192}$$

where $Q = P - \mathbf{e}\, d^\top$. Then the one-step contraction factor is $\beta = \alpha + \epsilon < 1$.

To generate ergodic matrices with dimension $n$, let $I_n$ be the identity matrix and $S_n$ be the cyclic shift matrix defined by $S_n e_i = e_{i+1 \bmod n}$. We provide the following four examples:

- $P_1 = 0.5\, I_5 + 0.5\, S_5$
- $P_2 = 0.6\, I_6 + 0.4\, S_6$
- $P_3 = 0.55\, I_7 + 0.45\, S_7$
- $P_4 = 0.6\, I_8 + 0.3\, S_8 + 0.1\, S_8^2$

We generate 1000 random unit vectors $x$ and compute each ratio $\dfrac{\|P_i x\|_{\mathsf{P}}}{\|x\|_{\mathsf{P}}}$. The empirical results are summarized below.

| **matrix** | $n$ | **max span ratio** | $\rho(Q)$ | $\alpha$ | $\epsilon$ | $\beta$ | $\text{ratio}_{\min}$ | $\text{ratio}_{\text{median}}$ | $\text{ratio}_{\text{p90}}$ | $\text{ratio}_{\max}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 5 | 1 | 0.8090 | 0.9045 | 0.0239 | 0.9284 | 0.3824 | 0.7950 | 0.8077 | 0.8090 |
| P2 | 6 | 1 | 0.8718 | 0.9359 | 0.0160 | 0.9519 | 0.5197 | 0.8510 | 0.8687 | 0.8718 |
| P3 | 7 | 1 | 0.9020 | 0.9510 | 0.0122 | 0.9633 | 0.5861 | 0.8799 | 0.8976 | 0.9014 |
| P4 | 8 | 1 | 0.8700 | 0.9350 | 0.0162 | 0.9513 | 0.4855 | 0.8226 | 0.8604 | 0.8685 |

Table 1: Empirical one-step contraction ratios for ergodic kernels using the fixed-kernel semi-norm $\|\cdot\|_{\mathsf{P}}$.

## E.2 Evaluations of Theorem 4.2

Lemma A.8 is the key step for Theorem 4.2, as Lemma A.8 proves a uniform one-step contraction across all $P$ in the uncertainty set. We therefore perform numerical evaluations on Lemma A.8 to demonstrate the one-step contraction property for Theorem 4.2 under contamination, total variation (TV), and Wasserstein-1 uncertainty. We select the same $P_1$, $P_2$, $P_3$, and $P_4$ in Appendix E.1 as four examples of the nominal model.

To numerically approximate $\|\cdot\|_{\mathcal{P}}$ defined in (82), we approximate $\|\cdot\|_{\mathcal{P}}$ by (i) discretizing the uncertainty set and (ii) using a finite product to approximate the extremal norm. First, we sample a family $\{P^{(i)}\}_{i=1}^m \subset \mathcal{P}$ of size $m$ and form

$$Q_i \;:=\; P^{(i)} - \mathbf{e}\, d_{P^{(i)}}^\top, \qquad \hat{r} \;=\; \max_i \rho(Q_i).$$

We set $\alpha = \min\{0.99, (1+\hat{r})/2\}$ and choose $\epsilon \in (0, 1-\alpha)$. To approximate the extremal norm, we build a library of scaled products of the $Q_i$'s up to maximum length $K$: for each $k = 0, 1, \ldots, K$ we draw products $M_{k,j} = Q_{i_k} \cdots Q_{i_1}$; the number of such draws at each $k$ is the "products per length" (denoted `samples_per_k` in the tables). This defines the surrogate

$$\|z\|_{\text{ext}}^{(K)} = \max_{0 \le k \le K, j} \alpha^{-k} \|M_{k,j} z\|_2. \tag{193}$$

We then set

$$\|x\|_{\mathcal{P}}^{(K)} = \max_{i=1,\dots,m} \|Q_i x\|_{\text{ext}}^{(K)} + \epsilon \min_{c\in\mathcal{C}} \|x - c\,\mathbf{e}\|_{\text{ext}}^{(K)}. \tag{194}$$

We generate 50 random unit vectors $x$ for each sampled uncertainty matrix in $\{P^{(i)}\}_{i=1}^m \subset \mathcal{P}$, and compute the ratios $\dfrac{\|P^{(i)}x\|_{\mathcal{P}}}{\|x\|_{\mathcal{P}}}$. The empirical results of the uncertainty sets studied in our settings are summarized below.

| **nominal** $\bar{\mathsf{P}}$ | $n$ | $\delta$ | $m$ | $K$ | `samples_per_k` | **max span ratio** | $\hat{r}$ | $\alpha$ | $\epsilon$ | $\gamma$ | $\text{ratio}_{\min}$ | $\text{ratio}_{\text{median}}$ | $\text{ratio}_{\text{p90}}$ | $\text{ratio}_{\max}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 5 | 0.15 | 30 | 3 | 25 | 1 | 0.8138 | 0.9069 | 0.0233 | 0.9302 | 0.2210 | 0.6396 | 0.8057 | 0.8134 |
| P2 | 6 | 0.15 | 30 | 3 | 25 | 1 | 0.8807 | 0.9403 | 0.0149 | 0.9553 | 0.2957 | 0.6581 | 0.8630 | 0.8785 |
| P3 | 7 | 0.15 | 30 | 3 | 25 | 1 | 0.9067 | 0.9534 | 0.0117 | 0.9650 | 0.3217 | 0.6683 | 0.8700 | 0.8901 |
| P4 | 8 | 0.15 | 30 | 3 | 25 | 1 | 0.8812 | 0.9406 | 0.0148 | 0.9555 | 0.3739 | 0.5964 | 0.8207 | 0.8624 |

Table 2: Empirical one-step contraction ratios under contamination uncertainty.

| **nominal** $\bar{\mathsf{P}}$ | $n$ | $\delta$ | $m$ | $K$ | `samples_per_k` | **max span ratio** | $\hat{r}$ | $\alpha$ | $\epsilon$ | $\gamma$ | $\text{ratio}_{\min}$ | $\text{ratio}_{\text{median}}$ | $\text{ratio}_{\text{p90}}$ | $\text{ratio}_{\max}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 5 | 0.15 | 30 | 3 | 25 | 1 | 0.8280 | 0.9140 | 0.0215 | 0.9355 | 0.3239 | 0.7609 | 0.8239 | 0.8280 |
| P2 | 6 | 0.15 | 30 | 3 | 25 | 1 | 0.9013 | 0.9507 | 0.0123 | 0.9630 | 0.4021 | 0.7904 | 0.8862 | 0.9006 |
| P3 | 7 | 0.15 | 30 | 3 | 25 | 1 | 0.9175 | 0.9588 | 0.0103 | 0.9691 | 0.4457 | 0.7918 | 0.8962 | 0.9162 |
| P4 | 8 | 0.15 | 30 | 3 | 25 | 1 | 0.8805 | 0.9403 | 0.0149 | 0.9552 | 0.4497 | 0.7503 | 0.8449 | 0.8739 |

Table 3: Empirical one-step contraction ratios under total variation (TV) uncertainty.

| **nominal** $\bar{\mathsf{P}}$ | $n$ | $\delta$ | $m$ | $K$ | `samples_per_k` | **max span ratio** | $\hat{r}$ | $\alpha$ | $\epsilon$ | $\gamma$ | $\text{ratio}_{\min}$ | $\text{ratio}_{\text{median}}$ | $\text{ratio}_{\text{p90}}$ | $\text{ratio}_{\max}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 5 | 0.15 | 30 | 3 | 25 | 1 | 0.8184 | 0.9092 | 0.0227 | 0.9319 | 0.3698 | 0.7569 | 0.8141 | 0.8188 |
| P2 | 6 | 0.15 | 30 | 3 | 25 | 1 | 0.8900 | 0.9450 | 0.0138 | 0.9587 | 0.4257 | 0.7889 | 0.8774 | 0.8894 |
| P3 | 7 | 0.15 | 30 | 3 | 25 | 1 | 0.9110 | 0.9555 | 0.0111 | 0.9666 | 0.4262 | 0.7818 | 0.8827 | 0.9080 |
| P4 | 8 | 0.15 | 30 | 3 | 25 | 1 | 0.8758 | 0.9379 | 0.0155 | 0.9534 | 0.4413 | 0.7224 | 0.8518 | 0.8689 |

Table 4: Empirical one-step contraction ratios under Wasserstein-1 uncertainty.

### E.3 Interpretations

Note that for the above tables, **max span ratio** denotes the largest one-step span contraction coefficient over the sampled families. This value equals to 1 for all the settings, meaning no strict one-step contraction in the span. In contrast, the quantities $\text{ratio}_{\min}$, $\text{ratio}_{median}$, $\text{ratio}_{\text{p90}}$, and $\text{ratio}_{\max}$ summarize the empirical one-step ratios $\frac{||Px||_{\mathcal{P}}}{||x||_{\mathcal{P}}}$ (in the robust case) and $\frac{||Px||_{\mathsf{P}}}{||x||_{\mathsf{P}}}$ (in the non-robust case) computed under the constructed semi-norms over all sampled kernels $P$ and random unit directions $x$. They report, respectively, the minimum, median, 90th percentile, and maximum observed value across those tests. In every table we have $\text{ratio}_{\max} < 1$ , so we observe empirical one-step contraction under our semi-norms even when span does not contract. Moreover, $\text{ratio}_{\max} \le \gamma$ (robust case) or $\text{ratio}_{\max} \le \beta$ (non-robust case), which is consistent with the corresponding theoretical contraction factor guaranteed by our constructions.

## F Some Auxiliary Lemmas for the Proofs

**Lemma F.1** (Theorem IV in [3])**.** *Let $\mathcal{Q}$ be a bounded set of square matrix such that $\rho(Q) < \infty$ for all $Q \in \mathcal{Q}$ where $\rho(\cdot)$ denotes the spectral radius. Then the joint spectral radius of $\mathcal{Q}$ can be defined as*

$$\hat{\rho}(\mathcal{Q}) \coloneqq \lim_{k\to\infty} \sup_{Q_i\in\mathcal{Q}} \rho(Q_k \dots Q_1)^{\frac{1}{k}} = \lim_{k\to\infty} \sup_{Q_i\in\mathcal{Q}} \|Q_k \dots Q_1\|^{\frac{1}{k}} \tag{195}$$

*where $\|\cdot\|$ is an arbitrary norm.*

**Lemma F.2** (Lemma 6 in [45])**.** *Under the setup and notation in Appendix B.1.1, if assuming the noise has bounded variance of $\mathbb{E}[\,\|w^t\|_{\mathcal{N},\overline{E}}^2 | \mathcal{F}^t] \le A + B\|x^t - x^*\|_{\mathcal{N},\overline{E}}^2$, we have*

$$\mathbb{E}\left[\|x^{t+1} - x^t\|_{s,\overline{E}}^2 | \mathcal{F}^t\right] \le (16 + 4B) c_u \rho_2 \eta_t^2 M_{\overline{E}}(x^t - x^*) + 2A\rho_2\eta_t^2. \tag{196}$$

**Lemma F.3** (Theorem D.1 in [39]). *The estimator $\hat{\sigma}_{\mathcal{P}_s^a}(V)$ obtained by* (18) *for contamination uncertainty sets is unbiased and has bounded variance as follows:*

$$\mathbb{E}\left[\hat{\sigma}_{\mathcal{P}_s^a}(V)\right] = \sigma_{\mathcal{P}_s^a}(V), \quad \textit{and} \quad \mathrm{Var}(\hat{\sigma}_{\mathcal{P}_s^a}(V)) \leq \|V\|^2 \tag{197}$$

**Lemma F.4** (Ergodic case of Lemma 9 in [34]). *For any average-reward MDP with stationary policy $\pi$ and the mixing time defined as*

$$\tau_{\mathrm{mix}} \coloneqq \arg\min_{t \geq 1} \left\{ \max_{\mu_0} \left\| (\mu_0 P_\pi^t)^\top - \nu^\top \right\|_1 \leq \frac{1}{2} \right\} \tag{198}$$

*where $P_\pi$ is the finite state Markov chain induced by $\pi$, $\mu_0$ is any initial probability distribution on $\mathcal{S}$ and $\nu$ is its invariant distribution. If $P_\pi$ is irreducible and aperiodic, then $\tau_{\mathrm{mix}} < +\infty$ and for the value function $V$, we have*

$$\|V\|_{\mathrm{sp}} \leq 4\tau_{\mathrm{mix}} \tag{199}$$