Efficient Q-Learning and Actor-Critic Methods for Robust Average-Reward Reinforcement Learning

Yang Xu[1] **Swetha Ganesh**[1] **Vaneet Aggarwal**[1]

[1]Purdue University, West Lafayette, Indiana, USA 47907

Abstract

We study model-free methods for distributionally robust infinite-horizon average-reward Markov decision processes (MDPs). We present non-asymptotic convergence analyses of Q-learning and actor–critic algorithms for robust average-reward MDPs under contamination, total-variation distance, and Wasserstein uncertainty sets. A key ingredient of our analysis is showing that the optimal robust Bellman operator is a strict contraction with respect to a carefully designed semi-norm. This property enables a stochastic approximation update that learns the optimal robust Q-function with $\tilde{\mathcal{O}}(\epsilon^{-2})$ dependence on the target accuracy. We also establish robust TD convergence bounds whose constants are uniform over all stationary policies, yielding an efficient data-driven routine for robust critic estimation. Building on this, we introduce an actor–critic algorithm that learns an ϵ-optimal robust policy with $\tilde{\mathcal{O}}(\epsilon^{-2})$ dependence on the target accuracy. We provide numerical simulations to illustrate the qualitative behavior of the proposed algorithms. Our results contribute to the theoretical foundations of robust planning under model misspecification, and to model-free approaches for building robust long-run policies directly from simulation data.

1 INTRODUCTION

Reinforcement learning (RL) has produced impressive results in fields such as robotics, finance, and health care by allowing agents to discover effective actions through interaction with their environments. Yet in many practical settings direct interaction is unsafe, prohibitively costly, or constrained by strict data budgets Sünderhauf et al. [2018], Höfer et al. [2021]. Practitioners therefore train agents within simulated environments before deploying them into the physical world. This approach is a common choice for robotic control and autonomous driving. The unavoidable gap between a simulator and reality can nonetheless erode performance once the policy encounters dynamics that were absent during training. Robust RL mitigates this risk by framing learning as an optimization problem over a family of transition models, aiming for reliable behaviour under the most adverse member of that family. In this work, we focus on the setting where a planner interacts only with a fixed nominal simulator but seeks a policy that is robust to transition uncertainty after deployment.

Reinforcement learning with an infinite horizon is usually studied under two reward formulations: the discounted formulation, which geometrically discounts future rewards, and the average-reward formulation, which maximizes the long-run mean return. Although the discounted objective has been widely used, its bias toward immediate gains can produce shortsighted policies in tasks that demand sustained efficiency. The average-reward formulation naturally fits these domains because each decision shapes cumulative performance over time. Typical examples include fleet management in ride-hailing, production scheduling in factories, and network resource allocation, where planners must optimize long-run throughput or service quality under uncertain dynamics. Yet the literature on robust reinforcement learning remains largely unexplored under the average-reward criterion.

Existing works on the robust average-reward RL is still sparse. Existing studies on analyzing Q-learning updates provide only asymptotic convergence results [Wang et al., 2023c, 2024b]. [Sun et al., 2024] examines the iteration complexity of vanilla policy gradient for the same problem but assumes an oracle that yields the exact robust Q functions. Sample complexity studies on model-based line of work converts the average-reward objective into an equivalent discounted task and then applies planning algorithms [Wang et al., 2023b, Chen et al., 2025]. [Roch et al., 2025] provides a model-free non-asymptotic convergence method

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:7613–7650.

by bounding the sample complexity of a robust Halpern iteration with recursive sampling. Non-asymptotic results for standard model-free control methods such as Q-learning or actor–critic therefore remain open.

Tackling the robust average-reward RL problem is significantly harder than solving its standard non-robust counterpart. In actor-critic and Q-learning algorithms, the learner needs to evaluate the value function, the Q function, and the long-run average reward, all under the most adverse transition kernel in the uncertainty set, while only observing samples from the single nominal kernel. We consider three common families of uncertainty sets: contamination sets, total-variation (TV) distance uncertainty sets, and Wasserstein distance uncertainty sets. In the non-robust setting, one can estimate values and rewards directly from trajectories, but robustness introduces an extra optimization layer in which an adversary reshapes the dynamics. Classical estimators [Wei et al., 2020, Agarwal et al., 2021, Jin et al., 2018] disregard this adversarial aspect and therefore fail. Progress requires new algorithms that can reason about the worst-case dynamics using only the data generated by the nominal model.

1.1 CHALLENGES AND CONTRIBUTIONS

Extending model-free methods such as Q-learning and actor-critic algorithms to the robust average-reward setting raises two principal obstacles. First, the optimal robust Bellman operator for Q–learning no longer inherits a contraction from a discount factor, and until now no norm nor semi-norm was known to shrink its one–step error. Second, actor–critic methods require accurate policy–gradient estimates, yet these depend on a robust Q function that cannot be obtained with finite–sample accuracy using naive rollouts. Although recent work of [Xu et al., 2025b] studies a value–based critic with non–asymptotic error bounds, its result is for a single fixed policy with policy-dependent constants. Thus, its results cannot directly address policy improvements and therefore leaves both obstacles open.

This paper overcomes these difficulties for both Q–learning and actor–critic control by developing a uniform contraction framework for robust average-reward Bellman operators and by integrating data-driven critic errors into robust mirror-descent policy optimization. Our main contributions are as follows:

- **Uniform contraction of the optimal robust Bellman operator.** Under Assumption 4.2 together with the policy-sequence overlap condition in Assumption A.1, we construct a semi-norm on the Q-function space and prove that the optimal robust average–reward Bellman operator for Q-learning is a strict contraction with a factor $\gamma_H < 1$ that is uniform over all deterministic policies and admissible uncertainty sets (Theorem 4.3).
- **Robust Q–learning with non-asymptotic sample complexity.** Leveraging this uniform contraction, we design a model–free robust Q–learning algorithm and show that the iterates converge to the optimal robust Q–function at a rate that yields a sample complexity of $\tilde{\mathcal{O}}\big(\epsilon^{-2}\big)$ for returning an ϵ–optimal robust Q function (Theorem 4.4). All constants in this bound are independent of the particular policy sequence, in contrast to prior robust average-reward evaluation results.
- **Robust actor–critic with noisy critic and uniform guarantees.** We develop a robust actor–critic algorithm that uses a critic subroutine to estimate the Q-function and updates the policy via mirror ascent. Different from [Xu et al., 2025b], our analysis is based on a one-step semi-norm contraction of the robust Bellman operator, with the contraction factor being uniform across all stationary policies (Lemma B.5). This yields critic error bounds whose rates and constants are independent of the policy sequence (Theorem B.9), which is essential to analyze actor–critic updates that change the policy over time. Incorporating our uniform critic bounds into the mirror-ascent analysis, we show that the robust actor converges to a near-optimal robust policy in $\mathcal{O}(\log T)$ iterations, yielding total sample complexity $\tilde{\mathcal{O}}(|\mathcal{S}||\mathcal{A}|\epsilon^{-2})$ for contamination uncertainty and $\tilde{\mathcal{O}}((|\mathcal{S}||\mathcal{A}|)^2\epsilon^{-2})$ for TV and Wasserstein uncertainty under the plain-average support estimator in Algorithm 3 (Theorem 5.6). This extends the average-reward robust mirror-descent framework of [Sun et al., 2024], which assumes exact robust gradients, to the practical setting with finite-sample critic errors.

Conceptually, our work provides the first model-free robust planners for average-reward MDPs whose entire pipeline admits non-asymptotic sample-complexity guarantees, without assuming an oracle for robust Q-functions or gradients.

2 RELATED WORK

The convergence guarantees of robust average reward RL have been studied by the following works. Wang et al. [2023b], Chen et al. [2025] both take 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. Roch et al. [2025] proposed a model-free robust Halpern iteration that alternates between a Halpern step in the quotient space and a recursively coupled Monte-Carlo sampler to estimate the worst-case Bellman operator. In particular, Chen et al. [2025], Roch et al. [2025] both obtained the sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2})$. Regarding Q-learning and actor-critic based methods, Wang et al. [2023c] proposed a model-free approach by developing a temporal difference learning framework for robust Q-learning algorithm, proving asymptotic convergence using ODE methods

in stochastic approximation, martingale theory, and multi-level Monte Carlo estimators to handle non-linearity in the robust Bellman operator. Sun et al. [2024], Wang et al. [2025] developed iteration complexities of $\tilde{\mathcal{O}}(\epsilon^{-1})$ for variants of vanilla policy gradient algorithm of average reward robust MDPs, both assuming direct queries of the robust Q functions. Thus, Sun et al. [2024], Wang et al. [2025] do not establish explicit convergence rate guarantees in terms of sample complexity.

Finite-sample guarantees are well established in distributionally robust RL under infinite horizon discounted reward settings. with the key recent works on actor-critic methods Wang and Zou [2022], Zhou et al. [2024], Li et al. [2022], Li and Lan [2023], Kumar et al. [2023], Kuang et al. [2022] and Q-learning methods Wang et al. [2023a, 2024a] focusing on solving the robust Bellman equation by finding the unique fixed-point solution. This approach is made feasible by the strict contraction property of the robust Bellman operator, 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, techniques that succeed in Q-learning and actor-critic algorithms for the discounted case do not transfer directly to robust average-reward RL. Our semi-norm construction can be viewed as recovering a contraction-type structure in the average-reward robust setting where discounting is unavailable.

There has been a rich literature on non-robust average-reward RL with finite-sample guarantees for Q-learning and actor–critic algorithms. These methods typically exploit the linearity of the non-robust Bellman operator [Zhang et al., 2021, Ganesh et al., 2025], span-contraction properties [Chen, 2025, Zhang et al., 2021], or direct sampling of policy gradients [Wei et al., 2020, Bai et al., 2024, Xu et al., 2025a, Ganesh and Aggarwal, 2025], together with ergodicity assumptions. In the robust setting, the Bellman operator becomes non-linear due to the inner maximization over uncertainty sets, and robustness is defined with respect to worst-case transitions that are not directly observed. Consequently, the existing techniques for non-robust average-reward RL do not directly carry over; our work can be viewed as extending contraction-based analyses to this more challenging robust regime.

3 SETTING

Consider a robust MDP with state space $\mathcal{S}$, action space $\mathcal{A}$, and deterministic reward function r with $R_{\text{sp}} := \max_{s,a} r(s,a) - \min_{s,a} r(s,a)$, where $|\mathcal{S}| = S$ and $|\mathcal{A}| = A$. We focus on the standard reward-span normalization of $r \in [0,1]$ and $R_{\text{sp}} \leq 1$. Throughout, $\Delta(\mathcal{S})$ denotes the probability simplex over $\mathcal{S}$, and $\mathbf{e}$ denotes the all-ones vector. The transition kernel is assumed to belong to an 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 compact (s,a)-rectangular uncertainty sets [Nilim and Ghaoui, 2003, Iyengar, 2005], i.e., $\mathcal{P} = \bigotimes_{s,a} \mathcal{P}_s^a$, where $\mathcal{P}_s^a \subseteq \Delta(\mathcal{S})$. This rectangular structure is standard in robust MDPs because it preserves dynamic programming and tractable Bellman operators, but it also rules out correlated perturbations across different state-action pairs. In the contraction proofs below we use convexity of each row set $\mathcal{P}_s^a$; this property holds for the contamination, total variation, and Wasserstein uncertainty sets studied in this paper. Popular uncertainty sets include those defined by the contamination model [Huber, 1965, Wang and Zou, 2022], total variation Lim et al. [2013], and Wasserstein distance Gao and Kleywegt [2022]. We consider the worst-case average-reward over the uncertainty set of MDPs. Specifically, we define the robust average-reward of a policy π as

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

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

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

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 π, and denote the set of all possible worst-case transition kernels by Ω_g^{π}, i.e., $\Omega_g^{\pi} \triangleq \{\mathsf{P} \in \mathcal{P} : g_{\mathsf{P}}^{\pi} = g_{\mathcal{P}}^{\pi}\}$, where g_{P}^{π} denotes the average reward of policy π under the single transition $\mathsf{P} \in \mathcal{P}$:

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

In this paper, we make similar assumptions for our algorithms regarding ergodicity, which are explicitly stated in Assumption 4.2 and Assumption 5.4. The ergodicity assumption is commonly used in the current robust average reinforcement learning literature Sun et al. [2024], Wang et al. [2025]. Under ergodicity of the nominal distribution, with suitable radius restrictions of the uncertainty set, we can derive ergodicity for all kernels in the entire uncertainty set. This indicates that 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 $g_{\mathsf{P}}^{\pi}(s) = g_{\mathsf{P}}^{\pi}(s')$. We thus drop the dependence on the initial state and denote g_{P}^{π} as the robust average reward. Furthermore, for facilitating analysis, define the mixing time of any $p \in \mathcal{P}$ under policy π by

$$t_{\text{mix}}^{p} := \arg\min_{t\geq 1}\left\{\max_{\mu_0} \left\|(\mu_0 p_{\pi}^{t})^{\top} - \nu^{\top}\right\|_1 \leq \frac{1}{2}\right\}, \quad (4)$$

where p_π is the finite state Markov chain induced by π, μ_0 is any initial probability distribution on $\mathcal{S}$, and ν is its invariant distribution. Define the span semi-norm as $\|v\|_{\mathrm{sp}} = \max_s v(s) - \min_s v(s)$, and let $h^{\pi,p}$ denote the average-reward relative value function of policy π under transition kernel p. By Lemma C.2, if p_π is irreducible and aperiodic, then

$$t^p_{\mathrm{mix}} < +\infty \quad \text{and} \quad \|h^{\pi,p}\|_{\mathrm{sp}} \le 4R_{\mathrm{sp}} t^p_{\mathrm{mix}} \le 4R_{\mathrm{sp}} t_{\mathrm{mix}}. \tag{5}$$

Under the reward normalization $R_{\mathrm{sp}} \le 1$, this gives $\|h^{\pi,p}\|_{\mathrm{sp}} \le 4t_{\mathrm{mix}}$. Here

$$t_{\mathrm{mix}} := \sup_{p\in\mathcal{P},\pi\in\Pi} t^p_{\mathrm{mix}}.$$

Under the ergodicity assumptions stated in Assumptions 4.2 and 5.4, together with the stated radius conditions, the chains are uniformly geometrically ergodic. Hence t_{mix} is finite and admits a uniform bound over the admissible pairs, which also justifies dropping the dependence on the initial state for the average reward.

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 now formally define the robust value function $V^\pi_{\mathsf{P}_V}$ by connecting it with the following robust Bellman equation:

Theorem 3.1 (Robust Bellman Equation, Theorem 3.1 in Wang et al. [2023c])**.** *For a fixed policy π and for each $s \in \mathcal{S}$, define the Robust Bellman operator with scalar g as follows*

$$\mathbf{T}^\pi_g(V)(s) = \sum_a \pi(a|s)\big[r(s,a) - g + \sigma_{\mathcal{P}^a_s}(V)\big], \tag{6}$$

where $\sigma_{\mathcal{P}^a_s}(V) := \min_{p\in\mathcal{P}^a_s} p^\top V$. If (g, V) is a solution to the robust Bellman equation:

$$V(s) = \mathbf{T}^\pi_g(V)(s), \quad \forall s \in \mathcal{S}, \tag{7}$$

then the scalar g corresponds to the robust average reward, i.e., $g = g^\pi_{\mathcal{P}}$, and the worst-case transition kernel P_V belongs to the set of minimizing transition kernels, i.e., $\mathsf{P}_V \in \Omega^\pi_g$, where $\Omega^\pi_g \triangleq \{\mathsf{P} \in \mathcal{P} : g^\pi_{\mathsf{P}} = g^\pi_{\mathcal{P}}\}$. Furthermore, $V = V^\pi_{\mathsf{P}_V} + c\mathbf{e}$ for some $c \in \mathbb{R}$ where $V^\pi_{\mathsf{P}_V}$ is defined as the relative value function of the policy π under the single transition P_V as follows:

$$V^\pi_{\mathsf{P}_V}(s) \triangleq \mathbb{E}_{\pi,\mathsf{P}_V}\left[\sum_{t=0}^{\infty}(r_t - g^\pi_{\mathsf{P}_V})|S_0 = s\right]. \tag{8}$$

where S_0 is the initial state.

Theorem 3.1 implies that the robust Bellman equation (7) 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 $\mathcal{P}^a_s$.

For the critic and actor analysis, we fix the representative V^π of the robust relative value function by the same anchor state s_0 used in Algorithm 5. Thus $V^\pi(s_0) = 0$. Given this representative, define the rowwise Bellman-minimizer set

$$\mathcal{M}_{s,a}(V^\pi) \triangleq \arg\min_{p\in\mathcal{P}^a_s} p^\top V^\pi. \tag{9}$$

Since each row set is compact and $p \mapsto p^\top V^\pi$ is continuous, $\mathcal{M}_{s,a}(V^\pi)$ is nonempty. By rectangularity, we may choose a transition kernel $\mathsf{P}^\pi_V \in \mathcal{P}$ whose row at every state-action pair satisfies

$$\mathsf{P}^\pi_V(\cdot|s,a) \in \mathcal{M}_{s,a}(V^\pi), \qquad \forall (s,a) \in \mathcal{S}\times\mathcal{A}. \tag{10}$$

Throughout the actor-critic section, P^π_V denotes such a rowwise Bellman-minimizing selector. This convention removes the ambiguity caused by nonunique worst-case kernels on rows that may not be used by π.

The robust one-step-deviation Q-function is defined by

$$\begin{aligned} Q^\pi(s,a) &\triangleq r(s,a) - g^\pi_{\mathcal{P}} + \sigma_{\mathcal{P}^a_s}(V^\pi) \\ &= r(s,a) - g^\pi_{\mathcal{P}} + \big(\mathsf{P}^\pi_V(\cdot|s,a)\big)^\top V^\pi. \end{aligned} \tag{11}$$

Equivalently, under the selector P^π_V in (10),

$$Q^\pi(s,a) = \mathbb{E}_{\pi,\mathsf{P}^\pi_V}\left[\sum_{t=0}^{\infty}(r_t - g^\pi_{\mathcal{P}})\,\middle|\, S_0 = s, A_0 = a\right]. \tag{12}$$

In the following, we present the explicit forms of $\sigma_{\mathcal{P}^a_s}(V)$ for different compact uncertainty sets, with $(s,a)-$rectangular structures.

Contamination Uncertainty Set The δ-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 [Huber, 1965]. Under this uncertainty set, the support function can be computed as

$$\sigma_{\mathcal{P}^a_s}(V) = (1-\delta)(\tilde{\mathsf{P}}^a_s)^\top V + \delta \min_s V(s), \tag{13}$$

and this is linear in the nominal transition kernel P^a_s. A direct approach [Wang et al., 2023c, Xu et al., 2025b] is to use the transition to the subsequent state to construct our estimator:

$$\hat{\sigma}_{\mathcal{P}^a_s}(V) \triangleq (1-\delta)V(s') + \delta \min_x V(x), \tag{14}$$

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

Total Variation Uncertainty Set. The total variation uncertainty set is $\mathcal{P}^a_s = \{q \in \Delta(\mathcal{S}) : \frac{1}{2}\|q - \tilde{\mathsf{P}}^a_s\|_1 \le \delta\}$, the

support function can be computed using its dual function Iyengar [2005]:

$$\sigma_{\mathcal{P}_s^a}(V) = \max_{\mu \in \mathbb{R}_+^S} \left((\tilde{\mathsf{P}}_s^a)^\top (V - \mu) - \delta \|V - \mu\|_{\mathrm{sp}} \right). \quad (15)$$

Here μ is a vector in $\mathbb{R}_+^S$. Equivalently, one may write the dual as a maximization over $z \leq V$ after setting $z = V - \mu$.

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

$$\mathcal{P}_s^a = \left\{ q \in \Delta(\mathcal{S}) : W_l(\tilde{\mathsf{P}}_s^a, q) \leq \delta \right\}. \quad (16)$$

The support function w.r.t. the Wasserstein distance set, can be calculated as follows Gao and Kleywegt [2023]:

$$\sigma_{\mathcal{P}_s^a}(V) = \sup_{\lambda \geq 0} \left(-\lambda \delta^l + \mathbb{E}_{\tilde{\mathsf{P}}_s^a}\left[\inf_y \left(V(y) + \lambda d(S, y)^l \right) \right] \right). \quad (17)$$

Our goal is to learn a policy that maximizes the worst-case long-run reward $\pi^\star = \arg\max_\pi g_{\mathcal{P}}^\pi$.

In the following two sections, we present two model-free algorithms: Robust Q-learning and Robust Actor-Critic, for solving the above problem with $\tilde{\mathcal{O}}(\epsilon^{-2})$ dependence on the target accuracy.

4 ROBUST Q-LEARNING

In this section, we formally study the convergence of the Q-learning algorithm for robust average-reward RL.

4.1 OPTIMAL ROBUST BELLMAN EQUATION

We begin by characterizing the optimal robust Bellman equation in the Q-function space and clarifying its solutions' properties. This foundation will support the contraction analysis and the robust Q-learning algorithm presented in the remainder of the section. For any $Q \in \mathbb{R}^{S \times A}$, write $V_Q(s) := \max_{b \in \mathcal{A}} Q(s, b)$. We now characterize the optimal robust Bellman operator along with the ergodicity assumption used in this setting.

Lemma 4.1 (Lemma 4.1 in Wang et al. [2023c])**.** *Consider the uncentered optimal robust Bellman operator*

$$\mathbf{H}[Q](s, a) = r(s, a) + \min_{p \in \mathcal{P}_s^a} \sum_{s'} p(s'|s, a) \max_{b \in \mathcal{A}} Q(s', b). \quad (18)$$

If (g, Q) is a solution to the optimal robust Bellman equation

$$g + Q(s, a) = \mathbf{H}[Q](s, a), \qquad \forall (s, a) \in \mathcal{S} \times \mathcal{A}, \quad (19)$$

then 1) $g = g_{\mathcal{P}}^*$ Wang et al. [2023b]; 2) the greedy policy w.r.t. Q, $\pi_Q(s) = \arg\max_a Q(s, a)$, is an optimal robust policy Wang et al. [2023b]; 3) $V_Q = V_{\mathsf{P}}^{\pi_Q} + c\mathbf{e}$ for some $\mathsf{P} \in \Omega_g^{\pi_Q}$ and $c \in \mathbb{R}$.

Assumption 4.2. Let Π_{det} be the finite set of deterministic policies. For each $\pi \in \Pi_{\mathrm{det}}$, the center kernel induced $\tilde{\mathsf{P}}^\pi$ is irreducible and aperiodic.

Lemma 4.1 states that any solution (g, Q) of (19) yields the optimal robust average reward $g_{\mathcal{P}}^*$ and an associated deterministic greedy optimal policy. Thus, Assumption 4.2 is stated only over deterministic stationary policies, which matches the policy class produced by Q-learning and incurs no loss for the optimal-control equation. For the contraction theorem below, Assumption 4.2 is supplemented by Assumption A.1 in the appendix. Assumption A.1 is an L-step uniform overlap condition for the nominal center kernels along arbitrary deterministic policy sequences. It is not needed for the Bellman characterization itself; it is used in Appendix A.1 to obtain a contraction modulus that is uniform over the changing greedy policies and admissible robust kernels encountered in the Q-learning analysis.

4.2 SEMI-NORM CONTRACTION OF THE OPTIMAL ROBUST BELLMAN OPERATOR

Lemma 4.1 characterizes the optimal value function as the unique Q (up to a constant shift) that satisfies the robust Bellman equation, yet it does not by itself tell us how to compute that solution. In classical discounted RL, the discounted Bellman operator is a contraction in the sup norm, so fixed-point iteration and Q-learning converge. In the robust average-reward setting no discount factor is present and the Bellman operator is non-linear, so contraction is far from obvious. Prior work Xu et al. [2025b] establishes a semi-norm contraction only for the policy-evaluation operator of a single fixed policy. In contrast, we show that the optimal-control operator $\mathbf{H}$, which acts on the larger space $\mathbb{R}^{S \times A}$ and includes a maximization over actions, is a strict contraction under a carefully constructed semi-norm.

To justify the non-asymptotic convergence of a standard Q–learning update, the following theorem illustrates the one-step contraction property of the optimal robust Bellman operator under a suitable semi-norm, which in turn enables stochastic iterations for solving the optimal robust Bellman equation.

Theorem 4.3. *Let Assumptions 4.2 and A.1 hold. Then, with certain restrictions on the radius of the uncertainty sets, there exists a semi-norm $\|\cdot\|_H$ and $\gamma_H \in (0, 1)$ with*

kernel being $\{c\mathbf{e} : c \in \mathbb{R}\}$ *such that for any* $Q_1, Q_2 \in \mathbb{R}^{SA}$*, we have that*

$$\|\mathbf{H}Q_1 - \mathbf{H}Q_2\|_H \leq \gamma_H \|Q_1 - Q_2\|_H \tag{20}$$

Proof sketch. Our argument proceeds in three steps. First, Appendix A.1 constructs a semi-norm $\|\cdot\|_H$ on $\mathbb{R}^{S\times A}$ that quotients out constant shifts and is built from the fluctuation matrices associated with all admissible deterministic policies and robust kernels. Under Assumption A.1 and the corresponding radius restrictions, these matrices contract by a common factor.

Second, for any Q_1, Q_2, let $\Delta Q = Q_1 - Q_2$ and $\Delta V = V_{Q_1} - V_{Q_2}$. For each state s, the scalar $\Delta V(s)$ lies in the convex hull of the action-wise differences $\{\Delta Q(s,a) : a \in \mathcal{A}\}$. Hence the constructed Q-space semi-norm controls the value-difference term that appears after taking the statewise maximum.

Third, Lemma A.7 shows that, for each action slice, $\big(\mathbf{H}[Q_1] - \mathbf{H}[Q_2]\big)(\cdot, a)$ can be represented by applying an admissible robust kernel and a deterministic policy to ΔV. Combining this representation with the uniform fluctuation contraction yields

$$\|\mathbf{H}Q_1 - \mathbf{H}Q_2\|_H \leq \gamma_H \|Q_1 - Q_2\|_H,$$

for some $\gamma_H < 1$. □

The concrete proof of Theorem 4.3 including the detailed construction of the semi-norm $\|\cdot\|_H$ is in Appendix A.1.

4.3 SIMULATION-BASED ROBUST Q-LEARNING

Equipped with the contraction property of $\mathbf{H}$, we now describe a simulation-based construction of $\hat{\mathbf{H}}$ that approximates the robust Q-Bellman operator as follows

$$\hat{\mathbf{H}}[Q](s,a) = r(s,a) + \hat{\sigma}_{\mathcal{P}_s^a}\big(\max_b Q(\cdot, b)\big), \tag{21}$$

where $\hat{\sigma}_{\mathcal{P}_s^a}$ is an estimator of $\sigma_{\mathcal{P}_s^a}$. Recall that Section 3 characterized the closed form expression of $\sigma_{\mathcal{P}_s^a}$ under contamination, TV distance, and Wasserstein distance uncertainty sets. Algorithm 4 further [Xu et al., 2025b] provides sampling methods for obtaining estimators of $\sigma_{\mathcal{P}_s^a}$ under all the above uncertainty sets with finite sampling complexities, bounded variances and decaying biases.

We now formally present our robust average-reward Q-learning procedure, described in Algorithm 1. Algorithm 1 uses Algorithm 4 to obtain $\hat{\mathbf{H}}[Q]$ in the form of (21), and adapts a synchronous TD learning approach to solve the optimal robust Bellman equation in (19). The algorithm assumes generative access to the nominal simulator $\tilde{\mathsf{P}}$ and synchronously updates all state-action pairs; this is the sampling model analyzed here rather than a trajectory-only online RL setting. An asynchronous variant would naturally update only visited pairs, but proving the same rate would require additional coverage and Markovian-noise arguments. Because the semi-norm in Theorem 4.3 ignores constant shifts, Line 10 of Algorithm 1 subtracts the current value at a fixed anchor state-action pair (s_0, a_0). This keeps the iterates inside the quotient space on which $\mathbf{H}$ contracts.

Algorithm 1 Robust Average Reward Q-Learning

Input: Initial values Q_0, Stepsizes η_t, Anchor state $s_0 \in \mathcal{S}$, Anchor action $a_0 \in \mathcal{A}$

1: **for** $t = 0, 1, \ldots, T-1$ **do**
2: $\quad V_{Q_t} \leftarrow \max_{b\in\mathcal{A}} Q_t(\cdot, b)$
3: $\quad$ **for** each $(s,a) \in \mathcal{S}\times\mathcal{A}$ **do**
4: $\quad\quad$ **if** Contamination **then** Sample $\hat{\sigma}_{\mathcal{P}_s^a}(V_{Q_t})$ according to (14)
5: $\quad\quad$ **else if** TV or Wasserstein **then** Sample $\hat{\sigma}_{\mathcal{P}_s^a}(V_{Q_t})$ according to Algorithm 4
6: $\quad\quad$ **end if**
7: $\quad$ **end for**
8: $\quad \hat{\mathbf{H}}[Q_t](s,a) \leftarrow r(s,a) + \hat{\sigma}_{\mathcal{P}_s^a}(V_{Q_t}), \forall (s,a) \in \mathcal{S}\times\mathcal{A}$
9: $\quad Q_{t+1}(s,a) \leftarrow Q_t(s,a) + \eta_t\left(\hat{\mathbf{H}}(Q_t)(s,a) - Q_t(s,a)\right),\ \forall (s,a) \in \mathcal{S}\times\mathcal{A}$
10: $\quad Q_{t+1}(s,a) = Q_{t+1}(s,a) - Q_{t+1}(s_0,a_0),\ \forall (s,a) \in \mathcal{S}\times\mathcal{A}$
11: **end for** **return** Q_T

The following theorem quantifies the sample complexity needed for Algorithm 1 to return an ϵ-accurate estimate of the robust optimal Q function.

Theorem 4.4. *If Q_t is generated by Algorithm 1 and Assumptions 4.2 and A.1 hold with the radius restrictions in Appendix A.1, then with stepsizes $\eta_t = \Theta(1/t)$ there are fixed problem-dependent constants, independent of ϵ and of the realized policy sequence, such that $\mathbb{E}[\|Q_T - Q^*\|_\infty^2] \leq \epsilon^2$ using $\tilde{\mathcal{O}}(\epsilon^{-2})$ dependence on the target accuracy. The state-action factor is SA for all three uncertainty models; the additional logarithmic factors for TV and Wasserstein come from the MLMC support-function estimator.*

Proof sketch. We view Algorithm 1 as a stochastic approximation to the fixed point of (19) in the quotient space induced by the semi-norm $\|\cdot\|_H$. The contraction property of $\mathbf{H}$ (Theorem 4.3), together with the bounded variance and decreasing bias of the estimator $\hat{\sigma}_{P_s^a}$, implies that the mean update has a negative drift under the smooth semi-Lyapunov function constructed in Appendix A.2, and that the noise term satisfies standard martingale-difference conditions with linear growth. Applying the stochastic approximation bound in Appendix A.2 then yields $\tilde{\mathcal{O}}(\epsilon^{-2})$ synchronous iterations, after absorbing the contraction and Moreau-smoothing constants into the fixed problem-dependent constants. With the MLMC truncation level chosen logarithmically for the TV and Wasserstein estimators, the anchored mean-square sup-norm error satisfies $\mathbb{E}[\|Q_T - Q^*\|_\infty^2] \leq \epsilon^2$. Counting the

state-action samples used in the synchronous updates gives the sample complexities stated in the theorem. □

The concrete proof of Theorem 4.4 is in Appendix A.2, where we also discuss the specific radius restrictions under contamination, TV distance, and Wasserstein distance uncertainty sets.

5 ACTOR-CRITIC

In this section we develop a robust average-reward actor-critic algorithm. Unlike the robust policy-gradient methods of [Sun et al., 2024, Wang et al., 2025], which assume access to the exact robust Q-function or its gradient, our method learns a robust critic solely from samples of the nominal kernel $\tilde{\mathsf{P}}$. On the critic side, we show how to construct, for any policy π, an estimate $\hat{Q}^\pi$ with $\|\hat{Q}^\pi - Q^\pi\|_\infty \le \epsilon$ using a number of nominal samples with $\tilde{\mathcal{O}}(\epsilon^{-2})$ dependence on the target accuracy, and how to obtain bounds whose constants are uniform over all stationary policies. On the actor side, we plug these data-driven critics into a robust mirror-ascent update and prove that the resulting actor-critic algorithm returns an ϵ-optimal robust policy with the same $\tilde{\mathcal{O}}(\epsilon^{-2})$ dependence on the target accuracy, with uncertainty-set-dependent state-action factors stated below. Specifically, given a policy π_k at iteration k, we first learn the robust critic by running the robust evaluation subroutine to obtain the robust value function and the robust average reward, and then reconstruct an estimate of the robust Q function, denoted as $\hat{Q}^{\pi_k}$. This design allows us to reuse the value-function machinery without assuming ergodicity of the Markov chain on the enlarged state-action space, which would be required by a direct Q-based stochastic approximation scheme.

Algorithm 2 Robust Average Reward Actor-Critic

Input: Initial policy π_0, step size sequence $\{\zeta_k\}$

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Obtain $\hat{Q}^{\pi_k}$ via Algorithm 3
3: **for** each $s \in \mathcal{S}$ **do**
4: Update policy: $\pi_{k+1}(\cdot|s) = \arg\max_{p\in\Delta(\mathcal{A})} \left\{\zeta_k \left\langle \hat{Q}^{\pi_k}(s,\cdot), p\right\rangle - \|p - \pi_k(\cdot|s)\|^2\right\}$.
5: **end for**
6: **end for**
7: **return** π_K

Critic analysis: We now formally describe the critic subroutine, the robust Q-function estimation for a fixed policy π, which is characterized in Algorithm 3. Given a policy π, Algorithm 3 constructs $\hat{Q}^\pi$ by first estimating the robust value function $\hat{V}^\pi$ and robust average reward $\hat{g}^\pi$, and then calculate $\hat{Q}^\pi$ by the Bellman equation (shown in Line 4 of Algorithm 3).

Proposition 5.1. *Let V^π be the anchored robust relative value function solving* (7)*, and let Q^π be the robust one-step-deviation Q-function defined by* (11)*. Then Q^π satisfies the robust Bellman equation*

$$Q^\pi(s,a) = r(s,a) - g^\pi_{\mathcal{P}} + \sigma_{\mathcal{P}^a_s}(V^\pi), \tag{22}$$

and

$$V^\pi(s) = \sum_a \pi(a|s) Q^\pi(s,a), \qquad \forall s \in \mathcal{S}. \tag{23}$$

Moreover, every selector P^π_V satisfying (10) *is a worst-case transition kernel for policy π, and the expectation representation* (12) *holds.*

Proposition 5.1 shows that Q^π can be reconstructed from V^π and $g^\pi_{\mathcal{P}}$, both of which are outputs of the policy evaluation routine.

Theorem 5.2. *For any fixed policy π, under Assumption 5.4, let $\hat{Q}^\pi$ be produced by Algorithm 3. Under the radius restrictions in Corollary B.3, there are choices of T, $N_{\max}$, and L_Q, detailed in Appendix B.3, such that*

$$\mathbb{E}\Big[\|\hat{Q}^\pi - Q^\pi\|^2_\infty\Big] \le \epsilon^2, \qquad \mathbb{E}\Big[\|\hat{Q}^\pi - Q^\pi\|_\infty\Big] \le \epsilon. \tag{24}$$

The total nominal transition-sample complexity is $\tilde{\mathcal{O}}(|\mathcal{S}||\mathcal{A}|\epsilon^{-2})$ for contamination uncertainty and $\tilde{\mathcal{O}}((|\mathcal{S}||\mathcal{A}|)^2\epsilon^{-2})$ for TV or Wasserstein uncertainty, with constants independent of π.

The proof in Appendix B.3 combines finite-sample error bounds for robust value and average-reward estimation with the Lipschitz stability of the Q-Bellman map (22). The larger TV and Wasserstein state-action factor comes from controlling the final support-function estimates uniformly over all (s,a) pairs.

Remark 5.3. Appendix B.3 proves policy-uniform critic constants. Thus the same critic parameters can be used along the finite policy sequence generated by Algorithm 2; this is the point that fixed-policy evaluation bounds do not provide by themselves.

Actor-critic sample complexity: Equipped with the robust critic with finite-sample guarantees, we now characterize the actor subroutine. The robust reward objective $g^\pi_{\mathcal{P}} = \min_{\mathsf{P}\in\mathcal{P}} g^\pi_{\mathsf{P}}$ may be nonsmooth because of the inner minimization over transition kernels. The following ergodicity condition is used for the actor analysis.

Assumption 5.4. For every policy $\pi \in \Pi := \{\pi | \pi : \mathcal{S} \to \Delta(\mathcal{A})\}$, the center kernel induced Markov chain $\tilde{\mathsf{P}}^\pi$ is irreducible and aperiodic.

We recall the definition of a Fréchet sub-gradient:

Algorithm 3 Robust Average Reward Q-function Estimation

Input: Policy π, number of critic iterations T, max level $N_{\max}$, support-estimation batch size L_Q

Obtain V_T^π and g_T^π via Algorithm 5
for each $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
  for $\ell = 1, \dots, L_Q$ **do**
    if Contamination **then** Sample $\hat{\sigma}^{(\ell)}_{\mathcal{P}^a_s}(V_T^\pi)$ according to (14)
    else if TV or Wasserstein **then** Sample $\hat{\sigma}^{(\ell)}_{\mathcal{P}^a_s}(V_T^\pi)$ according to Algorithm 4
    end if
  end for
  $\bar{\sigma}_{\mathcal{P}^a_s}(V_T^\pi) \leftarrow \frac{1}{L_Q}\sum_{\ell=1}^{L_Q} \hat{\sigma}^{(\ell)}_{\mathcal{P}^a_s}(V_T^\pi)$
  $\hat{Q}^\pi(s,a) \leftarrow r(s,a) - g_T^\pi + \bar{\sigma}_{\mathcal{P}^a_s}(V_T^\pi)$
end for
return $\hat{Q}^\pi$

Definition 5.5. For any function $f : \mathcal{X} \subseteq \mathbb{R}^N \to \mathbb{R}$, the Fréchet sub-gradient $u \in \mathbb{R}^N$ is a vector that satisfies

$$\liminf_{\delta \to 0,\, \delta \neq 0} \frac{f(x+\delta) - f(x) - \langle u, \delta \rangle}{\|\delta\|} \geq 0. \tag{25}$$

A vector u is called a Fréchet super-gradient of f at x if $-u$ is a Fréchet sub-gradient of $-f$ at x.

The result of Sun et al. [2024] is stated for robust average cost, where the adversary maximizes the cost and the policy performs mirror descent. Applying their result to the cost $c = -r$ gives the reward-side statement used here. Specifically, for the robust reward objective, the direction

$$G^\pi(s,a) = d^\pi_{\mathcal{P}}(s) Q^\pi(s,a) \tag{26}$$

is a Fréchet super-gradient of $g^\pi_{\mathcal{P}}$, where $d^\pi_{\mathcal{P}}$ is the stationary distribution under a worst-case kernel associated with π. Equivalently, $-G^\pi$ is a Fréchet sub-gradient of $-g^\pi_{\mathcal{P}}$. Thus the reward maximization update is a mirror-ascent update in the direction Q^π. The proof in Appendix B.2 only uses the corresponding reward-side performance-difference inequalities.

Let $D(\pi(\cdot|s), \pi'(\cdot|s))$ denote a Bregman divergence between policies at state s, and define the weighted divergence $D_d(\pi, \pi') = \sum_s d(s)\, D(\pi(\cdot|s), \pi'(\cdot|s))$. The exact robust policy mirror-ascent step takes the form

$$\pi_{k+1} = \arg\max_{\pi \in \Pi} \Big\{ \zeta_k \langle G^{\pi_k}, \pi \rangle - D_{d^{\pi_k}_{\mathcal{P}}}(\pi, \pi_k) \Big\}. \tag{27}$$

Choosing D to be the squared Euclidean distance yields the state-wise update below. Indeed, the factor $d^{\pi_k}_{\mathcal{P}}(s)$ multiplies both the linear term and the squared-distance term at each state, and all its entries are positive under the radius conditions. Hence it does not change the state-wise maximizer.

$$\pi_{k+1}(\cdot|s) = \arg\max_{p \in \Delta(\mathcal{A})} \Big\{ \zeta_k \langle Q^{\pi_k}(s,\cdot), p \rangle - \|p - \pi_k(\cdot|s)\|^2 \Big\}, \tag{28}$$

for all $s \in \mathcal{S}$.

The update (28) assumes access to the exact robust Q^{π_k}, which is the setting analyzed by Sun et al. [2024], Wang et al. [2025]. In our model-free setting, Q^{π_k} is unknown and must be estimated. Algorithm 2 therefore replaces Q^{π_k} by $\hat{Q}^{\pi_k}$ obtained from Algorithm 3, leading to the approximate update

$$\pi_{k+1}(\cdot|s) = \arg\max_{p \in \Delta(\mathcal{A})} \Big\{ \zeta_k \langle \hat{Q}^{\pi_k}(s,\cdot), p \rangle - \|p - \pi_k(\cdot|s)\|^2 \Big\}, \tag{29}$$

for all $s \in \mathcal{S}$. Our analysis shows that, as long as each critic estimate satisfies $\mathbb{E}\|\hat{Q}^{\pi_k} - Q^{\pi_k}\|_\infty \leq \epsilon$, the actor still converges to an ϵ-optimal robust policy.

To characterize the overall sample complexity of Algorithm 2, define

$$M := \sup_{\pi,\, P \in \mathcal{P}} \left\| \frac{d^{\pi^*}_P}{d^\pi_{P_\pi}} \right\|_\infty, \tag{30}$$

where $d^{\pi^*}_P$ and $d^\pi_{P_\pi}$ denote the stationary distributions under policies π^* and π with transition kernels $P \in \mathcal{P}$ and $P_\pi \in \Omega^\pi_g$, respectively. This quantity also appears in related analyses such as Sun et al. [2024]. Under Assumption 5.4 and the radius restrictions in Corollary B.3, it is finite; see Lemma B.4.

Theorem 5.6. *Let Assumption 5.4 hold and suppose the radius restrictions in Corollary B.3 hold. Consider Algorithm 2 with the number of policy iterations set to $K = \Theta\big(\log(1/\epsilon)\big)$ and stepsize sequence $\zeta_k \geq \zeta_{k-1}\left(1 - \frac{1}{M}\right)^{-1}$. Let $\mathcal{F}_k$ be the history before the k-th critic call. Suppose that the critic used at iteration k satisfies*

$$\mathbb{E}\left[\|\hat{Q}^{\pi_k} - Q^{\pi_k}\|_\infty \mid \mathcal{F}_k \right] \leq \epsilon, \qquad \forall k \geq 0. \tag{31}$$

This condition is guaranteed by Theorem 5.2 when Algorithm 3 is run with fresh nominal simulator samples and with the parameters stated there. Then the policy returned by Algorithm 2, denoted by π_K, satisfies

$$\mathbb{E}\left[g^{\pi^*}_{\mathcal{P}} - g^{\pi_K}_{\mathcal{P}} \right] \leq \mathcal{O}(\epsilon). \tag{32}$$

Proof sketch. We view (29) as an inexact mirror-ascent step for maximizing $g^\pi_{\mathcal{P}}$, where the exact reward-side supergradient direction Q^{π_k} is replaced by the noisy critic $\hat{Q}^{\pi_k}$. The conditional critic guarantee controls this noise at each random policy generated by the previous actor steps. Combining the reward-side performance-difference inequalities with

the first-order optimality condition of the Euclidean mirror-ascent step gives a one-step recursion for $g_{\mathcal{P}}^{\pi^*} - g_{\mathcal{P}}^{\pi_k}$. The recursion contracts by the factor $1 - 1/M$ up to the critic error and the mirror-ascent residual $2/(M\zeta_k)$. The geometrically increasing sequence $\{\zeta_k\}$ makes this residual summable at the same geometric rate. After $K = \Theta(\log(1/\epsilon))$ actor steps, the remaining error is $\mathcal{O}(\epsilon)$. Full details are given in Appendix B.2. □

Combining this actor iteration bound with Theorem 5.2 gives total nominal transition-sample complexity $\widetilde{\mathcal{O}}(|\mathcal{S}||\mathcal{A}|\epsilon^{-2})$ under contamination uncertainty and $\widetilde{\mathcal{O}}((|\mathcal{S}||\mathcal{A}|)^2\epsilon^{-2})$ under TV or Wasserstein uncertainty, up to the logarithmic number of actor iterations. This improves on the oracle-gradient setting of Sun et al. [2024], Wang et al. [2025], where access to Q^π (or its robust gradient) is assumed.

6 CONCLUSION

This paper presents the first Q-learning and actor–critic algorithms with finite-sample guarantees for distributionally robust average-reward MDPs. We leverage a semi-norm contraction of the optimal robust Bellman operator and establish uniform finite-sample convergence for the critic estimates across all policies. Together with existing tools, these results yield end-to-end $\tilde{\mathcal{O}}(\epsilon^{-2})$ dependence on the target accuracy for obtaining an ϵ-optimal robust policy, with the state-action dependence determined by the uncertainty model.

7 ACKNOWLEDGEMENT

We would like to thank Ankur Naskar of Purdue University for helpful discussions regarding Appendix A.

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

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.

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

Zaiwei Chen. Non-asymptotic guarantees for average-reward q-learning with adaptive stepsizes. *Advances in Neural Information Processing Systems*, 38, 2025.

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.

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

Swetha Ganesh and Vaneet Aggarwal. Regret analysis of average-reward unichain mdps via an actor-critic approach. *Advances in Neural Information Processing Systems*, 38, 2025.

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.

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

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

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.

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

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

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

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.

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.

Yan Li and Guanghui Lan. First-order policy optimization for robust policy evaluation. *arXiv preprint arXiv:2307.15890*, 2023.

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

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

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

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.

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.

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.

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.

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.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. A finite sample complexity bound for distributionally robust q-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3370–3398. PMLR, 2023a.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. Sample complexity of variance-reduced distributionally robust q-learning. *Journal of Machine Learning Research*, 25(341):1–77, 2024a.

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

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, 2023b.

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, 2023c.

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, 2024b.

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.

Yang Xu, Swetha Ganesh, Washim Mondal, Qinbo Bai, and Vaneet Aggarwal. Global convergence for average reward constrained mdps with primal-dual actor critic algorithm. *Advances in Neural Information Processing Systems*, 38, 2025a.

Yang Xu, Washim Uddin Mondal, and Vaneet Aggarwal. Finite-sample analysis of policy evaluation for robust average reward reinforcement learning. *Advances in Neural Information Processing Systems*, 38, 2025b.

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.

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.

Efficient Q-Learning and Actor-Critic Methods for Robust Average-Reward Reinforcement Learning (Supplementary Material)

Yang Xu[1] **Swetha Ganesh**[1] **Vaneet Aggarwal**[1]

[1]Purdue University, West Lafayette, Indiana, USA 47907

A MISSING PROOFS FOR SECTION 4

Recall that in the Q-learning algorithms, the policies are deterministic policies. Thus, we define Π_{det} to be the finite set of all deterministic stationary policies with $|\Pi_{\text{det}}| = A^S$.

Algorithm 4 Truncated MLMC Estimator, Algorithm 1 in Xu et al. [2025b]

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

Sample $N \sim \text{Geom}(1/2)$ on $\{0, 1, 2, \ldots\}$, i.e., $\mathbb{P}(N = n) = 2^{-(n+1)}$

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

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 (15)

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 (17)

end if

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

Let

$$p_{N'} \coloneqq \mathbb{P}(\min\{N, N_{\max}\} = N') = \begin{cases} 2^{-(N'+1)}, & 0 \le N' < N_{\max}, \\ 2^{-N_{\max}}, & N' = N_{\max}. \end{cases}$$

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

return $\hat{\sigma}_{\mathcal{P}_s^a}(V)$

A.1 PROOF OF THEOREM 4.3

For any $\mathsf{P} \in \mathcal{P}$, denote P^π as the transition matrix under policy $\pi \in \Pi_{\text{det}}$ and the unique stationary distribution d_{P}^π. We define the family of fluctuation matrices to be

$$\mathcal{F} \coloneqq \{F_{\mathsf{P}}^\pi = \mathsf{P}^\pi - E_{\mathsf{P}} : \pi \in \Pi_{\text{det}}, \mathsf{P} \in \mathcal{P}\} \tag{33}$$

where E_{P} is the matrix with all rows being identical to d_{P}^π.

To construct the desired one-step semi-norm contraction for $\mathbf{H}$, one key step is to show that the joint spectral radius of the family $\mathcal{F}$ is strictly less than 1. We introduce the Dobrushin's coefficient for an $n \times n$ Markov matrix P defined as follows:

$$\tau(P) \;:=\; 1 - \min_{i<j} \sum_{s=1}^{n} \min\big(P_{is}, P_{js}\big). \tag{34}$$

By Lemma C.5, denote the joint spectral radius of $\mathcal{F}$ as $\hat{\rho}(\mathcal{F})$ we have,

$$\hat{\rho}(\mathcal{F}) \coloneqq \lim_{k\to\infty} \sup_{F_i \in \mathcal{F}} \|F_k \dots F_1\|^{\frac{1}{k}} \le \inf_{m\ge 1} \left(\sup_{\substack{\pi_1,\dots,\pi_m \in \Pi \\ \mathsf{P}_1,\dots,\mathsf{P}_m \in \mathcal{P}}} \tau\big(\mathsf{P}_m^{\pi_m} \cdots \mathsf{P}_1^{\pi_1}\big) \right)^{1/m}. \tag{35}$$

Note that (35) relates the joint spectral radius of $\mathcal{F}$ to the Dobrushin's coefficient of product of $\mathsf{P}_i^{\pi_i}$. We now introduce a mild assumption to ensure a uniform "minimal mixing" across all deterministic policy sequences, which prevents the induced kernels from drifting arbitrarily far apart under adversarial policy switching. The assumption quantifies over deterministic policy sequences and does not require policy randomization.

Assumption A.1. Let $\{\tilde{\mathsf{P}}^{\pi} : \pi \in \Pi_{\mathrm{det}}\}$ be the centroid family. There exist an integer $L \ge 1$, a constant $\varepsilon_L \in (0, 1]$, and a distribution $\nu \in \Delta(\mathcal{S})$ such that for every L-length deterministic policy sequence $(\pi_1, \dots, \pi_L)$, for every $s \in \mathcal{S}$,

$$(\tilde{\mathsf{P}}^{\pi_L} \cdots \tilde{\mathsf{P}}^{\pi_1})(s, \cdot) \ge \varepsilon_L \nu(\cdot). \tag{36}$$

Remark A.2. Assumption A.1 is a sufficient proof condition for obtaining a policy-sequence-uniform contraction; it is not implied by fixed-policy irreducibility alone. It is, however, easy to check in common smoothed tabular models:

1. **Single-step minorization.** If there exist $\varepsilon > 0$ and $\nu \in \Delta(\mathcal{S})$ such that $\tilde{\mathsf{P}}(\cdot|s,a) \ge \varepsilon\nu(\cdot)$ for all (s,a), then Assumption A.1 holds with $L = 1$ and $\varepsilon_L = \varepsilon$.
2. **Rare reset or leakage.** If over every block of L steps, uniformly over the start state and action sequence, the nominal dynamics have probability at least $\beta > 0$ of drawing the state from a fixed distribution ν, then Assumption A.1 holds with $\varepsilon_L = \beta$.

Thus, small policy-independent jitter, smoothing, or rare reset events are sufficient to ensure the required uniform overlap.

Utilizing the ε_L defined in (36), we now formally discuss the uncertainty radius restrictions in the settings considered.

Lemma A.3 (Contamination uncertainty radius restrictions)**.** *Under Assumption 4.2 and Assumption A.1, consider the contamination uncertainty set*

$$\mathcal{P} \coloneqq \Big\{ \mathsf{P} : \forall (s,a), \mathsf{P}(\cdot|s,a) = (1-\delta)\tilde{\mathsf{P}}(\cdot|s,a) + \delta q(\cdot|s,a), q(\cdot|s,a) \in \Delta(\mathcal{S}) \Big\}, \quad 0 \le \delta < 1, \tag{37}$$

no smallness restriction on δ is needed beyond $\delta < 1$. Specifically,

$$\hat{\rho}(\mathcal{F}) \le \Big(1 - (1-\delta)^L \varepsilon_L\Big)^{1/L} \;<\; 1. \tag{38}$$

Moreover, for any π and $\mathsf{P} \in \mathcal{P}$, the induced kernel P^{π} is irreducible and aperiodic.

Proof. For a fixed policy π, the induced state–transition matrix P^{π} is expressed as

$$\mathsf{P}^{\pi}(s,s') \coloneqq \sum_a \pi(a|s)\mathsf{P}(s'|s,a) = (1-\delta)\sum_a \pi(a|s)\tilde{\mathsf{P}}(s'|s,a) + \delta \sum_a \pi(a|s) q(s'|s,a) \tag{39}$$

Define the induced uncertainty set $\mathcal{P}^{\pi} \coloneqq \{\mathsf{P}^{\pi} : \mathsf{P} \in \mathcal{P}, \pi \in \Pi_{\mathrm{det}}\}$ and define $\tilde{\mathsf{P}}^{\pi}(s,s') \coloneqq \sum_a \pi(a|s)\tilde{\mathsf{P}}(s'|s,a)$. Then (39) can be expressed as

$$\mathcal{P}^{\pi} \;= \Big\{ (1-\delta)\tilde{\mathsf{P}}^{\pi} + \delta q^{\pi} : q^{\pi} \text{ row–stochastic}, \pi \in \Pi_{\mathrm{det}} \Big\}. \tag{40}$$

Fix any length-L sequence $(\pi_1, \dots, \pi_L)$ and admissible contaminated kernels $\mathsf{P}_t^{\pi_t} = (1-\delta)\tilde{\mathsf{P}}^{\pi_t} + \delta q_t$. By nonnegativity, for each $i, j \in \mathcal{S}$

$$\mathsf{P}_L^{\pi_L} \cdots \mathsf{P}_1^{\pi_1}(i,j) \geq (1-\delta)^L \tilde{\mathsf{P}}^{\pi_L} \cdots \tilde{\mathsf{P}}^{\pi_1}(i,j).$$

Assumption A.1 implies for all $i \neq j$,

$$\sum_s \min\left\{ \left(\tilde{\mathsf{P}}^{\pi_L} \cdots \tilde{\mathsf{P}}^{\pi_1}\right)_{is}, \left(\tilde{\mathsf{P}}^{\pi_L} \cdots \tilde{\mathsf{P}}^{\pi_1}\right)_{js} \right\} \geq \varepsilon_L > 0. \tag{41}$$

By the fact that constant scaling preserves minima, hence

$$\sum_s \min\left\{ \left(\mathsf{P}_L^{\pi_L} \cdots \mathsf{P}_1^{\pi_1}\right)_{is}, \left(\mathsf{P}_L^{\pi_L} \cdots \mathsf{P}_1^{\pi_1}\right)_{js} \right\} \geq (1-\delta)^L \varepsilon_L > 0, \tag{42}$$

which is equivalent to the Dobrushin bound

$$\tau\left(\mathsf{P}_L^{\pi_L} \cdots \mathsf{P}_1^{\pi_1}\right) \leq 1 - (1-\delta)^L \varepsilon_L < 1. \tag{43}$$

Taking the supremum over all sequences and applying (35) implies

$$\hat{\rho}(\mathcal{F}) \leq \inf_{m \geq 1} \Big(\sup_{\substack{\pi_1, \dots, \pi_m \in \Pi \\ \mathsf{P}_1, \dots, \mathsf{P}_m \in \mathcal{P}}} \tau\left(\mathsf{P}_m^{\pi_m} \cdots \mathsf{P}_1^{\pi_1}\right) \Big)^{1/m} \leq \left(1 - (1-\delta)^L \varepsilon_L\right)^{1/L} < 1.$$

Regarding ergodicity, for each fixed π, since $\tilde{\mathsf{P}}^{\pi}$ is ergodic, there exists m_π with $(\tilde{\mathsf{P}}^{\pi})^{m_\pi} > 0$. Then $(\mathsf{P}^{\pi})^{m_\pi} \geq (1-\delta)^{m_\pi} (\tilde{\mathsf{P}}^{\pi})^{m_\pi} > 0$, so P_π is also irreducible and aperiodic. □

Lemma A.4 (TV distance uncertainty radius restrictions)**.** *Under Assumption 4.2 and Assumption A.1, consider the TV uncertainty set*

$$\mathcal{P} := \left\{ \mathsf{P} : \forall (s,a), \mathrm{TV}\left(\mathsf{P}(\cdot|s,a), \tilde{\mathsf{P}}(\cdot|s,a)\right) \leq \delta \right\}, \quad \delta \geq 0, \tag{44}$$

Because Π_{det} is finite, there exists $\tilde{m} \in \mathbb{N}$ such that $(\tilde{\mathsf{P}}^{\pi})^{\tilde{m}} > 0$ for all $\pi \in \Pi_{\mathrm{det}}$, set $b_0 = \min_{\pi \in \Pi_{\mathrm{det}}} \min_{i,s \in \mathcal{S}} \left((\tilde{\mathsf{P}}^{\pi})^{\tilde{m}}\right)_{is} \in (0,1]$, and define the TV radius threshold $\delta^\star_{\mathrm{TV}} = \min\left\{ \frac{\varepsilon_L}{2L}, \frac{b_0}{2\tilde{m}} \right\}$. Then if $0 \leq \delta < \delta^\star_{\mathrm{TV}}$, then the joint spectral radius of $\mathcal{F}$ is strictly less than 1. Specifically,

$$\hat{\rho}(\mathcal{F}) \leq \left(1 - (\varepsilon_L - 2L\delta)\right)^{1/L} < 1. \tag{45}$$

Moreover, for any π and $\mathsf{P} \in \mathcal{P}$, the induced kernel P^{π} is irreducible and aperiodic.

Proof. For a fixed policy π, the induced state–transition matrix P^{π} is expressed as

$$\mathsf{P}^{\pi}(s,s') := \sum_a \pi(a|s) \mathsf{P}(s'|s,a). \tag{46}$$

Then for each state s we have,

$$\mathrm{TV}\left(\mathsf{P}^{\pi}(s,\cdot), \tilde{\mathsf{P}}^{\pi}(s,\cdot)\right) = \mathrm{TV}\Big(\sum_a \pi(a|s)\mathsf{P}(\cdot|s,a), \sum_a \pi(a|s)\tilde{\mathsf{P}}(\cdot|s,a)\Big) \leq \sum_a \pi(a|s)\mathrm{TV}\left(\mathsf{P}(\cdot|s,a), \tilde{\mathsf{P}}(\cdot|s,a)\right) \leq \delta, \tag{47}$$

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

$$\mathcal{P}^{\pi} := \{\mathsf{P}^{\pi} : \mathsf{P} \in \mathcal{P}, \pi \in \Pi_{\mathrm{det}}\} \subseteq \left\{ M \text{ row–stochastic} : \forall s, \exists\, \pi \in \Pi_{\mathrm{det}}, \mathrm{TV}\left(M(s,\cdot), \tilde{\mathsf{P}}^{\pi}(s,\cdot)\right) \leq \delta \right\}. \tag{48}$$

For any L-length policy sequence $(\pi_1, \dots, \pi_L)$ and for any $\mathsf{P}_t \in \mathcal{P}$, by (48), for $t = 1, \dots, L$, we have

$$\sup_i \mathrm{TV}\left(\mathsf{P}_t^{\pi_t}(i,\cdot), \tilde{\mathsf{P}}^{\pi_t}(i,\cdot)\right) \leq \delta. \tag{49}$$

TV is a nonexpansion under right-multiplication by a stochastic matrix, hence the usual telescoping gives, for each start state i, by (49) we have

$$\mathrm{TV}\big(\mathsf{P}_1^{\pi_1}\cdots\mathsf{P}_L^{\pi_L}(i,\cdot),\tilde{\mathsf{P}}^{\pi_1}\cdots\tilde{\mathsf{P}}^{\pi_L}(i,\cdot)\big)\le\sum_{t=1}^{L}\sup_x\mathrm{TV}\big(\mathsf{P}_t^{\pi_t}(x,\cdot),\tilde{\mathsf{P}}^{\pi_t}(x,\cdot)\big)\ \le\ L\delta. \tag{50}$$

Then for $i\neq j$,

$$\begin{aligned}\sum_s\min\{\mathsf{P}_1^{\pi_1}\cdots\mathsf{P}_L^{\pi_L}(i,s),\mathsf{P}_1^{\pi_1}\cdots\mathsf{P}_L^{\pi_L}(j,s)\}\ge&\sum_s\min\{\tilde{\mathsf{P}}^{\pi_1}\cdots\tilde{\mathsf{P}}^{\pi_L}(i,s),\tilde{\mathsf{P}}^{\pi_1}\cdots\tilde{\mathsf{P}}^{\pi_L}(j,s)\}\\&-\mathrm{TV}\big(\mathsf{P}_1^{\pi_1}\cdots\mathsf{P}_L^{\pi_L}(i,\cdot),\tilde{\mathsf{P}}^{\pi_1}\cdots\tilde{\mathsf{P}}^{\pi_L}(i,\cdot)\big)\\&-\mathrm{TV}\big(\mathsf{P}_1^{\pi_1}\cdots\mathsf{P}_L^{\pi_L}(j,\cdot),\tilde{\mathsf{P}}^{\pi_1}\cdots\tilde{\mathsf{P}}^{\pi_L}(j,\cdot)\big).\end{aligned} \tag{51}$$

By Assumption A.1 and (50), $\sum_s\min\{\tilde{\mathsf{P}}^{\pi_1}\cdots\tilde{\mathsf{P}}^{\pi_L}(i,s),\tilde{\mathsf{P}}^{\pi_1}\cdots\tilde{\mathsf{P}}^{\pi_L}(j,s)\}\ge\varepsilon_L$, and the two TV terms are each $\le L\delta$, whence

$$\min_{i<j}\sum_s\min\{\mathsf{P}_1^{\pi_1}\cdots\mathsf{P}_L^{\pi_L}(i,s),\mathsf{P}_1^{\pi_1}\cdots\mathsf{P}_L^{\pi_L}(j,s)\}\ge\ \varepsilon_L-2L\delta.$$

Using the identity $\tau(M)=1-\min_{i<j}\sum_s\min\{M_{is},M_{js}\}$ for row-stochastic M further yields

$$\tau\big(\mathsf{P}_1^{\pi_1}\cdots\mathsf{P}_L^{\pi_L}\big)\le1-(\varepsilon_L-2L\delta)<1, \tag{52}$$

when $\delta<\varepsilon_L/(2L)$. Taking the supremum over sequences and applying (35) proves (45).

Regarding ergodicity, fix $\pi\in\Pi_{\mathrm{det}}$. Consider $(\mathsf{P}^\pi)^{\tilde m}$ versus $(\tilde{\mathsf{P}}^\pi)^{\tilde m}$. By the same telescoping as (50), for each start state i,

$$\mathrm{TV}\big((\mathsf{P}^\pi)^{\tilde m}(i,\cdot),(\tilde{\mathsf{P}}^\pi)^{\tilde m}(i,\cdot)\big)\le\tilde m\delta,$$

hence the ℓ_1 distance between those two state space distributions is at most $2\tilde m\delta$. Thus, for every (i,s),

$$(\mathsf{P}^\pi)^{\tilde m}_{is}\ \ge\ (\tilde{\mathsf{P}}^\pi)^{\tilde m}_{is}\ -\ \big\|(\mathsf{P}^\pi)^{\tilde m}(i,\cdot)-(\tilde{\mathsf{P}}^\pi)^{\tilde m}(i,\cdot)\big\|_1\ \ge\ b_0\ -\ 2\tilde m\,\delta.$$

If $\delta<b_0/(2\tilde m)$, then $(\mathsf{P}^\pi)^{\tilde m}_{is}>0$ for all i,s, i.e., $(\mathsf{P}^\pi)^{\tilde m}>0$ entrywise, so P^π is irreducible and aperiodic. Since the bound is uniform in π, the ergodicity holds for all $\pi\in\Pi_{\mathrm{det}}$. □

Lemma A.5 (Wasserstein distance uncertainty radius restrictions)**.** *Let $(\mathcal{S},d)$ be a finite metric space with $\delta_{\min}:=\min_{x\neq y}d(x,y)>0$, and $p\in[1,\infty)$. Under Assumption 4.2 and Assumption A.1, consider the Wasserstein uncertainty set*

$$\mathcal{P}:=\Big\{\mathsf{P}:\forall(s,a),W_p\big(\mathsf{P}(\cdot|s,a),\tilde{\mathsf{P}}(\cdot|s,a)\big)\le\delta\Big\},\qquad\delta\ge0.$$

Since Π_{det} is finite, there exists $\tilde m\in\mathbb{N}$ with $(\tilde{\mathsf{P}}^\pi)^{\tilde m}>0$ for all π, and define $b_0=\min_{\pi\in\Pi_{\mathrm{det}}}\min_{i,s\in\mathcal{S}}\big((\tilde{\mathsf{P}}^\pi)^{\tilde m}\big)_{is}\in(0,1]$. Set the Wasserstein radius threshold $\delta^\star_{\mathrm{W}}:=\min\big\{\frac{\delta_{\min}\varepsilon_L}{2L},\frac{\delta_{\min}b_0}{2\tilde m}\big\}$. Then if $0\le\delta<\delta^\star_{\mathrm{W}}$, then the joint spectral radius of $\mathcal{F}$ is strictly less than 1*. Specifically,*

$$\hat\rho(\mathcal{F})\le\Big(1-(\varepsilon_L-\frac{2L\delta}{\delta_{\min}})\Big)^{1/L}<1. \tag{53}$$

Moreover, for any π and $\mathsf{P}\in\mathcal{P}$, the induced kernel P^π is irreducible and aperiodic.

Proof. For a fixed policy π, the induced state–transition matrix P^π is expressed as

$$\mathsf{P}^\pi(s,s'):=\sum_a\pi(a|s)\mathsf{P}(s'|s,a) \tag{54}$$

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

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

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

$$\mathcal{P}^\pi \ \subseteq \big\{ M \text{ row–stochastic} : \ \forall s,\ W_p(M(s,\cdot), \tilde{\mathsf{P}}^\pi(s,\cdot); d) \le \delta \big\}. \tag{55}$$

We now draw connection between (55) and the TV version in (48). 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

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

which implies that

$$\mathrm{TV}(u,v) \ \le \ \frac{W_1(u,v;d)}{\delta_{\min}} \ \le \frac{W_p(u,v;d)}{\delta_{\min}}. \tag{56}$$

Therefore we can reduce (55) into a TV distance uncertainty set characterized as follows:

$$\mathcal{P}^\pi \subseteq \Big\{ M \text{ row–stochastic} : \ \forall s,\ \mathrm{TV}\big(M(s,\cdot), \tilde{\mathsf{P}}^\pi(s,\cdot)\big) \le \frac{\delta}{\delta_{\min}} \Big\}. \tag{57}$$

Invoking Lemma A.4 on (57) yields the results. □

Algorithm 5 Robust Average Reward TD, Algorithm 2 in [Xu et al., 2025b]

Input: Policy π, Initial values V_0, $g_0 = 0$, Stepsizes η_t, β_t, Max level $N_{\max}$, Anchor state $s_0 \in \mathcal{S}$

for $t = 0, 1, \ldots, T-1$ **do**
  for each $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
    if Contamination **then** Sample $\hat{\sigma}_{\mathcal{P}_s^a}(V_t)$ according to (14)
    else if TV or Wasserstein **then** Sample $\hat{\sigma}_{\mathcal{P}_s^a}(V_t)$ according to Algorithm 4
    end if
  end for
  $\hat{\mathbf{T}}_{g_0}(V_t)(s) \leftarrow \sum_a \pi(a|s)\big[r(s,a) - g_0 + \hat{\sigma}_{\mathcal{P}_s^a}(V_t)\big], \quad \forall s \in \mathcal{S}$
  $V_{t+1}(s) \leftarrow V_t(s) + \eta_t \Big(\hat{\mathbf{T}}_{g_0}(V_t)(s) - V_t(s)\Big), \quad \forall s \in \mathcal{S}$
  $V_{t+1}(s) = V_{t+1}(s) - V_{t+1}(s_0), \quad \forall s \in \mathcal{S}$
end for
for $t = 0, 1, \ldots, T-1$ **do**
  for each $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
    if Contamination **then** Sample $\hat{\sigma}_{\mathcal{P}_s^a}(V_T)$ according to (14)
    else if TV or Wasserstein **then** Sample $\hat{\sigma}_{\mathcal{P}_s^a}(V_T)$ according to Algorithm 4
    end if
  end for
  $\hat{\delta}_t(s) \leftarrow \sum_a \pi(a|s)\big[r(s,a) + \hat{\sigma}_{\mathcal{P}_s^a}(V_T)\big] - V_T(s), \quad \forall s \in \mathcal{S}$
  $\bar{\delta}_t \leftarrow \frac{1}{S}\sum_s \hat{\delta}_t(s)$
  $g_{t+1} \leftarrow g_t + \beta_t(\bar{\delta}_t - g_t)$
end for **return** V_T, g_T

Under the conditions of Lemma A.3-A.5, denote $r^* = \hat{\rho}(\mathcal{F})$, we construct the extremal norm $\|\cdot\|_{\mathrm{ext}}$ as follows:

$$\|x\|_{\mathrm{ext}} \coloneqq \sup_{k \ge 0} \sup_{F_1, \ldots, F_k \in \mathcal{F}} \alpha^{-k} \|F_k F_{k-1} \ldots F_1 x\|_2 \tag{58}$$

while α is chosen arbitrarily from $(r^*, 1)$ and we follow the convention that $\|F_k F_{k-1} \ldots F_1 x\|_2 = \|x\|_2$ when $k = 0$.

Lemma A.6. *Under the conditions of Lemma A.3-A.5, the operator $\|\cdot\|_{\mathrm{ext}}$ is a valid norm with $\|F_{\mathsf{P}}^\pi\|_{\mathrm{ext}} \le \alpha < 1$ for all $\mathsf{P} \in \mathcal{P}$ and $\pi \in \Pi_{\mathrm{det}}$.*

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

$$\|F_k F_{k-1} \ldots F_1\|_2 \le C\lambda^k \tag{59}$$

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

$$\alpha^{-k}\|F_kF_{k-1}\dots F_1x\|_2 \le \alpha^{-k}C\lambda^k\|x\|_2 = C\left(\frac{\lambda}{\alpha}\right)^k\|x\|_2 \longrightarrow 0 \quad \text{as} \quad k\to\infty \tag{60}$$

Thus the double supremum in (58) 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

$$\sup_{k=0}\sup_{F_1,\dots,F_k\in\mathcal{F}}\alpha^{-k}\|F_kF_{k-1}\dots F_1x\|_2 = \|x\|_2 = 0 \Rightarrow x = \mathbf{0} \tag{61}$$

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

$$\|cx\|_{\text{ext}} = \sup_{k\ge0}\sup_{F_1,\dots,F_k\in\mathcal{F}}\alpha^{-k}\|F_kF_{k-1}\dots F_1(cx)\|_2 = |c|\|x\|_{\text{ext}} \tag{62}$$

Regarding triangle inequality, using $\|F_k\dots F_1(x+y)\|_2 \le \|F_k\dots F_1x\|_2 + \|F_k\dots F_1y\|_2$ for any $x, y \in \mathbb{R}^S$, we obtain,

$$\|x+y\|_{\text{ext}} = \sup_{k\ge0}\sup_{F_1,\dots,F_k\in\mathcal{F}}\alpha^{-k}\|F_kF_{k-1}\dots F_1(x+y)\|_2 \le \|x\|_{\text{ext}} + \|y\|_{\text{ext}} \tag{63}$$

For any $F_{\mathsf{P}}^\pi \in \mathcal{F}$, we have

$$\begin{aligned}
\|F_{\mathsf{P}}^\pi x\|_{\text{ext}} &= \sup_{k\ge0}\sup_{F_1,\dots,F_k\in\mathcal{F}}\alpha^{-k}\|F_kF_{k-1}\dots F_1(F_{\mathsf{P}}^\pi x)\|_2 \\
&\le \sup_{k\ge1}\sup_{F_1,\dots,F_k\in\mathcal{F}}\alpha^{-(k-1)}\|F_kF_{k-1}\dots F_1x\|_2 \\
&= \alpha\sup_{k\ge1}\sup_{F_1,\dots,F_k\in\mathcal{F}}\alpha^{-k}\|F_kF_{k-1}\dots F_1x\|_2 \\
&\le \alpha\sup_{k\ge0}\sup_{F_1,\dots,F_k\in\mathcal{F}}\alpha^{-k}\|F_kF_{k-1}\dots F_1x\|_2 \\
&= \alpha\|x\|_{\text{ext}}
\end{aligned} \tag{64}$$

Since F_{P}^π is arbitrary, (64) implies that for any $F_{\mathsf{P}}^\pi \in \mathcal{F}$,

$$\|F_{\mathsf{P}}^\pi\|_{\text{ext}} = \sup_{x\neq\mathbf{0}}\frac{\|F_{\mathsf{P}}^\pi x\|_{\text{ext}}}{\|x\|_{\text{ext}}} \le \alpha < 1 \tag{65}$$

□

We are now ready to construct our desired $\|\cdot\|_H$. For any deterministic policy $\pi \in \Pi_{\text{det}}$ and any $Q \in \mathbb{R}^{S\times A}$, write

$$Q^\pi(s) := Q(s,\pi(s)), \qquad s \in \mathcal{S}.$$

Recall that

$$\|x\|_{\tilde{H}} \coloneqq \sup_{F_{\mathsf{P}}^\pi\in\mathcal{F}}\|F_{\mathsf{P}}^\pi x\|_{\text{ext}}, \qquad x \in \mathbb{R}^S. \tag{66}$$

Choose a scalar $\xi \in (0, 1-\alpha)$ and define

$$\|Q\|_H \coloneqq \sup_{\pi\in\Pi_{\text{det}}}\|Q^\pi\|_{\tilde{H}} + \xi\sup_{\pi\in\Pi_{\text{det}}}\inf_{c\in\mathbb{R}}\|Q^\pi - c\mathbf{e}\|_{\text{ext}}. \tag{67}$$

(i) $\|\cdot\|_H$ is a valid semi-norm. The first term in (67) is a supremum of semi-norms, and the second term is a supremum of quotient semi-norms. Hence $\|\cdot\|_H$ is nonnegative, homogeneous, and satisfies the triangle inequality.

(ii) Kernel of $\|\cdot\|_H$. Clearly every constant $Q = c\mathbf{e}$ satisfies $\|Q\|_H = 0$. Conversely, if $\|Q\|_H = 0$, then for every deterministic policy π, the vector Q^π is constant over states. Since deterministic policies can choose arbitrary actions state by state, this implies that all entries $Q(s,a)$ are equal to the same constant. Hence the kernel is exactly $\{c\mathbf{e} : c \in \mathbb{R}\}$.

(iii) One-Step Contraction Recall that the optimal robust Bellman operator is defined as follows

$$\mathbf{H}[Q](s,a) = r(s,a) + \min_{p\in\mathcal{P}_s^a} \sum_{s'} p(s'|s,a) \max_{b\in\mathcal{A}} Q(s',b) \tag{68}$$

Since $V_Q(s) = \max_{b\in\mathcal{A}} Q(s,b)$. For any $Q_1, Q_2 \in \mathbb{R}^{SA}$ and for each (s,a) pair,

$$\mathbf{H}[Q_1](s,a) - \mathbf{H}[Q_2](s,a) = \min_{p\in\mathcal{P}_s^a} p^\top V_{Q_1} - \min_{p\in\mathcal{P}_s^a} p^\top V_{Q_2} \le \max_{p\in\mathcal{P}_s^a} p^\top (V_{Q_1} - V_{Q_2}) \tag{69}$$

Likewise, we also have that

$$\begin{aligned}
\mathbf{H}[Q_2](s,a) - \mathbf{H}[Q_1](s,a) &\le \max_{p\in\mathcal{P}_s^a} p^\top (V_{Q_2} - V_{Q_1}) \\
&= \max_{p\in\mathcal{P}_s^a} p^\top (V_{Q_2} - V_{Q_1}) \\
&= \max_{p\in\mathcal{P}_s^a} -p^\top (V_{Q_1} - V_{Q_2}) \\
&= -\min_{p\in\mathcal{P}_s^a} p^\top (V_{Q_1} - V_{Q_2})
\end{aligned} \tag{70}$$

By (69)-(70), we thus have

$$\min_{p\in\mathcal{P}_s^a} p^\top (V_{Q_1} - V_{Q_2}) \le \mathbf{H}[Q_1](s,a) - \mathbf{H}[Q_2](s,a) \le \max_{p\in\mathcal{P}_s^a} p^\top (V_{Q_1} - V_{Q_2}) \tag{71}$$

Lemma A.7. *Fix $Q_1, Q_2 \in \mathbb{R}^{S\times A}$ and let $\Delta V := V_{Q_1} - V_{Q_2} \in \mathbb{R}^S$. Assume that each row set $\mathcal{P}_s^a$ is convex and compact. Then for every deterministic policy $\mu \in \Pi_{\mathrm{det}}$, there exists a transition kernel $\mathsf{P}_\mu \in \mathcal{P}$ such that*

$$\big(\mathbf{H}[Q_1] - \mathbf{H}[Q_2]\big)^\mu = \mathsf{P}_\mu^\mu \, \Delta V. \tag{72}$$

Proof. Fix $\mu \in \Pi_{\mathrm{det}}$. For each $s \in \mathcal{S}$, denote

$$h_s := \mathbf{H}[Q_1](s,\mu(s)) - \mathbf{H}[Q_2](s,\mu(s)),$$

and denote $m_s := \min_{p\in\mathcal{P}_s^{\mu(s)}} p^\top \Delta V$, $M_s := \max_{p\in\mathcal{P}_s^{\mu(s)}} p^\top \Delta V$. By the elementary inequality for differences of minima, $m_s \le h_s \le M_s$. Since $\mathcal{P}_s^{\mu(s)}$ is compact and $p \mapsto p^\top \Delta V$ is linear, we choose $p_s^{\min} \in \arg\min_{p\in\mathcal{P}_s^{\mu(s)}} p^\top \Delta V$, and $p_s^{\max} \in \arg\max_{p\in\mathcal{P}_s^{\mu(s)}} p^\top \Delta V$. If $M_s = m_s$, we set $\bar{p}_s = p_s^{\min}$. Otherwise we then define $\lambda_s := \frac{h_s - m_s}{M_s - m_s} \in [0,1]$, and $\bar{p}_s := (1-\lambda_s) p_s^{\min} + \lambda_s p_s^{\max}$.

By convexity, $\bar{p}_s \in \mathcal{P}_s^{\mu(s)}$, and by construction $\bar{p}_s^\top \Delta V = h_s$. By (s,a)-rectangularity, the rows $\{\bar{p}_s : s \in \mathcal{S}\}$ can be assembled into a kernel $\mathsf{P}_\mu \in \mathcal{P}$ by setting the $(s,\mu(s))$ row equal to $\bar{p}_s$ and choosing all other rows arbitrarily from their row uncertainty sets. Then, for every s,

$$[\mathsf{P}_\mu^\mu \Delta V](s) = \sum_{s'} \mathsf{P}_\mu(s'|s,\mu(s)) \Delta V(s') = \bar{p}_s^\top \Delta V = h_s. \tag{73}$$

This proves (72). □

For $x \in \mathbb{R}^S$, define $N_0(x) := \sup_{F\in\mathcal{F}} \|Fx\|_{\mathrm{ext}}$. Then the semi-norm (67) can be written as

$$\|Q\|_H = \sup_{\mu\in\Pi_{\mathrm{det}}} N_0(Q^\mu) + \xi \sup_{\mu\in\Pi_{\mathrm{det}}} \inf_{c\in\mathbb{R}} \|Q^\mu - c\mathbf{e}\|_{\mathrm{ext}}.$$

Let $\Delta Q = Q_1 - Q_2$ and $\Delta V = V_{Q_1} - V_{Q_2}$. For every state s,

$$\Delta V(s) = \max_a Q_1(s,a) - \max_a Q_2(s,a)$$

belongs to the convex hull of $\{\Delta Q(s,a) : a \in \mathcal{A}\}$. Hence there exists a randomized state-wise selector $\theta(\cdot|s)$ such that

$$\Delta V(s) = \sum_a \theta(a|s) \Delta Q(s,a).$$

Equivalently, ΔV belongs to the convex hull of the deterministic policy slices $\{\Delta Q^\mu : \mu \in \Pi_{\det}\}$. Since N_0 is a semi-norm,

$$N_0(\Delta V) \le \sup_{\mu \in \Pi_{\det}} N_0(\Delta Q^\mu) \le \|\Delta Q\|_H. \tag{74}$$

Now fix any deterministic policy μ. By Lemma A.7, there exists $\mathsf{P}_\mu \in \mathcal{P}$ such that

$$\big(\mathbf{H}[Q_1] - \mathbf{H}[Q_2]\big)^\mu = \mathsf{P}_\mu^\mu \Delta V.$$

Let E_μ be the rank-one matrix whose rows are the stationary distribution of P_μ^μ, and set $F_\mu \coloneqq \mathsf{P}_\mu^\mu - E_\mu \in \mathcal{F}$. Since $E_\mu \Delta V$ is a constant vector, $F(E_\mu \Delta V) = 0$ for all $F \in \mathcal{F}$. Therefore,

$$N_0(\mathsf{P}_\mu^\mu \Delta V) = \sup_{F \in \mathcal{F}} \|F F_\mu \Delta V\|_{\text{ext}} \le \alpha N_0(\Delta V), \tag{75}$$

where the last inequality follows from the extremal-norm contraction.

Moreover by $\inf_{c \in \mathbb{R}} \|\mathsf{P}_\mu^\mu \Delta V - c\mathbf{e}\|_{\text{ext}} \le \|F_\mu \Delta V\|_{\text{ext}} \le N_0(\Delta V)$, then combining the last two displays and taking the supremum over $\mu \in \Pi_{\det}$ gives

$$\|\mathbf{H}[Q_1] - \mathbf{H}[Q_2]\|_H \le (\alpha + \xi) N_0(\Delta V). \tag{76}$$

Using (74), we obtain $\|\mathbf{H}[Q_1] - \mathbf{H}[Q_2]\|_H \le (\alpha + \xi)\|Q_1 - Q_2\|_H$. Thus the contraction holds with $\gamma_H \coloneqq \alpha + \xi < 1$.

A.2 PROOF OF THEOREM 4.4

Based on the established semi-norm contraction property of the Bellman operator $\mathbf{H}$ in Theorem 4.3, we can formulate Algorithm 1 as a stochastic approximation framework as follows:

$$Q_{t+1} = Q_t + \eta_t \big[\hat{\mathbf{H}}(Q_t) - Q_t\big], \quad \text{where} \quad \hat{\mathbf{H}}(Q_t) = \mathbf{H}(Q_t) + w^t. \tag{77}$$

where $w^t \in \mathbb{R}^{SA}$ is the support-estimation noise with

$$w^t(s,a) = \hat{\sigma}_{\mathcal{P}_s^a}(V_{Q_t}) - \sigma_{\mathcal{P}_s^a}(V_{Q_t}), \qquad (s,a) \in \mathcal{S} \times \mathcal{A}.$$

The second moment of this noise need not be uniformly bounded over arbitrary iterates Q_t. Instead, Lemma C.4 and norm equivalence on the quotient space imply a linear-growth bound: there exist constants $A_0, B_0 < \infty$ and a bias level $\varepsilon_{\text{bias}}$ such that

$$\mathbb{E}\Big[\|w^t\|_{\mathcal{H},\overline{E}}^2 \mid \mathcal{F}^t\Big] \le A_0 + B_0 \|Q_t - Q^*\|_{\mathcal{H},\overline{E}}^2, \qquad \big\|\mathbb{E}[w^t \mid \mathcal{F}^t]\big\|_{\mathcal{H},\overline{E}} \le \varepsilon_{\text{bias}} \Big(1 + \|Q_t - Q^*\|_{\mathcal{H},\overline{E}}\Big). \tag{78}$$

For contamination uncertainty, $\varepsilon_{\text{bias}} = 0$. For TV and Wasserstein uncertainty, $\varepsilon_{\text{bias}} = O(2^{-N_{\max}/2})$.

The remaining is to analyze the stochastic approximation of the above setting in (77)-(78). We first express the semi-norm $\|\cdot\|_H$ as infimal convolution between the norm $\|\cdot\|_{\mathcal{H}}$ and the indicator function of constant vectors. For any norm $\|\cdot\|_c$ and the constant vector class $\overline{E} \coloneqq \{c\mathbf{e} : c \in \mathbb{R}\}$, define the indicator function $\delta_{\overline{E}}$ as

$$\delta_{\bar{E}}(Q) := \begin{cases} 0 & Q \in \bar{E} \subseteq \mathbb{R}^{SA}, \\ \infty & \text{otherwise.} \end{cases} \tag{79}$$

then by Zhang et al. [2021], the semi-norm induced by norm $\|\cdot\|_c$ and $\overline{E}$ can be expressed as the infimal convolution of $\|\cdot\|_c$ and the indicator function $\delta_{\overline{E}}$, defined as follows:

$$\|Q\|_{c,\overline{E}} \coloneqq (\|\cdot\|_c *_{\inf} \delta_{\overline{E}})(Q) = \inf_{Q'}(\|Q - Q'\|_c + \delta_{\overline{E}}(Q')) = \inf_{e \in \overline{E}} \|Q - e\|_c \quad \forall Q \in \mathbb{R}^{SA}. \tag{80}$$

Where $*_{\inf}$ denotes the infimal convolution operator. Since $\|\cdot\|_H$ constructed in (67) is a semi-norm with kernel being $\overline{E}$, we can construct a norm $\|\cdot\|_{\mathcal{H}}$ via reverse engineering via Lemma C.6 such that

$$\|Q\|_{\mathcal{H},\overline{E}} \coloneqq (\|\cdot\|_{\mathcal{H}} *_{\inf} \delta_{\overline{E}})(Q) = \|Q\|_H \tag{81}$$

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

$$\|\mathbf{H}(Q_1) - \mathbf{H}(Q_2)\|_{\mathcal{H},\overline{E}} \leq \gamma_H \|Q_1 - Q_2\|_{\mathcal{H},\overline{E}}, \quad \gamma_H \in (0,1), \quad \forall Q_1, Q_2 \in \mathbb{R}^{SA} \tag{82}$$

Under the above framework, we use a linear-growth noise bound rather than a uniformly bounded-noise assumption. Denote $D_t \coloneqq \|Q_t - Q^*\|_{\mathcal{H},\overline{E}}$, since both Q_t and Q^* are anchored at (s_0, a_0), Lemma C.7 implies

$$\|Q_t - Q^*\|_\infty \leq \frac{1}{c_H} \|Q_t - Q^*\|_{\mathcal{H},\overline{E}}.$$

Moreover, the max operator is 1-Lipschitz in $\|\cdot\|_\infty$, so

$$\|V_{Q_t} - V_{Q^*}\|_{\mathrm{sp}} \leq 2\|V_{Q_t} - V_{Q^*}\|_\infty \leq 2\|Q_t - Q^*\|_\infty \leq \frac{2}{c_H} D_t.$$

Let $B_\star \coloneqq \|V_{Q^*}\|_{\mathrm{sp}}$, by Lemma 4.1, V_{Q^*} is an average-reward relative value function up to an additive constant for some optimal greedy policy and some worst-case kernel in $\mathcal{P}$. Therefore, by Lemma C.2, we have

$$B_\star \leq 4 R_{\mathrm{sp}} t_{\mathrm{mix}}.$$

Under the reward-span normalization $R_{\mathrm{sp}} \leq 1$, this gives $B_\star \leq 4 t_{\mathrm{mix}}$. Then

$$\|V_{Q_t}\|_{\mathrm{sp}} \leq B_\star + \frac{2}{c_H} D_t. \tag{83}$$

Thus Lemma C.4, together with norm equivalence, implies that for each uncertainty model $U \in \{\mathrm{Cont}, \mathrm{TV}, \mathrm{Wass}\}$ there exist constants $A_0^U, B_0^U < \infty$ and a truncation-bias level ε_N^U such that

$$\mathbb{E}\Big[\|w^t\|_{\mathcal{H},\overline{E}}^2 \,\Big|\, \mathcal{F}^t\Big] \leq A_0^U + B_0^U D_t^2, \tag{84}$$

and

$$\big\|\mathbb{E}[w^t \mid \mathcal{F}^t]\big\|_{\mathcal{H},\overline{E}} \leq \varepsilon_N^U (1 + D_t). \tag{85}$$

Here

$$\varepsilon_N^{\mathrm{Cont}} = 0, \qquad \varepsilon_N^{\mathrm{TV}} = O(2^{-N_{\max}/2}), \qquad \varepsilon_N^{\mathrm{Wass}} = O(2^{-N_{\max}/2}),$$

where the hidden constants depend on the uncertainty radius, $B_\star$, and the norm-equivalence constants, but not on t. Moreover,

$$A_0^{\mathrm{Cont}} = O(B_\star^2), \qquad B_0^{\mathrm{Cont}} = O(1),$$

and

$$A_0^{\mathrm{TV}} = O(B_\star^2 N_{\max}), \qquad B_0^{\mathrm{TV}} = O(N_{\max}),$$

with the same orders for Wasserstein uncertainty:

$$A_0^{\mathrm{Wass}} = O(B_\star^2 N_{\max}), \qquad B_0^{\mathrm{Wass}} = O(N_{\max}).$$

A.2.1 Construction of Semi-Lyapunov $M_{\overline{E}}(\cdot)$, Smoothness, and Choice of the Moreau Parameter.

We make explicit the smoothing parameter used in the generalized Moreau envelope. For each $\mu > 0$, let $M_\mu(\cdot)$ be the Moreau envelope function in Definition 2.2 of Chen et al. [2020], applied with smoothing parameter μ. Following [Zhang et al., 2021, Propositions 1–2], we quotient out the constant-vector class $\overline{E}$ by defining

$$M_{\overline{E},\mu}(Q) = \big(M_\mu *_{\inf} \delta_{\overline{E}}\big)(Q). \tag{86}$$

For each $\mu > 0$, there exist constants $c_l(\mu), c_u(\mu) > 0$ such that

$$c_l(\mu) M_{\overline{E},\mu}(Q) \leq \frac{1}{2}\|Q\|_{\mathcal{H},\overline{E}}^2 \leq c_u(\mu) M_{\overline{E},\mu}(Q), \quad \forall Q \in \mathbb{R}^{SA}. \tag{87}$$

Moreover, $M_{\overline{E},\mu}$ is L_μ-smooth w.r.t. another semi-norm $\|\cdot\|_{s,\overline{E},\mu}$, and the Moreau approximation ratio satisfies

$$\lim_{\mu\downarrow 0}\frac{c_u(\mu)}{c_l(\mu)}=1. \tag{88}$$

Since Theorem 4.3 gives $\gamma_H<1$, we choose $\mu_H>0$ sufficiently small so that

$$\gamma_H\sqrt{\frac{c_u(\mu_H)}{c_l(\mu_H)}}\le\frac{1+\gamma_H}{2}<1. \tag{89}$$

This is only a choice of the smooth Lyapunov function and imposes no additional assumption on the MDP or on the uncertainty set. In the rest of Appendix A.2.2, we fix this μ_H and write $M_{\overline{E}}$, c_l, c_u, L, and $\|\cdot\|_{s,\overline{E}}$ for $M_{\overline{E},\mu_H}$, $c_l(\mu_H)$, $c_u(\mu_H)$, L_{μ_H}, and $\|\cdot\|_{s,\overline{E},\mu_H}$, respectively. Thus

$$c_l M_{\overline{E}}(Q)\le\frac{1}{2}\|Q\|^2_{\mathcal{H},\overline{E}}\le c_u M_{\overline{E}}(Q),\quad\forall Q\in\mathbb{R}^{SA}. \tag{90}$$

Concretely, L-smoothness means

$$M_{\overline{E}}(Q_2)\le M_{\overline{E}}(Q_1)+\langle\nabla M_{\overline{E}}(Q_1),Q_2-Q_1\rangle+\frac{L}{2}\|Q_2-Q_1\|^2_{s,\overline{E}},\quad\forall Q_1,Q_2\in\mathbb{R}^{SA}. \tag{91}$$

Furthermore, the gradient of $M_{\overline{E}}$ satisfies $\langle\nabla M_{\overline{E}}(Q),c\mathbf{e}\rangle=0$ for all Q and all $c\in\mathbb{R}$, and

$$\|\nabla M_{\overline{E}}(Q_1)-\nabla M_{\overline{E}}(Q_2)\|_{*,s,\overline{E}}\le L\|Q_2-Q_1\|_{s,\overline{E}},\quad\forall Q_1,Q_2\in\mathbb{R}^{SA}. \tag{92}$$

Define

$$\vartheta_H\coloneqq\gamma_H\sqrt{c_u/c_l},\qquad\alpha_2\coloneqq\frac{1-\vartheta_H}{2}. \tag{93}$$

By (89),

$$\alpha_2\ge\frac{1-\gamma_H}{4}>0. \tag{94}$$

This is the positive drift parameter used in the stepsize below.

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

$$\rho_1\|Q\|_{\mathcal{H},\overline{E}}\le\|Q\|_{s,\overline{E}}\le\rho_2\|Q\|_{\mathcal{H},\overline{E}},\quad\forall Q\in\mathbb{R}^{SA}. \tag{95}$$

Likewise, their dual norms, denoted $\|\cdot\|_{*,s,\overline{E}}$ and $\|\cdot\|_{*,\mathcal{H},\overline{E}}$, satisfy

$$\frac{1}{\rho_2}\|Q\|_{*,s,\overline{E}}\le\|Q\|_{*,\mathcal{H},\overline{E}}\le\frac{1}{\rho_1}\|Q\|_{*,s,\overline{E}},\quad\forall Q\in\mathbb{R}^{SA}. \tag{96}$$

A.2.2 Formal Derivation of Theorem 4.4

By L-smoothness w.r.t. $\|\cdot\|_{s,\overline{E}}$ in (91), for each t,

$$M_{\overline{E}}(Q_{t+1}-Q^*)\le M_{\overline{E}}(Q_t-Q^*)+\left\langle\nabla M_{\overline{E}}(Q_t-Q^*),Q_{t+1}-Q_t\right\rangle+\tfrac{L}{2}\|Q_{t+1}-Q_t\|^2_{s,\overline{E}} \tag{97}$$

where $Q_{t+1}-Q_t=\eta_t\big[\hat{\mathbf{H}}(Q_t)-Q_t\big]=\eta_t[\mathbf{H}(Q_t)+w^t-Q_t]$. Taking expectation of the second term of the RHS of (97) conditioned on the filtration $\mathcal{F}^t$ we obtain,

$$\begin{aligned}
&\mathbb{E}[\langle\nabla M_{\overline{E}}(Q_t-Q^*),Q_{t+1}-Q_t\rangle|\mathcal{F}^t]\\
&\quad=\eta_t\mathbb{E}[\langle\nabla M_{\overline{E}}(Q_t-Q^*),\mathbf{H}(Q_t)-Q_t+w^t\rangle|\mathcal{F}^t]\\
&\quad=\eta_t\langle\nabla M_{\overline{E}}(Q_t-Q^*),\mathbf{H}(Q_t)-Q_t\rangle+\eta_t\mathbb{E}[\langle\nabla M_{\overline{E}}(Q_t-Q^*),w^t\rangle|\mathcal{F}^t]\\
&\quad=\eta_t\langle\nabla M_{\overline{E}}(Q_t-Q^*),\mathbf{H}(Q_t)-Q_t\rangle+\eta_t\langle\nabla M_{\overline{E}}(Q_t-Q^*),\mathbb{E}[w^t|\mathcal{F}^t]\rangle.
\end{aligned} \tag{98}$$

To analyze the additional bias term $\langle \nabla M_{\overline{E}}(Q_t - Q^*), \mathbb{E}[w^t|\mathcal{F}^t]\rangle$, we use the Holder's inequality. Specifically, 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

$$\big\langle u,\, v\big\rangle \le \|u\|_*\, \|v\|, \quad \forall\, u, v. \tag{99}$$

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

$$\big\langle \nabla M_{\overline{E}}(Q_t - Q^*), \mathbb{E}[w^t|\mathcal{F}^t]\big\rangle \le \big\|\nabla M_{\overline{E}}(Q_t - Q^*)\big\|_{*,\,\mathcal{H},\overline{E}} \cdot \big\|\mathbb{E}[w^t|\mathcal{F}^t]\big\|_{\mathcal{H},\overline{E}}. \tag{100}$$

Using the linear-growth bias bound (85), it remains to bound $\|\nabla M_{\overline{E}}(Q_t - Q^*)\|_{*,\mathcal{H},\overline{E}}$. By setting $Q_2 = 0$ in (92), we get

$$\|\nabla M_{\overline{E}}(Q) - \nabla M_{\overline{E}}(0)\|_{*,s,\overline{E}} \le L\|Q\|_{s,\overline{E}}, \quad \forall\, Q \in \mathbb{R}^{SA}. \tag{101}$$

Thus,

$$\|\nabla M_{\overline{E}}(Q)\|_{*,s,\overline{E}} \le \|\nabla M_{\overline{E}}(0)\|_{*,s,\overline{E}} + L\|Q\|_{s,\overline{E}}, \quad \forall\, Q \in \mathbb{R}^{SA}. \tag{102}$$

By (96), we know that there exists $\frac{1}{\rho_2} \le \alpha \le \frac{1}{\rho_1}$ such that

$$\|\nabla M_{\overline{E}}(Q)\|_{*,\mathcal{H},\overline{E}} \le \alpha\|\nabla M_{\overline{E}}(Q)\|_{*,s,\overline{E}} \tag{103}$$

Thus, combining (102) and (103) would give:

$$\|\nabla M_{\overline{E}}(Q)\|_{*,\mathcal{H},\overline{E}} \le \alpha\big(\|\nabla M_{\overline{E}}(0)\|_{*,s,\overline{E}} + L\|Q\|_{s,\overline{E}}\big), \quad \forall\, Q \in \mathbb{R}^{SA}. \tag{104}$$

By (95), $\|Q\|_{s,\overline{E}} \le \rho_2\|Q\|_{\mathcal{H},\overline{E}}$, thus we have:

$$\|\nabla M_{\overline{E}}(Q)\|_{*,\mathcal{H},\overline{E}} \le \alpha\big(\|\nabla M_{\overline{E}}(0)\|_{*,s,\overline{E}} + L\rho_2\|Q\|_{\mathcal{H},\overline{E}}\big), \quad \forall\, Q \in \mathbb{R}^{SA}. \tag{105}$$

Hence, combining the above with (100), there exist some

$$G := \frac{1}{\rho_1}\max\Big\{L\rho_2,\ \|\nabla M_{\overline{E}}(0)\|_{*,s,\overline{E}}\Big\} = \mathcal{O}\big(\frac{1}{\rho_1}\max\{L\rho_2, \|\nabla M_{\overline{E}}(0)\|_{*,s,\overline{E}}\}\big) \tag{106}$$

such that

$$\mathbb{E}\Big[\langle \nabla M_{\overline{E}}(Q_t - Q^*),\, w^t\rangle|\mathcal{F}^t\Big] = \big\langle \nabla M_{\overline{E}}(Q_t - Q^*), \mathbb{E}[w^t|\mathcal{F}^t]\big\rangle. \tag{107}$$

Using (85), we obtain

$$\begin{aligned}\mathbb{E}\Big[\langle \nabla M_{\overline{E}}(Q_t - Q^*),\, w^t\rangle \,|\, \mathcal{F}^t\Big] &= \big\langle \nabla M_{\overline{E}}(Q_t - Q^*), \mathbb{E}[w^t|\mathcal{F}^t]\big\rangle \\ &\le G\,\varepsilon_N^U\big(1 + D_t\big)^2.\end{aligned} \tag{108}$$

By (90), $D_t^2 \le 2c_u M_{\overline{E}}(Q_t - Q^*)$. Hence there exists a constant $C_{\text{drift}} < \infty$ such that

$$\mathbb{E}\Big[\langle \nabla M_{\overline{E}}(Q_t - Q^*),\, w^t\rangle \,|\, \mathcal{F}^t\Big] \le C_{\text{drift}}\varepsilon_N^U\left(1 + M_{\overline{E}}(Q_t - Q^*)\right). \tag{109}$$

Combining (109) with (98) we obtain

$$\begin{aligned}\mathbb{E}\big[\langle \nabla M_{\overline{E}}(Q_t - Q^*), Q_{t+1} - Q_t\rangle \,\big|\, \mathcal{F}^t\big] &\le \eta_t\,\langle \nabla M_{\overline{E}}(Q_t - Q^*), \mathbf{H}(Q_t) - Q_t\rangle \\ &\quad + \eta_t C_{\text{drift}}\varepsilon_N^U\left(1 + M_{\overline{E}}(Q_t - Q^*)\right).\end{aligned} \tag{110}$$

To bound the first term in the RHS of (110), set $X_t := Q_t - Q^*$ and $Y_t := \mathbf{H}(Q_t) - Q^*$. Since Q^* solves the optimal Bellman equation up to an element of $\overline{E}$, we have $\mathbf{H}(Q^*) - Q^* \in \overline{E}$. Hence, because $\|\cdot\|_{\mathcal{H},\overline{E}}$ quotients out $\overline{E}$, Theorem 4.3 gives

$$\|Y_t\|_{\mathcal{H},\overline{E}} = \|\mathbf{H}(Q_t) - \mathbf{H}(Q^*)\|_{\mathcal{H},\overline{E}} \le \gamma_H\|X_t\|_{\mathcal{H},\overline{E}}. \tag{111}$$

By convexity of $M_{\overline{E}}$ and by (90),

$$\begin{aligned}\langle \nabla M_{\overline{E}}(X_t), \mathbf{H}(Q_t) - Q_t \rangle &= \langle \nabla M_{\overline{E}}(X_t), Y_t - X_t \rangle \\ &\le M_{\overline{E}}(Y_t) - M_{\overline{E}}(X_t) \\ &\le \frac{1}{2c_l}\|Y_t\|^2_{\mathcal{H},\overline{E}} - M_{\overline{E}}(X_t) \\ &\le \frac{\gamma_H^2}{2c_l}\|X_t\|^2_{\mathcal{H},\overline{E}} - M_{\overline{E}}(X_t) \\ &\le \left(\gamma_H^2 \frac{c_u}{c_l} - 1\right) M_{\overline{E}}(X_t) \\ &= -(1-\vartheta_H)(1+\vartheta_H) M_{\overline{E}}(X_t) \\ &\le -2\alpha_2 M_{\overline{E}}(X_t). \end{aligned} \tag{112}$$

The last inequality uses $\vartheta_H < 1$ and the definition $2\alpha_2 = 1 - \vartheta_H$ in (93). Combining (112), (110), and Lemma C.3 with (97), we arrive as follows:

$$\begin{aligned}\mathbb{E}\big[M_{\overline{E}}(Q_{t+1} - Q^*) \,\big|\, \mathcal{F}^t\big] \le \big(1 - 2\alpha_2\eta_t + \alpha_3^U \eta_t^2 + C_{\mathrm{drift}}\varepsilon_N^U \eta_t\big) M_{\overline{E}}(Q_t - Q^*) \\ + A_0^U \alpha_4 \eta_t^2 + C_{\mathrm{drift}}\varepsilon_N^U \eta_t. \end{aligned} \tag{113}$$

Here α_2 is the positive constant from (93), and

$$\ell_s \coloneqq \rho_2^2, \qquad \alpha_3^U \coloneqq (8 + 2B_0^U) c_u \ell_s L, \qquad \alpha_4 \coloneqq \ell_s L. \tag{114}$$

Choose

$$\eta_t \coloneqq \frac{2}{\alpha_2(t + K_U)}, \qquad K_U \coloneqq \max\left\{\frac{4\alpha_3^U}{\alpha_2^2}, 3\right\}. \tag{115}$$

Then $\alpha_3^U \eta_t^2 \le \frac{1}{t+K_U} = \frac{\alpha_2}{2}\eta_t$. For TV and Wasserstein uncertainty, choose $N_{\max}$ large enough so that $C_{\mathrm{drift}}\varepsilon_N^U \le \frac{\alpha_2}{2}$. This condition is used only for the drift inequality. For a target accuracy ϵ, we also choose $N_{\max}$ so that the final bias term in (120) is at most a constant multiple of ϵ^2. For contamination uncertainty, both conditions are automatic because $\varepsilon_N^{\mathrm{Cont}} = 0$. Therefore $C_{\mathrm{drift}}\varepsilon_N^U \eta_t \le \frac{1}{t+K_U}$, and (113) implies

$$\mathbb{E}\big[M_{\overline{E}}(Q_{t+1} - Q^*) \,\big|\, \mathcal{F}^t\big] \le \left(1 - \frac{2}{t + K_U}\right) M_{\overline{E}}(Q_t - Q^*) + A_0^U \alpha_4 \eta_t^2 + C_{\mathrm{drift}}\varepsilon_N^U \eta_t. \tag{116}$$

Taking total expectation in (116) and unrolling the resulting recursion gives

$$\mathbb{E}[M_{\overline{E}}(Q_T - Q^*)] \le \Gamma_T M_{\overline{E}}(Q_0 - Q^*) + \Gamma_T \sum_{t=0}^{T-1} \frac{A_0^U \alpha_4 \eta_t^2 + C_{\mathrm{drift}}\varepsilon_N^U \eta_t}{\Gamma_{t+1}}, \tag{117}$$

where $\Gamma_t \coloneqq \prod_{i=0}^{t-1}\left(1 - \frac{2}{i+K_U}\right)$. Since $K_U \ge 3$,

$$\Gamma_t = \frac{(K_U - 2)(K_U - 1)}{(t + K_U - 2)(t + K_U - 1)}. \tag{118}$$

Using $\eta_t = \frac{2}{\alpha_2(t+K_U)}$, a direct calculation gives

$$\Gamma_T \sum_{t=0}^{T-1} \frac{\eta_t^2}{\Gamma_{t+1}} \le \frac{4}{\alpha_2^2(T + K_U - 1)}, \qquad \Gamma_T \sum_{t=0}^{T-1} \frac{\eta_t}{\Gamma_{t+1}} \le \frac{2}{\alpha_2}. \tag{119}$$

Combining (117)–(119) with (90) yields

$$\mathbb{E}\Big[\|Q_T - Q^*\|^2_{\mathcal{H},\overline{E}}\Big] \le \frac{K_U^2 c_u}{(T + K_U - 2)^2 c_l}\|Q_0 - Q^*\|^2_{\mathcal{H},\overline{E}} + \frac{8A_0^U \alpha_4 c_u}{(T + K_U - 1)\alpha_2^2} + \frac{4c_u C_{\mathrm{drift}}\varepsilon_N^U}{\alpha_2}. \tag{120}$$

Theorem A.8 (Formal version of Theorem 4.4)**.** *Let Q_t be generated by Algorithm 1, and let Q^* be the anchored robust optimal Q-function satisfying* (19) *with $Q^*(s_0, a_0) = 0$. Suppose Assumptions 4.2 and A.1 hold, and suppose the radius conditions of Lemmas A.3–A.5 hold.*

For each uncertainty model $U \in \{\mathrm{Cont}, \mathrm{TV}, \mathrm{Wass}\}$, let A_0^U, B_0^U, and ε_N^U be the constants in (84) *and* (85)*. Choose the Moreau parameter as in* (89)*, and define ϑ_H and α_2 as in* (93)*. In particular, $\alpha_2 \geq \frac{1-\gamma_H}{4} > 0$. Define*

$$\ell_s \coloneqq \rho_2^2, \qquad \alpha_3^U \coloneqq (8 + 2B_0^U)c_u\ell_s L, \qquad \alpha_4 \coloneqq \ell_s L, \tag{121}$$

and

$$K_U \coloneqq \max\left\{\frac{4\alpha_3^U}{\alpha_2^2}, 3\right\}. \tag{122}$$

We choose $\eta_t = \frac{2}{\alpha_2(t+K_U)}$. For TV and Wasserstein uncertainty, first choose $N_{\max}$ large enough that $C_{\mathrm{drift}}\varepsilon_N^U \leq \frac{\alpha_2}{2}$. For a target accuracy ϵ, choose it large enough also that $\frac{4c_uC_{\mathrm{drift}}\varepsilon_N^U}{\alpha_2 c_H^2} \leq \epsilon^2/3$. For contamination uncertainty these conditions are automatic because $\varepsilon_N^{\mathrm{Cont}} = 0$. Then we have

$$\begin{aligned}\mathbb{E}\big[\|Q_T - Q^*\|_\infty^2\big] \leq \frac{K_U^2 c_u C_H^2}{(T + K_U - 2)^2 c_l c_H^2}\|Q_0 - Q^*\|_\infty^2 + \frac{8A_0^U\alpha_4 c_u}{(T + K_U - 1)\alpha_2^2 c_H^2} \\ + \frac{4c_uC_{\mathrm{drift}}\varepsilon_N^U}{\alpha_2 c_H^2}.\end{aligned} \tag{123}$$

Consequently, choosing $T = \Theta(\epsilon^{-2})$ for contamination uncertainty and choosing $T = \widetilde{\Theta}(\epsilon^{-2})$ with $N_{\max} = \Theta(\log(1/\epsilon))$ for TV and Wasserstein uncertainty yields $\mathbb{E}\big[\|Q_T - Q^*\|_\infty^2\big] \leq \epsilon^2$ after increasing the hidden constants if necessary.

Proof. The bound follows from (120) and the norm-translation Lemma C.7. The stated choices of T and $N_{\max}$ use $\varepsilon_N^{\mathrm{Cont}} = 0$ and $\varepsilon_N^{\mathrm{TV}}, \varepsilon_N^{\mathrm{Wass}} = O(2^{-N_{\max}/2})$. □

A.2.3 Sample Complexities for robust Q-Learning

To derive the sample complexities, we use the above results and set $\mathbb{E}\Big[\|Q_T - Q^*\|_\infty^2\Big] \leq \epsilon^2$. Let C_Q^U denote the finite problem-dependent constant obtained from (123) after absorbing $\alpha_2, \alpha_4, c_l, c_u, c_H, C_H, K_U, C_{\mathrm{drift}}$ and the fixed initialization. This constant may depend on the contraction and Moreau-smoothing construction, but it is independent of ϵ, T, $N_{\max}$, and the realized policy sequence.

For contamination uncertainty, $\varepsilon_N^{\mathrm{Cont}} = 0$ and $A_0^{\mathrm{Cont}} = \mathcal{O}(t_{\mathrm{mix}}^2)$. Hence it is enough to take $T = \mathcal{O}(C_Q^{\mathrm{Cont}} t_{\mathrm{mix}}^2 \epsilon^{-2})$, resulting in $\mathcal{O}(SAC_Q^{\mathrm{Cont}} t_{\mathrm{mix}}^2 \epsilon^{-2})$ nominal transition samples. For TV and Wasserstein uncertainty, take $N_{\max} = \mathcal{O}(\log(1/\epsilon))$ large enough so that the bias term in (123) is at most ϵ^2, and use $A_0^U = \mathcal{O}(t_{\mathrm{mix}}^2 N_{\max})$. It is then enough to take $T = \mathcal{O}(C_Q^U t_{\mathrm{mix}}^2 N_{\max}\epsilon^{-2})$. Since one call of Algorithm 4 has expected cost $\mathcal{O}(N_{\max})$ by Lemma C.4, the resulting sample complexity is $\tilde{\mathcal{O}}(SAC_Q^U t_{\mathrm{mix}}^2 \epsilon^{-2})$ for $U \in \{\mathrm{TV}, \mathrm{Wass}\}$.

B MISSING PROOFS FOR SECTION 5

B.1 PROOF OF PROPOSITION 5.1

Let V^π be the anchored solution of the robust Bellman equation (7). For each (s, a), compactness of $\mathcal{P}_s^a$ and continuity of $p \mapsto p^\top V^\pi$ imply that the set $\mathcal{M}_{s,a}(V^\pi)$ in (9) is nonempty. By rectangularity, choose $\mathsf{P}_V^\pi \in \mathcal{P}$ satisfying (10) for all (s, a).

By the definition of P_V^π, for every (s, a),

$$\sigma_{\mathcal{P}_s^a}(V^\pi) = \big(\mathsf{P}_V^\pi(\cdot|s, a)\big)^\top V^\pi. \tag{124}$$

Substituting (124) into the robust Bellman equation gives, for every s,

$$V^\pi(s) = \sum_a \pi(a|s) \left(r(s,a) - g^\pi_{\mathcal{P}} + \left(\mathsf{P}^\pi_V(\cdot|s,a)\right)^\top V^\pi \right)$$
$$= \sum_a \pi(a|s) r(s,a) - g^\pi_{\mathcal{P}} + \sum_{s'} \mathsf{P}^{\pi,\pi}_V(s,s') V^\pi(s'), \tag{125}$$

where $\mathsf{P}^{\pi,\pi}_V$ is the state transition matrix induced by policy π and kernel P^π_V.

Under the radius restrictions used in this section, $\mathsf{P}^{\pi,\pi}_V$ is irreducible and aperiodic. Let $d^\pi_{\mathsf{P}^\pi_V}$ be its stationary distribution. Multiplying (125) by $d^\pi_{\mathsf{P}^\pi_V}$ and summing over s cancels the value terms, since $(d^\pi_{\mathsf{P}^\pi_V})^\top \mathsf{P}^{\pi,\pi}_V = (d^\pi_{\mathsf{P}^\pi_V})^\top$. Hence

$$g^\pi_{\mathcal{P}} = \sum_s d^\pi_{\mathsf{P}^\pi_V}(s) \sum_a \pi(a|s) r(s,a) = g^\pi_{\mathsf{P}^\pi_V}. \tag{126}$$

Thus $\mathsf{P}^\pi_V \in \Omega^\pi_g$.

Now define Q^π by (11). The identity (22) is exactly the first line of (11). Moreover, using the Bellman equation,

$$\sum_a \pi(a|s) Q^\pi(s,a) = \sum_a \pi(a|s) \left(r(s,a) - g^\pi_{\mathcal{P}} + \sigma_{\mathcal{P}^a_s}(V^\pi) \right)$$
$$= \mathbf{T}^\pi_g(V^\pi)(s) = V^\pi(s). \tag{127}$$

Finally, since V^π solves the Poisson equation (125) for policy π and kernel P^π_V, the usual one-step decomposition of the relative value gives

$$\mathbb{E}_{\pi,\mathsf{P}^\pi_V}\left[\sum_{t=0}^{\infty} (r_t - g^\pi_{\mathcal{P}}) \,\middle|\, S_0 = s, A_0 = a \right] = r(s,a) - g^\pi_{\mathcal{P}} + \sum_{s'} \mathsf{P}^\pi_V(s'|s,a) V^\pi(s')$$
$$= r(s,a) - g^\pi_{\mathcal{P}} + \sigma_{\mathcal{P}^a_s}(V^\pi) = Q^\pi(s,a), \tag{128}$$

where the second equality uses (124). This proves the proposition.

B.2 PROOF OF THEOREM 5.6

Regarding the detailed radius restrictions for the uncertainty sets, we use Corollary B.3. For ease of notation, we write $g^\pi_{\mathcal{P}}$ as g^π in this proof.

The robust performance-difference inequalities used below are the reward-side version of the inequalities in Sun et al. [2024]. Their paper states the result for robust average cost. Applying it to $c = -r$ changes the robust cost objective into $-g^\pi$ and changes the cost Q-function into $-Q^\pi$. Therefore, for any policies π and π', and for rowwise Bellman-minimizing worst-case kernels $P_\pi \in \Omega^\pi_g$ and $P_{\pi'} \in \Omega^{\pi'}_g$, we have

$$g^{\pi'} - g^\pi \geq \mathbb{E}_{s\sim d^{\pi'}_{P_{\pi'}}} \left[\langle Q^\pi(s,\cdot), \pi'(\cdot|s) - \pi(\cdot|s) \rangle \right], \tag{129}$$
$$g^{\pi^\star} - g^\pi \leq \mathbb{E}_{s\sim d^{\pi^\star}_{P_\pi}} \left[\langle Q^\pi(s,\cdot), \pi^\star(\cdot|s) - \pi(\cdot|s) \rangle \right]. \tag{130}$$

Recall that Algorithm 2 uses the update

$$\pi_{k+1}(\cdot|s) = \arg\max_{p\in\Delta(\mathcal{A})} \{ \zeta_k \langle \hat{Q}^{\pi_k}(s,\cdot), p \rangle - \|p - \pi_k(\cdot|s)\|^2 \}. \tag{131}$$

For a fixed state s, the objective in (131) is concave in p over the simplex. Hence the first-order optimality condition implies that for any reference distribution $p(\cdot|s) \in \Delta(\mathcal{A})$,

$$\zeta_k \left\langle \hat{Q}^{\pi_k}(s,\cdot), p(\cdot|s) - \pi_{k+1}(\cdot|s) \right\rangle + \|\pi_{k+1}(\cdot|s) - \pi_k(\cdot|s)\|^2$$
$$\leq \|p(\cdot|s) - \pi_k(\cdot|s)\|^2 - \|p(\cdot|s) - \pi_{k+1}(\cdot|s)\|^2. \tag{132}$$

Taking $p = \pi_k$ in (132) gives

$$\zeta_k \left\langle \hat{Q}^{\pi_k}(s,\cdot), \pi_k(\cdot|s) - \pi_{k+1}(\cdot|s) \right\rangle \leq -2 \left\| \pi_{k+1}(\cdot|s) - \pi_k(\cdot|s) \right\|^2 \leq 0. \tag{133}$$

Taking $p = \pi^\star$ in (132) gives

$$\begin{aligned} \left\langle \hat{Q}^{\pi_k}(s,\cdot), \pi^\star(\cdot|s) - \pi_{k+1}(\cdot|s) \right\rangle \leq \frac{1}{\zeta_k} \Big[& \left\| \pi^\star(\cdot|s) - \pi_k(\cdot|s) \right\|^2 \\ & - \left\| \pi^\star(\cdot|s) - \pi_{k+1}(\cdot|s) \right\|^2 \Big]. \end{aligned} \tag{134}$$

Applying (129) with $(\pi, \pi') = (\pi_k, \pi_{k+1})$ gives

$$\begin{aligned} g^{\pi_k} - g^{\pi_{k+1}} &\leq \mathbb{E}_{s \sim d^{\pi_{k+1}}_{P_{\pi_{k+1}}}} \left[\langle Q^{\pi_k}(s,\cdot), \pi_k(\cdot|s) - \pi_{k+1}(\cdot|s) \rangle \right] \\ &= \mathbb{E}_{s \sim d^{\pi_{k+1}}_{P_{\pi_{k+1}}}} \left[\left\langle \hat{Q}^{\pi_k}(s,\cdot), \pi_k(\cdot|s) - \pi_{k+1}(\cdot|s) \right\rangle \right] \\ &\quad + \mathbb{E}_{s \sim d^{\pi_{k+1}}_{P_{\pi_{k+1}}}} \left[\left\langle Q^{\pi_k}(s,\cdot) - \hat{Q}^{\pi_k}(s,\cdot), \pi_k(\cdot|s) - \pi_{k+1}(\cdot|s) \right\rangle \right]. \end{aligned} \tag{135}$$

The first inner product on the right side of (135) is nonpositive for every state by (133). From the definition of M in (30),

$$\frac{d^{\pi^\star}_{P_{\pi_k}}(s)}{d^{\pi_{k+1}}_{P_{\pi_{k+1}}}(s)} \leq M, \qquad \forall s \in \mathcal{S}. \tag{136}$$

Since multiplying a nonpositive quantity by a larger nonnegative weight makes it smaller, we have

$$\begin{aligned} g^{\pi_k} - g^{\pi_{k+1}} \leq \frac{1}{M} & \mathbb{E}_{s \sim d^{\pi^\star}_{P_{\pi_k}}} \left[\left\langle \hat{Q}^{\pi_k}(s,\cdot), \pi_k(\cdot|s) - \pi_{k+1}(\cdot|s) \right\rangle \right] \\ & + \mathbb{E}_{s \sim d^{\pi_{k+1}}_{P_{\pi_{k+1}}}} \left[\left\langle Q^{\pi_k}(s,\cdot) - \hat{Q}^{\pi_k}(s,\cdot), \pi_k(\cdot|s) - \pi_{k+1}(\cdot|s) \right\rangle \right]. \end{aligned} \tag{137}$$

Applying (130) with $\pi = \pi_k$ gives

$$\begin{aligned} g^{\pi^\star} - g^{\pi_k} \leq & \mathbb{E}_{s \sim d^{\pi^\star}_{P_{\pi_k}}} \left[\left\langle \hat{Q}^{\pi_k}(s,\cdot), \pi^\star(\cdot|s) - \pi_k(\cdot|s) \right\rangle \right] \\ & + \mathbb{E}_{s \sim d^{\pi^\star}_{P_{\pi_k}}} \left[\left\langle Q^{\pi_k}(s,\cdot) - \hat{Q}^{\pi_k}(s,\cdot), \pi^\star(\cdot|s) - \pi_k(\cdot|s) \right\rangle \right]. \end{aligned} \tag{138}$$

Multiplying (137) by M and adding (138), we obtain

$$\begin{aligned} & \left(g^{\pi^\star} - g^{\pi_k} \right) + M \left(g^{\pi_k} - g^{\pi_{k+1}} \right) \\ & \leq \mathbb{E}_{s \sim d^{\pi^\star}_{P_{\pi_k}}} \left[\left\langle \hat{Q}^{\pi_k}(s,\cdot), \pi^\star(\cdot|s) - \pi_{k+1}(\cdot|s) \right\rangle \right] \\ & \quad + \mathbb{E}_{s \sim d^{\pi^\star}_{P_{\pi_k}}} \left[\left\langle Q^{\pi_k}(s,\cdot) - \hat{Q}^{\pi_k}(s,\cdot), \pi^\star(\cdot|s) - \pi_k(\cdot|s) \right\rangle \right] \\ & \quad + M \mathbb{E}_{s \sim d^{\pi_{k+1}}_{P_{\pi_{k+1}}}} \left[\left\langle Q^{\pi_k}(s,\cdot) - \hat{Q}^{\pi_k}(s,\cdot), \pi_k(\cdot|s) - \pi_{k+1}(\cdot|s) \right\rangle \right]. \end{aligned} \tag{139}$$

For any two distributions $u, v \in \Delta(\mathcal{A})$, $\|u - v\|_1 \leq 2$. Therefore Hölder's inequality gives

$$\left| \left\langle Q^{\pi_k}(s,\cdot) - \hat{Q}^{\pi_k}(s,\cdot), u - v \right\rangle \right| \leq 2 \| Q^{\pi_k} - \hat{Q}^{\pi_k} \|_\infty. \tag{140}$$

Using this bound in the two error terms of (139), and then using (134), yields

$$\begin{aligned} & \left(g^{\pi^\star} - g^{\pi_k} \right) + M \left(g^{\pi_k} - g^{\pi_{k+1}} \right) \\ & \leq \frac{1}{\zeta_k} \mathbb{E}_{s \sim d^{\pi^\star}_{P_{\pi_k}}} \left[\left\| \pi^\star(\cdot|s) - \pi_k(\cdot|s) \right\|^2 - \left\| \pi^\star(\cdot|s) - \pi_{k+1}(\cdot|s) \right\|^2 \right] \\ & \quad + 2(M+1) \| Q^{\pi_k} - \hat{Q}^{\pi_k} \|_\infty. \end{aligned} \tag{141}$$

The left side can be rewritten as

$$\left(g^{\pi^\star} - g^{\pi_k}\right) + M\left(g^{\pi_k} - g^{\pi_{k+1}}\right) = M\left(g^{\pi^\star} - g^{\pi_{k+1}}\right) - (M-1)\left(g^{\pi^\star} - g^{\pi_k}\right). \tag{142}$$

Combining (141) and (142) gives the samplewise recursion

$$\begin{aligned} g^{\pi^\star} - g^{\pi_{k+1}} &\le \left(1 - \frac{1}{M}\right)\left(g^{\pi^\star} - g^{\pi_k}\right) + \frac{1}{M\zeta_k}\mathbb{E}_{s\sim d^{\pi^\star}_{P^{\pi_k}}}\Big[\|\pi^\star(\cdot|s) - \pi_k(\cdot|s)\|^2 \\ &\quad - \|\pi^\star(\cdot|s) - \pi_{k+1}(\cdot|s)\|^2\Big] + 2\left(1 + \frac{1}{M}\right)\|Q^{\pi_k} - \hat{Q}^{\pi_k}\|_\infty \\ &\le \left(1 - \frac{1}{M}\right)\left(g^{\pi^\star} - g^{\pi_k}\right) + \frac{2}{M\zeta_k} + 2\left(1 + \frac{1}{M}\right)\|Q^{\pi_k} - \hat{Q}^{\pi_k}\|_\infty. \end{aligned} \tag{143}$$

The last inequality uses $\|u - v\|^2 \le 2$ for probability vectors under the Euclidean norm.

Let $\mathcal{F}_k$ be the history before the k-th critic call. The policy π_k is $\mathcal{F}_k$-measurable, and Algorithm 3 uses fresh nominal simulator samples. By the assumed conditional critic guarantee,

$$\mathbb{E}\left[\|Q^{\pi_k} - \hat{Q}^{\pi_k}\|_\infty \mid \mathcal{F}_k\right] \le \epsilon. \tag{144}$$

Taking conditional expectation in (143) and then total expectation gives

$$\mathbb{E}[g^{\pi^\star} - g^{\pi_{k+1}}] \le \left(1 - \frac{1}{M}\right)\mathbb{E}[g^{\pi^\star} - g^{\pi_k}] + \frac{2}{M\zeta_k} + 2\left(1 + \frac{1}{M}\right)\epsilon. \tag{145}$$

Unrolling (145) gives

$$\begin{aligned} \mathbb{E}[g^{\pi^\star} - g^{\pi_K}] &\le \left(1 - \frac{1}{M}\right)^K \mathbb{E}[g^{\pi^\star} - g^{\pi_0}] \\ &\quad + \sum_{k=0}^{K-1}\left(1 - \frac{1}{M}\right)^{K-1-k}\left[\frac{2}{M\zeta_k} + 2\left(1 + \frac{1}{M}\right)\epsilon\right]. \end{aligned} \tag{146}$$

Since $\zeta_k \ge \zeta_{k-1}(1 - 1/M)^{-1}$, we have $\zeta_k \ge \zeta_0(1 - 1/M)^{-k}$ and hence $1/\zeta_k \le (1/\zeta_0)(1 - 1/M)^k$. Substituting this into (146) yields

$$\mathbb{E}[g^{\pi^\star} - g^{\pi_K}] \le \left(1 - \frac{1}{M}\right)^K \mathbb{E}[g^{\pi^\star} - g^{\pi_0}] + \frac{2K}{M\zeta_0}\left(1 - \frac{1}{M}\right)^{K-1} + 2(M+1)\epsilon. \tag{147}$$

Choosing the hidden constant in $K = \Theta(\log(1/\epsilon))$ large enough makes the first two terms in (147) at most a constant multiple of ϵ. Therefore $\mathbb{E}[g^{\pi^\star} - g^{\pi_K}] \le \mathcal{O}(\epsilon)$.

Remark B.1. We note that the geometrically increasing actor parameter in Theorem 5.6 is not essential for the stated sample complexity. Starting from the one-step recursion (145), one may instead choose a fixed actor parameter $\zeta_k \equiv \zeta$. If $\mathbb{E}[\|\hat{Q}^{\pi_k} - Q^{\pi_k}\|_\infty \mid \mathcal{F}_k] \le \delta$ for every generated policy, then unrolling the recursion gives

$$\mathbb{E}[g^{\pi^\star} - g^{\pi_K}] \le \left(1 - \frac{1}{M}\right)^K \mathbb{E}[g^{\pi^\star} - g^{\pi_0}] + \frac{2}{\zeta} + 2(M+1)\delta. \tag{148}$$

Thus, for a prescribed target accuracy ϵ, taking $\zeta = \Theta(\epsilon^{-1})$, $\delta = \Theta(\epsilon)$, and $K = \Theta(\log(1/\epsilon))$ again yields $\mathbb{E}[g^{\pi^\star} - g^{\pi_K}] \le \mathcal{O}(\epsilon)$.

Since Theorem 5.2 achieves $\delta = \Theta(\epsilon)$ with $\widetilde{O}(|\mathcal{S}||\mathcal{A}|\epsilon^{-2})$ samples under contamination uncertainty and $\widetilde{O}((|\mathcal{S}||\mathcal{A}|)^2\epsilon^{-2})$ samples under TV or Wasserstein uncertainty, the same uncertainty-set-dependent sample complexities apply to the actor-critic method. Here the actor parameter is fixed across actor iterations but depends on the target accuracy; an accuracy-independent constant $\zeta = O(1)$ would leave an $O(1/\zeta)$ residual term under this proof.

B.3 PROOF OF THEOREM 5.2

We first show the contraction property of the Bellman operator for the policy evaluation in the following lemma, different from the semi-norm used in Xu et al. [2025b], our setting requires the contraction to hold uniformly for all $\pi \in \Pi$. We first construct the family fluctuation matrices for each fixed π as follows:

$$\mathcal{F}^\pi \coloneqq \{F_{\mathsf{P}}^\pi = \mathsf{P}^\pi - E_{\mathsf{P}} : \mathsf{P} \in \mathcal{P}\} \tag{149}$$

where E_{P} and P^π are defined in (33).

To establish uniform contraction for all $\pi \in \Pi$, we need to find the radius conditions such that $\sup_{\pi\in\Pi} \hat{\rho}(\mathcal{F}^\pi) < 1$. We start by characterizing the ergodicity behavior of P^π for all π:

Lemma B.2. *Under Assumption 5.4, there exist a finite integer $\bar{m} \in \mathbb{N}$ and a constant $\bar{a}_0 \in (0, 1]$ such that, for all $\pi \in \Pi$,*

$$(\tilde{\mathsf{P}}^\pi)^{\bar{m}} > 0 \quad \textit{(for each entry)}, \qquad \min_{i<j} \sum_{s\in\mathcal{S}} \min\left\{(\tilde{\mathsf{P}}^\pi)^{\bar{m}}_{is}, (\tilde{\mathsf{P}}^\pi)^{\bar{m}}_{js}\right\} \geq \bar{a}_0. \tag{150}$$

Equivalently, $\tau\big((\tilde{\mathsf{P}}^\pi)^{\bar{m}}\big) \leq 1 - \bar{a}_0$ for all $\pi \in \Pi$, where τ is the Dobrushin's coefficient defined in (34).

Proof. For each fixed $\pi \in \Pi$, since $\tilde{\mathsf{P}}^\pi$ is irreducible and aperiodic, there exists an integer $m_\pi \in \mathbb{N}$ with $(\tilde{\mathsf{P}}^\pi)^{m_\pi} > 0$ for each entry. Define $\varepsilon_\pi(\pi') = \min_{i,j}(\tilde{\mathsf{P}}^{\pi'})^{m_\pi}_{ij}$. Since the map $\pi' \mapsto (\tilde{\mathsf{P}}^{\pi'})^{m_\pi}$ is continuous due to each entry being a polynomial in the entries of $\tilde{\mathsf{P}}^{\pi'}$, ε_π is continuous; denote $\varepsilon_\pi(\pi) =: \eta_\pi > 0$. Therefore, there exists an open neighborhood $U_\pi \subset \Pi$ such that for each $s, s' \in \mathcal{S}$,

$$(\tilde{\mathsf{P}}^{\pi'})^{m_\pi}(s, s') \geq \tfrac{\eta_\pi}{2} > 0 \qquad \forall\, \pi' \in U_\pi. \tag{151}$$

For each simplex $\Delta_{|\mathcal{A}|} = \{x \in \mathbb{R}^{|\mathcal{A}|} : x \geq 0,\ \mathbf{1}^\top x = 1\}$, it is closed and bounded. Thus, Π is compact and the open cover $\{U_\pi : \pi \in \Pi\}$ admits a finite subcover. We can accordingly choose $\pi^1, \ldots, \pi^K$ with $\Pi \subset \bigcup_{k=1}^K U_{\pi^k}$. Define the common exponent $\bar{m} = \max_{1\leq k\leq K} m_{\pi^k} < \infty$. Then for any $\pi \in \Pi$, pick k such that $\pi \in U_{\pi^k}$. Then $(\tilde{\mathsf{P}}^\pi)^{m_{\pi^k}} \geq \eta_{\pi^k}/2 > 0$ entrywise. Since all later powers of a stochastic matrix with one positive power remain positive, we thus obtain $(\tilde{\mathsf{P}}^\pi)^{\bar{m}} > 0$ entrywise.

Define the continuous function

$$f(\pi) \coloneqq \min_{i<j} \sum_{s\in\mathcal{S}} \min\left\{(\tilde{\mathsf{P}}^\pi)^{\bar{m}}_{is}, (\tilde{\mathsf{P}}^\pi)^{\bar{m}}_{js}\right\}. \tag{152}$$

We then have $f(\pi) > 0$ for every $\pi \in \Pi$. Since f is continuous and Π is compact, the minimum $\bar{a}_0 \coloneqq \min_{\pi\in\Pi} f(\pi)$ is attained and satisfies $\bar{a}_0 > 0$. The stated Dobrushin bound follows from the identity $\tau(P) = 1 - \min_{i<j} \sum_s \min\{P_{is}, P_{js}\}$. □

Equipped with Lemma B.2, we are now ready to quantify the radius restrictions in the critic settings as follows:

Corollary B.3 (Uniform $\bar{m}$-step contraction for the critic)**.** *Let Assumption 5.4 hold and let $\bar{m}, \bar{a}_0$ be given by Lemma B.2. Define also the uniform $\bar{m}$-step entrywise lower bound at the center,*

$$\bar{b}_0 \coloneqq \min_{\pi\in\Pi} \min_{i,s\in\mathcal{S}} \big((\tilde{\mathsf{P}}^\pi)^{\bar{m}}\big)_{is} \in (0, 1]. \tag{153}$$

Fix any policy $\pi \in \Pi$ and the fluctuation family $\mathcal{F}^\pi = \{F_{\mathsf{P}}^\pi = \mathsf{P}^\pi - E_{\mathsf{P}}^\pi : \mathsf{P} \in \mathcal{P}\}$. Under the following radius conditions on $\mathcal{P}$, there exists $\beta \in (0, 1)$, independent of π, such that for every $\pi \in \Pi$ and every $\mathsf{P}_1, \ldots, \mathsf{P}_{\bar{m}} \in \mathcal{P}$,

$$\tau\big(\mathsf{P}_{\bar{m}}^\pi \mathsf{P}_{\bar{m}-1}^\pi \cdots \mathsf{P}_1^\pi\big) \leq 1 - \beta. \tag{154}$$

Consequently,

$$\sup_{\pi\in\Pi} \hat{\rho}(\mathcal{F}^\pi) \leq (1 - \beta)^{1/\bar{m}} < 1. \tag{155}$$

Moreover, for each admissible $\mathsf{P} \in \mathcal{P}$ and $\pi \in \Pi$, the matrix P^π is irreducible and aperiodic. The radius and the resulting β are:

- ***Contamination uncertainty:*** *Let* $\delta \in [0, 1)$*, then* (154) *holds with* $\beta = (1-\delta)^{\bar{m}}\bar{a}_0$.
- ***TV distance uncertainty:*** *Let* $\delta \in [0, \min\{\bar{a}_0/(2\bar{m}), \bar{b}_0/(2\bar{m})\})$*, then* (154) *holds with* $\beta = \bar{a}_0 - 2\bar{m}\delta$.
- ***Wasserstein-p uncertainty set:*** *Let* $\delta \in [0, \min\{d_{\min}\bar{a}_0/(2\bar{m}), d_{\min}\bar{b}_0/(2\bar{m})\})$*, where* $d_{\min} := \min_{x\neq y} d(x, y) > 0$. *Then* (154) *holds with* $\beta = \bar{a}_0 - 2\bar{m}\delta/d_{\min}$.

Proof. Fix $\pi \in \Pi$. For notational simplicity write $M_i = \mathsf{P}_i^\pi$ and $\widetilde{M} = \tilde{\mathsf{P}}^\pi$. We first prove the $\bar{m}$-step product bound in (154).

For contamination uncertainty, each M_i can be written as $M_i = (1-\delta)\widetilde{M} + \delta Q_i$, where Q_i is row-stochastic. Therefore, by nonnegativity,

$$M_{\bar{m}}M_{\bar{m}-1}\cdots M_1 \geq (1-\delta)^{\bar{m}}\widetilde{M}^{\bar{m}} \tag{156}$$

entrywise. Hence, for any two rows r, r',

$$\sum_{s\in\mathcal{S}} \min\{(M_{\bar{m}}\cdots M_1)_{rs}, (M_{\bar{m}}\cdots M_1)_{r's}\} \geq (1-\delta)^{\bar{m}}\bar{a}_0. \tag{157}$$

This proves (154) with $\beta = (1-\delta)^{\bar{m}}\bar{a}_0$.

For TV uncertainty, convexity of TV gives $TV(M_i(r,\cdot), \widetilde{M}(r,\cdot)) \leq \delta$ for every row r and every i. By a standard telescoping argument and the fact that TV distance is non-expansive under multiplication by a stochastic matrix, for each $1 \leq \ell \leq \bar{m}$,

$$\sup_r TV\big((M_\ell\cdots M_1)(r,\cdot), \widetilde{M}^\ell(r,\cdot)\big) \leq \ell\delta. \tag{158}$$

Using the identity $\sum_s \min\{p_s, q_s\} = 1 - TV(p, q)$ and the triangle inequality for TV, for any two rows r, r',

$$\begin{aligned}
&\sum_{s\in\mathcal{S}} \min\{(M_{\bar{m}}\cdots M_1)_{rs}, (M_{\bar{m}}\cdots M_1)_{r's}\} \\
&\quad\geq \sum_{s\in\mathcal{S}} \min\{(\widetilde{M}^{\bar{m}})_{rs}, (\widetilde{M}^{\bar{m}})_{r's}\} - 2\bar{m}\delta \\
&\quad\geq \bar{a}_0 - 2\bar{m}\delta.
\end{aligned} \tag{159}$$

Thus (154) holds with $\beta = \bar{a}_0 - 2\bar{m}\delta$ when $\delta < \bar{a}_0/(2\bar{m})$. The same telescoping bound also gives, for every r, s,

$$(M_{\bar{m}}\cdots M_1)_{rs} \geq (\widetilde{M}^{\bar{m}})_{rs} - 2\bar{m}\delta \geq \bar{b}_0 - 2\bar{m}\delta > 0 \tag{160}$$

when $\delta < \bar{b}_0/(2\bar{m})$. Applying this last display with the constant sequence $M_1 = \cdots = M_{\bar{m}} = \mathsf{P}^\pi$ shows that $(\mathsf{P}^\pi)^{\bar{m}} > 0$ for every $\mathsf{P} \in \mathcal{P}$ and $\pi \in \Pi$.

For Wasserstein uncertainty, the finite-state metric satisfies $TV(u, v) \leq W_p(u, v)/d_{\min}$ for all distributions u, v on $\mathcal{S}$. Hence the Wasserstein case reduces to the TV case with δ replaced by $\delta/d_{\min}$, giving the stated radius condition and $\beta = \bar{a}_0 - 2\bar{m}\delta/d_{\min}$.

Finally, Lemma C.5 applied to the family $\{\mathsf{P}^\pi : \mathsf{P} \in \mathcal{P}\}$ gives, for each fixed π,

$$\hat{\rho}(\mathcal{F}^\pi) \leq \left(\sup_{\mathsf{P}_1,\ldots,\mathsf{P}_{\bar{m}}\in\mathcal{P}} \tau(\mathsf{P}_{\bar{m}}^\pi\cdots\mathsf{P}_1^\pi)\right)^{1/\bar{m}} \leq (1-\beta)^{1/\bar{m}}. \tag{161}$$

Taking the supremum over $\pi \in \Pi$ proves (155). The entrywise positivity above implies irreducibility and aperiodicity of each induced kernel. □

Lemma B.4. *Under Assumption 5.4 and the radius restrictions in Corollary B.3, there exists* $b_\star > 0$ *such that for every policy* $\pi \in \Pi$*, every admissible kernel* $\mathsf{P} \in \mathcal{P}$*, and every state* $s \in \mathcal{S}$*,*

$$d_{\mathsf{P}}^\pi(s) \geq b_\star. \tag{162}$$

Consequently, the mismatch constant M *in* (30) *is finite.*

Proof. By Corollary B.3, there are $\bar{m} \in \mathbb{N}$ and $b > 0$, independent of π and P, such that

$$(\mathsf{P}^\pi)^{\bar{m}}(i,j) \geq b, \qquad \forall i,j \in \mathcal{S}. \tag{163}$$

For contamination uncertainty, one can take $b = (1-\delta)^{\bar{m}}\bar{b}_0$. For TV uncertainty, one can take $b = \bar{b}_0 - 2\bar{m}\delta$. For Wasserstein uncertainty, one can take $b = \bar{b}_0 - 2\bar{m}\delta/d_{\min}$. Let d_{P}^π be the stationary distribution of P^π. Then, for any $j \in \mathcal{S}$,

$$\begin{aligned} d_{\mathsf{P}}^\pi(j) &= \sum_i d_{\mathsf{P}}^\pi(i)(\mathsf{P}^\pi)^{\bar{m}}(i,j) \\ &\geq \sum_i d_{\mathsf{P}}^\pi(i) b = b. \end{aligned} \tag{164}$$

Thus $b_\star = b$ works. Since every numerator in (30) is at most 1 and every denominator is at least $b_\star$, we have $M \leq 1/b_\star < \infty$. □

Lemma B.5. *Under Assumption 5.4, and the radius conditions in Corollary B.3, for any fixed policy $\pi \in \Pi$, there exists a corresponding semi-norm $\|\cdot\|_{\pi,\mathcal{P}}$ with kernel $\{c\mathbf{e} : c \in \mathbb{R}\}$ such that the robust Bellman operator $\mathbf{T}_g^\pi$ under π is a one-step contraction with a uniform contraction factor. Specifically, there exist $\gamma \in (0,1)$ independent of π such that*

$$\|\mathbf{T}_g^\pi(V_1) - \mathbf{T}_g^\pi(V_2)\|_{\pi,\mathcal{P}} \leq \gamma \|V_1 - V_2\|_{\pi,\mathcal{P}}, \forall\, V_1, V_2 \in \mathbb{R}^S,\ g \in \mathbb{R}. \tag{165}$$

Proof. By Corollary B.3, for every π the fluctuation family $\mathcal{F}^\pi = \{F_{\mathsf{P}}^\pi = \mathsf{P}^\pi - E_{\mathsf{P}}^\pi : \mathsf{P} \in \mathcal{P}\}$ has joint spectral radius

$$\widehat{\rho}(\mathcal{F}^\pi) \leq \gamma_* \coloneqq (1-\beta)^{1/\bar{m}} < 1, \tag{166}$$

where γ_* is independent of π. Choose α such that $\gamma_* < \alpha < 1$. Define

$$\|x\|_{\pi,\mathrm{ext}} \coloneqq \sup_{k \geq 0} \sup_{F_1,\ldots,F_k \in \mathcal{F}^\pi} \alpha^{-k} \|F_k F_{k-1} \ldots F_1 x\|_2. \tag{167}$$

As in Lemma A.6, this is a norm and it satisfies

$$\|Fx\|_{\pi,\mathrm{ext}} \leq \alpha \|x\|_{\pi,\mathrm{ext}}, \qquad \forall F \in \mathcal{F}^\pi. \tag{168}$$

Choose $\xi \in (0, 1-\alpha)$ and define

$$\|x\|_{\pi,\mathcal{P}} \coloneqq \sup_{F \in \mathcal{F}^\pi} \big\|Fx\big\|_{\pi,\mathrm{ext}} + \xi \inf_{c \in \mathbb{R}} \big\|x - c\mathbf{e}\big\|_{\pi,\mathrm{ext}}. \tag{169}$$

This is a semi-norm with kernel $\{c\mathbf{e} : c \in \mathbb{R}\}$.

Fix $V_1, V_2 \in \mathbb{R}^S$ and set $\Delta = V_1 - V_2$. The scalar g cancels in the difference $\mathbf{T}_g^\pi(V_1) - \mathbf{T}_g^\pi(V_2)$. For every (s,a), the elementary inequality for differences of minima gives

$$\min_{p \in \mathcal{P}_s^a} p^\top \Delta \leq \sigma_{\mathcal{P}_s^a}(V_1) - \sigma_{\mathcal{P}_s^a}(V_2) \leq \max_{p \in \mathcal{P}_s^a} p^\top \Delta. \tag{170}$$

Since $\mathcal{P}_s^a$ is compact and convex, and since $p \mapsto p^\top \Delta$ is linear, there exists $p_{s,a} \in \mathcal{P}_s^a$ such that

$$p_{s,a}^\top \Delta = \sigma_{\mathcal{P}_s^a}(V_1) - \sigma_{\mathcal{P}_s^a}(V_2). \tag{171}$$

Indeed, choose minimizer and maximizer rows in (170), and take the convex combination that produces the middle value. By rectangularity, the rows $p_{s,a}$ can be assembled into a kernel $\mathsf{P}_\Delta \in \mathcal{P}$. Therefore,

$$\begin{aligned} \mathbf{T}_g^\pi(V_1)(s) - \mathbf{T}_g^\pi(V_2)(s) &= \sum_a \pi(a|s) \left(\sigma_{\mathcal{P}_s^a}(V_1) - \sigma_{\mathcal{P}_s^a}(V_2)\right) \\ &= \sum_a \pi(a|s) p_{s,a}^\top \Delta = (\mathsf{P}_\Delta^\pi \Delta)(s). \end{aligned} \tag{172}$$

Let E_Δ be the rank-one matrix whose rows equal the stationary distribution of P_Δ^π, and let $F_\Delta = \mathsf{P}_\Delta^\pi - E_\Delta \in \mathcal{F}^\pi$. Since $E_\Delta \Delta$ is a constant vector and every $F \in \mathcal{F}^\pi$ annihilates constant vectors,

$$\begin{aligned}\sup_{F\in\mathcal{F}^\pi} \|F\mathsf{P}_\Delta^\pi \Delta\|_{\pi,\mathrm{ext}} &= \sup_{F\in\mathcal{F}^\pi} \|FF_\Delta \Delta\|_{\pi,\mathrm{ext}} \\ &\le \alpha \sup_{F\in\mathcal{F}^\pi} \|F\Delta\|_{\pi,\mathrm{ext}}. \end{aligned} \tag{173}$$

For the quotient term, choose the constant vector $E_\Delta \Delta$. Then

$$\begin{aligned}\inf_{c\in\mathbb{R}} \|\mathsf{P}_\Delta^\pi \Delta - c\mathbf{e}\|_{\pi,\mathrm{ext}} &\le \|\mathsf{P}_\Delta^\pi \Delta - E_\Delta \Delta\|_{\pi,\mathrm{ext}} \\ &= \|F_\Delta \Delta\|_{\pi,\mathrm{ext}} \le \sup_{F\in\mathcal{F}^\pi} \|F\Delta\|_{\pi,\mathrm{ext}}. \end{aligned} \tag{174}$$

Combining (173) and (174) with the definition (169) gives

$$\begin{aligned}\|\mathbf{T}_g^\pi(V_1) - \mathbf{T}_g^\pi(V_2)\|_{\pi,\mathcal{P}} &= \|\mathsf{P}_\Delta^\pi \Delta\|_{\pi,\mathcal{P}} \\ &\le (\alpha+\xi) \sup_{F\in\mathcal{F}^\pi} \|F\Delta\|_{\pi,\mathrm{ext}} \\ &\le (\alpha+\xi)\|\Delta\|_{\pi,\mathcal{P}}. \end{aligned} \tag{175}$$

Thus (165) holds with $\gamma = \alpha + \xi < 1$, and this γ is independent of π. □

Unlike the policy-evaluation settings in [Xu et al., 2025b], the actor–critic procedure repeatedly re-solves the critic at changing policies. Hence all constants used in the finite-sample analysis must be independent of the π. Lemma B.2-B.5 provide uniform $\bar{m} \in \mathbb{N}$ and $\bar{a}_0, \bar{b}_0 > 0$ such there exist a semi-norm $\|\cdot\|_{\pi,\mathcal{P}}$ for each π in which the robust Bellman operator contracts with the same factor γ_*. In order to carry out the remaining analysis of the critic, we now provide the uniform sample complexity bound (independent of the policy) for estimating the robust value function $V_\mathcal{P}^\pi$ and robust average reward $g_\mathcal{P}^\pi$ for any π.

Lemma B.6. *Under the norm $\|\cdot\|_{\pi,\mathrm{ext}}$ constructed in* (167)*, there exists a constant $\bar{B} < \infty$, independent of π, such that for all $\pi \in \Pi, x \in \mathbb{R}^S$*

$$\|x\|_{\pi,\mathrm{ext}} \;\le\; \bar{B}\|x\|_2. \tag{176}$$

Proof. Fix $\pi \in \Pi$ and a product $F_k \cdots F_1$ with $F_i = F_{\mathsf{P}_i}^\pi \in \mathcal{F}^\pi$. Write $M_i = \mathsf{P}_i^\pi$. Since $F_i\mathbf{e} = 0$ and $E_{\mathsf{P}_i}^\pi z$ is a constant vector for every z, an induction gives

$$F_k \cdots F_1 x = M_k \cdots M_1 x - c_k \mathbf{e} \tag{177}$$

for some scalar c_k. Hence $\|F_k \cdots F_1 x\|_{\mathrm{sp}} = \|M_k \cdots M_1 x\|_{\mathrm{sp}}$. By Corollary B.3, every product of $\bar{m}$ admissible induced kernels has Dobrushin coefficient at most $1-\beta$. Since Dobrushin coefficients are submultiplicative and at most one, if $k = q\bar{m} + r$ with $0 \le r < \bar{m}$, then

$$\|M_k \cdots M_1 x\|_{\mathrm{sp}} \le (1-\beta)^q \|x\|_{\mathrm{sp}}. \tag{178}$$

For $k \ge 1$, the vector $F_k \cdots F_1 x$ has zero mean under the stationary distribution of M_k, which has strictly positive entries by Corollary B.3. Therefore 0 lies between its minimum and maximum, so $\|F_k \cdots F_1 x\|_2 \le \sqrt{S}\|F_k \cdots F_1 x\|_{\mathrm{sp}}$. Using $\|x\|_{\mathrm{sp}} \le 2\|x\|_2$, for $k \ge 1$ we obtain

$$\|F_k \cdots F_1 x\|_2 \le 2\sqrt{S}(1-\beta)^q \|x\|_2. \tag{179}$$

Since $\alpha > (1-\beta)^{1/\bar{m}}$, we have $(1-\beta)/\alpha^{\bar{m}} < 1$. Thus, for $k = q\bar{m} + r \ge 1$,

$$\alpha^{-k}\|F_k \cdots F_1 x\|_2 \le 2\sqrt{S}\alpha^{-r}\left(\frac{1-\beta}{\alpha^{\bar{m}}}\right)^q \|x\|_2 \le 2\sqrt{S}\alpha^{-(\bar{m}-1)}\|x\|_2. \tag{180}$$

For $k = 0$, the term in (167) equals $\|x\|_2$. Taking the supremum over k and over all admissible products gives

$$\|x\|_{\pi,\mathrm{ext}} \le \max\{1, 2\sqrt{S}\alpha^{-(\bar{m}-1)}\}\|x\|_2. \tag{181}$$

The constant $\bar{B} := \max\{1, 2\sqrt{S}\alpha^{-(\bar{m}-1)}\}$ is finite and independent of π. □

Lemma B.7. *Let $\bar{B} < \infty$ be the policy-uniform constant from Lemma B.6, i.e. $\|x\|_{\pi,\mathrm{ext}} \le \bar{B}\|x\|_2$ for all π, x. Then, for all $\pi \in \Pi$ and $x \in \mathbb{R}^{\mathcal{S}}$,*

$$\xi\|x\|_{\infty,0} \le \|x\|_{\pi,\mathcal{P}} \le (\alpha+\xi)\bar{B}\sqrt{S}\|x\|_{\infty,0}, \tag{182}$$

where $\|x\|_{\infty,0} := \min_{c\in\mathbb{R}} \|x - c\mathbf{e}\|_\infty$ is the centered ℓ_∞ semi-norm. Equivalently,

$$\frac{\xi}{2}\|x\|_{\mathrm{sp}} \le \|x\|_{\pi,\mathcal{P}} \le \frac{(\alpha+\xi)\bar{B}\sqrt{S}}{2}\|x\|_{\mathrm{sp}}. \tag{183}$$

Here ξ and α are chosen in Lemma B.5 and are independent of π.

Proof. Let $c_\infty \in \arg\min_{c\in\mathbb{R}} \|x - c\mathbf{e}\|_\infty$ and write $z := x - c_\infty\mathbf{e}$. Since $F\mathbf{e} = \mathbf{0}$ for all $F \in \mathcal{F}^\pi$, we have $Fx = Fz$. By the definition of $\|x\|_{\pi,\mathcal{P}}$ in (169), the extremal norm contraction, and the choice of c_∞,

$$\begin{aligned}
\|x\|_{\pi,\mathcal{P}} &\le \sup_{F\in\mathcal{F}^\pi} \|Fz\|_{\pi,\mathrm{ext}} + \xi\|z\|_{\pi,\mathrm{ext}} \\
&\le (\alpha+\xi)\|z\|_{\pi,\mathrm{ext}} \\
&\le (\alpha+\xi)\bar{B}\|z\|_2 \\
&\le (\alpha+\xi)\bar{B}\sqrt{S}\|z\|_\infty \\
&= (\alpha+\xi)\bar{B}\sqrt{S}\|x\|_{\infty,0}.
\end{aligned} \tag{184}$$

This proves the upper bound in (182). For the lower bound, drop the first term in (169) and use the $k = 0$ term in $\|\cdot\|_{\pi,\mathrm{ext}}$:

$$\begin{aligned}
\|x\|_{\pi,\mathcal{P}} &\ge \xi \inf_{c\in\mathbb{R}} \|x - c\mathbf{e}\|_{\pi,\mathrm{ext}} \\
&\ge \xi \inf_{c\in\mathbb{R}} \|x - c\mathbf{e}\|_2 \\
&\ge \xi \inf_{c\in\mathbb{R}} \|x - c\mathbf{e}\|_\infty \\
&= \xi\|x\|_{\infty,0}.
\end{aligned} \tag{185}$$

Finally, $\|x\|_{\infty,0} = \frac{1}{2}(\max_i x_i - \min_i x_i) = \frac{1}{2}\|x\|_{\mathrm{sp}}$ yields (183). □

Remark B.8. By the above lemmas, the sampling chains are geometrically ergodic with a policy-uniform mixing time

$$\bar{t}_{\mathrm{mix}} := \bar{m}\left\lceil \frac{\log 4}{-\log(1-\beta)} \right\rceil. \tag{186}$$

and stationary distributions admit a policy-uniform lower bound. Consequently, the semi–Lyapunov constants (Section B.1.2 in [Xu et al., 2025b]) in the policy evaluation analysis now admit policy-uniform versions. As in Appendix A.2.2, choose the Moreau smoothing parameter in this policy-evaluation Lyapunov construction sufficiently small, using the uniform contraction factor $\gamma < 1$, so that the resulting policy-uniform constants satisfy

$$\bar{\vartheta} := \gamma\sqrt{\bar{c}_u/\bar{c}_\ell} \le \frac{1+\gamma}{2} < 1, \qquad \bar{\alpha}_2 := \frac{1-\bar{\vartheta}}{2} \ge \frac{1-\gamma}{4} > 0. \tag{187}$$

This is again only a choice of the smooth Lyapunov function. With this choice, there exist $\bar{c}_\ell > 0$ and $\bar{c}_u < \infty$ such that the inequalities of the form $\bar{c}_\ell\, M_{\overline{E}}(x) \le \frac{1}{2}\|x\|^2_{\mathcal{N},\overline{E}} \le \bar{c}_u\, M_{\overline{E}}(x)$ hold uniformly over π (notations borrowed from Section B.1.2 of Xu et al. [2025b]). In addition, by setting $\bar{c}_{\mathcal{P}} := \xi$ and $\bar{C}_{\mathcal{P}} := (\alpha+\xi)\bar{B}\sqrt{|\mathcal{S}|}$, the coefficients $\bar{c}_{\mathcal{P}}$ and $\bar{C}_{\mathcal{P}}$ also recover the coefficients $c_{\mathcal{P}}$ and $C_{\mathcal{P}}$ from Section B.1.2 of Xu et al. [2025b] uniformly independent of π.

We are now able to present the formal results for the critic estimation bounds.

Theorem B.9. *Let $\pi \in \Pi$ and $\mathsf{P} \in \mathcal{P}$. Consider the iterates (V_t^π, g_t^π) from Algorithm 5 with stepsizes $\beta_t := \frac{1}{t+1}$ and $\eta_t := \frac{2}{\bar{\alpha}_2(t+K)}$. Under Assumption 5.4 and the radii of Corollary B.3, define the policy-uniform constants as follows. Let $\gamma := \alpha + \xi$ be the contraction factor from Lemma B.5. Let $\bar{t}_{\mathrm{mix}}$ be as in* (186)*. Set*

$$\bar{c}_{\mathcal{P}} := \xi, \qquad \bar{C}_{\mathcal{P}} := (\alpha+\xi)\bar{B}\sqrt{|\mathcal{S}|}. \tag{188}$$

where $\bar{B}$ is the policy-uniform constant in Lemma B.6. Let $\bar{c}_\ell$ and $\bar{c}_u$ denote the policy-uniform semi–Lyapunov constants from Remark B.8 (notations borrowed from Section B.1.2 of Xu et al. [2025b]). Let $\bar{\rho}_2, \bar{L}, \bar{G}$ denote the corresponding policy-uniform constants in Section B.1.2 of Xu et al. [2025b]. Define the following

$$\begin{aligned}
&\bar{\vartheta} := \gamma\sqrt{\bar{c}_u/\bar{c}_\ell}, && \bar{\alpha}_2 := \frac{1-\bar{\vartheta}}{2}, && \bar{\alpha}_4 := \bar{\rho}_2\,\bar{L},\\
&\bar{\alpha}_3 := 8\,\bar{c}_u\,\bar{\alpha}_4, && K := \max\{4\bar{\alpha}_3/\bar{\alpha}_2^2, 3\}, && \bar{C}_2 := \frac{1}{K} + \log\Big(\frac{T-1+K}{K}\Big),\\
&\bar{C}_3 := \bar{G}\big(1 + 8\bar{C}_{\mathcal{P}}\bar{t}_{\mathrm{mix}}\big).
\end{aligned} \tag{189}$$

By Remark B.8, the Moreau smoothing parameter is chosen so that $\bar{\vartheta} \le (1+\gamma)/2$, hence $\bar{\alpha}_2 \ge (1-\gamma)/4 > 0$. Finally, let $(\bar{\varepsilon}_{\mathrm{bias}}, \bar{A})$ be the (policy-uniform) bias and second-moment constants of the support-function estimator from Algorithm 4 (cf. Lemma D.1 of Xu et al. [2025b]), i.e., under contamination uncertainty,

$$(\bar{A}, \bar{\varepsilon}_{\mathrm{bias}}) = \big(32\bar{C}_{\mathcal{P}}^2\bar{t}_{\mathrm{mix}}^2,\ 0\big), \tag{190}$$

under TV uncertainty,

$$\bar{A} = 2\bar{C}_{\mathcal{P}}^2\big(24(1+\tfrac{1}{\delta})\sqrt{|\mathcal{S}|\,2^{-N_{\max}}}\,\bar{t}_{\mathrm{mix}}\big)^2 + 96\bar{C}_{\mathcal{P}}^2\bar{t}_{\mathrm{mix}}^2 + 4608\bar{C}_{\mathcal{P}}^2(1+\tfrac{1}{\delta})^2|\mathcal{S}|\,\bar{t}_{\mathrm{mix}}^2 N_{\max}, \tag{191}$$

$$\bar{\varepsilon}_{\mathrm{bias}} = 48\bar{C}_{\mathcal{P}}(1+\tfrac{1}{\delta})\sqrt{|\mathcal{S}|\,2^{-N_{\max}}}\,\bar{t}_{\mathrm{mix}}, \tag{192}$$

and under Wasserstein uncertainty,

$$(\bar{A}, \bar{\varepsilon}_{\mathrm{bias}}) = \Big(2\bar{C}_{\mathcal{P}}^2\big(24\sqrt{|\mathcal{S}|\,2^{-N_{\max}}}\,\bar{t}_{\mathrm{mix}}\big)^2 + 96\bar{C}_{\mathcal{P}}^2\bar{t}_{\mathrm{mix}}^2 + 4608\bar{C}_{\mathcal{P}}^2|\mathcal{S}|\,\bar{t}_{\mathrm{mix}}^2 N_{\max},\ 48\bar{C}_{\mathcal{P}}\sqrt{|\mathcal{S}|\,2^{-N_{\max}}}\,\bar{t}_{\mathrm{mix}}\Big). \tag{193}$$

Then, for all $T \ge 1$,

$$\mathbb{E}\big[\|V_T^\pi - V_{\mathcal{P}}^\pi\|_{\infty,0}^2\big] \le \frac{4K^2\bar{c}_u\bar{C}_{\mathcal{P}}^2}{(T+K)^2\bar{c}_\ell\bar{c}_{\mathcal{P}}^2}\|V_0^\pi - V_{\mathcal{P}}^\pi\|_{\infty,0}^2 + \frac{8\bar{A}\bar{\alpha}_4\bar{c}_u}{(T+K)\bar{\alpha}_2^2\bar{c}_{\mathcal{P}}^2} + \frac{\bar{c}_u\bar{C}_3\bar{C}_2\bar{\varepsilon}_{\mathrm{bias}}}{\bar{\alpha}_2\bar{c}_{\mathcal{P}}^2},$$

$$\mathbb{E}\big[|g_T^\pi - g_{\mathcal{P}}^\pi|^2\big] \le \frac{\bar{B}_1}{T+K} + \frac{\bar{B}_2\log^2(T)}{(T+K)^2} + \frac{\bar{B}_3\bar{t}_{\mathrm{mix}}^2 N_{\max}\log^2(T)}{(T+K)(1-\gamma)^2} + \frac{\bar{B}_4\bar{t}_{\mathrm{mix}}^2 2^{-N_{\max}/2}\log^3(T)}{1-\gamma} + \bar{B}_5\bar{t}_{\mathrm{mix}}^2 2^{-N_{\max}}\log^2(T),$$

for some absolute constants $\bar{B}_1, \bar{B}_2, \bar{B}_3, \bar{B}_4, \bar{B}_5 > 0$ depending only on the policy-uniform quantities above.

Proof. The proof is identical to the robust value bound and robust average reward bound in the policy evaluation analysis (Theorem D.2 and D.3 of Xu et al. [2025b]): use Lemma B.5 for the one-step contraction in $\|\cdot\|_{\pi,\mathcal{P}}$ with the uniform modulus $\gamma = \alpha + \xi$; translate all norm occurrences to the centered ℓ_∞ semi-norm using Lemma B.7 and Remark B.8 (uniform $\bar{c}_{\mathcal{P}}, \bar{C}_{\mathcal{P}}$); and replace $(c_\ell, c_u, t_{\mathrm{mix}})$ by $(\bar{c}_\ell, \bar{c}_u, \bar{t}_{\mathrm{mix}})$ per Remark B.8. □

We now start the formal proof of Theorem 5.2. Define the error decomposition

$$\|\hat{Q}^\pi - Q^\pi\|_\infty \le |g_T^\pi - g_{\mathcal{P}}^\pi| + \max_{s,a}\big|\bar{\sigma}_{\mathcal{P}_s^a}(V_T^\pi) - \sigma_{\mathcal{P}_s^a}(V_{\mathcal{P}}^\pi)\big|. \tag{194}$$

Squaring and taking expectation gives

$$\mathbb{E}\big[\|\hat{Q}^\pi - Q^\pi\|_\infty^2\big] \le 2\mathbb{E}\big[|g_T^\pi - g_{\mathcal{P}}^\pi|^2\big] + 2\mathbb{E}\Big[\max_{s,a}\big(\bar{\sigma}_{\mathcal{P}_s^a}(V_T^\pi) - \sigma_{\mathcal{P}_s^a}(V_{\mathcal{P}}^\pi)\big)^2\Big]. \tag{195}$$

By Theorem B.9 (the second inequality), we have the following direct second-moment bound:

$$\begin{aligned}
\mathbb{E}\Big[|g_T^\pi - g_{\mathcal{P}}^\pi|^2\Big] \le{}& \frac{\bar{B}_1}{T+K} + \frac{\bar{B}_2\log^2(T+K)}{(T+K)^2} + \frac{\bar{B}_3\,\bar{t}_{\mathrm{mix}}^2 N_{\max}\log^2(T+K)}{(T+K)(1-\gamma)^2}\\
&+ \frac{\bar{B}_4\,\bar{t}_{\mathrm{mix}}^2 2^{-N_{\max}/2}\log^3(T+K)}{1-\gamma} + \bar{B}_5\,\bar{t}_{\mathrm{mix}}^2 2^{-N_{\max}}\log^2(T+K),
\end{aligned} \tag{196}$$

where $\bar{B}_1, \ldots, \bar{B}_5 > 0$ are absolute constants independent of π.

Regarding the second term on the RHS of (195), write

$$\bar{\sigma}_{\mathcal{P}_s^a}(V_T^\pi) = \frac{1}{L_Q} \sum_{\ell=1}^{L_Q} \hat{\sigma}_{\mathcal{P}_s^a}^{(\ell)}(V_T^\pi). \tag{197}$$

For any (s, a),

$$\begin{aligned} \left|\bar{\sigma}_{\mathcal{P}_s^a}(V_T^\pi) - \sigma_{\mathcal{P}_s^a}(V_{\mathcal{P}}^\pi)\right|^2 \le 2\left|\bar{\sigma}_{\mathcal{P}_s^a}(V_T^\pi) - \sigma_{\mathcal{P}_s^a}(V_T^\pi)\right|^2 \\ + 2\left|\sigma_{\mathcal{P}_s^a}(V_T^\pi) - \sigma_{\mathcal{P}_s^a}(V_{\mathcal{P}}^\pi)\right|^2. \end{aligned} \tag{198}$$

The support map is 1-Lipschitz in $\|\cdot\|_\infty$, hence

$$\left|\sigma_{\mathcal{P}_s^a}(V_T^\pi) - \sigma_{\mathcal{P}_s^a}(V_{\mathcal{P}}^\pi)\right| \le \|V_T^\pi - V_{\mathcal{P}}^\pi\|_\infty. \tag{199}$$

Both V_T^π and $V_{\mathcal{P}}^\pi$ are anchored at s_0, so $\|V_T^\pi - V_{\mathcal{P}}^\pi\|_\infty \le \|V_T^\pi - V_{\mathcal{P}}^\pi\|_{\mathrm{sp}} = 2\|V_T^\pi - V_{\mathcal{P}}^\pi\|_{\infty,0}$. Conditional on V_T^π, the samples $\{\hat{\sigma}_{\mathcal{P}_s^a}^{(\ell)}(V_T^\pi)\}_{\ell=1}^{L_Q}$ are independent. Therefore Lemma C.4 gives

$$\mathbb{E}\left[\left|\bar{\sigma}_{\mathcal{P}_s^a}(V_T^\pi) - \sigma_{\mathcal{P}_s^a}(V_T^\pi)\right|^2 \,\middle|\, V_T^\pi\right] \le \frac{\bar{A}_\sigma}{L_Q} + \bar{\varepsilon}_\sigma^2, \tag{200}$$

where $\bar{A}_\sigma = O(1)$ for contamination uncertainty and $\bar{A}_\sigma = O(N_{\max})$ for TV and Wasserstein uncertainty, while $\bar{\varepsilon}_\sigma = 0$ for contamination and $\bar{\varepsilon}_\sigma = O(2^{-N_{\max}/2})$ for TV and Wasserstein.

We now bound the maximum over all state-action pairs. Let $n = |\mathcal{S}||\mathcal{A}|$. Under contamination uncertainty, conditional on V_T^π, the centered variables $\hat{\sigma}_{\mathcal{P}_s^a}^{(\ell)}(V_T^\pi) - \sigma_{\mathcal{P}_s^a}(V_T^\pi)$ are mean-zero and bounded in absolute value by $\|V_T^\pi\|_{\mathrm{sp}}$. Therefore, the standard maximal bound for bounded empirical averages gives

$$\mathbb{E}\left[\max_{s,a}\left|\bar{\sigma}_{\mathcal{P}_s^a}(V_T^\pi) - \sigma_{\mathcal{P}_s^a}(V_T^\pi)\right|^2 \,\middle|\, V_T^\pi\right] \le \frac{C\log(2n)}{L_Q}\|V_T^\pi\|_{\mathrm{sp}}^2, \tag{201}$$

for a universal constant $C > 0$. Also, $\|V_T^\pi\|_{\mathrm{sp}}^2 \le 2\|V_T^\pi - V_{\mathcal{P}}^\pi\|_{\mathrm{sp}}^2 + 2\|V_{\mathcal{P}}^\pi\|_{\mathrm{sp}}^2$, and Lemma C.2 gives $\|V_{\mathcal{P}}^\pi\|_{\mathrm{sp}} \le 4\bar{t}_{\mathrm{mix}}$ under the reward-span normalization. Since $\|x\|_{\mathrm{sp}} = 2\|x\|_{\infty,0}$, (201), Theorem B.9, and the Lipschitz bound above imply that $L_Q = \widetilde{\Theta}(\epsilon^{-2})$ is sufficient for the contamination case.

For TV and Wasserstein uncertainty, Lemma C.4 gives the second-moment and bias bound above, and we use the deterministic inequality $\max_i z_i^2 \le \sum_i z_i^2$. This yields

$$\mathbb{E}\left[\max_{s,a}\left|\bar{\sigma}_{\mathcal{P}_s^a}(V_T^\pi) - \sigma_{\mathcal{P}_s^a}(V_{\mathcal{P}}^\pi)\right|^2\right] \le 2n\left(\frac{\bar{A}_\sigma}{L_Q} + \bar{\varepsilon}_\sigma^2\right) + 2\mathbb{E}\left[\|V_T^\pi - V_{\mathcal{P}}^\pi\|_\infty^2\right]. \tag{202}$$

For TV and Wasserstein uncertainty, $\bar{A}_\sigma = \mathcal{O}(N_{\max})$ and $\bar{\varepsilon}_\sigma = \mathcal{O}(2^{-N_{\max}/2})$. Thus, choosing $L_Q = \Theta(nN_{\max}\epsilon^{-2})$ and $N_{\max} = \Theta(\log(n/\epsilon))$ makes the RHS of (202) $\widetilde{O}(\epsilon^2)$.

Now set $T = \widetilde{\Theta}(\epsilon^{-2})$ large enough to absorb the logarithmic factors in (196). For TV and Wasserstein uncertainty, choose $N_{\max} = \Theta(\log(n/\epsilon))$, with a sufficiently large absolute constant so that $2^{-N_{\max}/2} \lesssim \epsilon^2$ and $n2^{-N_{\max}} \lesssim \epsilon^2$. Then (196) gives

$$\mathbb{E}\left[\left|g_T^\pi - g_{\mathcal{P}}^\pi\right|^2\right] \le \mathcal{O}(\epsilon^2) \tag{203}$$

after increasing the hidden constants if necessary.

Combining (195), (196), the contamination bound (201) or the TV/Wasserstein bound (202), and the value-function bound from Theorem B.9, we obtain

$$\mathbb{E}\left[\|\hat{Q}^\pi - Q^\pi\|_\infty^2\right] = \widetilde{O}(\epsilon^2).$$

Increasing the constants in the choices of T, L_Q, and $N_{\max}$ when applicable gives

$$\mathbb{E}\left[\|\hat{Q}^\pi - Q^\pi\|_\infty^2\right] \le \epsilon^2.$$

The sample-complexity claim follows because Algorithm 5 uses $\widetilde{O}(|\mathcal{S}||\mathcal{A}|\epsilon^{-2})$ samples under contamination uncertainty and $\widetilde{O}(|\mathcal{S}||\mathcal{A}|N_{\max}\epsilon^{-2})$ samples under TV or Wasserstein uncertainty. The final batched support-estimation step uses nL_Q nominal samples under contamination uncertainty, which is $\widetilde{O}(|\mathcal{S}||\mathcal{A}|\epsilon^{-2})$, and has expected cost $\mathcal{O}(nL_QN_{\max})$ under TV or Wasserstein uncertainty, which is $\widetilde{O}((|\mathcal{S}||\mathcal{A}|)^2\epsilon^{-2})$.

C SOME AUXILIARY LEMMAS FOR THE PROOFS

Lemma C.1 (Theorem IV in Berger and Wang [1992]). *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*

$$\hat{\rho}(\mathcal{Q}) \coloneqq \lim_{k\to\infty} \sup_{Q_i\in\mathcal{Q}} \rho(Q_k \dots Q_1)^{\frac{1}{k}} = \lim_{k\to\infty} \sup_{Q_i\in\mathcal{Q}} \|Q_k \dots Q_1\|^{\frac{1}{k}} \tag{204}$$

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

Lemma C.2 (Ergodic case of Lemma 9 in Wang et al. [2022]). *Consider a finite average-reward Markov reward process induced by a stationary policy π and transition kernel P, with transition matrix P_π and stationary distribution ν. Define*

$$\tau_{\mathrm{mix}} \coloneqq \arg\min_{t\geq 1} \left\{ \max_{\mu_0} \left\| (\mu_0 P_\pi^t)^\top - \nu^\top \right\|_1 \leq \frac{1}{2} \right\}. \tag{205}$$

Let $\bar{r}_\pi(s) \coloneqq \sum_{a\in\mathcal{A}} \pi(a|s) r(s,a)$, $g_P^\pi \coloneqq \nu^\top \bar{r}_\pi$, and let $R_{\mathrm{sp}} \coloneqq \max_{s,a} r(s,a) - \min_{s,a} r(s,a)$. If P_π is irreducible and aperiodic, then $\tau_{\mathrm{mix}} < +\infty$. Moreover, for every $T \geq 1$, the centered finite-horizon value defined as

$$h_T^{\pi,P}(s) \coloneqq \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{T-1} \left(\bar{r}_\pi(S_t) - g_P^\pi \right) \middle| S_0 = s \right] \tag{206}$$

satisfies

$$\|h_T^{\pi,P}\|_{\mathrm{sp}} \leq 4R_{\mathrm{sp}}\tau_{\mathrm{mix}}. \tag{207}$$

Consequently, any average-reward relative value function $h^{\pi,P}$ obtained as the limit of $h_T^{\pi,P}$, equivalently any solution to the Poisson equation up to an additive constant, satisfies

$$\|h^{\pi,P}\|_{\mathrm{sp}} \leq 4R_{\mathrm{sp}}\tau_{\mathrm{mix}}. \tag{208}$$

In particular, under the normalization $R_{\mathrm{sp}} \leq 1$, we have $\|h^{\pi,P}\|_{\mathrm{sp}} \leq 4\tau_{\mathrm{mix}}$.

Lemma C.3 (Linear-growth version of Lemma 6 in Zhang et al. [2021]). *Under the setup and notation in Appendix A.2, suppose we have*

$$\mathbb{E}\Big[\|w^t\|_{\mathcal{H},\overline{E}}^2 \,\Big|\, \mathcal{F}^t \Big] \leq A_0 + B_0 \|x^t - x^*\|_{\mathcal{H},\overline{E}}^2.$$

Let $\ell_s \coloneqq \rho_2^2$, so that $\|x\|_{s,\overline{E}}^2 \leq \ell_s \|x\|_{\mathcal{H},\overline{E}}^2$. Then we have

$$\mathbb{E}\Big[\|x^{t+1} - x^t\|_{s,\overline{E}}^2 \,\Big|\, \mathcal{F}^t \Big] \leq (16 + 4B_0) c_u \ell_s \eta_t^2 M_{\overline{E}}(x^t - x^*) + 2A_0 \ell_s \eta_t^2. \tag{209}$$

Lemma C.4 (Theorem 5.2-5.4 in Xu et al. [2025b], Lemma 12 in Wang et al. [2023c]). *Let $\hat{\sigma}_{\mathcal{P}_s^a}(V)$ be the estimator of $\sigma_{\mathcal{P}_s^a}(V)$ obtained from Algorithm 4, then the following properties hold:*

1. *Denote M as the number of state-action pair samples needed. Then $M = 1$ under contamination uncertainty set. Furthermore, under TV and Wasserstein distance uncertainty sets, $M = 2^{N'+1}$ (where $N' = \min\{N, N_{\max}\}$), which then implies,*

$$\mathbb{E}[M] = N_{\max} + 2 = \mathcal{O}(N_{\max}). \tag{210}$$

2. *Under contamination uncertainty set, $\hat{\sigma}_{\mathcal{P}_s^a}(V)$ is unbiased satisfying*

$$\mathbb{E}\left[\hat{\sigma}_{\mathcal{P}_s^a}(V) \right] = \sigma_{\mathcal{P}_s^a}(V). \tag{211}$$

Under TV uncertainty set, we have a bias that is exponentially decaying with respect to the maximum level $N_{\max}$ as follows:

$$\left|\mathbb{E}\left[\hat{\sigma}_{\mathcal{P}_s^a}(V) - \sigma_{\mathcal{P}_s^a}(V)\right]\right| \leq 6(1+\frac{1}{\delta})2^{-\frac{N_{\max}}{2}}\|V\|_{\text{sp}} \tag{212}$$

where δ denotes the radius of TV distance and $\|\cdot\|_{\text{sp}}$ denotes the span semi-norm. Similarly, under Wasserstein uncertainty set, we have:

$$\left|\mathbb{E}\left[\hat{\sigma}_{\mathcal{P}_s^a}(V) - \sigma_{\mathcal{P}_s^a}(V)\right]\right| \leq 6 \cdot 2^{-\frac{N_{\max}}{2}}\|V\|_{\text{sp}} \tag{213}$$

3. *Under contamination uncertainty set, the variance of $\hat{\sigma}_{\mathcal{P}_s^a}(V)$ is bounded by the l_2 norm of V as*

$$\text{Var}(\hat{\sigma}_{\mathcal{P}_s^a}(V)) \leq \|V\|^2. \tag{214}$$

Under TV uncertainty set, the variance of $\hat{\sigma}_{\mathcal{P}_s^a}(V)$ is bounded by is as follows:

$$\text{Var}(\hat{\sigma}_{\mathcal{P}_s^a}(V)) \leq 3\|V\|_{\text{sp}}^2 + 144(1+\frac{1}{\delta})^2\|V\|_{\text{sp}}^2 N_{\max} \tag{215}$$

Similarly, under Wasserstein uncertainty set, we have:

$$\text{Var}(\hat{\sigma}_{\mathcal{P}_s^a}(V)) \leq 3\|V\|_{\text{sp}}^2 + 144\|V\|_{\text{sp}}^2 N_{\max} \tag{216}$$

Lemma C.5 (Lemma A.3 in Xu et al. [2025b])**.** *Define the Dobrushin's coefficient of an n dimensional Markov matrix P as*

$$\tau(P) \coloneqq 1 - \min_{i<j}\sum_{s=1}^{n}\min(P_{is}, P_{js}), \tag{217}$$

Let $\mathcal{P}$ be a family of Markov matrix with the same dimension with each having a unique stationary distribution. We define the family of fluctuation matrices to be

$$\mathcal{F} \coloneqq \{\mathsf{P} - E_{\mathsf{P}} : \mathsf{P} \in \mathcal{P}\} \tag{218}$$

where E_{P} is the matrix with all rows being identical to the stationary distribution of P. Then the joint spectral radius of the family $\mathcal{F}$ is upper bounded by the following:

$$\hat{\rho}(\mathcal{F}) \leq \inf_{m\geq 1}\left(\sup_{\mathsf{P}_i\in\mathcal{P}}\tau(\mathsf{P}_1\cdot\ldots\cdot\mathsf{P}_m)\right)^{\frac{1}{m}} \tag{219}$$

Lemma C.6. *For any $n \in \mathbb{N}$, let $X = \mathbb{R}^n$, let $\overline{E} = \{c\mathbf{e} : c \in \mathbb{R}\}$, and let $\|\cdot\|_{sn}$ be a semi-norm with kernel being $\overline{E}$. Define the linear functional $\ell(x) = \frac{1}{n}\sum_{i=1}^n x_i$, and then define $\|x\|_n = \|x - \ell(x)\mathbf{e}\|_{sn} + \left|\ell(x)\right|$. Then $\|\cdot\|_n$ is a norm on X, and moreover*

$$\left(\|\cdot\|_n *_{\inf} \delta_{\overline{E}}\right)(x) = \inf_{c\in\mathbb{R}}\|x - c\mathbf{e}\|_n = \|x\|_{sn} \quad \textit{for all } x \in X. \tag{220}$$

Proof. We first show that $\|\cdot\|_n$ is a genuine norm.

- Positive homogeneity. For any $\lambda \in \mathbb{R}$,

$$\|\lambda x\|_n = \|\lambda x - \ell(\lambda x)\mathbf{e}\|_{sn} + \left|\ell(\lambda x)\right| = |\lambda|\|x - \ell(x)\mathbf{e}\|_{sn} + |\lambda||\ell(x)| = |\lambda|\|x\|_n.$$

- Triangle inequality. For $x, y \in X$,

$$\begin{aligned}\|x+y\|_n &= \|(x+y) - \ell(x+y)\mathbf{e}\|_{sn} + \left|\ell(x+y)\right| \\ &= \|(x-\ell(x)\mathbf{e}) + (y-\ell(y)\mathbf{e})\|_{sn} + \left|\ell(x)+\ell(y)\right| \\ &\leq \|x-\ell(x)\mathbf{e}\|_{sn} + \|y-\ell(y)\mathbf{e}\|_{sn} + |\ell(x)| + |\ell(y)| \\ &= \|x\|_n + \|y\|_n.\end{aligned}$$

- Positive definiteness. If $\|x\|_n = 0$, then $|\ell(x)| = 0$ and $\|x - \ell(x)\mathbf{e}\|_{sn} = 0$. Hence $\ell(x) = 0$ and $x - \ell(x)\mathbf{e} \in \overline{E}$. Thus $x \in \overline{E}$, so $x = c\mathbf{e}$ for some $c \in \mathbb{R}$. Since $\ell(x) = c = 0$, we obtain $x = 0$.

Regarding the infimal convolution, by definition,

$$(\|\cdot\|_n *_{\inf} \delta_{\overline{E}})(x) = \inf_{c\in\mathbb{R}} \|x - c\mathbf{e}\|_n.$$

Then

$$\|x - c\mathbf{e}\|_n = \|(x - c\mathbf{e}) - \ell(x - c\mathbf{e})\mathbf{e}\|_{sn} + \big|\ell(x - c\mathbf{e})\big|.$$

But $\ell(x - c\mathbf{e}) = \ell(x) - c$ and

$$(x - c\mathbf{e}) - \ell(x - c\mathbf{e})\mathbf{e} = x - \ell(x)\mathbf{e}$$

independent of c. Hence

$$\|x - c\mathbf{e}\|_n = \|x - \ell(x)\mathbf{e}\|_{sn} + |\ell(x) - c|.$$

Minimizing over $c \in \mathbb{R}$ by taking $c = \ell(x)$ gives

$$\inf_{c\in\mathbb{R}} \|x - c\mathbf{e}\|_n = \|x - \ell(x)\mathbf{e}\|_{sn} = \|x\|_{sn},$$

since $\ell(x)\mathbf{e} \in \ker(\|\cdot\|_{sn})$. This completes the proof. □

Lemma C.7. *Let $x \in \mathbb{R}^{SA}$ satisfy $x_i = 0$ for some fixed index i. Then*

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

Moreover, since all semi-norms with the same kernel spaces are equivalent, there are constants $c_H, C_H > 0$ such that for the semi-norm $\|\cdot\|_H$ defined in (20)*,*

$$c_H\|x\|_{\mathrm{sp}} \le \|x\|_H \le\ C_H\|x\|_{\mathrm{sp}} \quad \forall\, x \in \mathbb{R}^n,$$

then

$$c_H\|x\|_\infty \le c_H\|x\|_{\mathrm{sp}} \le \|x\|_H \le C_H\|x\|_{\mathrm{sp}} \le 2C_H\|x\|_\infty. \tag{221}$$

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

$$\max_j x_j \ \le\ \|x\|_\infty, \quad \min_j x_j \ \ge\ -\|x\|_\infty,$$

and so

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

Since $\|\cdot\|_{\mathrm{sp}}$ and $\|\cdot\|_H$ 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_H and C_H such that

$$c_H\|x\|_\infty \le c_H\|x\|_{\mathrm{sp}} \le \|x\|_H \le C_H\|x\|_{\mathrm{sp}} \le 2C_H\|x\|_\infty$$

as claimed. □

D NUMERICAL EVALUATIONS

In this section, we provide some numerical evaluation results to demonstrate the performance of our algorithms. We evaluate our Algorithm 1 on a ride-hailing domain with five zones and Algorithm 2 on a minimal three-state control loop. Both environments are average-reward problems with two actions and ergodic dynamics. We apply robust evaluation with contamination, TV distance, and Wasserstein-1 distance uncertainty sets around the nominal transition models; the corresponding robustness parameters are reported with each experiment.

We note that Figures 1-2 show convergence end-to-end across contamination, TV distance and Wasserstein-1 uncertainty sets with various radius. In the figure, the anchored sup-norm error $\|Q_t - Q^*\|_\infty$ decays for the Q-learning algorithm , while the robust average reward g_t rises toward g^* and stabilizes for the actor-critic algorithm. We note that the Q^* and g^* are computed via a coarse grid search over stationary policies with a discretized uncertainty set, so any small terminal gap may reflect reference discretization rather than algorithmic bias. Overall, the trends align with our $\tilde{O}(\varepsilon^{-2})$ dependence on the target accuracy. The remaining section provides the detailed settings of the numerical experiments.

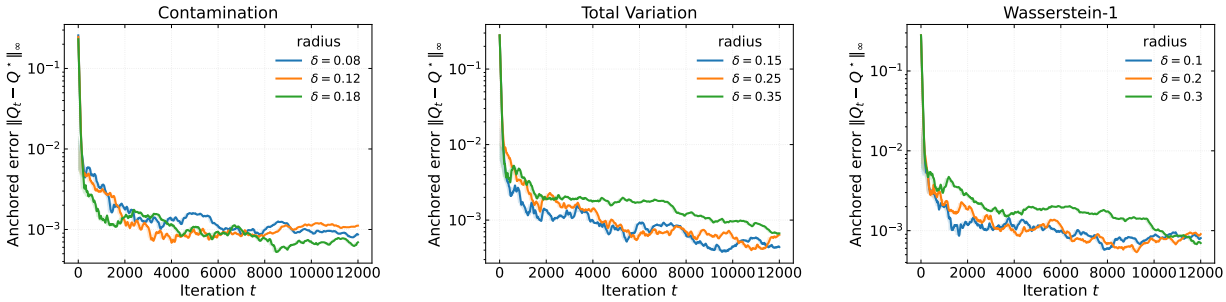

Figure 1: Ride-hailing via robust Q-learning

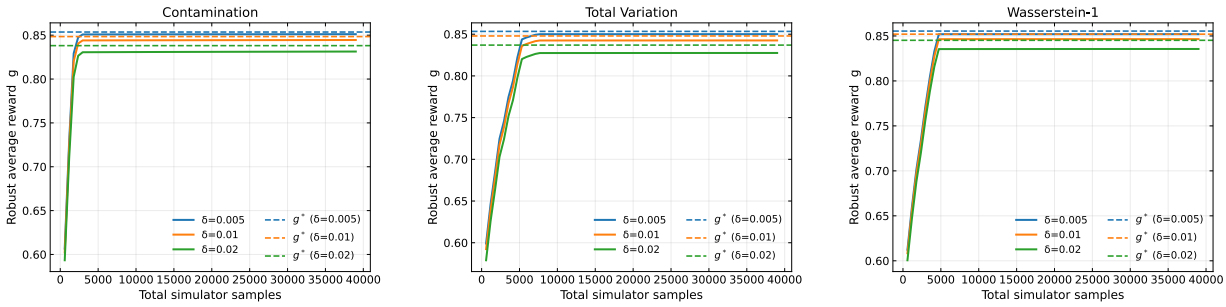

Figure 2: Control loop via robust actor-critic

D.1 THREE-STATE "OPERATE VS. PROTECT" LOOP

This environment abstracts a system that trades off performance against resource protection (e.g., server throughput vs. thermal throttling, or battery discharge vs. charge). The state is a coarse operating level with three values: *low*, *medium*, and *high*. Action *boost* drives performance upward; action *conserve* protects resources and favors staying or moving left. Rewards increase with the operating level; *conserve* adds a small immediate penalty.

Rewards. Let the base per-step utilities by level be $(0.20, 0.55, 0.90)$. The two actions incur penalties $(0.00, 0.05)$ respectively, so the realized rewards are:

State	Boost	Conserve
Low	0.20	0.15
Medium	0.55	0.50
High	0.90	0.85

Transitions. Rows index the current state; columns index the next state in the order Low, Medium, High. The *boost* action pushes right (≈80% right, 17% self, 3% left in the interior), while *conserve* pushes left (symmetrically). At the boundaries we use the exact probabilities induced by the code's normalization.[1]

Boost

	Low	Medium	High
Low	0.036	0.964	0.000
Medium	0.030	0.170	0.800
High	0.000	0.100	0.900

[1] At the edges, the "left" or "right" assignment can coincide with the self index; rows are normalized afterwards.

Conserve

	Low	Medium	High
Low	0.900	0.100	0.000
Medium	0.800	0.170	0.030
High	0.000	0.964	0.036

D.2 FIVE-STATE RIDE-HAILING DOMAIN

This environment models a single driver operating across five zones: Downtown, University, Airport, Suburbs, and a Depot (garage/charging). The state is the driver's current zone. The action is an operating mode: *aggressive* (accept surge/long hauls; proactive repositioning toward hubs) or *conservative* (prefer short local rides; stay nearby). Immediate rewards represent net earnings by zone; hubs pay more, Depot is low, and aggressive typically yields higher per-step revenue. The nominal transition model captures destination mixes induced by each mode: aggressive drifts the driver toward major hubs and the Depot, whereas conservative increases self-loops and nearby moves. Zones are arranged along a line to define geographic closeness for Wasserstein robustness.

Rewards.

Zone	Aggressive	Conservative
Downtown	2.0	1.2
University	1.6	1.1
Airport	2.2	1.4
Suburbs	1.2	0.9
Depot	0.8	0.5

Transitions. Rows index the current zone; columns index the next zone in the order Downtown, University, Airport, Suburbs, Depot.

Aggressive

	DT	Univ	Airp	Sub	Depot
Downtown	0.20	0.20	0.30	0.10	0.20
University	0.30	0.15	0.25	0.10	0.20
Airport	0.35	0.10	0.20	0.10	0.25
Suburbs	0.25	0.15	0.20	0.20	0.20
Depot	0.30	0.20	0.25	0.10	0.15

Conservative

	DT	Univ	Airp	Sub	Depot
Downtown	0.40	0.20	0.10	0.20	0.10
University	0.20	0.45	0.05	0.20	0.10
Airport	0.15	0.10	0.45	0.15	0.15
Suburbs	0.10	0.25	0.10	0.45	0.10
Depot	0.15	0.15	0.15	0.15	0.40

In both environments, the learning objective is to find a stationary policy that maximizes long-run average performance while remaining reliable under plausible misspecification of the nominal transition model. The concrete rewards and transitions above yield ergodic chains and cleanly expose the effects of contamination, total-variation, and geography-aware Wasserstein uncertainty.