Primal-Only Actor Critic Algorithm for Robust Constrained Average Cost MDPs

Anirudh Satheesh[*]
University of Maryland

ANIRUDHS@TERPMAIL.UMD.EDU

Sooraj Sathish[*]
IIIT Bangalore

SOORAJ.SATHISH@IIITB.AC.IN

Swetha Ganesh
Purdue University

GANESH49@PURDUE.EDU

Keenan Powell
University of Maryland

KPOWELL1@TERPMAIL.UMD.EDU

Vaneet Aggarwal
Purdue University

VANEET@PURDUE.EDU

Abstract

In this work, we study the problem of finding robust and safe policies in Robust Constrained Average-Cost Markov Decision Processes (RCMDPs). A key challenge in this setting is the lack of strong duality, which prevents the direct use of standard primal-dual methods for constrained RL. Additional difficulties arise from the average-cost setting, where the Robust Bellman operator is not a contraction under any norm. To address these challenges, we propose an actor-critic algorithm for Average-Cost RCMDPs. We show that our method achieves both ϵ-feasibility and ϵ-optimality, and we establish a sample complexities of $\tilde{O}\left(\epsilon^{-4}\right)$ and $\tilde{O}\left(\epsilon^{-6}\right)$ with and without slackness assumption, which is comparable to the discounted setting.

1 Introduction

Reinforcement Learning (RL) has achieved remarkable success across domains such as robotics (Chen et al., 2023), transportation (Al-Abbasi et al., 2019), and large language model fine-tuning (Gaur et al., 2025). However, most approaches assume that training and deployment occur under identical conditions. In practice, real-world experiments are costly and risky, necessitating reliance on simulators. Additionally, even the most detailed simulators cannot fully capture the variability, noise, and stochasticity of real-world environments. This mismatch, known as the sim2real gap, can lead to severe performance degradation and, in safety-critical systems, catastrophic failures or equipment damage.

Many real-world applications also impose strict safety or resource constraints: autonomous vehicles must guarantee human safety, communication networks must respect bandwidth limits, and transportation systems must meet time constraints, and the stochastic nature of RL policies makes consistently satisfying these constraints challenging. The need for both robustness against environmental shifts and adherence to strict constraints

---

[*]. These authors contributed equally.

motivates the Robust Constrained Markov Decision Process (RCMDP) framework, where policies must guarantee worst-case performance under an uncertainty set of transitions while satisfying constraints.

In the RCMDP framework, distributional robustness is modeled by defining an uncertainty set of environments that captures potential distribution shifts in transition dynamics, with the objective of optimizing the worst-case performance within this set. Constraint satisfaction is incorporated by augmenting the reward function with additional constraint functions that must remain below specified thresholds. Our setting adds further complexity by focusing on the infinite-horizon average reward, rather than the discounted return. This formulation is more suitable for capturing long-term objectives and is particularly relevant in applications requiring persistent and consistent performance over extended time horizons.

The literature on robust and constrained MDPs is still limited, especially in the average-reward case. Strong duality does not hold in RCMDPs (Wang et al., 2022; Ma et al., 2025), which prevents extending sample-efficient primal-dual algorithms to the robust MDP formulation. Thus, recent work has opted for primal-only methods. Kitamura et al. (2025) propose an epigraph formulation for discounted reward RCMDP, but the necessity of binary search increases the sample complexity to $\tilde{O}(\epsilon^{-6}\log(\epsilon^{-1}))$. Additionally, this work requires perfect estimation of the robust value function, which may not be tractable for large state spaces. Ma et al. (2025) and Ganguly et al. (2025) develop primal-only algorithms that achieve sample complexities of $O(\epsilon^{-6})$, while also in the discounted setting.

Challenges and Contributions Two primary challenges motivated our specific solution method. Firstly, the average reward setting's Bellman operator does not satisfy a trivial contraction property like we have in the discounted setting. Secondly, most works on Constrained MDPs utilize the Primal-Dual method and propose algorithms that alternatively update the respective Lagrangian multipliers. However, Wang et al. (2022); Ma et al. (2025) shows that strong duality does not hold in the robust setting, which motivates the choice of primal-only algorithms.

We summarize our work and contributions as follows:

- We present the first formulation and analysis of the Average Cost RCMDPs, extending beyond the discounted reward setting and addressing the major problem of the lack of strong duality.

- We propose an actor critic algorithm that theoretically guarantees ϵ-feasibility and optimality for different uncertainty sets (Contamination, TV Distance, Wasserstein).

- We show sample complexity guarantees of $O(\epsilon^{-4})$ with the slackness assumption and $O(\epsilon^{-6})$ without the slackness assumption.

2 Related Work

2.1 Constrained Reward MDPs

Constrained Reward MDPs (CMDPs) have been studied extensively in the literature, both in the discounted reward and average reward setting. The model-based approaches (Chen et al., 2022; Agarwal et al., 2022b,a) construct estimates of the transition probabilities

and then derive *safe* policies. These approaches often involve continuously solving Linear Programs (LPs) as estimated models are updated, leading to computational inefficiency and a need for substantial memory. Model-free algorithms (Wei et al., 2022; Bai et al., 2024; Xu et al., 2025b) learn the optimal policy or value function directly from sampling of the environment. This is generally more compute efficient and requires less memory. Owing to the clear advantages and the real-world applicability of model-free algorithms, we also focus our attention on this setting.

Constrained RL problems have been addressed using various model-free solution methods, but the most common approach has been the primal-dual method (Altman, 2021; Paternain et al., 2022; Bai et al., 2022; Wang et al., 2022; Bai et al., 2023, 2024; Mondal and Aggarwal, 2024; Xu et al., 2025b). Here, the constrained problem is converted into its dual formulation, where the objective is a weighted sum of the reward and the constraints. These weights are Lagrangian multipliers, which are updated alternatively until convergence. Paternain et al. (2019) show that strong duality holds in the non-robust constrained RL setting, and this primal-dual method attains zero duality gap. The less-studied counterpart are the works on primal-only solutions (Dalal et al., 2018; Liu et al., 2020; Yang et al., 2020). These works ensure that the constraints are not violated (or violation is bounded) without the use of Lagrange multipliers. For example, CRPO (Xu et al., 2021) ensures convergence to an optimal safe policy by only updating the reward when no constraint is violated. They leverage a novel combinatorial bucketing approach to show the convergence even when the objective being optimised switches every iteration. Since strong duality does not hold in the distributionally robust setting (Wang et al., 2022; Ma et al., 2025), we look to design a primal-only algorithm.

Method	Setting	Sample Complexity
(Xu et al., 2025b)	Constrained, Average	$O(\epsilon^{-2})$
(Li et al., 2022)	Robust, Discounted	$\widetilde{O}(\epsilon^{-2})$
(Xu et al., 2025a)	Robust, Average	$\widetilde{O}(\epsilon^{-2})$
(Kitamura et al., 2025)	Robust, Constrained, Discounted	$\widetilde{O}(\epsilon^{-6})$
Ma et al. (2025)	Robust, Constrained, Discounted	$O(\epsilon^{-6})$
Ganguly et al. (2025) (w/o Slackness)	Robust, Constrained, Discounted	$O(\epsilon^{-6})$
Ganguly et al. (2025) (w/ Slackness)	Robust, Constrained, Discounted	$O(\epsilon^{-4})$
Our work (w/o Slackness)	Robust, Constrained, Average	$O(\epsilon^{-6})$
Our work (w/ Slackness)	Robust, Constrained, Average	$O(\epsilon^{-4})$

Table 1: Comparison of sample complexities of different methods to solve Robust and Constrained MDPs. Our work achieves state of the art sample complexity over existing robust constrained MDP methods.

2.2 Robust RL

Dynamic programming approaches to solve model-based robust RL problems have been explored extensively in the past (Nilim and Ghaoui, 2003; Iyengar, 2005; Wiesemann et al., 2013; Tamar et al., 2014). More recent studies on the discounted reward setting have focused their efforts on problems where the uncertainty set is unknown and only samples

from the nominal distribution can be collected (Zhou et al., 2021; Panaganti and Kalathil, 2022; Wang et al., 2022, 2023a).

Most studies on robust RL have primarily considered the infinite-horizon discounted reward setting, where the Bellman operator always satisfies a contraction property it inherits from the discount factor. Since no such trivial contraction property exists in the average reward setting, the literature on robust average reward (Wang et al., 2023c; Chen et al., 2025) consider approaches where the results from discounted reward could be converted to that for average reward, while in the absence of constraints. Some average-reward works explore model-free solutions through Halpern iteration (Roch et al., 2025), while others exploit ODE methods in stochastic approximation to prove convergence (Wang et al., 2023d).

More recently, a novel semi-norm with the contraction property has been found (Xu et al., 2025c), and a corresponding Actor-Critic approach to robust average reward unconstrained RL has been proposed (Xu et al., 2025a). In our work, we leverage this semi-norm and propose a similar Actor-Critic approach to Robust Constrained Average cost RL.

2.3 Robust Constrained MDPs

The literature on robust constrained MDPs (RCMDPs) is limited because strong duality does not hold in this setting (Wang et al., 2022). Some studies (Russel et al., 2020; Mankowitz et al., 2020; Wang et al., 2022; Zhang et al., 2024) have tried to address this problem through primal-dual methods by quantifying and tracking the duality gap or restricting to certain policy classes that satisfy strong duality. However, these works do not provide explicit iteration and sample complexity guarantees.

More recently, Kitamura et al. (2025) propose an epigraph formulation of the discounted primal problem and provide explicit sample complexity guarantees. Unfortunately, the binary search employed in this solution elicits a very high sample complexity. Furthermore, it is known that the binary search approaches fail when the robust value estimates are noisy (Horstein, 2003).

To address the above shortcomings, Ganguly et al. (2025) propose a unique formulation of the discounted RCMDP problem without the use of epigraphs and binary search. They show ϵ-feasibility and optimality of their mirror-descent algorithm, offering improved iteration complexity guarantees. A parallel work by Ma et al. (2025) takes inspiration from CRPO (Xu et al. (2021)) and achieves the same sample complexity guarantees as Ganguly et al. (2025).

We emphasize that, to the best of our knowledge, this is the first work to address the RCMDP problem in the average reward/cost setting and provide optimal sample complexity guarantees. Table 1 is a concise comparison of the various methods in the literature and demonstrates the optimal performance of our algorithm.

3 Formulation

3.1 Robust Average Cost MDPs

An infinite horizon robust average cost Markov Decision Process (MDP) can be defined by the tuple $(\mathcal{S}, \mathcal{A}, r, \mathcal{P}, \rho)$, where $\mathcal{S}$ is the state space, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the cost function, $\mathcal{P}$ is an uncertainty set of transition kernels, and $\rho : \mathcal{S} \to [0, 1]$ is the initial state

distribution. At each timestep, a transition kernel P is randomly selected from $\mathcal{P}$ and is used to transition the environment to the next state. We focus on the (s, a)-rectangular uncertainty set $\mathcal{P} = \otimes_{s,a} \mathcal{P}_s^a$ (Nilim and Ghaoui, 2003; Iyengar, 2005), where

$$\mathcal{P}_s^a = \{P \in \Delta(\mathcal{S}) : D(P, P^\circ) \leq R\} \tag{1}$$

and P° is the nominal transition kernel. The goal of the policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is to maximize the worst-case average cost over the set of transitions $\mathcal{P}$

$$g_{\mathcal{P}}^\pi(s) = \max_{\kappa \in \otimes_{k \geq 0} \mathcal{P}} \lim_{T \to \infty} \mathbb{E}_{\kappa, \pi} \left[\frac{1}{T} \sum_{t=0}^{T-1} r_t \middle| s_0 = s \right] \tag{2}$$

Wang et al. (2024) showed this objective is the same under the stationary model

$$g_{\mathcal{P}}^\pi(s) = \max_{P \in \mathcal{P}} \lim_{T \to \infty} \mathbb{E}_{P, \pi} \left[\frac{1}{T} \sum_{t=0}^{T-1} r_t \middle| s_0 = s \right] \tag{3}$$

Thus, we focus solely on the stationary case. We denote the maximizers of Eq (3) as the worst-case transition kernels and $\Omega_g^\pi = \{P \in \mathcal{P} : g_P^\pi = g_{\mathcal{P}}^\pi\}$, where

$$g_P^\pi = \lim_{T \to \infty} \mathbb{E}_{P, \pi} \left[\frac{1}{T} \sum_{t=0}^{T-1} r_t \middle| s_0 = s \right] \tag{4}$$

is the average cost of π with transition kernel P.

We also focus on the model-free setting, where samples can only be accessed from the nominal transition kernel P°. We are interested in estimating both the robust value function $V_{P_V}^\pi$ and the robust average cost $g_{P_V}^\pi$. The robust value function can be defined through the robust Bellman equation in Theorem 1.

Theorem 1 (Robust Bellman Equation, Theorem 3.1 in (Wang et al., 2023d)) *If (g, V) is a solution to the robust Bellman equation*

$$V(s) = \sum_a \pi(a|s)(r(s, a) - g + \sigma_{\mathcal{P}_s^a}(V)), \forall s \in \mathcal{S} \tag{5}$$

where $\sigma_{\mathcal{P}_s^a} = \min_{P \in \mathcal{P}_s^a}$ is denoted as the support function, then the scalar g corresponds to the robust average cost, i.e., $g = g_{\mathcal{P}}^\pi$, and the worst-case transition kernel P_V belongs to the set of minimizing transition kernels, i.e., $P_V \in \Omega_{\mathcal{P}}^\pi$ where $\Omega_g^\pi = \{P \in \mathcal{P} : g_P^\pi = g_{\mathcal{P}}^\pi\}$. Furthermore, the function V is unique up to an additive constant, where if V is a solution to the Bellman equation, then we have $V = V_{P_V}^\pi + c\mathbf{e}$, where $c \in \mathbb{R}$ and $\mathbf{e}$ is the all-ones vector in $\mathbb{R}^{|\mathcal{S}|}$, and $V_{P_V}^\pi$ is defined as the relative value function of the policy π under the single transition P_V as follows:

$$V_{P_V}^\pi(s) = \mathbb{E}_{P_V, \pi} \left[\sum_{t=0}^\infty (r_t - g_{P_V}^\pi) \middle| s_0 = s \right] \tag{6}$$

Using Theorem 1, we can define $\sigma_{\mathcal{P}_s^a}(V)$ for different uncertainty sets.

Contamination Uncertainty Set The R-contamination uncertainty set is $\mathcal{P}_s^a = \{(1 - R)P_{s,a}^\circ + Rq \,|\, q \in \Delta(\mathcal{S})\}$, where $R \in (0, 1)$ is the radius of the uncertainty set. The support function of the R-contamination set can be directly computed as

$$\sigma_{\mathcal{P}_s^a}(V) = (1 - R)P_{s,a}^\circ V + R \max_s V(s) \tag{7}$$

We can also use this formulation to construct the estimator of the worst case transition effect

$$\hat{\sigma}_{\mathcal{P}_s^a}(V) = (1 - R)V(s') + R \max_x V(x) \tag{8}$$

where s' is the next state using the nominal transition kernel.

Total Variation Uncertainty Set The total variation uncertainty set is $\mathcal{P}_s^a = \left\{\frac{1}{2}\|q - P_{s,a}^\circ\|_1 \leq R \,|\, q \in \Delta(\mathcal{S})\right\}$. We can define the support function using its dual formulation

$$\sigma_{\mathcal{P}_s^a}(V) = \min_{\mu \geq \mathbf{0}} \left(P_{s,a}^\circ(V - \mu) - R\|V - \mu\|_{\mathrm{sp}}\right) \tag{9}$$

where $\|\cdot\|_{\mathrm{sp}}$ is the span semi-norm (Iyengar, 2005).

Wasserstein Uncertainty Sets We consider the l-Wasserstein distance $W_l(q, p) = \inf_{\mu \in \Gamma(p,q)} \|d\|_{\mu,l}$, where $l \in [1, \infty)$, $p, q \in \Delta(\mathcal{S})$, $\Gamma(p, q)$ is the distributions over $\mathcal{S} \times \mathcal{S}$ with marginal distributions p, q, and $\|d\|_{\mu,l} = \left(\mathbb{E}_{(X,Y) \sim \mu}\left[d(X, Y)^l\right]\right)^{\frac{1}{l}}$. The Wasserstein distance uncertainty set is then defined as $\mathcal{P}_s^a = \{W_l(P_{s,a}^\circ, q) \leq R \,|\, q \in \Delta(\mathcal{S})\}$. Then we can define the support function for the Wasserstein uncertainty set (Gao and Kleywegt, 2023) as

$$\sigma_{\mathcal{P}_s^a} = \inf_{\lambda \geq 0} \left(-\lambda \delta^l + \mathbb{E}_{s \sim \mathcal{S}, a \sim \pi(s)}\left[\sup_y V(y) + \lambda d(\mathcal{S}, y)^l\right]\right) \tag{10}$$

Following Theorem 1, we can define the robust Bellman operator in Theorem 11.

Theorem 2 (Robust Bellman Operator (Wang et al., 2024)) *The robust Bellman Operator* $\mathbf{T}_g$ *is defined as*

$$\mathbf{T}_g(V)(s) = \sum_a \pi(a|s)\left[r(s, a) - g + \sigma_{\mathcal{P}_s^a}(V)\right], \forall s \in \mathcal{S} \tag{11}$$

The main challenge with the robust Bellman operator is that it does not satisfy a contraction under standard norms. Thus, we leverage the contraction under the semi-norm introduced in Xu et al. (2025c) for our stochastic approximation algorithms. We also make the assumption throughout this work that the induced Markov Chain from the policy π is irreducible and aperiodic (Assumption 1).

Assumption 1 (Ergodicity) *The Markov chain induced by every policy* $\pi \in \Pi$ *is irreducible and aperiodic for all* $P \in \mathcal{P}$, *where* $\Pi = \{\pi | \pi : \mathcal{S} \to \Delta(\mathcal{A})\}$.

Assumption 1 is widely used in the robust average reward reinforcement learning literature (Wang et al., 2023d,b; Sun et al., 2024; Xu et al., 2025c). This ensures that from any state, it is possible to eventually reach any other state, and the system does not get stuck in deterministic cycles. This guarantees the existence of a unique stationary distribution for a given policy, which is fundamental to the average-cost setting. Under this assumption, the average cost is independent of the starting state, so we can write the robust average cost as $g_{\mathcal{P}}^{\pi}$.

3.2 Robust Constrained Average Cost MDPs

We extend robust average cost MDPs to robust constrained average cost MDPs by including I constraint functions $c_i : \mathcal{S} \times \mathcal{A} \to [0, 1]$ and corresponding thresholds $b_i \in \mathbb{R}$ for each $i = 1, 2, \cdots, I$ (we keep the constraint values bounded between 0 and 1 for simplicity). Thus, the worst-case average constraint value with policy π on constraint i is

$$g_{\mathcal{P}}^{\pi,i} = \max_{P \in \mathcal{P}} \lim_{T \to \infty} \mathbb{E}_{P,\pi} \left[\frac{1}{T} \sum_{t=0}^{T-1} c_{i,t} \right] \tag{12}$$

where $c_{i,t}$ is the constraint value of constraint i at time t. Additionally, let $g_{\mathcal{P}}^{\pi,0} = g_{\mathcal{P}}^{\pi}$ be the worst case average cost. The goal of the robust constrained average cost MDP is to find a policy that minimizes the worst-case average cost while ensuring each constraint is satisfied under the worst-case transition kernel:

$$\pi^* = \arg\min_{\pi} g_{\mathcal{P}}^{\pi,0} \quad \text{s.t.}$$
$$g_{\mathcal{P}}^{\pi,i} \leq b_i, i = 1, 2, \cdots, I \tag{13}$$

Issues with Primal-Dual methods Many existing works approach constrained reinforcement learning problem via primal-dual algorithms using Lagrange multipliers. However, this approach faces two fundamental obstacles in the robust setting.

First, strong duality is not guaranteed. While (Paternain et al., 2019) established that the duality gap is often zero for standard (non-robust) CMDPs, this result relies on Slater's condition to ensure a strictly feasible policy. In the robust case, however, the set of achievable robust state-action occupancy measures (i.e., those under the worst-case models) is not necessarily convex (Wang et al., 2022). This breakdown of the underlying convexity means that Slater's condition is no longer sufficient to guarantee a zero duality gap and allow us to use the dual formulation.

$$g_{\mathcal{P}}^{\pi^*} = \min_{\lambda \in \mathbb{R}_+^N} \min_{\pi \in \Pi} \left(g_{\mathcal{P}}^{\pi} + \sum_{i=1}^{I} \lambda_i (g_{\mathcal{P}}^{\pi,i} - b_i) \right) \tag{14}$$

Secondly, the formulation of the Lagrangian is difficult to solve in the robust case due to the maximization over the transition kernels in the uncertainty set (Kitamura et al., 2025). This motivates the need for approaches that do not rely on strong duality.

4 Proposed Algorithm: Robust Constrained Average Cost Actor Critic

We can avoid the non-convexity and intractability issues with primal dual methods that require strong duality by focusing solely on primal methods. For our problem formulation

we take inspiration from a recent work by Ganguly et al. (2025) and adapt it for the average reward case,

$$F_{\mathcal{P}}^{\pi} = \min_{\pi} \max \left\{ \frac{g_{\mathcal{P}}^{\pi,0}}{\lambda}, \; \max_{i} \left\{ g_{\mathcal{P}}^{\pi,i} - b_i + \zeta \right\} \right\} \tag{15}$$

where ζ is the slackness term. Here, the intuition is to focus on the largest constraint violation and optimize for it in each update. If no constraints are violated ($g_{\mathcal{P}}^{\pi,i} - b_i \leq 0, \forall i \in [1, \ldots I]$), then we optimize for the cost function $g_{\mathcal{P}}^{\pi,0}$. Here, λ is introduced to regulate the trade-off between optimizing the primary cost and mitigating constraint violations. A sufficiently large λ ensures that constraint violations cannot be ignored, while feasibility shifts the focus back to minimizing the cost objective.

It is worth noting that our work is not a direct extension of Ganguly et al. (2025) to the average cost setting, as their approach relies solely on mirror descent. Our choice of an actor-critic (AC) framework is necessitated by a core challenge in the average-cost setting: the robust Bellman operator is not a contraction under standard norms. Consequently, standard gradient-based methods like Online Mirror Descent are not directly applicable, as the iterative processes needed to estimate their required Q-functions would diverge. Our AC approach resolves this directly: the critic uses a specialized algorithm that converges under a specific semi-norm (Xu et al. (2025c)), providing the stable Q-function estimates the actor requires for a provably convergent update. Furthermore, although we draw inspiration from Xu et al. (2025a) and Sun et al. (2024), we cleverly utilise their results on critic estimation sample complexity and robust performance difference (respectively) for the constrained setting, which they did not tackle.

Instead of solving a convex optimization problem in the dual formulation, we look at Eq. (15) and perform gradient descent in the direction of ∇F. However, F is a function of non linear robust average costs calculated by taking the maximum over all transition kernels in the uncertainty set for each constraint / objective. Thus, we cannot directly take the gradient, as this objective is not differentiable everywhere. To circumvent this, we can employ subgradient methods, which have been heavily used in non-differentiable optimization.

Definition 3 (Definition 3.1 in (Sun et al., 2024)) *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*

$$\liminf_{\substack{\delta \to 0 \\ \delta \neq 0}} \frac{f(x) - f(x) - \langle u, \delta \rangle}{\|\delta\|} \geq 0 \tag{16}$$

When f is differentiable, the subgradient of f is the same as the gradient. Leveraging subgradient methods, we can find the subgradient for the robust average cost MDP.

Lemma 4 (Lemma 3.2 in (Sun et al., 2024)) *Let $d_{\mathcal{P}}^{\pi}$ denote the stationary distribution of the state under the worst-case transition kernel of policy π. Denote the robust Q-function as under policy π as*

$$Q^\pi(s,a) = \max_{\kappa \in \otimes_{t \geq 0} \mathcal{P}} \mathbb{E}\left[\sum_{t=0}^{\infty} \left(r(s_t, a_t) - g_{\mathcal{P}}^\pi \right) \ \Big| \right.$$
$$\left. s_0 = s, \ a_0 = a, \ \pi \right].$$

(17)

Then let $\nabla g_{\mathcal{P}}^\pi$ be the subgradient of $g_{\mathcal{P}}^\pi$, we have

$$\nabla g_{\mathcal{P}}^\pi(s,a) = d_{\mathcal{P}}^\pi Q_{\mathcal{P}}^\pi(s,a) \tag{18}$$

Theorem 5 (Theorem 5.3 in (Xu et al., 2025a)) *Let the robust Q-function under policy π be defined by Eq. (17), then Q^π satisfies the robust Bellman equation*

$$Q^\pi(s,a) = r(s,a) - g_{\mathcal{P}}^\pi + \sigma_{\mathcal{P}_s^a}(V^\pi) \tag{19}$$

where $V^\pi = \sum_a \pi(a|s)Q^\pi(s,a)$ is the robust relative value function, and $g_{\mathcal{P}}^\pi$ is the robust average cost.

Before proving convergence, we need to show that solving our formulation in Eq. (15) is equivalent to solving for the original robust constrained MDP problem (Eq. (13)). This fundamental result is shown in Lemma 6. In Lemma 6, we have two cases: with and without the slackness assumption (Assumption 2).

Assumption 2 (Slackness Assumption) *We assume that $\max_{i \in [1,I]} g_{\mathcal{P}}^{\pi^*,i} - b_i \leq -\zeta$, for some $\zeta > 0$.*

Assumption 2 allows us to ensure exact feasibility of the optimal policy that minimizes $F_{\mathcal{P}}^\pi$ instead of ϵ-feasibility. Additionally, as shown in the proof of Lemma 6, it allows us to decouple λ from ϵ, which improves the sample complexity.

Lemma 6 *If $\hat{\pi}^*$ is the optimal policy of Eq. (15), then $\hat{\pi}^*$ is an $\frac{\epsilon}{2}$-feasible policy and $\frac{\epsilon}{2}$-optimal to the optimal policy π^* of Eq. (13), when $\lambda = 4/\max\{\epsilon, \zeta\}$.*

We show that our formulation's optimal policy cost objective is ϵ-close to the cost objective of the policy optimized for the original RCMDP problem. Furthermore, we are able to ensure that any constraint is violated only by ϵ at the maximum. The proof utilizes the general properties of the objective to show ϵ-optimality and then leverages a proof by contradiction to show that we achieve ϵ-feasibility with and without slackness. Assuming a constraint is violated by more than ϵ is shown to contradict the premise that $\hat{\pi}^*$ is the optimal policy for our objective.

Next, we need a way to attain the optimal policy $\hat{\pi}^*$ through gradient descent, which requires the (sub)gradient of the objective $F_{\mathcal{P}}^\pi$.

Lemma 7 *We can rewrite $\nabla F_{\mathcal{P}}^\pi$ as*

$$\nabla F_{\mathcal{P}}^\pi(s,a) = \tilde{d}_{\mathcal{P}}^\pi Q_{\mathcal{P}}^\pi(s,a) = d_{\mathcal{P}}^{\pi,i_{\max}^\pi} Q_{\mathcal{P}}^{\pi,i_{\max}^\pi}(s,a) \tag{20}$$

9

The proofs of Lemma 2 and 3 are given in Appendix B.

We now possess the required tools and provide the proposed algorithm in Algorithm 1.

Each iteration performs a gradient-based policy update, but the core challenge lies in estimating the gradient itself. Since the theoretical subgradient (Lemma 4) is proportional to the robust Q function, estimating this Q function becomes the most critical task. This is a non-trivial task because the robust Bellman operator is not a contraction under standard norms. Our Actor-Critic framework becomes vital for this, where the critic's role is to produce a stable Q-function estimate.

1. In each iteration of our algorithm, we calculate estimates $g_N^{\pi_t,i}$ and $V_N^{\pi_t,i}$ for the worst case average cost and worst case value function respectively by running Algorithm 2 (our critic) for $N = O(\epsilon^{-2})$ iterations for each of the cost and constraints.

2. Next, we compute the worst-case value function (or support function, $\hat{\sigma}_{\mathcal{P}}$) over the uncertainty set (Contamination, TV and Wasserstein) via Algorithm 3, which implements a variance-reducing Truncated MLMC estimator.

3. Finally, we apply the worst-case Bellman operator (Theorem 5) to obtain the Q-function for each of these components.

With the Q-function for each component estimated, the actor performs the policy update. It uses Lemma 7 to identify the active objective (the most violated constraint or the cost) and selects its corresponding Q-function for the update step .

It is to be noted that since the task of policy evaluation is identical in both constrained and unconstrained settings, we can directly employ these established, sample-efficient algorithms(2, 3) for our critic. Algorithms 2 and 3 are presented in Appendix A.

Algorithm 1 Average-Cost Robust Constrained Actor Critic

1: **Input:** Initial policy π_0; iterations T; learning rate η
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3:      Robust evaluation: estimate $g_N^{\pi_t,i}$, $V_N^{\pi_t,i}(s,a)$ using Algorithm 2 for $i = 0, 1, \cdots I$.
4:      Obtain $\hat{\sigma}_{\mathcal{P}_s^a}\left(V_N^{\pi_t,i}\right)$ using Algorithm 3 for $i = 0, 1, \cdots I$
5:      **for** $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
6:          $\hat{Q}_{\mathcal{P}}^{\pi_t,0}(s,a) = r(s,a) - g_N^{\pi_t,0} + \hat{\sigma}_{\mathcal{P}_s^a}\left(V_N^{\pi_t,0}\right)$
7:          **for** $i \in \{1, 2, \cdots I\}$ **do**
8:              $\hat{Q}_{\mathcal{P}}^{\pi_t,i}(s,a) = c_i(s,a) - g_N^{\pi_t,i} + \hat{\sigma}_{\mathcal{P}_s^a}\left(V_N^{\pi_t,i}\right)$
9:          **end for**
10:      **end for**
11:      Calculate $\hat{Q}_{\mathcal{P}}^{\pi_t}$ from $g_N^{\pi_t,i}$, $\hat{Q}_{\mathcal{P}}^{\pi_t,i}$, for $i \in \{0, 1, \cdots I\}$
12:      $\pi_{t+1} \leftarrow \arg\min_{p \in \Delta(\mathcal{A})} \{\eta \left\langle \hat{Q}_{\mathcal{P}}^{\pi_t}, p \right\rangle + \|p - \pi_t(\cdot|s)\|^2\}$
13: **end for**
14: **return** $\hat{\pi} = \arg\min_{t=0,\cdots,T-1} F_{\mathcal{P}}^{\pi_t}$

---

5 Theoretical Analysis

The critic's role is to estimate the Q-function, $Q_{\mathcal{P}}^{\pi_t}$. Per the analysis of Xu et al. (2025a), Lemma 11 (Appendix C) provides a guarantee that we can obtain an ε accurate estimate of this Q-function with $\tilde{O}(\epsilon^{-2})$ number of samples.

A common method to prove the convergence of policy optimization methods is to form an average of the performance differences between the current policy and the optimal policy,

$$\underset{t=0,1,\dots,T}{\arg\min} \left(F_{\mathcal{P}}^{\pi_t} - F_{\mathcal{P}}^{\hat{\pi}^*} \right) \leq \frac{1}{T} \sum_{t=0}^{T-1} \left(F_{\mathcal{P}}^{\pi_{t+1}} - F_{\mathcal{P}}^{\hat{\pi}^*} \right),$$

which allows us to find an upper bound on the performance difference. However, the performance difference lemma (Lemma 12 in Appendix C) given by Sun et al. (2024) expresses each difference in terms of an expectation under the stationary distribution $d_{\mathcal{P}}^{\pi_t}$ of the current policy. Since the stationary distribution (and hence the expectation) changes with π_t at every step, these terms do not align across iterations and our desired telescoping structure breaks down.

To overcome this difficulty, we introduce a regularity assumption linking performance gaps under the worst-case transition kernel $\mathcal{P}$ to those under the nominal kernel P°:

Assumption 3 *For all policies $\pi \in \Pi$)*

$$g_{\mathcal{P}}^{\pi} - g_{\mathcal{P}}^{\hat{\pi}^*} \leq C\mathbb{E}_{s \sim d_{P^\circ}^{\pi}} \left[\langle Q_P^{\pi}(s, \cdot), \pi(\cdot|s) - \hat{\pi}^*(\cdot|s) \rangle \right] \tag{21}$$

This assumption extends the robust performance difference lemma by relating the worst-case performance gap to the nominal kernel's stationary distribution. Intuitively, it asserts that the degradation in performance under the worst-case model cannot exceed that under the nominal model by more than a constant multiplicative factor C. A related assumption is common in the discounted robust MDP setting (Tamar et al. (2014); Zhou et al. (2023); Ganguly et al. (2025)), which states: $\gamma p(s'|s, a) = \beta p_0(s'|s, a)$ for some $\beta \in (0, 1)$. We notice that if $\gamma = 1$, the assumption does not hold anymore (a trivial counterexample is when $s' = s$). Thus, we cannot leverage an assumption of the same form for our average cost setting. However, it is to be noted that our assumption is not completely arbitrary and is grounded in the assumption made by the discounted RMDP literature. A detailed equivalence relation is provided in Appendix C. We now have the required tools to state and prove our main theorem on convergence:

Theorem 8 *Using a stepsize of $\eta = O(\epsilon)$, Algorithm 1 returns a policy $\hat{\pi}$ that is both ϵ-feasible and ϵ-optimal after $T = \tilde{O}(\epsilon^{-2}\lambda^2)$ iterations.*

We know from Xu et al. (2025a) that the critic requires $\tilde{O}(\epsilon^{-2})$ samples. Therefore, if we assume the slackness condition, we obtain an iteration complexity of $T = O(\epsilon^{-2}\zeta^{-2})$ and a corresponding sample complexity of $O(\epsilon^{-4}\zeta^{-2})$. If we do not assume slackness, we obtain $T = O(\epsilon^{-4})$ and a sample complexity of $O(\epsilon^{-6})$.

Theorem 8 establishes that Algorithm 1 converges to a policy that is simultaneously ϵ-optimal and ϵ-feasible with sample complexities comparable to existing discounted reward settings.

The convergence analysis is composed of three key steps:

- The proof begins by invoking our critical Assumption 3. This step allows us to shift the analysis to the more tractable non-robust setting where the stationary distribution is fixed with respect to a single optimal policy. We then utilize the standard result of the non-robust performance difference lemma, yielding an expectation over the Q-function.

- We decompose the inner product into two parts corresponding to the policy change in one step and the remaining gap to the optimal policy. We then leverage the three-point lemma of Bregman Divergence and Holder's inequality to obtain a telescoping sum of the form

$$\cdots \leq \frac{1}{\eta}(\|\hat{\pi}^*(\cdot|s) - \pi_t(\cdot|s)\|^2 - \|\hat{\pi}^*(\cdot|s) - \pi_{t+1}(\cdot|s)\|^2)$$
$$+ \text{error-terms}$$

- Finally, we sum the inequality over all iterations $t = 0, \cdots, T - 1$. The telescoping terms cancel out, leaving a bound on the average performance gap that depends on the initial policy distance, the step size η, and the critic estimation error ε. Using the fact that the minimum of a distribution is at most the average, we obtain the above mentioned iteration complexities.

6 Conclusion

In this work, we present an actor critic algorithm to solve the robust constrained average cost MDP problem. We show that our algorithm outputs an ϵ-feasible and ϵ-optimal policy with a sample complexity of $O(\epsilon^{-4})$ when using the slackness assumption and $O(\epsilon^{-6})$ when not using the slackness assumption. Not only are we the first algorithm to tackle this specific setting, but we also obtain equal sample complexity guarantees with existing discounted RCMDP algorithms.

Weakening the assumptions in this work while preserving the same sample complexity is an important avenue for future research. Furthermore, a gap persists between the sample complexity results for robust constrained and robust unconstrained settings in both average- and discounted-reward cases. Closing this gap, either through the development of improved algorithms and refined analyses or by establishing tighter lower bounds on sample complexity, constitutes another significant direction for future work.

References

Mridul Agarwal, Qinbo Bai, and Vaneet Aggarwal. Concave utility reinforcement learning with zero-constraint violations. *Transactions on Machine Learning Research*, 2022a. ISSN 2835-8856.

Mridul Agarwal, Qinbo Bai, and Vaneet Aggarwal. Regret guarantees for model-based reinforcement learning with long-term average constraints. In *Uncertainty in Artificial Intelligence*, pages 22–31. PMLR, 2022b.

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

Eitan Altman. *Constrained Markov decision processes.* Routledge, 2021.

Qinbo Bai, Amrit Singh Bedi, Mridul Agarwal, Alec Koppel, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3682–3689, 2022.

Qinbo Bai, Amrit Singh Bedi, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via conservative natural policy gradient primal-dual algorithm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6737–6744, 2023.

Qinbo Bai, Washim Mondal, and Vaneet Aggarwal. Learning general parameterized policies for infinite horizon average reward constrained mdps via primal-dual policy gradient algorithm. *Advances in Neural Information Processing Systems*, 37:108566–108599, 2024.

Jiayu Chen, Tian Lan, and Vaneet Aggarwal. Option-aware adversarial inverse reinforcement learning for robotic control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5902–5908. IEEE, 2023.

Liyu Chen, Rahul Jain, and Haipeng Luo. Learning infinite-horizon average-reward markov decision process with constraints. In *International Conference on Machine Learning*, pages 3246–3270. PMLR, 2022.

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

Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*, 2018.

Sourav Ganguly, Arnob Ghosh, Kishan Panaganti, and Adam Wierman. Efficient policy optimization in robust constrained mdps with iteration complexity guarantees. *arXiv preprint arXiv:2505.19238*, 2025.

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

Mudit Gaur, Utsav Singh, Amrit Singh Bedi, Raghu Pasupathu, and Vaneet Aggarwal. On the sample complexity bounds in bilevel reinforcement learning. *arXiv preprint arXiv:2503.17644*, 2025.

Michael Horstein. Sequential transmission using noiseless feedback. *IEEE Transactions on Information Theory*, 9(3):136–143, 2003.

Garud N Iyengar. Robust dynamic programming. *Math. Oper. Res.*, 30(2):257–280, May 2005.

Toshinori Kitamura, Tadashi Kozuno, Wataru Kumagai, Kenta Hoshino, Yohei Hosoe, Kazumi Kasaura, Masashi Hamaya, Paavo Parmas, and Yutaka Matsuo. Near-optimal policy identification in robust constrained markov decision processes via epigraph form. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

Yongshuai Liu, Jiaxin Ding, and Xin Liu. Ipo: Interior-point policy optimization under constraints. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4940–4947, 2020.

Shaocong Ma, Ziyi Chen, Yi Zhou, and Heng Huang. Rectified robust policy optimization for model-uncertain constrained reinforcement learning without strong duality. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Daniel J Mankowitz, Dan A Calian, Rae Jeong, Cosmin Paduraru, Nicolas Heess, Sumanth Dathathri, Martin Riedmiller, and Timothy Mann. Robust constrained reinforcement learning for continuous control with model misspecification. *arXiv preprint arXiv:2010.10644*, 2020.

Washim U Mondal and Vaneet Aggarwal. Sample-efficient constrained reinforcement learning with general parameterization. *Advances in Neural Information Processing Systems*, 37:68380–68405, 2024.

Arnab Nilim and Laurent Ghaoui. Robustness in markov decision problems with uncertain transition matrices. In *Advances in Neural Information Processing Systems*, 2003.

Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pages 9582–9602. PMLR, 2022.

Santiago Paternain, Luiz Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. *Advances in Neural Information Processing Systems*, 32, 2019.

Santiago Paternain, Miguel Calvo-Fullana, Luiz FO Chamon, and Alejandro Ribeiro. Safe policies for reinforcement learning via primal-dual methods. *IEEE Transactions on Automatic Control*, 68(3):1321–1336, 2022.

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.

Reazul Hasan Russel, Mouhacine Benosman, and Jeroen Van Baar. Robust constrained-mdps: Soft-constrained robust policy optimization under model uncertainty. *arXiv preprint arXiv:2010.04870*, 2020.

Zhongchang Sun, Sihong He, Fei Miao, and Shaofeng Zou. Policy optimization for robust average reward mdps. In *Advances in Neural Information Processing Systems*, pages 17348–17372, 2024.

Aviv Tamar, Shie Mannor, and Huan Xu. Scaling up robust mdps using function approximation. In *International conference on machine learning*, pages 181–189. PMLR, 2014.

Qiuhao Wang, Chin Pang Ho, and Marek Petrik. Policy gradient in robust mdps with global convergence guarantee, 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, 2024.

Yue Wang, Fei Miao, and Shaofeng Zou. Robust constrained reinforcement learning. *arXiv preprint arXiv:2209.06866*, 2022.

Yue Wang, Alvaro Velasquez, George Atia, Ashley Prater-Bennette, and Shaofeng Zou. Robust average-reward markov decision processes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):15215–15223, Jun. 2023b.

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, 2023c.

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, 2023d.

Honghao Wei, Xin Liu, and Lei Ying. A provably-efficient model-free algorithm for infinite-horizon average-reward constrained markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3868–3876, 2022.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Tengyu Xu, Yingbin Liang, and Guanghui Lan. Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *Proceedings of the 38th International Conference on Machine Learning*, pages 11480–11491, 2021.

Yang Xu, Swetha Ganesh, and Vaneet Aggarwal. Efficient q-learning and actor-critic methods for robust average reward reinforcement learning. *arXiv preprint arXiv:2506.07040*, 2025a.

Yang Xu, Swetha Ganesh, Washim Uddin Mondal, Qinbo Bai, and Vaneet Aggarwal. Global convergence for average reward constrained mdps with primal-dual actor critic algorithm. *arXiv preprint arXiv:2505.15138*, 2025b.

Yang Xu, Washim Uddin Mondal, and Vaneet Aggarwal. Finite-sample analysis of policy evaluation for robust average reward reinforcement learning. *arXiv preprint arXiv:2502.16816*, 2025c.

Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J Ramadge. Projection-based constrained policy optimization. *arXiv preprint arXiv:2010.03152*, 2020.

Zhengfei Zhang, Kishan Panaganti, Laixi Shi, Yanan Sui, Adam Wierman, and Yisong Yue. Distributionally robust constrained reinforcement learning under strong duality. In *Reinforcement Learning Conference*, 2024.

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*, 36:97–133, 2023.

Zhengqing Zhou, Zhengyuan Zhou, Qinxun Bai, Linhai Qiu, Jose Blanchet, and Peter Glynn. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3331–3339. PMLR, 2021.

Appendix A. Missing Algorithms from Section 4

Algorithm 2 Robust average cost TD (Algorithm 2 in Xu et al. (2025c))

1: **Input:** Policy π, Initial values V_0, $g_0 = 0$, Stepsizes η_t, β_t, Max level $N_{\max}$, Anchor state $s_0 \in \mathcal{S}$
2: **for** $t = 0, 1, \ldots, T - 1$ **do**
3:     **for** each $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**
4:         **if** Contamination **then**
5:             Sample $\hat{\sigma}_{P_{s,a}}(V_t)$ according to Eq. 8
6:         **else if** TV or Wasserstein **then**
7:             Sample $\hat{\sigma}_{P_{s,a}}(V_t)$ according to Algorithm 3
8:         **end if**
9:     **end for**
10:     $\hat{T}_{g_0}(V_t)(s) \leftarrow \sum_a \pi(a|s) \left[r(s, a) - g_0 + \hat{\sigma}_{P_{s,a}}(V_t) \right], \quad \forall s \in \mathcal{S}$
11:     $V_{t+1}(s) \leftarrow V_t(s) + \eta_t \left(\hat{T}_{g_0}(V_t)(s) - V_t(s) \right), \quad \forall s \in \mathcal{S}$
12:     $V_{t+1}(s) \leftarrow V_{t+1}(s) - V_{t+1}(s_0), \quad \forall s \in \mathcal{S}$
13: **end for**
14: **for** $t = 0, 1, \ldots, T - 1$ **do**
15:     **for** each $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**
16:         **if** Contamination **then**
17:             Sample $\hat{\sigma}_{P_{s,a}}(V_t)$ according to Eq. 8
18:         **else if** TV or Wasserstein **then**
19:             Sample $\hat{\sigma}_{P_{s,a}}(V_t)$ according to Algorithm 3
20:         **end if**
21:     **end for**
22:     $\delta_t(s) \leftarrow \sum_a \pi(a|s) \left[r(s, a) + \hat{\sigma}_{P_{s,a}}(V_T) \right] - V_T(s), \quad \forall s \in \mathcal{S}$
23:     $\bar{\delta}_t \leftarrow \frac{1}{S} \sum_s \delta_t(s)$
24:     $g_{t+1} \leftarrow g_t + \beta_t(\bar{\delta}_t - g_t)$
25: **end for**
26: **return** V_T, g_T

Algorithm 3 Truncated MLMC Estimator, Algorithm 1 in Xu et al. (2025c)

1: **Input:** $s \in \mathcal{S}$, $a \in \mathcal{A}$, Max level $N_{\max}$, Value function V
2: Sample $N \sim \text{Geom}(0.5)$
3: $N' \leftarrow \min\{N, N_{\max}\}$
4: Collect $2^{N'+1}$ i.i.d. samples of $\{s_i'\}_{i=1}^{2^{N'+1}}$ with $s_i' \sim \tilde{P}_s^a$ for each i
5: $\hat{P}_{s,N'+1}^{a,E} \leftarrow \frac{1}{2^{N'}} \sum_{i=1}^{2^{N'}} \mathbf{1}\{s_{2i}'\}$
6: $\hat{P}_{s,N'+1}^{a,O} \leftarrow \frac{1}{2^{N'}} \sum_{i=1}^{2^{N'}} \mathbf{1}\{s_{2i-1}'\}$
7: $\hat{P}_{s,N'+1}^{a} \leftarrow \frac{1}{2^{N'+1}} \sum_{i=1}^{2^{N'+1}} \mathbf{1}\{s_i'\}$
8: $\hat{P}_{s,1}^{a} \leftarrow \mathbf{1}\{s_1'\}$
9: **if** TV **then**
10:     Obtain $\sigma_{\hat{P}_{s,1}^a}(V)$, $\sigma_{\hat{P}_{s,N'+1}^a}(V)$, $\sigma_{\hat{P}_{s,N'+1}^{a,E}}(V)$, $\sigma_{\hat{P}_{s,N'+1}^{a,O}}(V)$ from Eq. 9
11: **else if** Wasserstein **then**
12:     Obtain $\sigma_{\hat{P}_{s,1}^a}(V)$, $\sigma_{\hat{P}_{s,N'+1}^a}(V)$, $\sigma_{\hat{P}_{s,N'+1}^{a,E}}(V)$, $\sigma_{\hat{P}_{s,N'+1}^{a,O}}(V)$ from Eq. 10
13: **end if**
14: $\Delta_{N'}(V) \leftarrow \sigma_{\hat{P}_{s,N'+1}^a}(V) - \frac{1}{2}\left[\sigma_{\hat{P}_{s,N'+1}^{a,E}}(V) + \sigma_{\hat{P}_{s,N'+1}^{a,O}}(V)\right]$
15: $\hat{\sigma}_{\mathcal{P}_s^a}(V) \leftarrow \sigma_{\hat{P}_{s,1}^a}(V) + \frac{\Delta_{N'}(V)}{\mathbb{P}(N'=n)}$, where $p'(n) = \mathbb{P}(N'=n)$
16: **return** $\hat{\sigma}_{\mathcal{P}_s^a}(V)$

Appendix B. Missing Proofs for Section 4

Lemma 9 (Proof of Lemma 6) Proof *Let $\hat{\pi}^*$ be the policy that minimizes our smoothed objective F. Then we have that the optimality difference between π^* (true optimal policy of Eq 13) and $\hat{\pi}^*$ is*

$$\frac{g_{\mathcal{P}}^{\hat{\pi}^*,0}}{\lambda} - \frac{g_{\mathcal{P}}^{\pi^*,0}}{\lambda} \overset{(a)}{\leq} \max_i G_{\mathcal{P}}^{\hat{\pi}^*} - \max_i G_{\mathcal{P}}^{\pi^*} \tag{22}$$

$$\overset{(b)}{\leq} 0 \tag{23}$$

where (a) is from the definition of $\max$ and because π^* is a feasible policy, and (b) is by the optimality of $\hat{\pi}^*$. Then to prove feasibility, we have two cases.

Case 1 (No slackness): *By contradiction, assume optimal policy $\hat{\pi}^*$ violates the constraints by more than $\frac{\epsilon}{2}$, where we set the slackness coefficient ζ to 0:*

$$\max_{i \in \{1, \cdots I\}} \left\{ g_{\mathcal{P}}^{\hat{\pi}^*,i} - b_i + \zeta \right\} = \max_{i \in \{1, \cdots I\}} \left\{ g_{\mathcal{P}}^{\hat{\pi}^*,i} - b_i \right\} > \frac{\epsilon}{2} \tag{24}$$

We set the hyperparameter $\lambda = \frac{4}{\epsilon}$. The maximum average cost for the objective, $g_{\mathcal{P}}^{\hat{\pi}^*,i}$, is bounded by 1 because the cost function is bounded by 1. Similar logic holds for π^*. Therefore, the objective cost term satisfies:

$$\frac{g_{\mathcal{P}}^{\hat{\pi}^*,0}}{\lambda} \leq \frac{1}{\lambda} = \frac{\epsilon}{4}, \quad \frac{g_{\mathcal{P}}^{\pi^*,0}}{\lambda} \leq \frac{1}{\lambda} = \frac{\epsilon}{4} \tag{25}$$

Combining (24) and (25) yields

$$\max \left\{ \frac{g_{\mathcal{P}}^{\hat{\pi}^*,0}}{\lambda}, \max_i \left\{ g_{\mathcal{P}}^{\hat{\pi}^*,i} - b_i \right\} \right\} = \max_i G_{\mathcal{P}}^{\hat{\pi}^*} > \frac{\epsilon}{2}. \tag{26}$$

Thus, we have

$$\max_i G_{\mathcal{P}}^{\pi^*} = \max \left\{ \frac{g_{\mathcal{P}}^{\pi^*,0}}{\lambda}, \max_i \left\{ g_{\mathcal{P}}^{\pi^*,i} - b_i \right\} \right\} \leq \frac{\epsilon}{4} < \frac{\epsilon}{2} = \max_i G_{\mathcal{P}}^{\hat{\pi}^*} \tag{27}$$

which is a contradiction as by definition, $G_{\mathcal{P}}^{\hat{\pi}^*} \leq G_{\mathcal{P}}^{\pi^*}$. Therefore, the maximum violation is at most $\epsilon/2$ with $\lambda = \frac{4}{\epsilon}$.

Case 2 (With Slackness): *Assumption 2 gives us a way to prove exact feasibility of $\hat{\pi}^*$ by choosing $\lambda = \frac{4}{\zeta}$, assuming that $\zeta > \epsilon$. We again assume by contradiction that the optimal policy $\hat{\pi}^*$ violates the constraints by more than $\frac{\epsilon}{2}$:*

$$\max_{i \in \{1, \cdots I\}} \left\{ g_{\mathcal{P}}^{\hat{\pi}^*,i} - b_i + \zeta \right\} > \frac{\zeta}{2} \tag{28}$$

Then following the same logic in Case 1, we have $\frac{g_{\mathcal{P}}^{\hat{\pi}^*,0}}{\lambda} \leq \frac{\zeta}{4}$ and

$$\max \left\{ \frac{g_{\mathcal{P}}^{\hat{\pi}^*,0}}{\lambda}, \max_i \left\{ g_{\mathcal{P}}^{\hat{\pi}^*,i} - b_i + \zeta \right\} \right\} = \max_i G_{\mathcal{P}}^{\hat{\pi}^*} > \frac{\zeta}{2}. \tag{29}$$

This yields

$$\max_i G_{\mathcal{P}}^{\pi^*} = \max \left\{ \frac{g_{\mathcal{P}}^{\pi^*,0}}{\lambda}, \max_i \left\{ g_{\mathcal{P}}^{\pi^*,i} - b_i \right\} \right\} \le \frac{\zeta}{4} < \frac{\zeta}{2} = \max_i G_{\mathcal{P}}^{\hat{\pi}^*} \tag{30}$$

which again is a contradiction. Thus, the original assumption is false and

$$\max_{i \in \{1, \cdots I\}} \left\{ g_{\mathcal{P}}^{\hat{\pi}^*,i} - b_i + \zeta \right\} \le \frac{\zeta}{2} \implies \max_{i \in \{1, \cdots I\}} \left\{ g_{\mathcal{P}}^{\hat{\pi}^*,i} - b_i \right\} \le -\frac{\zeta}{2} \le 0 \tag{31}$$

Thus, with the slackness assumption, we obtain exact feasibility. ∎

Lemma 10 (Restatement of Lemma 7) *We can rewrite $\nabla F_{\mathcal{P}}^{\pi}$ as*

$$\nabla F_{\mathcal{P}}^{\pi}(s,a) = \tilde{d}_{\mathcal{P}}^{\pi} Q_{\mathcal{P}}^{\pi}(s,a) = d_{\mathcal{P}}^{\pi,i_{\max}^{\pi}} Q_{\mathcal{P}}^{\pi,i_{\max}^{\pi}}(s,a) \tag{32}$$

Proof *We first look at the inner product between the subgradient of $F_{\mathcal{P}}^{\pi}$ and π*

$$\langle \nabla F_{\mathcal{P}}^{\pi}, \pi \rangle \overset{(a)}{=} \sum_{i=0}^{I} w_i^{\pi} \left\langle \nabla G_{\mathcal{P}}^{\pi,j}, \pi \right\rangle \tag{33}$$

$$\overset{(b)}{=} \sum_{j=0}^{I} w_j^{\pi} \sum_s d_{\mathcal{P}}^{\pi,j}(s) \sum_a Q_{\mathcal{P}}^{\pi,j}(s,a) \, \pi(a \mid s) \tag{34}$$

where (a) uses the definition of the derivative of the objective, and in (b), we use the definition of the subgradient from Lemma 4. Fixing a state s and looking only at the coefficient of $\pi(a|s)$ yields

$$\sum_{i=0}^{I} w_i^{\pi} d_{\mathcal{P}}^{\pi,i}(s) Q_{\mathcal{P}}^{\pi,i}(s,a) = \sum_{i=0}^{I} w_i^{\pi} d_{\mathcal{P}}^{\pi,i}(s) Q_{\mathcal{P}}^{\pi,i}(s,a) \tag{35}$$

$$= \sum_{i=0}^{I} d_{\mathcal{P}}^{\pi,i_{\max}^{\pi}} 1_{i=i_{\max}^{\pi}} Q_{\mathcal{P}}^{\pi,i_{\max}^{\pi}}(s,a) \tag{36}$$

$$= d_{\mathcal{P}}^{\pi,i_{\max}^{\pi}} Q_{\mathcal{P}}^{\pi,i_{\max}^{\pi}}(s,a) \tag{37}$$

which is the desired result. ∎

Appendix C. Missing Lemmas and Proofs for Section 5

Lemma 11 (Xu et al. (2025a)) *We have that the expected estimation error between the true smoothed Q-function and our estimate is bounded by ε with a sample complexity of $O(\epsilon^{-2})$.*

$$\mathbb{E}\left[\left\| Q_{\mathcal{P}}^{\pi_t}(s,\cdot) - \hat{Q}_{\mathcal{P}}^{\pi_t}(s,\cdot) \right\|_{\infty} \right] \le \varepsilon \tag{38}$$

Next, we define the performance difference lemma for robust average cost MDPs from (Sun et al., 2024).

Lemma 12 (Lemma 4.1 in (Sun et al., 2024)) *For any two policies π, π', we have that*

$$g_{\mathcal{P}}^{\pi} - g_{\mathcal{P}}^{\pi'} \geq \mathbb{E}_{s \sim d_{\mathcal{P}}^{\pi'}} \left[\langle Q_{\mathcal{P}}^{\pi}(s, \cdot), \pi(\cdot|s) - \pi'(\cdot|s) \rangle \right] \tag{39}$$

where $d_{\mathcal{P}}^{\pi'}$ denotes the stationary distribution under the worst-case transition kernel of policy π'.

C.1 A detailed motivation of Assumption 3

We draw a direct connection between our Assumption 3 and the corresponding assumption commonly employed in the *discounted robust MDP* literature.

- In the discounted setting, **Lemma B.3** of Ganguly et al. (2025) is central to the convergence analysis and follows as a *direct consequence* of the assumption $\gamma p(s' \mid s, a) \leq \beta p_0(s' \mid s, a)$ for all s', s, a, where $\beta \in (0, 1)$. This multiplicative dominance condition induces a contraction in the state-transition dynamics, which leads to the bound

$$\Phi(\pi) - \Phi(\pi^*) \leq \frac{1}{1 - \beta} \mathbb{E}_{s \sim d_{P^\circ}^{\pi^*}} \left[\langle Q_{\mathcal{P}}^{\pi_t}(s, \cdot), \pi_t(\cdot \mid s) - \pi^*(\cdot \mid s) \rangle \right], \tag{40}$$

  where Φ denotes their discounted objective function. The factor $\frac{1}{1-\beta}$ quantifies the effective discount-induced dependence between states.

- In contrast, in the *average reward* robust MDP setting there is no discount factor, making the bound in (40) inapplicable since no contraction holds. To address this, we introduce Assumption 3, which serves as an analogous regularity condition:

$$g_{\mathcal{P}}^{\pi} - g_{\mathcal{P}}^{\hat{\pi}^*} \leq C \, \mathbb{E}_{s \sim d_{P^\circ}^{\pi}} \left[\langle Q_P^{\pi}(s, \cdot), \pi(\cdot \mid s) - \hat{\pi}^*(\cdot \mid s) \rangle \right]. \tag{41}$$

  Here, the constant $C \geq 1$ replaces the role of the geometric factor $\frac{1}{1-\beta}$ from the discounted setting. Intuitively, C captures the extent to which the stationary distribution under the worst-case transition kernel $\mathcal{P}$ can differ from that under the nominal kernel P°, thus providing a measure of distributional regularity in the absence of discounting.

  This substitution generalizes the discounted assumption to the average-reward regime by replacing the explicit contraction (through β) with a bounded performance coupling constant C, ensuring that the worst-case performance degradation remains controlled relative to the nominal dynamics.

Lemma 13 (Restatement of Theorem 4) *Using a stepsize of $\eta = O(\epsilon)$, Algorithm 1 returns a policy $\hat{\pi}$ that is both ϵ-feasible and ϵ-optimal after $T = \frac{18C^2 Q_{\max}^2 \Delta}{\epsilon^2} \lambda^2$ iterations.*

Proof

By the non-robust performance difference lemma between the optimal policy and the current policy, we have

$$F_{\mathcal{P}}^{\pi_t} - F_{\mathcal{P}}^{\hat{\pi}^*} = g_{\mathcal{P}}^{\pi_t,i} - g_{\mathcal{P}}^{\hat{\pi}^*,i} \tag{42}$$

$$\leq g_{P_{\pi_t}}^{\pi_t,i} - g_{P_{\pi_t}}^{\hat{\pi}^*,i} \tag{43}$$

$$\overset{(a)}{=} \mathbb{E}_{s \sim d_{P^\circ}^{\hat{\pi}^*}} \left[\langle Q_{\mathcal{P}}^{\pi_t}(s,\cdot), \pi_t(\cdot|s) - \hat{\pi}^*(\cdot|s) \rangle \right] \tag{44}$$

$$= \mathbb{E}_{s \sim d_{P^\circ}^{\hat{\pi}^*}} \left[\langle Q_{\mathcal{P}}^{\pi_t}(s,\cdot), \pi_t(\cdot|s) - \pi_{t+1}(\cdot|s) \rangle \right] + \mathbb{E}_{s \sim d_{P^\circ}^{\hat{\pi}^*}} \left[\langle Q_{\mathcal{P}}^{\pi_t}(s,\cdot), \pi_{t+1}(\cdot|s) - \hat{\pi}^*(\cdot|s) \rangle \right] \tag{45}$$

where (a) uses Assumption 3. From the three point lemma of Bregman Divergence

$$F_{\mathcal{P}}^{\pi_t} - F_{\mathcal{P}}^{\hat{\pi}^*} \leq \mathbb{E}_{s \sim d_{P^\circ}^{\hat{\pi}^*}} \Big[\langle Q_{\mathcal{P}}^{\pi_t}(s,\cdot), \pi_t(\cdot|s) - \pi_{t+1}(\cdot|s) \rangle - \tfrac{1}{\eta}\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|^2$$
$$+ \tfrac{1}{\eta}\|\hat{\pi}^*(\cdot|s) - \pi_t(\cdot|s)\|^2 - \tfrac{1}{\eta}\|\hat{\pi}^*(\cdot|s) - \pi_{t+1}(\cdot|s)\|^2 + 2\varepsilon \Big] \tag{46}$$

From Holder's inequality, we have

$$\langle Q_{\mathcal{P}}^{\pi_t}(s,\cdot), \pi_t(\cdot|s) - \pi_{t+1}(\cdot|s) \rangle - \tfrac{1}{\eta}\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|^2 \leq \|Q_{\mathcal{P}}^{\pi_t}(s,\cdot)\|_\infty \|\pi_t(\cdot|s) - \pi_{t+1}(\cdot|s)\|_1$$
$$- \tfrac{1}{2\eta}\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_2^2 \tag{47}$$

By adding and subtracting $\frac{\eta}{2}\|Q_{\mathcal{P}}^{\pi_t}(s,\cdot)\|_\infty^2$ and using the fact that $\|\pi(\cdot|s) - \pi'(\cdot|s)\|_1 \geq \|\pi(\cdot|s) - \pi'(\cdot|s)\|_2$, we get

$$\langle Q_{\mathcal{P}}^{\pi_t}(s,\cdot), \pi_t(\cdot|s) - \pi_{t+1}(\cdot|s) \rangle - \tfrac{1}{\eta}\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|^2 \leq \frac{-1}{2\eta}(\eta\|Q_{\mathcal{P}}^{\pi_t}(s,\cdot)\|_\infty - \|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|)^2$$
$$+ \frac{\eta}{2}\|Q_{\mathcal{P}}^{\pi_t}(s,\cdot)\|_\infty^2$$
$$\leq \frac{\eta}{2}\|Q_{\mathcal{P}}^{\pi_t}(s,\cdot)\|_\infty^2 \tag{48}$$

Now summing equation 46 over t and taking the average, we have

$$\sum_t (F_{\mathcal{P}}^{\pi_t} - F_{\mathcal{P}}^{\hat{\pi}^*}) \leq \sum_{t=0}^{T-1} \left(\mathbb{E}_{s \sim d_{P^\circ}^{\hat{\pi}^*}} \frac{\eta}{2}\|Q_{\mathcal{P}}^{\pi_t}(s,\cdot)\|_\infty^2 + \frac{1}{\eta}\mathbb{E}_{s \sim d_{P^\circ}^{\hat{\pi}^*}} \left[\|\hat{\pi}^*(\cdot|s) - \pi_t(\cdot|s)\|^2 - \|\hat{\pi}^*(\cdot|s) - \pi_{t+1}(\cdot|s)\|^2 \right] + 2\varepsilon \right) \tag{49}$$

$$\leq \frac{CT\eta Q_{\max}^2}{2} + \frac{C}{\eta}\mathbb{E}_{s \sim d_{P^\circ}^{\hat{\pi}^*}} \|\hat{\pi}^*(\cdot|s) - \pi_0(\cdot|s)\|^2 + 2CT\varepsilon \tag{50}$$

where C is the distribution mismatch coefficient. Let $\hat{\pi}$ be the output of our algorithm where $\hat{\pi} = \arg\min_{t=0,\cdots,T-1} F_{\mathcal{P}}^{\pi_t}$. Then we have

$$F_{\mathcal{P}}^{\hat{\pi}} - F_{\mathcal{P}}^{\hat{\pi}^*} \leq \frac{1}{T}\sum_t (F_{\mathcal{P}}^{\pi_t} - F_{\mathcal{P}}^{\hat{\pi}^*}) \tag{51}$$

$$\leq \frac{C\eta Q_{\max}^2}{2} + \frac{C}{T\eta}\mathbb{E}_{s \sim d_{P^\circ}^{\hat{\pi}^*}} \|\hat{\pi}^*(\cdot|s) - \pi_0(\cdot|s)\|^2 + 2C\varepsilon \leq \frac{\epsilon}{2} \tag{52}$$

22

First, we note that ε can be made arbitrarily small. From Xu et al. (2025a) and running the critic estimate (Algorithm 2) with $O(\epsilon^{-2})$ samples, we can obtain $\varepsilon = O(\epsilon)$.

Solving for T, by equating each of the three terms to $\epsilon/6$, yields $T = \frac{18C^2 Q_{\max}^2 \Delta}{\epsilon^2}$ using a step size $\eta = \frac{\epsilon}{2CQ_{\max}^2}$, where $\Delta = \|\hat{\pi}^*(\cdot|s) - \pi_0(\cdot|s)\|^2$.

However, to achieve both ϵ-optimality and ϵ-feasibility, we must run the algorithm until $g_{\mathcal{P}}^{\hat{\pi},0} - g_{\mathcal{P}}^{\hat{\pi}^*,0} \leq \epsilon$. Now, we have

$$\frac{g_{\mathcal{P}}^{\hat{\pi},0}}{\lambda} - \frac{g_{\mathcal{P}}^{\hat{\pi}^*,0}}{\lambda} \leq F_{\mathcal{P}}^{\hat{\pi}} - F_{\mathcal{P}}^{\hat{\pi}^*} \leq \frac{\epsilon}{\lambda} \tag{53}$$

Thus, number of iterations needed is $\frac{18C^2 Q_{\max}^2 \Delta}{\epsilon^2} \lambda^2$.
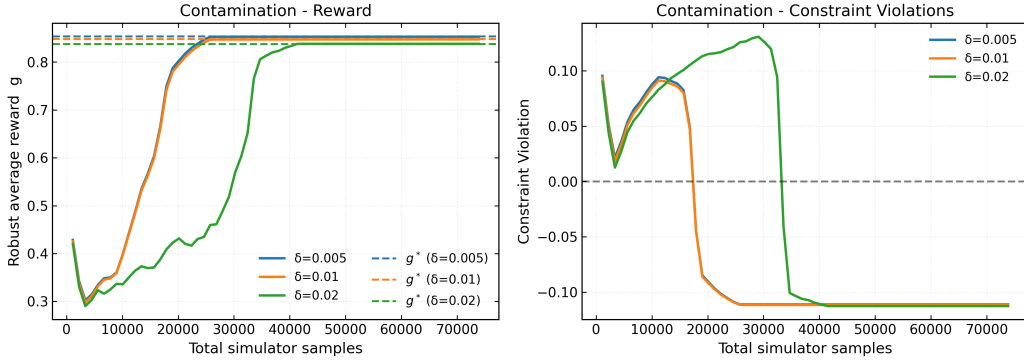
∎

Appendix D. Numerical Experiments

Figure 1: Performance of the Robust Constrained Average-Cost Actor-Critic algorithm under the Contamination uncertainty set.
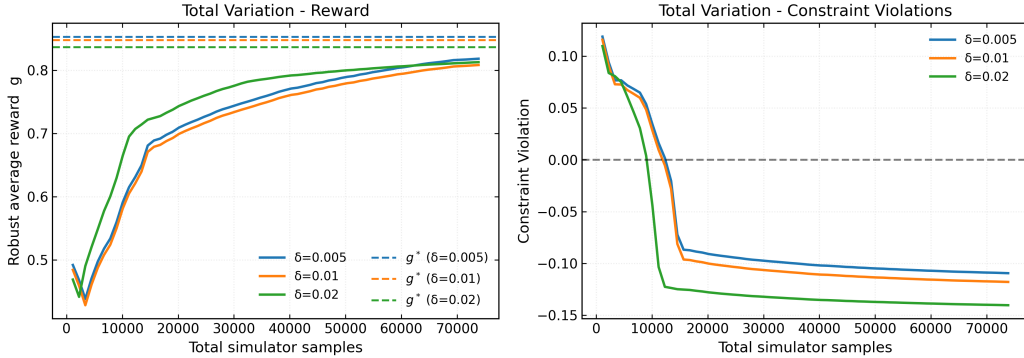

Figure 2: Performance of the Robust Constrained Average-Cost Actor-Critic algorithm under the Total Variation (TV) uncertainty set.
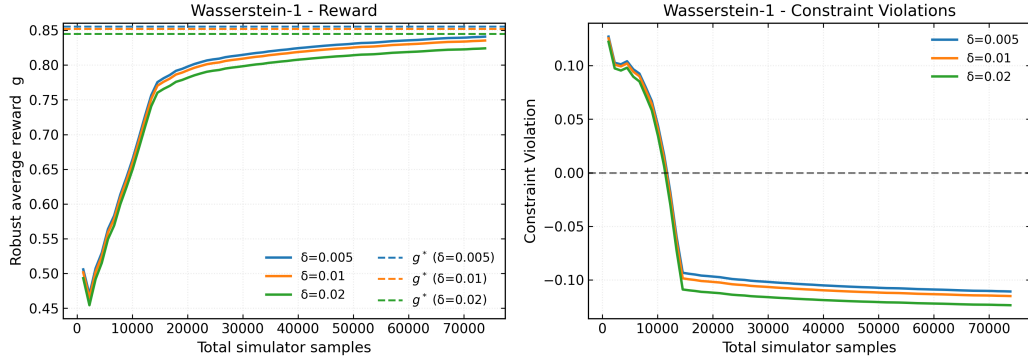
Figure 3: Performance of the Robust Constrained Average-Cost Actor-Critic algorithm under the Wasserstein uncertainty set.
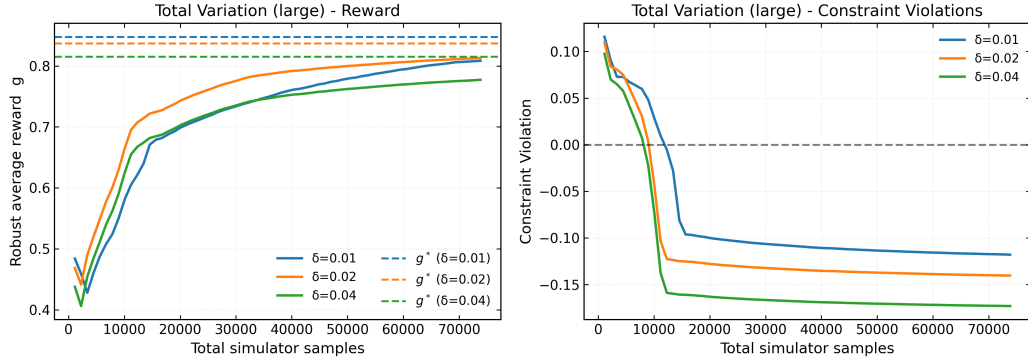

Figure 4: Performance of the Robust Constrained Average-Cost Actor-Critic algorithm under the TV uncertainty set with larger uncertainty set radii.
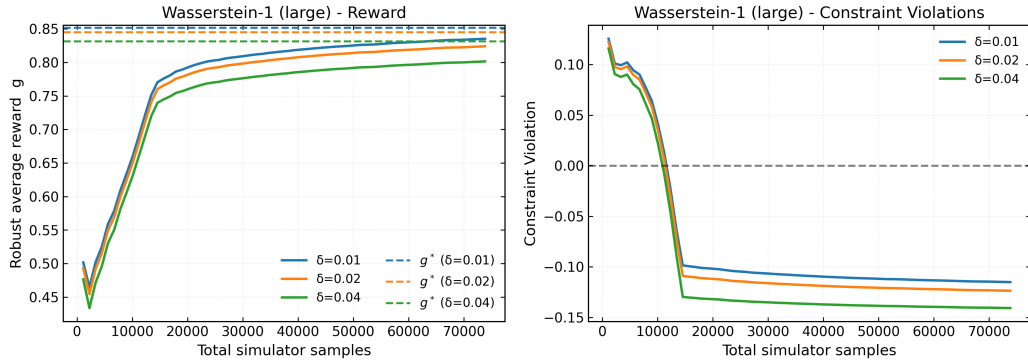

Figure 5: Performance of the Robust Constrained Average-Cost Actor-Critic algorithm under the Wasserstein uncertainty set with larger uncertainty set radii.