Policy Gradient Method For Robust Reinforcement Learning

Yue Wang [1] **Shaofeng Zou** [1]

Abstract

This paper develops the first policy gradient method with global optimality guarantee and complexity analysis for robust reinforcement learning under model mismatch. Robust reinforcement learning is to learn a policy robust to model mismatch between simulator and real environment. We first develop the robust policy (sub-)gradient, which is applicable for any differentiable parametric policy class. We show that the proposed robust policy gradient method converges to the global optimum asymptotically under direct policy parameterization. We further develop a smoothed robust policy gradient method, and show that to achieve an ϵ-global optimum, the complexity is $\mathcal{O}(\epsilon^{-3})$. We then extend our methodology to the general model-free setting, and design the robust actor-critic method with differentiable parametric policy class and value function. We further characterize its asymptotic convergence and sample complexity under the tabular setting. Finally, we provide simulation results to demonstrate the robustness of our methods.

1. Introduction

In practical reinforcement learning (RL) (Sutton & Barto, 2018) applications, the training environment may often times deviate from the test environment, resulting in a model mismatch between the two. Such model mismatch could be because of, e.g., modeling error between simulator and real-world applications, model deviation due to non-stationarity of the environment, unexpected perturbation and potential adversarial attacks. This may lead to a significant performance degradation in the testing environment.

To solve the issue of model mismatch, a framework of robust Markov decision process (MDP) was introduced in (Bagnell et al., 2001; Nilim & El Ghaoui, 2004; Iyengar, 2005), where the MDP model is not fixed but comes from some uncertainty set. The goal of robust RL is to find a policy that optimize the worst-case performance over all possible MDP models in the uncertainty set. Value-based approaches have been extensively studied under the tabular setting and with function approximation, e.g., (Iyengar, 2005; Nilim & El Ghaoui, 2004; Badrinath & Kalathil, 2021; Wiesemann et al., 2013; Roy et al., 2017; Tamar et al., 2014; Lim et al., 2013; Bagnell et al., 2001; Satia & Lave Jr, 1973; Xu & Mannor, 2010; Wang & Zou, 2021). There are also other approaches that are shown to be successful empirically, e.g., based on adversarial training, (Vinitsky et al., 2020; Pinto et al., 2017; Abdullah et al., 2019; Hou et al., 2020; Rajeswaran et al., 2017; Atkeson & Morimoto, 2003; Morimoto & Doya, 2005; Huang et al., 2017; Kos & Song, 2017; Lin et al., 2017; Pattanaik et al., 2018; Mandlekar et al., 2017), which however lack theoretical robustness and optimality guarantee.

The policy gradient method (Williams, 1992; Sutton et al., 1999; Konda & Tsitsiklis, 2000; Kakade, 2001), which models and optimizes the policy directly, has been widely used in RL thanks to its ease of implementation in model-free setting, scalability to large/continuous state and action spaces, and applicability to any differentiable policy parameterization. Despite a large body of empirical and theoretical work on policy gradient method, development of policy gradient approach for robust RL with provable robustness to model mismatch and optimality guarantee still remains largely open in the literature.

In this paper, we develop the first policy gradient method for robust RL under model mismatch with provable robustness, global optimality and complexity analysis. We focus on the R-contamination uncertainty set model (Huber, 1965; Du et al., 2018; Huber & Ronchetti, 2009; Wang & Zou, 2021; Nishimura & Ozaki, 2004; Prasad et al., 2020a;b). Our robust policy gradient method inherits advantages of vanilla policy gradient methods and their variants, and provide provable guarantee on global optimality and robustness. In particular, the challenges and our major contributions are summarized as follows.

- Robust RL aims to optimize the worst-case performance, named robust value function, where the worst-case is

[1]Department of Electrical Engineering, University at Buffalo, New York, USA. Correspondence to: Shaofeng Zou <szou3@buffalo.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

taken over some uncertainty set of MDPs. However, the robust value function involves a "max" and thus may not be differentiable in the policy. Our first contribution in this paper is the development of robust policy gradient, where we derive the Fréchet sub-gradient of the robust value function, and further show that it is the gradient almost everywhere. We would like to highlight that our robust policy gradient applies to any differentiable and Lipschitz policy class.

- Motivated by recent advancements on the global optimality of vanilla policy gradient methods, we are interested in a natural question that whether the global optimum of robust RL can be attained by our robust policy gradient method. The major challenge lies in that the robust value function involves a "max" over the uncertainty set, and thus has a much more complicated landscape than the vanilla value function. We consider the direct parametric policy class and show that the robust value function satisfies the Polyak-Łojasiewicz (PL) condition (Polyak, 1963; Lojasiewicz, 1963), and our robust policy gradient method converges to a global optimum almost surely.

- The robust value function may not be differentiable everywhere, which is the major challenge in the convergence rate analysis. We then design a smoothed robust policy gradient method as an approximation, where the corresponding smoothed objective function is differentiable. We show that smoothed robust policy gradient method converges to an ϵ-global optimum of the original non-differentiable robust RL problem with a complexity of $\mathcal{O}(\epsilon^{-3})$.

- To understand the fundamentals of our robust policy gradient method, our results discussed so far focus on the ideal setting assuming perfect knowledge of the (smoothed) robust policy gradient. Although this is a commonly used setting in recent studies of vanilla policy gradient, e.g., (Agarwal et al., 2021; Cen et al., 2021; Mei et al., 2020; Bhandari & Russo, 2019), such knowledge is typically unknown in practice and need to be estimated from samples. We then focus on the model-free setting, where only samples from the centroid of the uncertainty set are available, and we design a model-free robust actor-critic algorithm. Our robust actor-critic can be applied with arbitrary differential parametric policy class and value function approximation in practice. Theoretically, we prove the global optimality of our robust actor critic method under the tabular setting with direct parametric policy class.

1.1. Related Works

Global optimality of vanilla policy gradient method. In the non-robust setting, policy gradient methods (Williams, 1992; Sutton et al., 1999) as well as their extensions (Sutton et al., 1999; Konda & Tsitsiklis, 2000; Kakade, 2001; Schulman et al., 2015; 2017) have been successful in various applications, e.g., (Schulman et al., 2015; 2017). Despite the surge of interest in policy gradient methods, theoretical understanding remains limited to convergence to local optimum and stationary points. It was not until recently that the global optimality of various policy gradient methods was established (Bhandari & Russo, 2021; 2019; Agarwal et al., 2021; Mei et al., 2020; Li et al., 2021; Laroche & des Combes, 2021; Zhang et al., 2021a; Cen et al., 2021; Zhang et al., 2020a; Lin, 2022). In this paper, we focus on policy gradient methods for robust RL. The major new challenge lies in that the robust value function is not differentiable everywhere, and the landscape of the robust value function is much more complicated than the vanilla value function.

Value-based approach for robust RL. Robust MDP was introduced and studied in (Iyengar, 2005; Nilim & El Ghaoui, 2004; Bagnell et al., 2001; Satia & Lave Jr, 1973; Wiesemann et al., 2013; Lim & Autef, 2019; Xu & Mannor, 2010; Yu & Xu, 2015; Lim et al., 2013; Tamar et al., 2014), where the uncertainty set is assumed to be *known* to the learner, and the problem can be solved using dynamic programming. Later, the studies were generalized to the model-free setting where stochastic samples from the central MDP of the uncertainty set are available in an online fashion (Roy et al., 2017; Badrinath & Kalathil, 2021; Wang & Zou, 2021; Tessler et al., 2019) and an offline fashion (Zhou et al., 2021; Yang et al., 2021; Panaganti & Kalathil, 2021; Goyal & Grand-Clement, 2018; Kaufman & Schaefer, 2013; Ho et al., 2018; 2021; Si et al., 2020). In this paper, we focus on approaches that model and optimize the policy directly, and develop robust policy gradient method. Our method inherits advantages of policy gradient, and has a broader applicability than value-based method for large-scale problems.

Direct policy search for robust RL. Robust policy gradient method for constrained MDP was studied in (Russel et al., 2020), however, there are mistakes in the gradient derivation. More specifically, the fact that the worst-case transition kernel is a function of the policy was ignored when deriving the gradient. A recent paper (Derman et al., 2021) showed the equivalence between robust MDP and regularized MDP, and developed a policy gradient method for the case with only reward uncertainty. In (Derman et al., 2021), it was discussed that it is difficult to their methods extend to problems with uncertain transition kernel because of the dependency between the worst-case transition kernel and the policy. (Eysenbach & Levine, 2021) showed a similar result that maximum entropy regularized MDP is robust to model mismatch. In this paper, we focus on the challenging problem with uncertain transition kernel, and derive the robust policy gradient. We also note that a separate line of work (Zhang et al., 2021c;b) studies the corruption-robust RL problems, where the goal is to learn a robust policy to

data corruption, and is fundamentally different from the problem in this paper.

2. Preliminaries

Markov Decision Process. An MDP $(\mathcal{S}, \mathcal{A}, \mathsf{P}, c, \gamma)$ is specified by: a state space $\mathcal{S}$, an action space $\mathcal{A}$, a transition kernel $\mathsf{P} = \{p_s^a \in \Delta(\mathcal{S}), a \in \mathcal{A}, s \in \mathcal{S}\}$[1], where p_s^a is the distribution of the next state over $\mathcal{S}$ upon the agent taking action a in state s, a cost function $c : \mathcal{S} \times \mathcal{A} \to [0, 1]$, and a discount factor $\gamma \in [0, 1)$. At each time step t, the agent in state s_t takes an action a_t. The environment then transits to the next state s_{t+1} according to the distribution $p_{s_t}^{a_t}$, and provides a cost signal $c(s_t, a_t) \in [0, 1]$ to the agent.

A stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ maps any state to a distribution over the action space $\mathcal{A}$. More specifically, in state s, the agent takes action a with probability $\pi(a|s)$. The value function of a stationary policy π starting from $s \in \mathcal{S}$ measures the expected accumulated discounted cost by following policy π: $\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t c(S_t, A_t)|S_0 = s, \pi\right]$, and the goal is to find a policy that minimizes the value function for any initial state $s \in \mathcal{S}$.

Robust MDP. The transition kernel of the robust MDP is not fixed but is from some uncertainty set $\mathcal{P}$. In this paper, we focus on the (s, a)-rectangular uncertainty set (Nilim & El Ghaoui, 2004; Iyengar, 2005), i.e., $\mathcal{P} = \bigotimes_{s,a} \mathcal{P}_s^a$, where $\mathcal{P}_s^a \subseteq \Delta(\mathcal{S})$. At each time step, after the agent takes an action, the environment transits to the next state following any transition kernel $\mathsf{P} \in \mathcal{P}$, and the choice of kernels can be time-varying. A sequence of transition kernel $\kappa = (\mathsf{P}_0, \mathsf{P}_1 ...) \in \bigotimes_{t\geq 0} \mathcal{P}$ can be viewed as a policy chosen by the nature and is referred to as the nature's policy.

The robust value function of a policy π is defined as the worst-case expected accumulated discounted cost over κ when following π and starting from s:

$$V^\pi(s) \triangleq \max_{\kappa \in \bigotimes_{t\geq 0} \mathcal{P}} \mathbb{E}_\kappa \left[\sum_{t=0}^{\infty} \gamma^t c(S_t, A_t)|S_0 = s, \pi\right]. \quad (1)$$

The robust action-value function can be defined: $Q^\pi(s, a) \triangleq \max_{\kappa \in \bigotimes_{t\geq 0} \mathcal{P}} \mathbb{E}_\kappa \left[\sum_{t=0}^{\infty} \gamma^t c(S_t, A_t)|S_0 = s, A_0 = a, \pi\right]$. It has been shown that $V^\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s) Q^\pi(s, a)$ (Nilim & El Ghaoui, 2004; Iyengar, 2005).

The robust Bellman operator of a policy π is defined as

$$\mathbf{T}_\pi V(s) \triangleq \sum_{a\in\mathcal{A}} \pi(a|s) \left(c(s, a) + \gamma \sigma_{\mathcal{P}_s^a}(V)\right), \quad (2)$$

where $\sigma_{\mathcal{P}_s^a}(V) \triangleq \max_{p\in\mathcal{P}_s^a} p^\top V$ is the support function of V on $\mathcal{P}_s^a$. $\mathbf{T}_\pi$ is a contraction and V^π is the unique fixed point (Nilim & El Ghaoui, 2004; Iyengar, 2005; Puterman, 2014).

Define the expected worst-case total cost function under the initial distribution ρ as $J_\rho(\pi) \triangleq \mathbb{E}_{S\sim\rho}[V^\pi(S)]$. The goal of the agent is to find an optimal policy that minimizes $J_\rho(\pi)$: $\min_\pi J_\rho(\pi)$.

R-Contamination Uncertainty Set. In this paper, we focus on an adversarial model of the uncertainty set, R-contamination, where the nature could arbitrarily perturb the state transition of the MDP with a small probability. Let $\mathsf{P} = \bigotimes_{s\in\mathcal{S},a\in\mathcal{A}} p_s^a$ be a transition kernel. The R-contamination uncertainty set centered at P is defined as $\mathcal{P} \triangleq \bigotimes_{s,a} \mathcal{P}_s^a$, where

$$\mathcal{P}_s^a \triangleq \{(1 - R)p_s^a + Rq | q \in \Delta(\mathcal{S})\}, s \in \mathcal{S}, a \in \mathcal{A}. \quad (3)$$

This uncertainty set model is widely used in the literature of robust learning and optimization, e.g., (Huber, 1965; Du et al., 2018; Wang et al., 2021; Huber & Ronchetti, 2009; Nishimura & Ozaki, 2004; 2006; Prasad et al., 2020a;b). The R-contamination set models the scenario where the state transition could be arbitrarily perturbed with a small probability R, hence is more suitable for systems suffering from random perturbations, adversarial attacks, and outliers in sampling. R-contamination set can also be connected to uncertainty sets defined by total variation, KL-divergence and Hellinger distance via inequalities, e.g., Pinsker's inequality. On the other hand, the R-contamination model is more clean and straightforward, which makes the derivation of the robust policy gradient, and the convergence and complexity analyses tractable.

Under the R-contamination model, the support function can be easily computed as follows:

$$\sigma_{\mathcal{P}_s^a}(V) = (1 - R) \sum_{s'\in\mathcal{S}} p_{s,s'}^a V(s') + R \max_{s'} V(s'), \quad (4)$$

where $p_{s,s'}^a = p_s^a(s')$.

3. Robust Policy Gradient

Consider a parametric policy class $\Pi_\Theta = \{\pi_\theta : \theta \in \Theta\}$. Denote $J_\rho(\pi_\theta)$ by $J_\rho(\theta)$. Robust RL aims to find an optimal policy $\pi_{\theta^*} \in \Pi_\Theta$ that minimizes the expected worst-case accumulated discounted cost:

$$\theta^* \in \arg\min_{\theta\in\Theta} J_\rho(\theta). \quad (5)$$

Let $J_\rho^* \triangleq \min_{\theta\in\Theta} J_\rho(\theta)$. Recall the definition of $V^\pi(s)$ in (1). Due to the max over κ, $V^\pi(s)$ may not be differentiable. To solve this issue, we adopt the Fréchet sub-differential (Kruger, 2003).

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

Definition 3.1. For a function $f : \mathcal{X} \subseteq \mathbb{R}^N \to \mathbb{R}$, a vector $u \in \mathbb{R}^N$ is called a Fréchet sub-differential of f at x, if

$$\lim_{h\to 0}\inf_{h\neq 0} \frac{f(x+h)-f(x)-\langle h,u\rangle}{\|h\|} \geq 0. \tag{6}$$

The set of all the sub-differential of f at x is denoted by $\partial f(x)$.

Clearly, when $f(x)$ is differentiable at x, $\partial f(x) = \{\nabla f(x)\}$.

Without loss of generality, we assume that the parametric policy class is differentiable and Lipschitz.

Assumption 3.2. The policy class Π_Θ is differentiable and k_π-Lipschitz, i.e., for any $s \in \mathcal{S}, a \in \mathcal{A}$ and $\theta \in \Theta$, there exists a universal constant k_π, such that $\|\nabla \pi_\theta(a|s)\| \leq k_\pi$.

This assumption can be easily satisfied by many policy classes, e.g., direct parameterization (Agarwal et al., 2021), soft-max (Mei et al., 2020; Li et al., 2021; Zhang et al., 2021a; Wang & Zou, 2020), or neural network with Lipschitz and smooth activation functions (Du et al., 2019; Neyshabur, 2017; Miyato et al., 2018).

Define the discounted visitation distribution as $d_s^\pi(s') \triangleq (1-\gamma+\gamma R)\sum_{t=0}^\infty \gamma^t(1-R)^t \cdot \mathbb{P}(S_t = s'|S_0 = s, \pi)$, and let $d_\rho^\pi(s') = \mathbb{E}_{S\sim\rho}[d_S^\pi(s')]$. Denote $s_\theta \triangleq \arg\max_s V^{\pi_\theta}(s)$. Then, the (sub-)gradient of $J_\rho(\theta)$ can be computed as the following theorem.

Theorem 3.3 (Robust Policy Gradient)**.** *Consider a class of policies Π_Θ satisfying Assumption 3.2. For any distribution ρ, denote*

$$\psi_\rho(\theta) \triangleq \frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)} \sum_{s\in\mathcal{S}} d_{s_\theta}^{\pi_\theta}(s) \sum_{a\in\mathcal{A}} \nabla\pi_\theta(a|s) Q^{\pi_\theta}(s,a) + \frac{1}{1-\gamma+\gamma R} \sum_{s\in\mathcal{S}} d_\rho^{\pi_\theta}(s) \sum_{a\in\mathcal{A}} \nabla\pi_\theta(a|s) Q^{\pi_\theta}(s,a), \tag{7}$$

then, (1) almost everywhere in Θ, $J_\rho(\theta)$ is differentiable in θ and $\psi_\rho(\theta) = \nabla J_\rho(\theta)$; (2) at non-differentiable θ, $\psi_\rho(\theta) \in \partial J_\rho(\theta)$.

If we set $R = 0$, i.e., the uncertainty set $\mathcal{P}$ reduces to a singleton $\{\mathsf{P}\}$ and there is no robustness, then $\psi_\rho(\theta)$ in (7) reduces to the vanilla policy gradient (Sutton et al., 1999).

As can be seen from (7), the robust policy (sub)-gradient is a function of the robust Q-function. Note that in robust RL, $J_\rho(\theta)$ may not be differentiable everywhere, and therefore, the sub-gradient is needed. The mistake in (Russel et al., 2020) is due to the ignorance of the dependence of the worst-case kernel on θ, and their robust policy gradient is a function of the vanilla Q-function not the robust Q-function, which should not be the case.

In policy gradient approaches, the agent often has the challenge of exploration. If ρ highly concentrates on a subset of $\mathcal{S}$, then the agent may not be able to explore all the states and may end up with a sub-optimal solution. To solve this issue, we introduce an optimization measure μ satisfying $\mu_{\min} \triangleq \min_s \mu(s) > 0$ (Agarwal et al., 2021). The initial state distribution ρ is called the performance measure. As we will show in the next section, although we want to minimize $\mathbb{E}_\rho[V^\pi]$, we can perform sub-gradient descent with respect to $J_\mu \triangleq \mathbb{E}_\mu[V^\pi]$ and the algorithm can still find an optimum of $\mathbb{E}_\rho[V^\pi]$.

Given the robust policy gradient in Theorem 3.3, we design our robust policy gradient algorithm in Algorithm 1. We note that our Algorithm 1 can be applied to any arbitrarily parameterized policy class that is differentiable and Lipschitz. Here $\prod_\Theta$ denotes the projection onto Θ.

Algorithm 1 Robust Policy Gradient

Input: T, α_t
Initialization: θ_0
  for $t = 0, 1, ..., T-1$ **do**
    $\theta_{t+1} \leftarrow \prod_\Theta(\theta_t - \alpha_t \psi_\mu(\theta_t))$
  end for
Output: θ_T

4. Global Optimality: Direct Parameterization

In this section, we show that the robust objective function $J_\rho(\theta)$ satisfies the Polyak-Łojasiewicz (PL) condition when the direct parametric policy class is used, i.e., $\Theta = (\Delta(\mathcal{A}))^{|\mathcal{S}|}$ and $\pi_\theta(a|s) = \theta_{s,a}$, and we further show the global optimality of Algorithm 1.

Algorithm 1 is in fact a sub-gradient descent algorithm for a non-convex function J_μ. Following classic results from stochastic approximation and optimization (Beck, 2017; Borkar, 2009; Borkar & Meyn, 2000), Algorithm 1 is expected to converge to stationary points only. Showing the global optimality requires further characterization of J_μ, which involves the "max" over the transition kernels, and is thus more challenging than the vanilla non-robust case.

We first show that the robust objective function J_ρ satisfies the PL-condition under the direct parameterization. Informally, a function $f(\theta)$ is said to satisfy the PL condition if $f(\theta) - f(\theta^*) \leq \mathcal{O}(F(\nabla f(\theta)))$, for some suitable scalar notion of first-order stationarity F, which measures how large the gradient is (Karimi et al., 2016; Bolte et al., 2007a). This condition implies that if a solution is close to some first-order stationary point, the function value is then close to the global optimum.

Theorem 4.1 (PL-Condition)**.** *Under direct policy parameterization and for any optimization measure $\mu \in \Delta(\mathcal{S})$ and*

performance measure $\rho \in \Delta(\mathcal{S})$,

$$J_\rho(\theta) - J_\rho^* \leq C_{PL} \max_{\hat{\pi} \in (\Delta(\mathcal{A}))^{|\mathcal{S}|}} \langle \pi_\theta - \hat{\pi}, \psi_\mu(\theta) \rangle, \quad (8)$$

where $C_{PL} = \frac{1}{(1-\gamma)\mu_{\min}}$.

Note that $\psi_\mu(\theta)$ on the right hand side of (8) is a sub-gradient of $J_\mu(\theta)$, and $J_\rho(\theta)$ on the left hand side of (8) is the objective function with respect to ρ. Therefore, for any optimization measure μ, a stationary point of J_μ is a global optimum of J_ρ. Thus the PL-condition in Theorem 4.1 with results from stochastic approximation will lead to the global optimality of Algorithm 1 in the following theorem.

Theorem 4.2. *If* $\alpha_t > 0$, $\sum_{t=0}^{\infty} \alpha_t = \infty$ *and* $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$, *then Algorithm 1 converges to a global optimum of* $J_\rho(\theta)$ *almost surely.*

Theorem 4.2 suggests that our robust policy gradient algorithm converges to a global optimum, which matches the global optimality results for vanilla policy gradients in e.g., (Agarwal et al., 2021; Mei et al., 2020). However, the analysis here is much more challenging due to the non-differentiable objective function and min-max problem structure.

5. Smoothed Robust Policy Gradient

It is in general challenging to analyze the convergence rate of Algorithm 1, which is a projected sub-gradient descent algorithm for non-differentiable non-convex function. In this section, we construct a smoothed approximation $J_{\sigma,\rho}$, which converges to J_ρ as $\sigma \to \infty$. We develop a smoothed robust policy gradient, and show that the smoothed $J_{\sigma,\rho}$ satisfies the PL-condition. We characterize its global optimality and show that to achieve an ϵ-global optimal, the sample complexity is $\mathcal{O}(\epsilon^{-3})$.

For convenience, in the remaining of this paper we assume that $\rho = \mu$ and $\rho_{\min} > 0$, and we omit the subscript ρ in J_ρ and $J_{\sigma,\rho}$. The algorithm design and theoretical results can be similarly extended to the general setting with $\rho \neq \mu$ as in Section 3 and Section 4.

We use the LogSumExp (LSE) operator to approximate the $\max$ operator, where

$$\text{LSE}(\sigma, V) = \frac{\log(\sum_{i=1}^{d} e^{\sigma V(i)})}{\sigma}$$

for $V \in \mathbb{R}^d$ and some $\sigma > 0$. The approximation error $|\text{LSE}(\sigma, V) - \max V| \to 0$ as $\sigma \to \infty$. By replacing $\max$ in (2) using LSE, the smoothed Bellman operator is

$$\mathbf{T}_\sigma^\pi V(s) = \mathbb{E}_{A \sim \pi(\cdot|s)} \Big[c(s, A) + \gamma(1-R) \sum_{s' \in \mathcal{S}} p_{s,s'}^A V(s') + \gamma R \cdot \text{LSE}(\sigma, V) \Big]. \quad (9)$$

The reason why we do not use soft-max is because with soft-max, the induced Bellman operator is not a contraction anymore (Asadi & Littman, 2017; Wang & Zou, 2021). With LSE, $\mathbf{T}_\sigma^\pi$ is a contraction and has a unique fixed point (Wang & Zou, 2021), which we denote by V_σ^π and name as the smoothed robust value function. We can also define the smoothed robust action-value function $Q_\sigma^\pi(s,a) \triangleq c(s,a) + \gamma(1-R) \sum_{s' \in \mathcal{S}} p_{s,s'}^a V_\sigma^\pi(s') + \gamma R \cdot \text{LSE}(\sigma, V_\sigma^\pi)$.

It can be shown that V_σ^π is differentiable in θ and it converges to V^π as $\sigma \to \infty$. We then define the smoothed objective function as

$$J_\sigma(\theta) = \sum_{s \in \mathcal{S}} \rho(s) V_\sigma^{\pi_\theta}(s), \quad (10)$$

and let $J_\sigma^* = \min_\theta J_\sigma(\theta)$. Note $J_\sigma(\theta)$ is an approximation of $J(\theta)$, and J_σ^* also converges to J^* as $\sigma \to \infty$. In the following theorem, we derive the gradient of J_σ, which holds for any differentiable policy class.

Theorem 5.1. *Consider a policy class* Π_Θ *that is differentiable. The gradient of* $J_\sigma(\theta)$ *is*

$$\nabla J_\sigma(\theta) = B(\rho, \theta) + \frac{\gamma R \sum_{s \in \mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)} B(s, \theta)}{(1-\gamma) \sum_{s \in \mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}}, \quad (11)$$

where $B(s,\theta) \triangleq \frac{1}{1-\gamma+\gamma R} \sum_{s' \in \mathcal{S}} d_s^\pi(s') \sum_{a \in \mathcal{A}} \nabla \pi_\theta(a|s') \cdot Q_\sigma^{\pi_\theta}(s', a)$, *and* $B(\rho, \theta) \triangleq \mathbb{E}_{S \sim \rho}[B(S, \theta)]$.

We then design the smoothed robust policy gradient algorithm in Algorithm 2.

Algorithm 2 Smoothed Robust Policy Gradient

Input: T, σ, α_t
Initialization: θ_0
  for $t = 0, 1, ..., T-1$ **do**
    $\theta_{t+1} \leftarrow \prod_\Theta (\theta_t - \alpha_t \nabla J_\sigma(\theta_t))$
  end for
Output: θ_T

5.1. Global Optimality Under Direct Parameterization

We focus on direct policy parameterization. We show that the smoothed objective function J_σ satisfies the PL-condition and develop the global optimality of Algorithm 2.

Theorem 5.2 (PL-Condition)**.** *Consider direct policy parameterization. Then*

$$J_\sigma(\theta) - J_\sigma^* \leq C_{PL} \max_{\hat{\pi} \in (\Delta(|\mathcal{A}|))^{|\mathcal{S}|}} \langle \pi_\theta - \hat{\pi}, \nabla J_\sigma(\theta) \rangle + \left(\frac{\gamma R}{1-\gamma} \right) \frac{2 \log |\mathcal{S}|}{\sigma}. \quad (12)$$

This theorem implies that if the gradient $\nabla J_\sigma(\theta)$ is small, then θ falls into a small neighbour (radius of $\mathcal{O}(\sigma^{-1})$) of the global optimum J_σ^*. By choosing a large σ, the difference between $J_\sigma(\theta) - J_\rho^*$ can be made arbitrarily small, and thus the global optimality with respect to J_ρ^* can be established.

5.2. Convergence Rate

We then study the convergence rate of Algorithm 2 under direct policy parameterization. We first make an additional smoothness assumption on Π_Θ.

Assumption 5.3. Π_Θ is l_π-smooth, i.e., for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$ and for any $\theta_1, \theta_2 \in \Theta$,

$$\|\nabla \pi_{\theta_1}(a|s) - \nabla \pi_{\theta_2}(a|s)\| \leq l_\pi \|\theta_1 - \theta_2\|. \tag{13}$$

Then under Assumption 5.3, we show that J_σ is L_σ-smooth.

Lemma 5.4. *Under Assumptions 3.2 and 5.3, for any θ_1, θ_2,*

$$\|\nabla J_\sigma(\theta_1) - \nabla J_\sigma(\theta_2)\| \leq L_\sigma \|\theta_1 - \theta_2\|, \tag{14}$$

where $L_\sigma = \mathcal{O}(\sigma)$ and its exact definition is in (89).

The value of σ controls the tradeoff between the smoothness of J_σ and the approximation error between J_σ and J. By choosing σ carefully, we have the following theorem.

Theorem 5.5. *For any $\epsilon > 0$, set $\sigma = \frac{2\gamma R \log|\mathcal{S}|}{\epsilon(1-\gamma)} = \mathcal{O}(\epsilon^{-1})$ and $T = \frac{64|\mathcal{S}|C_{PL}^2 L_\sigma C_\sigma}{\epsilon^2} = \mathcal{O}(\epsilon^{-3})$ in Algorithm 2 [2], then*

$$\min_{t \leq T-1} J(\theta_t) - J^* \leq 3\epsilon. \tag{15}$$

Theorem 5.5 shows that Algorithm 2 converges to an ϵ-global optimum within $\mathcal{O}(\epsilon^{-3})$ steps. This rate is slower than the one of non-robust policy gradient in (Agarwal et al., 2021) by a factor of $\mathcal{O}(\epsilon^{-1})$, which is due to the additional robustness requirement and the smoothing technique using σ. If we set $R = 0$, i.e., no robustness, the value of σ will then be irrelevant, and our algorithm reduces to non-robust policy gradient algorithm. With $R = 0$, $L_\sigma = \mathcal{O}(1)$ and $C_\sigma = \mathcal{O}(1)$, then our complexity also reduces to $\mathcal{O}(\epsilon^{-2})$, which is the same as the one in (Agarwal et al., 2021).

6. Robust Actor-Critic

The results discussed so far assume full knowledge of robust value functions and visitation distributions, and thus the (smoothed) robust policy gradient is exactly known. Although this is a commonly used setting for theoretical analysis e.g., in (Agarwal et al., 2021; Mei et al., 2020; Bhandari & Russo, 2019; Cen et al., 2021), such knowledge is usually unknown in practice. In this section, we focus on the practical *model-free* setting where only training samples from the centroid transition kernel P can be obtained.

As can be seen from (7) and (11), to obtain the (smoothed) robust policy (sub)-gradient, we first need to estimate the robust value function. Robust value function measures the performance on the worst-case transition kernel which is typically different from the one that generates samples. However, Monte Carlo (MC) method can only be used to estimate the value function on the kernel that generates the samples.

To solve this issue, we design a robust TD algorithm, and combine it with our robust policy gradient descent to design the robust actor-critic algorithm. Consider a parametric robust action value function Q_ζ, e.g., linear function approximation or neural network. The robust TD algorithm is given in Algorithm 3. Note that by replacing $\max$ in the algorithm by LSE we can also get the smoothed robust TD algorithm to estimate V_σ^π.

Algorithm 3 Robust TD

Input: T_c, π, β_t
Initialization: ζ, s_0
  Choose $a_0 \sim \pi(\cdot|s_0)$
  for $t = 0, 1, ..., T_c - 1$ **do**
    Observe c_t, s_{t+1}
    Choose $a_{t+1} \sim \pi(\cdot|s_{t+1})$
    $V_t^* \leftarrow \max_s \{\sum_{a\in\mathcal{A}} \pi(a|s) Q_\zeta(s,a)\}$
    $\delta_t \leftarrow Q_\zeta(s_t, a_t) - c_t - \gamma(1-R) Q_\zeta(s_{t+1}, a_{t+1}) - \gamma R V_t^*$
    $\zeta \leftarrow \zeta - \beta_t \delta_t \nabla_\zeta Q_\zeta(s_t, a_t)$
  end for
Output: ζ

We provide the convergence proof of robust TD under the tabular setting in Appendix C.1. For convergence under general function approximation, additional regularity conditions might be needed (Korda & La, 2015; Dalal et al., 2018; Bhandari et al., 2018; Cai et al., 2019; Roy et al., 2017).

With the robust TD algorithm, we then develop our robust actor-critic algorithm in Algorithm 4. The algorithm can be applied with any differentiable value function approximation and parametric policy class.

We then smooth and specialize Algorithm 4 to the tabular setting with direct policy parameterization (see Algorithm 6 in Appendix C.2 for the details). We derive the global optimality and convergence rate for smoothed robust actor-critic in the following theorem. In the algorithm, we set T_c large enough so that $\|Q_{T_c} - Q_\sigma^\pi\|_\infty \leq \epsilon_{\text{est}}$, where ϵ_{est} denotes the estimate error of robust value function. We note that the smoothed robust TD algorithm in Algorithm 5 can be shown to converges to an ϵ_{est}-global optimum with $\mathcal{O}(\epsilon_{\text{est}}^{-2})$ samples following similar methods as in (Wang &

[2] $C_\sigma = \frac{1}{1-\gamma}(1 + 2\gamma R \frac{\log|\mathcal{S}|}{\sigma})$ denotes the upper bound of Q_σ^π.

Algorithm 4 Robust Actor-Critic

Input: T, T_c, σ, α_t, M
Initialization: θ_0
  for $t = 0, 1, ..., T-1$ **do**
    Run Algorithm 3 for T_c times
    $Q_t \leftarrow Q_{\zeta_{T_c}}$
    $V_t(s) \leftarrow \sum_{a\in\mathcal{A}} \pi_\theta(a|s) Q_t(s,a)$ for all $s \in \mathcal{S}$
    for $j = 1, ..., M$ **do**
      Sample $T^j \sim$ **Geom**$(1-\gamma+\gamma R)$
      Sample $s_0^j \sim \rho$
      Sample trajectory from s_0^j: $(s_0^j, a_0^j, ..., s_{T^j}^j)$ following π_{θ_t}
      $B_t^j \leftarrow \frac{1}{1-\gamma+\gamma R} \sum_{a\in\mathcal{A}} \nabla\pi_\theta(a|s_{T^j}^j) Q_t(s_{T^j}^j, a)$
      $x_0^j \leftarrow \arg\max_s V_t(s)$
      Sample trajectory from x_0^j: $(x_0^j, b_0^j, ..., x_{T^j}^j)$ following π_{θ_t}
      $D_t^j \leftarrow \frac{1}{1-\gamma+\gamma R} \sum_{a\in\mathcal{A}} \nabla\pi_\theta(a|x_{T^j}^j) Q_t(x_{T^j}^j, a)$
      $g_t^j \leftarrow B_t^j + \frac{\gamma R}{1-\gamma} D_t^j$
    end for
    $g_t \leftarrow \frac{\sum_{j=1}^M g_t^j}{M}$
    $\theta_{t+1} \leftarrow \prod_\Theta(\theta_t - \alpha_t g_t)$
  end for
Output: θ_T

Zou, 2021).

Theorem 6.1. *For the smoothed robust actor-critic algorithm under the tabular setting with direct policy parameterization, if we set* $\epsilon_{est} = \mathcal{O}(\epsilon^2)$, $M = \mathcal{O}(\epsilon^{-2})$ *and* $T = \mathcal{O}(\epsilon^{-3})$, *then,*

$$\min_{t\le T} \mathbb{E}[J(\theta_t) - J^*] \le 7\epsilon. \tag{16}$$

An explicit bound can be found in (161) in the Appendix. The sample complexity of Robust TD (Algorithm 5) is $\mathcal{O}(\epsilon_{\text{est}}^{-2})$, then the robust TD requires $T_c = \mathcal{O}(\epsilon^{-4})$ samples. Hence the overall sample complexity of Algorithm 6 is $\mathcal{O}(T(M+T_c)) = \mathcal{O}(\epsilon^{-7})$ to find an ϵ-global optimum.

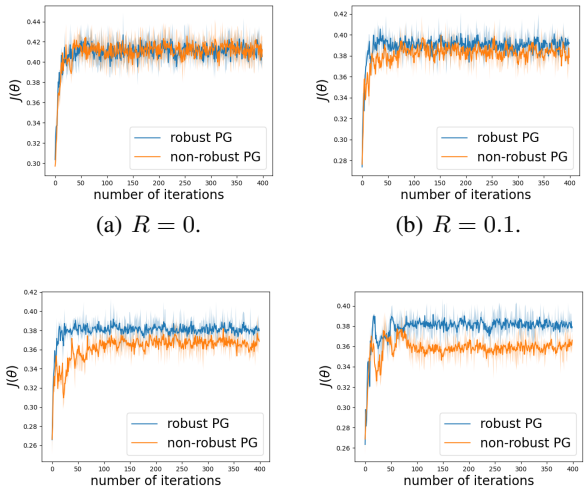
7. Numerical Results

In this section, we demonstrate the convergence and robustness of our algorithms using numerical experiments. We also include several additional experiments in Section F in Appendix. We test our algorithms on the Garnet problem (Archibald et al., 1995) and the Taxi environment from OpenAI (Brockman et al., 2016). The Garnet problem can be specified by $\mathcal{G}(S_n, A_n)$, where the state space $\mathcal{S}$ has S_n states $(s_1, ..., s_{S_n})$ and action space has A_n actions $(a_1, ..., a_{A_n})$. The agent can take any actions in any state, but only gets reward $r = 1$ if it takes a_1 in s_1 or takes a_2 in other states (it will receive 0 reward in other cases). The transition kernels are randomly generated.

7.1. Robust v.s. Non-robust Policy Gradient

We first compare our robust policy gradient method with vanilla policy gradient. To show the robustness of our algorithm over the vanilla policy gradient method, we compare their robust value functions, i.e., worst-case performance, for different values of R. We first train the robust policy gradient algorithm and store the obtained policy θ_t at each time step. At each time step, we run robust TD in Algorithm 5 with a sample size 200 for 30 times to estimate the average objective function value $J(\theta_t)$. We then plot $J(\theta_t)$ v.s. the number of iterations t on the Garnet problems $\mathcal{G}(12, 6)$ and $\mathcal{G}(20, 10)$ in Figure 1 and Figure 2, respectively, and plot results on the Taxi environment from OpenAI in Figure 3. We do the same for the vanilla policy gradient method. The upper and lower envelopes of the curves correspond to the 95 and 5 percentiles of the 30 curves, respectively.

As can be seen from Figure 1, when $R = 0$, the robust policy gradient method reduces to the non-robust vanilla policy gradient, and our results show both algorithms have the same performance. When $R > 0$, the robust policy gradient obtains a policy that performs much better than the non-robust vanilla policy gradient, which demonstrates the robustness of our method.

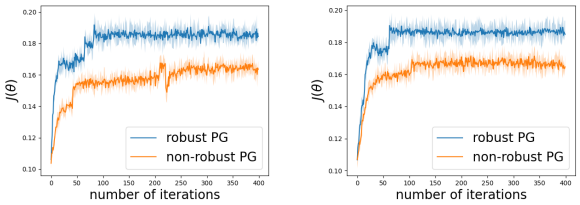


(a) $R = 0$. (b) $R = 0.1$. (c) $R = 0.15$. (d) $R = 0.25$.

Figure 1. Robust Policy Gradient v.s. Non-robust Policy Gradient on Garnet Problem $\mathcal{G}(12, 6)$.

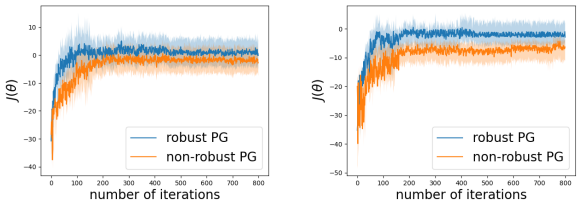
7.2. Smoothed Robust Policy Gradient

In this section, we demonstrate the performance of our smoothed robust policy gradient method on the Garnet problem $\mathcal{G}(12, 6)$. As we showed in Section 5, the smoothed algorithm approximate the robust policy gradient algorithm as $\sigma \to \infty$. Here, we set different values of σ in Algorithm 2 and plot their objective functions v.s. number of iterations



(a) $R = 0.2$. (b) $R = 0.25$.

Figure 2. Robust Policy Gradient v.s. Non-robust Policy Gradient on Garnet Problem $\mathcal{G}(20, 10)$.

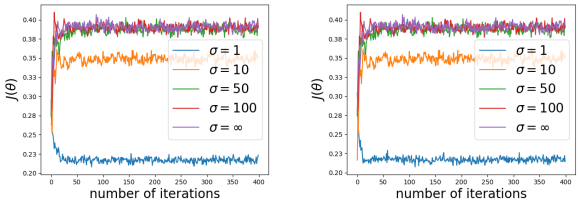


(a) $R = 0.1$. (b) $R = 0.15$.

Figure 3. Robust Policy Gradient v.s. Non-robust Policy Gradient on Taxi Problem.

to demonstrate such an approximation.

As shown in Figure 4, when σ is small (e.g., $\sigma = 1$), the performance of smoothed robust policy gradient is poor. As σ increases, smoothed robust policy gradient behaves similarly to the robust policy gradient, which corresponds to the curve with $\sigma = \infty$. This experiment hence verifies our theoretical results that we can approximate the robust policy gradient by choosing a suitably large σ.

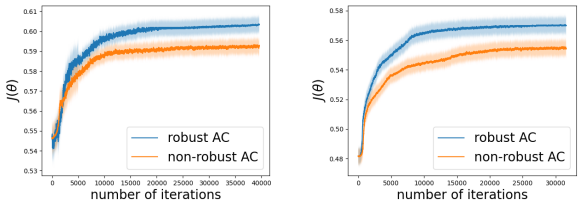


(a) $R = 0.1$. (b) $R = 0.15$.

Figure 4. Smoothed Robust Policy Gradient.

7.3. Robust Actor-Critic

In Figure 5, we consider Garnet problem $\mathcal{G}(30, 20)$ using neural network parameterized policy, where we use a two-layer neural network with 15 neurons in the hidden layer to parameterize the policy π_θ. We then use a two-layer neural network (with 20 neurons in the hidden layer) in the critic. At each time step, we run Algorithm 3 for 30 times to estimate the robust value function. We then use the estimate to simulate Algorithm 4. We plot $J(\theta_t)$ v.s. the number of iterations in Figure 5, and the upper and lower envelopes of the curves correspond to the 95 and 5 percentiles of the 30 trajectories. As the results show, our robust actor-critic algorithm finds a policy that achieves a higher accumulated discounted reward on the worst-case transition kernel than the vanilla actor-critic algorithm (Sutton & Barto, 2018).



(a) $R = 0.15$. (b) $R = 0.2$.

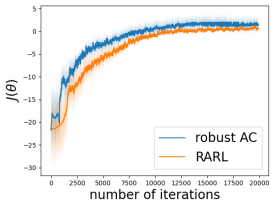
Figure 5. Robust Actor-Critic (AC) v.s. Non-robust Actor-Critic on Garnet Problem $\mathcal{G}(30, 20)$.

7.4. Comparison with Other Robust Algorithms

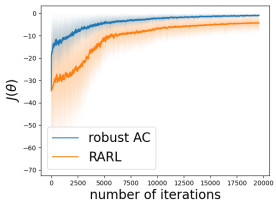
We compare our robust algorithms with two other robust RL approahces, including the robust adversarial reinforcement learning (RARL) approach in (Pinto et al., 2017), and the adversarially robust policy learning (ARPL) approach in (Mandlekar et al., 2017).

7.4.1. Comparison with RARL

The basic idea of the RARL approach is to introduce an adversary that perturbs the state transition to minimize the accumulated discounted reward. Then the agent and the adversary are trained alternatively using adversarial training. To apply their algorithm to our problem setting, we set an adversarial player, whose goal is to minimize the accumulated discounted reward that the agent receives. The action space $\mathcal{A}_{ad}$ of the adversary is the state space $\mathcal{A}_{ad} \triangleq \mathcal{S}$. The agent and the adversary take actions a_a, a_{ad}, then the environment will transit to state a_{ad} with probability R or transit following the unperturbed MDP $p_s^{a_a}$ with probability $1 - R$. We compare our robust actor-critic algorithm with the RARL algorithm on the Taxi environment. Similarly, at each time step, we run Algorithm 3 with neural function approximation for 30 times to estimate the robust value function. We then use the results to simulate Algorithm 4 and RARL. We plot the robust value function $J(\theta_t)$ v.s. the number of iterations in Figure 6. The upper and lower envelops correspond the 95 and 5 percentiles of the 30 trajectories. As Figure 6 shows, our robust actor-critic algorithm achieves a much higher accumulative discounted reward than the RARL approach under the worst-case transition kernel, and thus is more robust to model mismatch.

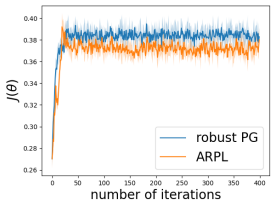


(a) $R = 0.1$.

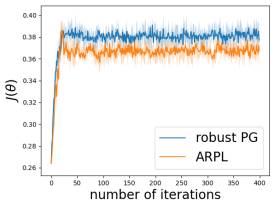


(b) $R = 0.15$.

Figure 6. Robust Actor-Critic v.s. RARL on Taxi Environment.



(a) $R = 0.2$.



(b) $R = 0.3$.

Figure 7. Robust PG v.s. ARPL on $\mathcal{G}(10, 8)$ Environment.

7.4.2. Comparison with ARPL

The idea of the ARPL approach (Mandlekar et al., 2017) can be summarized as follows. During the training, an adversary randomly perturb the observation. Specifically, at each time step, with a small probability p_{per}, the observed state is perturbed arbitrarily. The algorithm update the policy using policy gradient method with these perturbed samples .

Consider the $\mathcal{G}(10, 8)$ problem. We set $p_{\text{per}} = R$. At each time step, we first generate $n = 20$ trajectories $\mathcal{T}^i, i = 1, ..., 20$ starting at $s_0^i \sim \rho$ following π_{θ_t}, and the length of $\mathcal{T}^i$ is generated according to **Geometric**$(1 - \gamma)$. Then, with probability R the state observed by the agent is perturbed to an arbitrary state randomly. We then use these perturbed trajectories and Monte Carlo method to estimate the action-value functions $Q_t(s, a)$. We also use these perturbed trajectories to estimate the visitation distribution. Then we perform a sample-based policy gradient step using Q_t and these trajectories. We compare our robust policy gradient algorithm with the ARPL algorithm. At each time step, we update the policies θ_t according to the two algorithms, and use 30 trajectories to estimate the robust value function $J(\theta_t)$, which measures the worst-case performance over the uncertainty set. We then plot the robust value function $J(\theta_t)$ for the two approaches v.s. the number of iterations in Figure 7. The upper and lower envelops correspond the 95 and 5 percentiles of the 30 trajectories. As Figure 7 shows, our robust PG algorithm finds a policy that achieves a higher reward under the worst-case transition kernel than the ARPL approach, and hence is more robust to the model uncertainty.

8. Discussions

In this paper, we develop a direct policy search method for robust RL. Our robust algorithms can be applied with arbitrary differentiable value function approximation and policy parameterization, and thus is scalable to problems with large state and action spaces. In this paper, the analysis is for the direct policy parameterization, and our approach can also be extended to establish the global optimality under other policy parameterizations and develop robust natural policy gradient approaches for robust RL. We focus on the R-contamination model for the uncertainty set in this paper, which can be closely related to sets defined by total variation and Kullback-Leibler divergence using Pinsker's inequality. It is also of future interest to investigate model-free approaches for other uncertainty sets defined via e.g., total variation, Kullback-Leibler divergence, Wasserstain distance.

9. Acknowledgment

The work was supported by the National Science Foundation under Grants CCF-2106560 and CCF-2007783.

References

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

Achiam, J., Held, D., Tamar, A., and Abbeel, P. Constrained policy optimization. In *Proc. International Conference on Machine Learning (ICML)*, pp. 22–31. PMLR, 2017.

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

Archibald, T., McKinnon, K., and Thomas, L. On the generation of Markov decision processes. *Journal of the Operational Research Society*, 46(3):354–361, 1995.

Asadi, K. and Littman, M. L. An alternative softmax operator for reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, volume 70, pp. 243–252. JMLR, 2017.

Atkeson, C. G. and Morimoto, J. Nonparametric representation of policies and value functions: A trajectory-based approach. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 1643–1650, 2003.

Badrinath, K. P. and Kalathil, D. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *Proc. International Conference on Machine Learning (ICML)*, pp. 511–520. PMLR, 2021.

Bagnell, J. A., Ng, A. Y., and Schneider, J. G. Solving uncertain Markov decision processes. 09 2001.

Beck, A. *First-order methods in optimization*. SIAM, 2017.

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Bhandari, J. and Russo, D. On the linear convergence of policy gradient methods for finite MDPs. In *Proc. International Conference on Artifical Intelligence and Statistics (AISTATS)*, pp. 2386–2394. PMLR, 2021.

Bhandari, J., Russo, D., and Singal, R. A finite time analysis of temporal difference learning with linear function approximation. In *Proc. Annual Conference on Learning Theory (CoLT)*, pp. 1691–1692. PMLR, 2018.

Bolte, J., Daniilidis, A., and Lewis, A. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007a.

Bolte, J., Daniilidis, A., Lewis, A., and Shiota, M. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007b.

Borkar, V. S. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

Borkar, V. S. and Meyn, S. P. The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38 (2):447–469, 2000.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016.

Cai, Q., Yang, Z., Lee, J. D., and Wang, Z. Neural temporal-difference learning converges to global optima. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 11312–11322, 2019.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 2021.

Clarke, F. H. *Optimization and nonsmooth analysis*. SIAM, 1990.

Dalal, G., Szörényi, B., Thoppe, G., and Mannor, S. Finite sample analyses for TD(0) with function approximation. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pp. 6144–6160, 2018.

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

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1675–1685. PMLR, 2019.

Du, S. S., Wang, Y., Balakrishnan, S., Ravikumar, P., and Singh, A. Robust nonparametric regression under huber's ϵ-contamination model. *arXiv preprint arXiv:1805.10406*, 2018.

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

Federer, H. *Geometric measure theory*. Springer, 2014.

Ghadimi, S. and Lan, G. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.

Ghadimi, S., Lan, G., and Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155 (1-2):267–305, 2016.

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

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

Ho, C. P., Petrik, M., and Wiesemann, W. Partial policy iteration for l_1-robust Markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46, 2021.

Hou, L., Pang, L., Hong, X., Lan, Y., Ma, Z., and Yin, D. Robust reinforcement learning with Wasserstein constraint. *arXiv preprint arXiv:2006.00945*, 2020.

Huang, S., Papernot, N., Goodfellow, I., Duan, Y., and Abbeel, P. Adversarial attacks on neural network policies. In *Proc. International Conference on Learning Representations (ICLR)*, 2017.

Huber, P. and Ronchetti, E. *Robust Statistics*. John Wiley & Sons, Inc, 2009.

Huber, P. J. A robust version of the probability ratio test. *Ann. Math. Statist.*, 36:1753–1758, 1965.

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

Kakade, S. M. A natural policy gradient. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, volume 14, pp. 1531–1538, 2001.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the Polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

Kaufman, D. L. and Schaefer, A. J. Robust modified policy iteration. *INFORMS Journal on Computing*, 25(3):396–410, 2013.

Konda, V. R. and Tsitsiklis, J. N. Actor-critic algorithms. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 1008–1014, 2000.

Korda, N. and La, P. On TD(0) with function approximation: Concentration bounds and a centered variant with exponential convergence. In *Proc. International Conference on Machine Learning (ICML)*, pp. 626–634, 2015.

Kos, J. and Song, D. Delving into adversarial attacks on deep policies. In *Proc. International Conference on Learning Representations (ICLR)*, 2017.

Kruger, A. Y. On Fréchet subdifferentials. *Journal of Mathematical Sciences*, 116(3):3325–3358, 2003.

Laroche, R. and des Combes, R. T. Dr jekyll & mr hyde: the strange case of off-policy policy updates. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. Softmax policy gradient methods can take exponential time to converge. *arXiv preprint arXiv:2102.11270*, 2021.

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

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

Lin, X. On the convergence rates of policy gradient methods. *arXiv preprint arXiv:2201.07443*, 2022.

Lin, Y.-C., Hong, Z.-W., Liao, Y.-H., Shih, M.-L., Liu, M.-Y., and Sun, M. Tactics of adversarial attack on deep reinforcement learning agents. In *Proc. International Joint Conferences on Artificial Intelligence (IJCAI)*, pp. 3756–3762, 2017.

Lojasiewicz, S. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117(87-89):2, 1963.

Majewski, S., Miasojedow, B., and Moulines, E. Analysis of nonsmooth stochastic approximation: the differential inclusion approach. *arXiv preprint arXiv:1805.01916*, 2018.

Mandlekar, A., Zhu, Y., Garg, A., Fei-Fei, L., and Savarese, S. Adversarially robust policy learning: Active construction of physically-plausible perturbations. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3932–3939. IEEE, 2017.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. On the global convergence rates of softmax policy gradient methods. In *Proc. International Conference on Machine Learning (ICML)*, pp. 6820–6829. PMLR, 2020.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *Proc. International Conference on Learning Representations (ICLR)*, 2018.

Morimoto, J. and Doya, K. Robust reinforcement learning. *Neural computation*, 17(2):335–359, 2005.

Neyshabur, B. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*, 2017.

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

Nishimura, K. G. and Ozaki, H. Search and knightian uncertainty. *Journal of Economic Theory*, 119(2):299–333, 2004.

Nishimura, K. G. and Ozaki, H. An axiomatic approach to ϵ-contamination. *Economic Theory*, 27(2):333–340, 2006.

Panaganti, K. and Kalathil, D. Sample complexity of robust reinforcement learning with a generative model. *arXiv preprint arXiv:2112.01506*, 2021.

Pattanaik, A., Tang, Z., Liu, S., Bommannan, G., and Chowdhary, G. Robust deep reinforcement learning with adversarial attacks. In *Proc. International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2040–2042, 2018.

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

Polyak, B. T. Gradient methods for minimizing functionals. *Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki*, 3(4):643–653, 1963.

Prasad, A., Srinivasan, V., Balakrishnan, S., and Ravikumar, P. On learning Ising models under Huber's contamination model. *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020a.

Prasad, A., Suggala, A. S., Balakrishnan, S., and Ravikumar, P. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):601–627, 2020b.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Rajeswaran, A., Ghotra, S., Ravindran, B., and Levine, S. Epopt: Learning robust neural network policies using model ensembles. In *Proc. International Conference on Learning Representations (ICLR)*, 2017.

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

Russel, R. H., Benosman, M., and Van Baar, J. Robust constrained-MDPs: Soft-constrained robust policy optimization under model uncertainty. *arXiv preprint arXiv:2010.04870*, 2020.

Ruszczyński, A. Convergence of a stochastic subgradient method with averaging for nonsmooth nonconvex constrained optimization. *Optimization Letters*, pp. 1–11, 2020.

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

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1889–1897. PMLR, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, Massachusetts, 2018.

Sutton, R. S., McAllester, D. A., Singh, S. P., Mansour, Y., et al. Policy gradient methods for reinforcement learning with function approximation. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, volume 99, pp. 1057–1063. Citeseer, 1999.

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

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

Touati, A., Zhang, A., Pineau, J., and Vincent, P. Stable policy optimization via off-policy divergence regularization. In *Proc. International Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 1328–1337. PMLR, 2020.

Vinitsky, E., Du, Y., Parvate, K., Jang, K., Abbeel, P., and Bayen, A. Robust reinforcement learning using adversarial populations. *arXiv preprint arXiv:2008.01825*, 2020.

Wang, Y. and Zou, S. Finite-sample analysis of Greedy-GQ with linear function approximation under Markovian noise. In *Proc. International Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 11–20. PMLR, 2020.

Wang, Y. and Zou, S. Online robust reinforcement learning with model uncertainty. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Wang, Y., Zou, S., and Zhou, Y. Finite-sample analysis for two time-scale non-linear TDC with general smooth function approximation. *arXiv preprint arXiv:2104.02836*, 2021.

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

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Xu, H. and Mannor, S. Distributionally robust Markov decision processes. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 2505–2513, 2010.

Yang, W., Zhang, L., and Zhang, Z. Towards theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics. *arXiv preprint arXiv:2105.03863*, 2021.

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

Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., and Wang, M. Variational policy gradient method for reinforcement learning with general utilities. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 4572–4583, 2020a.

Zhang, J., Lin, H., Jegelka, S., Jadbabaie, A., and Sra, S. Complexity of finding stationary points of nonsmooth nonconvex functions. *arXiv preprint arXiv:2002.04130*, 2020b.

Zhang, S., Tachet, R., and Laroche, R. Global optimality and finite sample analysis of softmax off-policy actor critic under state distribution mismatch. *arXiv preprint arXiv:2111.02997*, 2021a.

Zhang, X., Chen, Y., Zhu, J., and Sun, W. Corruption-robust offline reinforcement learning. *arXiv preprint arXiv:2106.06630*, 2021b.

Zhang, X., Chen, Y., Zhu, X., and Sun, W. Robust policy gradient against strong data corruption. *arXiv preprint arXiv:2102.05800*, 2021c.

Zhou, Z., Bai, Q., Zhou, Z., Qiu, L., Blanchet, J., and Glynn, P. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *Proc. International Conference on Artifical Intelligence and Statistics (AISTATS)*, pp. 3331–3339. PMLR, 2021.

Appendix

In the following proofs, for a vector $v \in \mathbb{R}^d$, $\|v\|$ denotes its l_2 norm, and $\max v \triangleq \max_i v(i)$ denotes its largest entry. For a matrix A, $\|A\|$ denotes its operator norm. For a policy $\pi_\theta \in \mathit{\Pi}_\Theta$, denote $s_\theta \triangleq \arg\max_s V^{\pi_\theta}(s)$, where V^{π_θ} is the robust value function.

A. Robust Policy Gradient

A.1. Robust Value Function Is Lipschitz

In this section, we show that the robust value function is Lipschitz.

Lemma A.1. *Under Assumption 3.2, the robust value function is Lipschitz in θ, i.e., for any $s \in \mathcal{S}$, any $\theta_1, \theta_2 \in \Theta$,*

$$|V^{\pi_{\theta_1}}(s) - V^{\pi_{\theta_2}}(s)| \leq \frac{k_\pi |\mathcal{A}|}{(1-\gamma)^2}\|\theta_1 - \theta_2\| \triangleq L_V \|\theta_1 - \theta_2\|, \tag{17}$$

where $L_V = \frac{k_\pi|\mathcal{A}|}{(1-\gamma)^2}$.

Proof. For any $\theta_1, \theta_2 \in \Theta$, we have that

$$\begin{aligned}
&V^{\pi_{\theta_1}}(s) - V^{\pi_{\theta_2}}(s) \\
&= \sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s)Q^{\pi_{\theta_1}}(s,a)) - \sum_{a\in\mathcal{A}}(\pi_{\theta_2}(a|s)Q^{\pi_{\theta_2}}(s,a)) \\
&= \sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s)Q^{\pi_{\theta_1}}(s,a)) - \sum_{a\in\mathcal{A}}(\pi_{\theta_2}(a|s)Q^{\pi_{\theta_1}}(s,a)) + \sum_{a\in\mathcal{A}}(\pi_{\theta_2}(a|s)Q^{\pi_{\theta_1}}(s,a)) - \sum_{a\in\mathcal{A}}(\pi_{\theta_2}(a|s)Q^{\pi_{\theta_2}}(s,a)) \\
&= \sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s) - \pi_{\theta_2}(a|s))Q^{\pi_{\theta_1}}(s,a) + \sum_{a\in\mathcal{A}}\pi_{\theta_2}(a|s)(Q^{\pi_{\theta_1}}(s,a) - Q^{\pi_{\theta_2}}(s,a)) \\
&\overset{(a)}{=} \sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s) - \pi_{\theta_2}(a|s))Q^{\pi_{\theta_1}}(s,a) \\
&\quad + \sum_{a\in\mathcal{A}}\pi_{\theta_2}(a|s)\left(\gamma(1-R)\sum_{s'\in\mathcal{S}}p^a_{s,s'}\left(V^{\pi_{\theta_1}}(s') - V^{\pi_{\theta_2}}(s')\right) + \gamma R\left(\max V^{\pi_{\theta_1}} - \max V^{\pi_{\theta_2}}\right)\right) \\
&\overset{(b)}{=} \sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s) - \pi_{\theta_2}(a|s))Q^{\pi_{\theta_1}}(s,a) \\
&\quad + \gamma(1-R)\sum_{s'\in\mathcal{S}}\mathbb{P}(S_1 = s'|S_0 = s, \pi_{\theta_2})\underbrace{\left(V^{\pi_{\theta_1}}(s') - V^{\pi_{\theta_2}}(s')\right)}_{(c)} + \gamma R\left(\max V^{\pi_{\theta_1}} - \max V^{\pi_{\theta_2}}\right),
\end{aligned} \tag{18}$$

where the equation (a) is from the fact that Q^π is the fixed point of the robust Bellman operator $\mathbf{T}_\pi$, i.e., $Q^\pi(s,a) = c(s,a) + \gamma(1-R)\sum_{s'\in\mathcal{S}}p^a_{s,s'}V^\pi(s') + \gamma R \max V^\pi$, where $V^\pi(s) = \sum_{a\in\mathcal{A}}\pi(a|s)Q^\pi(s,a)$, and equation (b) is because $\mathbb{P}(S_1 = s'|S_0 = s, \pi_{\theta_2}) = \sum_{a\in\mathcal{A}}\pi_{\theta_2}(a|s)p^a_{s,s'}$. Note that the term (c) $V^{\pi_{\theta_1}}(s') - V^{\pi_{\theta_2}}(s')$ is again the difference between robust value functions, hence apply (18) recursively, and we obtain that

$$\begin{aligned}
&V^{\pi_{\theta_1}}(s) - V^{\pi_{\theta_2}}(s) \\
&= \sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s) - \pi_{\theta_2}(a|s))Q^{\pi_{\theta_1}}(s,a) \\
&\quad + \gamma(1-R)\sum_{s'\in\mathcal{S}}\mathbb{P}(S_1 = s'|S_0 = s, \pi_{\theta_2})(V^{\pi_{\theta_1}}(s') - V^{\pi_{\theta_2}}(s')) + \gamma R\left(\max V^{\pi_{\theta_1}} - \max V^{\pi_{\theta_2}}\right) \\
&= \sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s) - \pi_{\theta_2}(a|s))Q^{\pi_{\theta_1}}(s,a) + \gamma R\left(\max V^{\pi_{\theta_1}} - \max V^{\pi_{\theta_2}}\right)
\end{aligned}$$

$$
\begin{aligned}
&+\gamma(1-R)\sum_{s'\in\mathcal{S}}\mathbb{P}(S_1=s'|S_0=s,\pi_{\theta_2})\Bigg(\sum_{a'}(\pi_{\theta_1}(a'|s')-\pi_{\theta_2}(a'|s'))Q^{\pi_{\theta_1}}(s',a')\\
&+\gamma R\left(\max V^{\pi_{\theta_1}}-\max V^{\pi_{\theta_2}}\right)+\gamma(1-R)\sum_{s''\in\mathcal{S}}\mathbb{P}(S_1=s''|S_0=s',\pi_{\theta_2})(V^{\pi_{\theta_1}}(s'')-V^{\pi_{\theta_2}}(s''))\Bigg)\\
&=\sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s)-\pi_{\theta_2}(a|s))Q^{\pi_{\theta_1}}(s,a)+\gamma R\left(\max V^{\pi_{\theta_1}}-\max V^{\pi_{\theta_2}}\right)\\
&+\gamma(1-R)\sum_{s'\in\mathcal{S}}\mathbb{P}(S_1=s'|S_0=s,\pi_{\theta_2})\sum_{a'}(\pi_{\theta_1}(a'|s')-\pi_{\theta_2}(a'|s'))Q^{\pi_{\theta_1}}(s',a')\\
&+\gamma R\gamma(1-R)\left(\max V^{\pi_{\theta_1}}-\max V^{\pi_{\theta_2}}\right)\\
&+\gamma^2(1-R)^2\sum_{s'\in\mathcal{S}}\mathbb{P}(S_1=s'|S_0=s,\pi_{\theta_2})\sum_{s''\in\mathcal{S}}\mathbb{P}(S_1=s''|S_0=s',\pi_{\theta_2})(V^{\pi_{\theta_1}}(s'')-V^{\pi_{\theta_2}}(s''))\\
&=\sum_{s'\in\mathcal{S}}\left(\mathbb{P}(S_0=s'|S_0=s,\pi_{\theta_2})+\gamma(1-R)\mathbb{P}(S_1=s'|S_0=s,\pi_{\theta_2})\right)\left(\sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s')-\pi_{\theta_2}(a|s'))Q^{\pi_{\theta_1}}(x,a)\right)\\
&+\gamma R(1+\gamma(1-R))\left(\max V^{\pi_{\theta_1}}-\max V^{\pi_{\theta_2}}\right)\\
&+\gamma^2(1-R)^2\sum_{s'\in\mathcal{S}}\mathbb{P}(S_1=s'|S_0=s,\pi_{\theta_2})\sum_{s''\in\mathcal{S}}\mathbb{P}(S_1=s''|S_0=s',\pi_{\theta_2})(V^{\pi_{\theta_1}}(s'')-V^{\pi_{\theta_2}}(s'')).
\end{aligned}
\tag{19}
$$

Note that in the last term, $\sum_{s'\in\mathcal{S}}\mathbb{P}(S_1=s'|S_0=s,\pi_{\theta_2})\mathbb{P}(S_1=s''|S_0=s',\pi_{\theta_2})=\mathbb{P}(S_2=s''|S_0=s,\pi_{\theta_2})$, hence the last term can be written as $\gamma^2(1-R)^2\sum_{s'\in\mathcal{S}}\mathbb{P}(S_2=s'|S_0=s,\pi_{\theta_2})(V^{\pi_{\theta_1}}(s')-V^{\pi_{\theta_2}}(s'))$. Then the difference between robust value functions can be further written as

$$
\begin{aligned}
&V^{\pi_{\theta_1}}(s)-V^{\pi_{\theta_2}}(s)\\
&=\sum_{s'\in\mathcal{S}}\left(\mathbb{P}(S_0=s'|S_0=s,\pi_{\theta_2})+\gamma(1-R)\mathbb{P}(S_1=s'|S_0=s,\pi_{\theta_2})\right)\left(\sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s')-\pi_{\theta_2}(a|s'))Q^{\pi_{\theta_1}}(s',a)\right)\\
&+\gamma R(1+\gamma(1-R))\left(\max V^{\pi_{\theta_1}}-\max V^{\pi_{\theta_2}}\right)\\
&+\gamma^2(1-R)^2\sum_{s'\in\mathcal{S}}\mathbb{P}(S_2=s'|S_0=s,\pi_{\theta_2})(V^{\pi_{\theta_1}}(s')-V^{\pi_{\theta_2}}(s')).
\end{aligned}
\tag{20}
$$

Recursively applying this equation, we have that

$$
\begin{aligned}
&V^{\pi_{\theta_1}}(s)-V^{\pi_{\theta_2}}(s)\\
&=\sum_{s'\in\mathcal{S}}\left(\sum_{k=0}^{\infty}(\gamma(1-R))^k\mathbb{P}(S_k=s'|S_0=s,\pi_{\theta_2})\left(\sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s')-\pi_{\theta_2}(a|s'))Q^{\pi_{\theta_1}}(s',a)\right)\right)\\
&\quad+\gamma R(\sum_{k=0}^{\infty}(\gamma(1-R))^k)\left(\max V^{\pi_{\theta_1}}-\max V^{\pi_{\theta_2}}\right)\\
&=\sum_{s'\in\mathcal{S}}\left(\sum_{k=0}^{\infty}(\gamma(1-R))^k\mathbb{P}(S_k=s'|S_0=s,\pi_{\theta_2})\left(\sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s')-\pi_{\theta_2}(a|s'))Q^{\pi_{\theta_1}}(s',a)\right)\right)\\
&\quad+\frac{\gamma R}{1-\gamma+\gamma R}\left(\max V^{\pi_{\theta_1}}-\max V^{\pi_{\theta_2}}\right).
\end{aligned}
\tag{21}
$$

The first term can be treated as an expectation under the discounted visitation distribution:

$$
\begin{aligned}
V^{\pi_{\theta_1}}(s)-V^{\pi_{\theta_2}}(s)=&\frac{1}{1-\gamma(1-R)}\sum_{s'\in\mathcal{S}}\left(d_s^{\pi_{\theta_2}}(s')\left(\sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s')-\pi_{\theta_2}(a|s'))Q^{\pi_{\theta_1}}(s',a)\right)\right)\\
&+\frac{\gamma R}{1-\gamma+\gamma R}\left(\max V^{\pi_{\theta_1}}-\max V^{\pi_{\theta_2}}\right),
\end{aligned}
\tag{22}
$$

where $d_s^{\pi_{\theta_2}}(s') = (1-\gamma(1-R))\sum_{k=0}^{\infty}\gamma^k(1-R)^k\left(\mathbb{P}(S_k = s'|S_0 = s_0, \pi)\right)$.

Taking absolute value on both sides, we then have that

$$|V^{\pi_{\theta_1}}(s) - V^{\pi_{\theta_2}}(s)| = \left|\frac{1}{1-\gamma(1-R)}\sum_{s'\in\mathcal{S}}\left(d_s^{\pi_{\theta_2}}(s')\left(\sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s') - \pi_{\theta_2}(a|s'))Q^{\pi_{\theta_1}}(s',a)\right)\right) + \frac{\gamma R}{1-\gamma+\gamma R}\left(\max V^{\pi_{\theta_1}} - \max V^{\pi_{\theta_2}}\right)\right|. \tag{23}$$

We then consider the second term $\max V^{\pi_{\theta_1}} - \max V^{\pi_{\theta_2}}$ in (23). Without loss of generality, assume that $V^{\pi_{\theta_1}}(s_{\theta_1}) \geq V^{\pi_{\theta_2}}(s_{\theta_2})$. Then $|\max V^{\pi_{\theta_1}} - \max V^{\pi_{\theta_2}}| = V^{\pi_{\theta_1}}(s_{\theta_1}) - V^{\pi_{\theta_2}}(s_{\theta_2}) = V^{\pi_{\theta_1}}(s_{\theta_1}) - V^{\pi_{\theta_2}}(s_{\theta_1}) + V^{\pi_{\theta_2}}(s_{\theta_1}) - V^{\pi_{\theta_2}}(s_{\theta_2}) \leq V^{\pi_{\theta_1}}(s_{\theta_1}) - V^{\pi_{\theta_2}}(s_{\theta_1})$ because $V^{\pi_{\theta_2}}(s_{\theta_1}) \leq V^{\pi_{\theta_2}}(s_{\theta_2})$. Then (23) can be written as

$$|V^{\pi_{\theta_1}}(s) - V^{\pi_{\theta_2}}(s)| \leq \left|\frac{1}{1-\gamma(1-R)}\sum_{s'\in\mathcal{S}}\left(d_s^{\pi_{\theta_2}}(s')\left(\sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s') - \pi_{\theta_2}(a|s'))Q^{\pi_{\theta_1}}(s',a)\right)\right)\right| + \frac{\gamma R}{1-\gamma+\gamma R}\left|\left(V^{\pi_{\theta_1}}(s_{\theta_1}) - V^{\pi_{\theta_2}}(s_{\theta_1})\right)\right|. \tag{24}$$

Note that this inequality holds for any state s. By setting $s = s_{\theta_1}$ we have

$$|V^{\pi_{\theta_1}}(s_{\theta_1}) - V^{\pi_{\theta_2}}(s_{\theta_1})| \leq \left|\frac{1}{1-\gamma(1-R)}\sum_{s'\in\mathcal{S}}\left(d_{s_{\theta_1}}^{\pi_{\theta_2}}(s')\left(\sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s') - \pi_{\theta_2}(a|s'))Q^{\pi_{\theta_1}}(s',a)\right)\right)\right| + \frac{\gamma R}{1-\gamma+\gamma R}\left|\left(V^{\pi_{\theta_1}}(s_{\theta_1}) - V^{\pi_{\theta_2}}(s_{\theta_1})\right)\right|, \tag{25}$$

which implies that

$$|V^{\pi_{\theta_1}}(s_{\theta_1}) - V^{\pi_{\theta_2}}(s_{\theta_1})| \leq \left|\frac{1}{1-\gamma}\sum_{s'\in\mathcal{S}}\left(d_{s_{\theta_1}}^{\pi_{\theta_2}}(s')\left(\sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s') - \pi_{\theta_2}(a|s'))Q^{\pi_{\theta_1}}(s',a)\right)\right)\right|. \tag{26}$$

Plugging this inequality in (24) implies that

$$\begin{aligned}
&|V^{\pi_{\theta_1}}(s) - V^{\pi_{\theta_2}}(s)| \\
&\leq \left|\frac{1}{1-\gamma(1-R)}\sum_{s'\in\mathcal{S}}\left(d_s^{\pi_{\theta_2}}(s')\left(\sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s') - \pi_{\theta_2}(a|s'))Q^{\pi_{\theta_1}}(s',a)\right)\right)\right| \\
&\quad + \left|\frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)}\sum_{s'\in\mathcal{S}}\left(d_{s_{\theta_1}}^{\pi_{\theta_2}}(s')\left(\sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s') - \pi_{\theta_2}(a|s'))Q^{\pi_{\theta_1}}(s',a)\right)\right)\right|.
\end{aligned} \tag{27}$$

Note that from $c(s,a) \in [0,1]$, $Q^{\pi}(s,a) = \max_{\kappa}\mathbb{E}_{\kappa}[\sum_{t=0}^{\infty}\gamma^t c(S_t,A_t)|S_0 = s, A_0 = a, \pi] \leq \frac{1}{1-\gamma}$ for any π, s, a; and from Assumption 3.2, $|\pi_{\theta_1}(a|s') - \pi_{\theta_2}(a|s')| \leq k_{\pi}\|\theta_1 - \theta_2\|$, hence we have that

$$\begin{aligned}
|V^{\pi_{\theta_1}}(s) - V^{\pi_{\theta_2}}(s)| &\leq \left(\frac{1}{1-\gamma(1-R)} + \frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)}\right)\frac{k_{\pi}|\mathcal{A}|}{1-\gamma}\|\theta_1 - \theta_2\| \\
&= \frac{k_{\pi}|\mathcal{A}|}{(1-\gamma)^2}\|\theta_1 - \theta_2\|,
\end{aligned} \tag{28}$$

which completes the proof. □

The following corollary is straightforward hence the proof is omitted:

Corollary A.2. *The robust action-value functions $Q^{\pi_\theta}(s,a)$ for any $s \in \mathcal{S}, a \in \mathcal{A}$ and the objective function $J_\rho(\theta) = \sum_{s\in\mathcal{S}}\rho(s)V^{\pi_\theta}(s)$ are Lipschitz in θ with constant L_V, i.e., for any $\theta_1, \theta_2 \in \Theta$,*

$$|Q^{\pi_{\theta_1}}(s,a) - Q^{\pi_{\theta_2}}(s,a)| \leq L_V\|\theta_1 - \theta_2\|, \tag{29}$$

$$|J_\rho(\theta_1) - J_\rho(\theta_2)| \leq L_V\|\theta_1 - \theta_2\|. \tag{30}$$

A.2. Proof of Theorem 3.3: Sub-gradient of Robust Value Function

In this section, we derive the sub-gradient of robust value function $J_\rho(\theta)$ and prove Theorem 3.3.

Denote $s_\theta = \arg\max_s V^{\pi_\theta}(s)$, and define

$$\hat{\phi}_s(\theta) \triangleq \frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)} \sum_{s' \in \mathcal{S}} d_{s_\theta}^{\pi_\theta}(s') \sum_{a \in \mathcal{A}} \nabla \pi_\theta(a|s') Q^{\pi_\theta}(s', a) + \frac{1}{1-\gamma+\gamma R} \sum_{s' \in \mathcal{S}} d_s^{\pi_\theta}(s') \sum_{a \in \mathcal{A}} \nabla \pi_\theta(a|s') Q^{\pi_\theta}(s', a), \tag{31}$$

Recall from (7) that $\psi_\rho(\theta)$ is the average of $\hat{\phi}_s(\theta)$ under distribution ρ. In the following, we show that $\hat{\phi}_s(\theta) \in \partial V^{\pi_\theta}(s)$, and hence naturally, $\psi_\rho(\theta) \in \partial J_\rho(\theta)$. In Section A.2.1, we will demonstrate how we derive the expression of $\hat{\phi}_s(\theta)$.

From (22) we can show that

$$\begin{aligned}
V^{\pi_{\theta_1}}(s) - V^{\pi_{\theta_2}}(s) &= \frac{1}{1-\gamma(1-R)} \sum_{s' \in \mathcal{S}} \left(d_s^{\pi_{\theta_2}}(s') \left(\sum_{a \in \mathcal{A}} (\pi_{\theta_1}(a|s') - \pi_{\theta_2}(a|s')) Q^{\pi_{\theta_1}}(s', a) \right) \right) \\
&\quad + \frac{\gamma R}{1-\gamma+\gamma R} \left(V^{\pi_{\theta_1}}(s_{\theta_1}) - V^{\pi_{\theta_2}}(s_{\theta_2}) \right) \\
&= \frac{1}{1-\gamma(1-R)} \sum_{s' \in \mathcal{S}} \left(d_s^{\pi_{\theta_2}}(s') \left(\sum_{a \in \mathcal{A}} (\pi_{\theta_1}(a|s') - \pi_{\theta_2}(a|s')) Q^{\pi_{\theta_1}}(s', a) \right) \right) \\
&\quad + \frac{\gamma R}{1-\gamma+\gamma R} \left(V^{\pi_{\theta_1}}(s_{\theta_1}) - V^{\pi_{\theta_1}}(s_{\theta_2}) + V^{\pi_{\theta_1}}(s_{\theta_2}) - V^{\pi_{\theta_2}}(s_{\theta_2}) \right) \\
&\geq \frac{1}{1-\gamma(1-R)} \sum_{s' \in \mathcal{S}} \left(d_s^{\pi_{\theta_2}}(s') \left(\sum_{a \in \mathcal{A}} (\pi_{\theta_1}(a|s') - \pi_{\theta_2}(a|s')) Q^{\pi_{\theta_1}}(s', a) \right) \right) \\
&\quad + \frac{\gamma R}{1-\gamma+\gamma R} \left(V^{\pi_{\theta_1}}(s_{\theta_2}) - V^{\pi_{\theta_2}}(s_{\theta_2}) \right),
\end{aligned} \tag{32}$$

where the last inequality is from $V^{\pi_{\theta_1}}(s_{\theta_1}) - V^{\pi_{\theta_1}}(s_{\theta_2}) \geq 0$. Set $s = s_{\theta_2}$ in the LHS of (32), we further have that

$$V^{\pi_{\theta_1}}(s_{\theta_2}) - V^{\pi_{\theta_2}}(s_{\theta_2}) \geq \frac{1}{(1-\gamma)} \sum_{s' \in \mathcal{S}} \left(d_{s_{\theta_2}}^{\pi_{\theta_2}}(s') \left(\sum_{a \in \mathcal{A}} (\pi_{\theta_1}(a|s') - \pi_{\theta_2}(a|s')) Q^{\pi_{\theta_1}}(s', a) \right) \right). \tag{33}$$

Plug (33) back into (32), it follows that

$$\begin{aligned}
V^{\pi_{\theta_1}}(s) - V^{\pi_{\theta_2}}(s) &\geq \frac{1}{1-\gamma(1-R)} \sum_{s' \in \mathcal{S}} \left(d_s^{\pi_{\theta_2}}(s') \left(\sum_{a \in \mathcal{A}} (\pi_{\theta_1}(a|s') - \pi_{\theta_2}(a|s')) Q^{\pi_{\theta_1}}(s', a) \right) \right) \\
&\quad + \frac{\gamma R}{(1-\gamma+\gamma R)(1-\gamma)} \sum_{s' \in \mathcal{S}} \left(d_{s_{\theta_2}}^{\pi_{\theta_2}}(s') \left(\sum_{a \in \mathcal{A}} (\pi_{\theta_1}(a|s') - \pi_{\theta_2}(a|s')) Q^{\pi_{\theta_1}}(s', a) \right) \right).
\end{aligned} \tag{34}$$

Hence for any $\theta, \theta + h \in \Theta$, we have that

$$\begin{aligned}
V^{\pi_{\theta+h}}(s) - V^{\pi_\theta}(s) &\geq \frac{1}{1-\gamma(1-R)} \sum_{s' \in \mathcal{S}} d_s^{\pi_\theta}(s') \sum_{a \in \mathcal{A}} \left(\pi_{\theta+h}(a|s') - \pi_\theta(a|s') \right) Q^{\pi_{\theta+h}}(s', a) \\
&\quad + \frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)} \sum_{s' \in \mathcal{S}} d_{s_\theta}^{\pi_\theta}(s') \sum_{a \in \mathcal{A}} \left(\pi_{\theta+h}(a|s') - \pi_\theta(a|s') \right) Q^{\pi_{\theta+h}}(s', a).
\end{aligned} \tag{35}$$

We then can show that

$$V^{\pi_{\theta+h}}(s) - V^{\pi_\theta}(s) - \langle \hat{\phi}_s(\theta), h \rangle \geq \frac{1}{1-\gamma(1-R)} \sum_{s' \in \mathcal{S}} d_s^{\pi_\theta}(s') \sum_{a \in \mathcal{A}} \left(\pi_{\theta+h}(a|s') - \pi_\theta(a|s') \right) Q^{\pi_{\theta+h}}(s', a)$$

$$
\begin{aligned}
&- \frac{1}{1-\gamma+\gamma R} \sum_{s' \in \mathcal{S}} d_s^{\pi_\theta}(s') \sum_{a \in \mathcal{A}} \langle h, \nabla \pi_\theta(a|s') \rangle Q^{\pi_\theta}(s', a) \\
&+ \frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)} \sum_{s' \in \mathcal{S}} d_{s_\theta}^{\pi_\theta}(s') \sum_{a \in \mathcal{A}} \left(\pi_{\theta+h}(a|s') - \pi_\theta(a|s')\right) Q^{\pi_{\theta+h}}(s', a) \\
&- \frac{\gamma R}{1-\gamma+\gamma R} \frac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} d_{s_\theta}^{\pi_\theta}(s') \sum_{a \in \mathcal{A}} \langle h, \nabla \pi_\theta(a|s') \rangle Q^{\pi_\theta}(s', a). \qquad (36)
\end{aligned}
$$

The first two terms in (36) can be rewritten as follows:

$$
\begin{aligned}
&\frac{1}{1-\gamma(1-R)} \sum_{s' \in \mathcal{S}} d_s^{\pi_\theta}(s') \sum_{a \in \mathcal{A}} \left(\pi_{\theta+h}(a|s') - \pi_\theta(a|s')\right) Q^{\pi_{\theta+h}}(s', a) \\
&\quad - \frac{1}{1-\gamma+\gamma R} \sum_{s' \in \mathcal{S}} d_s^{\pi_\theta}(s') \sum_{a \in \mathcal{A}} \langle h, \nabla \pi_\theta(a|s') \rangle Q^{\pi_\theta}(s', a) \\
&= \frac{1}{1-\gamma(1-R)} \sum_{s' \in \mathcal{S}} d_s^{\pi_\theta}(s') \sum_{a \in \mathcal{A}} \left(\left(\pi_{\theta+h}(a|s') - \pi_\theta(a|s')\right) Q^{\pi_{\theta+h}}(s', a) - \langle h, \nabla \pi_\theta(a|s') \rangle Q^{\pi_\theta}(s', a) \right). \qquad (37)
\end{aligned}
$$

Moreover note that

$$
\begin{aligned}
&\left(\pi_{\theta+h}(a|s') - \pi_\theta(a|s')\right) Q^{\pi_{\theta+h}}(s', a) - \langle h, \nabla \pi_\theta(a|s') \rangle Q^{\pi_\theta}(s', a) \\
&= \left(\pi_{\theta+h}(a|s') - \pi_\theta(a|s')\right) Q^{\pi_{\theta+h}}(s', a) - \langle h, \nabla \pi_\theta(a|s') \rangle Q^{\pi_{\theta+h}}(s', a) \\
&\quad + \langle h, \nabla \pi_\theta(a|s') \rangle Q^{\pi_{\theta+h}}(s', a) - \langle h, \nabla \pi_\theta(a|s') \rangle Q^{\pi_\theta}(s', a). \qquad (38)
\end{aligned}
$$

Since π_θ is differentiable, hence

$$
\lim_{\|h\| \to 0} \frac{\pi_{\theta+h}(a|s') - \pi_\theta(a|s') - \langle h, \nabla \pi_\theta(a|s') \rangle}{\|h\|} = 0. \qquad (39)
$$

Because $Q^{\pi_\theta} \geq 0$ for any θ, we further have that

$$
\lim_{\|h\| \to 0} \frac{\left(\pi_{\theta+h}(a|s') - \pi_\theta(a|s')\right) Q^{\pi_{\theta+h}}(s', a) - \langle h, \nabla \pi_\theta(a|s') \rangle Q^{\pi_{\theta+h}}(s', a)}{\|h\|} \geq 0. \qquad (40)
$$

For the remaining term in (38), note that $\lim_{h \to 0} Q^{\pi_{\theta+h}}(s', a) - Q^{\pi_\theta}(s', a) = 0$ as Q^{π_θ} is Lipschitz in θ, thus

$$
\begin{aligned}
&\lim_{\|h\| \to 0} \frac{\langle h, \nabla \pi_\theta(a|s') (Q^{\pi_{\theta+h}}(s', a) - Q^{\pi_\theta}(s', a)) \rangle}{\|h\|} \\
&= \lim_{\|h\| \to 0} \left\langle \frac{h}{\|h\|}, \nabla \pi_\theta(a|s') (Q^{\pi_{\theta+h}}(s', a) - Q^{\pi_\theta}(s', a)) \right\rangle \\
&= \lim_{\|h\| \to 0} \langle e_h, \nabla \pi_\theta(a|s') (Q^{\pi_{\theta+h}}(s', a) - Q^{\pi_\theta}(s', a)) \rangle, \qquad (41)
\end{aligned}
$$

where e_h is the normalized vector of h, and $\nabla \pi_\theta(a|s')(Q^{\pi_{\theta+h}}(s', a) - Q^{\pi_\theta}(s', a))$ is a vector of dimension $|\mathcal{S}| \times |\mathcal{A}|$.

Clearly, $\lim_{\|h\| \to 0} \langle e_h, \nabla \pi_\theta(a|s') (Q^{\pi_{\theta+h}}(s', a) - Q^{\pi_\theta}(s', a)) \rangle = 0$, which is also due to the Lipschitz of Q^{π_θ}. Hence combining all these inequalities in (40) and (41), we have that

$$
\lim_{\|h\| \to 0} \inf_{h \neq 0} \frac{\sum_{s' \in \mathcal{S}} d_s^{\pi_\theta}(s') \sum_{a \in \mathcal{A}} \left(\left(\pi_{\theta+h}(a|s') - \pi_\theta(a|s')\right) Q^{\pi_{\theta+h}}(s', a) - \langle h, \nabla \pi_\theta(a|s') \rangle Q^{\pi_\theta}(s', a) \right)}{\|h\|} \geq 0. \qquad (42)
$$

Similarly, we can also show that for the remaining terms in (36),

$$
\lim_{\|h\| \to 0} \inf_{h \neq 0} \frac{\sum_{s' \in \mathcal{S}} d_{s_\theta}^{\pi_\theta}(s') \sum_{a \in \mathcal{A}} \left(\left(\pi_{\theta+h}(a|s') - \pi_\theta(a|s')\right) Q^{\pi_{\theta+h}}(s', a) - \langle h, \nabla \pi_\theta(a|s') \rangle Q^{\pi_\theta}(s', a) \right)}{\|h\|} \geq 0. \qquad (43)
$$

Hence

$$\lim_{\|h\|\to 0}\inf_{h\neq 0}\frac{V^{\pi_{\theta+h}}(s)-V^{\pi_\theta}(s)-\langle\hat{\phi}_s(\theta),h\rangle}{\|h\|}\geq 0, \tag{44}$$

and this implies that $\hat{\phi}_s(\theta)\in\partial V^{\pi_\theta}(s)$ for any θ.

Now we consider J_ρ. From the definition, $J_\rho(\theta)=\sum_{s\in\mathcal{S}}\rho(s)V^{\pi_\theta}(s)$. Hence we have that

$$\begin{aligned}
\partial J_\rho(\theta) &\supseteq \sum_{s\in\mathcal{S}}\rho(s)\partial V^{\pi_\theta}(s)\\
&\ni \frac{\gamma R}{1-\gamma+\gamma R}\frac{1}{1-\gamma}\sum_{s'\in\mathcal{S}}d_{s_\theta}^{\pi_\theta}(s')\sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s')Q^{\pi_\theta}(s',a)\\
&\quad+\frac{1}{1-\gamma+\gamma R}\sum_{s\in\mathcal{S}}\rho(s)\sum_{s'\in\mathcal{S}}d_s^{\pi_\theta}(s')\sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s')Q^{\pi_\theta}(s',a)\\
&= \frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)}\sum_{s'\in\mathcal{S}}d_{s_\theta}^{\pi_\theta}(s')\sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s')Q^{\pi_\theta}(s',a)\\
&\quad+\frac{1}{1-\gamma+\gamma R}\sum_{s'\in\mathcal{S}}d_\rho^{\pi_\theta}(s')\sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s')Q^{\pi_\theta}(s',a)\\
&\triangleq \psi_\rho(\theta),
\end{aligned} \tag{45}$$

which implies $\psi_\rho(\theta)\in\partial J_\rho(\theta)$.

A.2.1. Derivation of The Sub-gradient $\hat{\phi}_s(\theta)$

In this section we show how we derive the expression of the sub-gradient $\hat{\phi}_s$.

Recall that $\max_s V^{\pi_\theta}(s)$ is Lipschitz in θ as shown in Section A. From the Rademacher's theorem (Federer, 2014), we know that $\max_s V^{\pi_\theta}(s)$ is differentiable almost everywhere. Hence at those differentiable θ, we define $\phi(\theta)=\nabla\max_s V^{\pi_\theta}(s)$, which can be also viewed as a sub-gradient of $\max_s V^{\pi_\theta}(s)$, i.e., $\phi(\theta)\in\partial\max_s V^{\pi_\theta}(s)$.

It is known from (Kruger, 2003) that $\partial(f)+\partial(g)\subseteq\partial(f+g)$, hence we have that for any $s\in\mathcal{S}$,

$$\begin{aligned}
\partial V^{\pi_\theta}(s) &\supseteq \sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s)Q^{\pi_\theta}(s,a)+\sum_{a\in\mathcal{A}}\pi_\theta(a|s)\partial Q^{\pi_\theta}(s,a)\\
&\supseteq \sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s)Q^{\pi_\theta}(s,a)+\sum_{a\in\mathcal{A}}\pi_\theta(a|s)\partial\left(c(s,a)+\gamma(1-R)\sum_{s'\in\mathcal{S}}p_{s,s'}^a V^{\pi_\theta}(s')+\gamma R\max_s V^{\pi_\theta}(s)\right)\\
&\supseteq \sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s)Q^{\pi_\theta}(s,a)+\sum_{a\in\mathcal{A}}\pi_\theta(a|s)\left(\gamma(1-R)\sum_{s'\in\mathcal{S}}p_{s,s'}^a\partial V^{\pi_\theta}(s')+\gamma R\phi(\theta)\right)\\
&= \sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s)Q^{\pi_\theta}(s,a)+\sum_{a\in\mathcal{A}}\pi_\theta(a|s)\left(\gamma(1-R)\sum_{s'\in\mathcal{S}}p_{s,s'}^a\partial V^{\pi_\theta}(s')\right)+\gamma R\phi(\theta)\\
&= \gamma R\phi(\theta)+\sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s)Q^{\pi_\theta}(s,a)+\gamma(1-R)\sum_{s'\in\mathcal{S}}\mathbb{P}(S_1=s'|S_0=s,\pi_\theta)\partial V^{\pi_\theta}(s').
\end{aligned} \tag{46}$$

Recursively applying (46), we have that

$$\begin{aligned}
\partial V^{\pi_\theta}(s) &\supseteq \gamma R\phi(\theta)+\sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s)Q^{\pi_\theta}(s,a)+\gamma(1-R)\sum_{s'\in\mathcal{S}}\mathbb{P}(S_1=s'|S_0=s,\pi_\theta)\partial V^{\pi_\theta}(s')\\
&\supseteq \gamma R\phi(\theta)+\sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s)Q^{\pi_\theta}(s,a)\\
&\quad+\gamma(1-R)\sum_{s'\in\mathcal{S}}\mathbb{P}(S_1=s'|S_0=s,\pi_\theta)\left(\sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s')Q^{\pi_\theta}(s',a)+\sum_{a\in\mathcal{A}}\pi_\theta(a|s')\partial Q^{\pi_\theta}(s',a)\right)
\end{aligned}$$

$$
\begin{aligned}
&\supseteq \gamma R\phi(\theta) + \sum_{a\in\mathcal{A}} \nabla\pi_\theta(a|s) Q^{\pi_\theta}(s,a) + \gamma(1-R)\sum_{s'\in\mathcal{S}} \mathbb{P}(S_1=s'|S_0=s,\pi_\theta)\sum_{a\in\mathcal{A}} \nabla\pi_\theta(a|s') Q^{\pi_\theta}(s',a)\\
&\quad + \gamma(1-R)\sum_{s'\in\mathcal{S}} \mathbb{P}(S_1=s'|S_0=s,\pi_\theta)\sum_{a\in\mathcal{A}} \pi_\theta(a|s')\partial Q^{\pi_\theta}(s',a)\\
&\supseteq \gamma R\phi(\theta) + \sum_{a\in\mathcal{A}} \nabla\pi_\theta(a|s) Q^{\pi_\theta}(s,a) + \gamma(1-R)\sum_{s'\in\mathcal{S}} \mathbb{P}(S_1=s'|S_0=s,\pi_\theta)\sum_{a\in\mathcal{A}} \nabla\pi_\theta(a|s') Q^{\pi_\theta}(s',a)\\
&\quad + \gamma(1-R)\sum_{s'\in\mathcal{S}} \mathbb{P}(S_1=s'|S_0=s,\pi_\theta)\sum_{a\in\mathcal{A}} \pi_\theta(a|s')\left(\gamma(1-R)\sum_{s''\in\mathcal{S}} p^a_{s',s''}\partial V^{\pi_\theta}(s'') + \gamma R\phi(\theta)\right)\\
&= \gamma R\phi(\theta) + \sum_{a\in\mathcal{A}} \nabla\pi_\theta(a|s) Q^{\pi_\theta}(s,a) + \gamma(1-R)\sum_{s'\in\mathcal{S}} \mathbb{P}(S_1=s'|S_0=s,\pi_\theta)\sum_{a\in\mathcal{A}} \nabla\pi_\theta(a|s') Q^{\pi_\theta}(s',a)\\
&\quad + \gamma^2(1-R)^2\sum_{s''\in\mathcal{S}} \mathbb{P}(S_2=s''|S_0=s,\pi_\theta)\partial V^{\pi_\theta}(s'') + \gamma^2 R(1-R)\phi(\theta)\\
&\supseteq \ldots\\
&\ni \gamma R\left(\sum_{k=0}^{\infty}\gamma^k(1-R)^k\right)\phi(\theta) + \sum_k\left(\gamma^k(1-R)^k\sum_{s'\in\mathcal{S}}\mathbb{P}(S_k=s'|S_0=s,\pi_\theta)\sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s')Q^{\pi_\theta}(s',a)\right)\\
&= \frac{\gamma R}{1-\gamma+\gamma R}\phi(\theta) + \frac{1}{1-\gamma+\gamma R}\sum_{s'\in\mathcal{S}} d_s^{\pi_\theta}(s')\sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s')Q^{\pi_\theta}(s',a). \qquad (47)
\end{aligned}
$$

From Lemma D.1, we have that $\left\{\partial V^\theta(s)|_{s=s_\theta} : s_\theta \in \arg\max_s V^{\pi_\theta}(s)\right\} \subseteq \partial \max V^\theta(s)$. Set $s = s_\theta$ in (47), we have that

$$
\begin{aligned}
\partial \max_s V^{\pi_\theta}(s) = \{\phi(\theta)\} \supseteq \partial V^{\pi_\theta}|_{s_\theta} &\ni \frac{\gamma R}{1-\gamma+\gamma R}\phi(\theta) + \frac{1}{1-\gamma+\gamma R}\sum_{s'\in\mathcal{S}} d_{s_\theta}^{\pi_\theta}(s')\sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s')Q^{\pi_\theta}(s',a)\\
&\triangleq c_1\phi(\theta) + c_2, \qquad (48)
\end{aligned}
$$

where $c_1 = \frac{\gamma R}{1-\gamma+\gamma R}$ and $c_2 = \frac{1}{1-\gamma+\gamma R}\sum_{s'\in\mathcal{S}} d_{s_\theta}^{\pi_\theta}(s')\sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s')Q^{\pi_\theta}(s',a)$. Hence we have that

$$
\phi(\theta) = c_1\phi(\theta) + c_2, \qquad (49)
$$

and

$$
\phi(\theta) = \frac{c_2}{1-c_1} = \frac{1}{1-\gamma}\sum_{s'\in\mathcal{S}} d_{s_\theta}^{\pi_\theta}(s')\sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s')Q^{\pi_\theta}(s',a). \qquad (50)
$$

Hence we get an explicit expression of the gradient of $\max_s V^{\pi_\theta}(s)$. We then plug it in (47), we further have that

$$
\begin{aligned}
&\frac{\gamma R}{1-\gamma+\gamma R}\frac{1}{1-\gamma}\sum_{s'\in\mathcal{S}} d_{s_\theta}^{\pi_\theta}(s')\sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s')Q^{\pi_\theta}(s',a) + \frac{1}{1-\gamma+\gamma R}\sum_{s'\in\mathcal{S}} d_s^{\pi_\theta}(s')\sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s')Q^{\pi_\theta}(s',a)\\
&\in \partial V^{\pi_\theta}(s), \qquad (51)
\end{aligned}
$$

for any θ such that $\max_s V^{\pi_\theta}(s)$ is differentiable.

As we showed in the last section, at any non-differentiable θ, (51) is also a sub-gradient of $V^{\pi_\theta}(s)$.

A.3. Proof of Theorem 4.1: Global Optimality: PL-Condition under Direct Parametrization

In this section, we show that the sub-gradient in Theorem 3.3 satisfies the PL-condition under the direct parametrization. Note that each entry of $\psi_\mu(\theta)$ can be written as

$$
\psi_\mu(\theta)_{s,b} = \frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)}\sum_{s'\in\mathcal{S}} d_{s_\theta}^{\pi_\theta}(s')\sum_{a\in\mathcal{A}}(\nabla\pi_\theta(a|s'))_{s,b}Q^{\pi_\theta}(s',a)
$$

$$+\frac{1}{1-\gamma+\gamma R}\sum_{s'\in\mathcal{S}}d_{\mu}^{\pi_\theta}(s')\sum_{a\in\mathcal{A}}(\nabla\pi_\theta(a|s'))_{s,b}Q^{\pi_\theta}(s',a). \tag{52}$$

When π_θ is directly parameterized, i.e., $\pi_\theta(a|s)=\theta_{s,a}$, $(\nabla\pi_\theta(a|s'))_{s,b}=\mathbb{1}_{(s,b)=(s',a)}$. Hence, $\psi(\theta)_{s,b}$ for any $s\in\mathcal{S},b\in\mathcal{A}$ can be further written as

$$\psi_\mu(\theta)_{s,b}=\frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)}d_{s_\theta}^{\pi_\theta}(s)Q^{\pi_\theta}(s,b)+\frac{1}{1-\gamma+\gamma R}d_{\mu}^{\pi_\theta}(s)Q^{\pi_\theta}(s,b). \tag{53}$$

Similar to (23) to (27) without taking the absolute value, we have that

$$\begin{aligned}
&V^{\pi_{\theta_1}}(s)-V^{\pi_{\theta_2}}(s)\\
&\le\frac{1}{1-\gamma(1-R)}\sum_{s'\in\mathcal{S}}\left(d_s^{\pi_{\theta_2}}(s')\left(\sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s')-\pi_{\theta_2}(a|s'))Q^{\pi_{\theta_1}}(s',a)\right)\right)\\
&\quad+\frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)}\sum_{s'\in\mathcal{S}}\left(d_{s_{\theta_1}}^{\pi_{\theta_2}}(s')\left(\sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s')-\pi_{\theta_2}(a|s'))Q^{\pi_{\theta_1}}(s',a)\right)\right).
\end{aligned} \tag{54}$$

Set $\theta_1=\theta$ and $\theta_2=\theta^*$, where $\theta^*\in\arg\min_{(\Delta(\mathcal{A}))^{|\mathcal{S}|}}J_\rho(\theta)$, then we have

$$\begin{aligned}
&V^{\pi_\theta}(s)-V^{\pi_{\theta^*}}(s)\\
&\le\frac{1}{1-\gamma(1-R)}\sum_{s'\in\mathcal{S}}\left(d_s^{\pi_{\theta*}}(s')\left(\sum_{a\in\mathcal{A}}(\pi_\theta(a|s')-\pi_{\theta*}(a|s'))Q^{\pi_\theta}(s',a)\right)\right)\\
&\quad+\frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)}\sum_{s'\in\mathcal{S}}\left(d_{s_\theta}^{\pi_{\theta*}}(s')\left(\sum_{a\in\mathcal{A}}(\pi_\theta(a|s')-\pi_{\theta*}(a|s'))Q^{\pi_\theta}(s',a)\right)\right).
\end{aligned} \tag{55}$$

Note that $J_\rho(\theta)=\sum_{s\in\mathcal{S}}\rho(s)V^{\pi_\theta}(s)$, hence

$$\begin{aligned}
J_\rho(\theta)-J^*&=\sum_{s\in\mathcal{S}}\rho(s)(V^{\pi_\theta}(s)-V^{\pi_{\theta^*}}(s))\\
&\le\sum_{s\in\mathcal{S}}\rho(s)\left(\frac{1}{1-\gamma(1-R)}\sum_{s'\in\mathcal{S}}\left(d_s^{\pi_{\theta*}}(s')\left(\sum_{a\in\mathcal{A}}(\pi_\theta(a|s')-\pi_{\theta*}(a|s'))Q^{\pi_\theta}(s',a)\right)\right)\right.\\
&\quad\left.+\frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)}\sum_{s'\in\mathcal{S}}\left(d_{s_\theta}^{\pi_{\theta*}}(s')\left(\sum_{a\in\mathcal{A}}(\pi_\theta(a|s')-\pi_{\theta*}(a|s'))Q^{\pi_\theta}(s',a)\right)\right)\right)\\
&=\sum_{s'\in\mathcal{S}}\left(\frac{1}{1-\gamma(1-R)}d_\rho^{\pi_{\theta^*}}(s')+\frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)}d_{s_\theta}^{\pi_{\theta^*}}(s')\right)\left(\sum_{a\in\mathcal{A}}(\pi_\theta(a|s')-\pi_{\theta*}(a|s'))Q^{\pi_\theta}(s',a)\right)\\
&=\sum_{s'\in\mathcal{S}}\frac{\frac{1}{1-\gamma(1-R)}d_\rho^{\pi_{\theta^*}}(s')+\frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)}d_{s_\theta}^{\pi_{\theta^*}}(s')}{\frac{1}{1-\gamma+\gamma R}d_\mu^{\pi_\theta}(s')+\frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)}d_{s_\theta}^{\pi_\theta}(s')}\left(\frac{1}{1-\gamma+\gamma R}d_\mu^{\pi_\theta}(s')+\frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)}d_{s_\theta}^{\pi_\theta}(s')\right)\\
&\quad\cdot\left(\sum_{a\in\mathcal{A}}(\pi_\theta(a|s')-\pi_{\theta*}(a|s'))Q^{\pi_\theta}(s',a)\right)\\
&=\sum_{s'\in\mathcal{S}}\frac{l_\theta^*(s')}{l_\theta(s')}l_\theta(s')\cdot\left(\sum_{a\in\mathcal{A}}(\pi_\theta(a|s')-\pi_{\theta*}(a|s'))Q^{\pi_\theta}(s',a)\right),
\end{aligned} \tag{56}$$

where $l_\theta(s)\triangleq\left(\frac{1}{1-\gamma+\gamma R}d_\mu^{\pi_\theta}(s)+\frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)}d_{s_\theta}^{\pi_\theta}(s)\right)$ and $l_\theta^*(s)\triangleq\frac{1}{1-\gamma(1-R)}d_\rho^{\pi_{\theta^*}}(s)+\frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)}d_{s_\theta}^{\pi_{\theta^*}}(s)$. Recall (53), then $\psi_\mu(\theta)_{s,b}=l_\theta(s)Q^{\pi_\theta}(s,b)$.

The ratio of distribution $\frac{l_\theta^*(s)}{l_\theta(s)}$ can be bounded as follows. Note that

$$\frac{1}{1-\gamma+\gamma R}d_\mu^{\pi_\theta}(s)+\frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)}d_{s_\theta}^{\pi_\theta}(s)\ge\frac{1}{1-\gamma+\gamma R}d_\mu^{\pi_\theta}(s)$$

$$\geq \frac{1}{1-\gamma+\gamma R}(1-\gamma+\gamma R)\mu(s)$$
$$\geq \mu_{\min}, \tag{57}$$

which is from $d_\mu^{\pi_\theta}(s) \geq (1-\gamma+\gamma R)\mu(s)$. Hence the ratio can be bounded as

$$\frac{\frac{1}{1-\gamma(1-R)}d_\rho^{\pi_{\theta*}}(s) + \frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)}d_{s_\theta}^{\pi_{\theta*}}(s)}{\frac{1}{1-\gamma+\gamma R}d_\mu^{\pi_\theta}(s) + \frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)}d_{s_\theta}^{\pi_\theta}(s)} \leq \frac{1}{(1-\gamma)\mu_{\min}} \triangleq C_{PL}. \tag{58}$$

Note that (56) can be further bounded as

$$\begin{aligned}
&\sum_{s'\in\mathcal{S}} \frac{l_\theta^*(s')}{l_\theta(s')} l_\theta(s') \cdot \left(\sum_{a\in\mathcal{A}} (\pi_\theta(a|s') - \pi_{\theta*}(a|s'))Q^{\pi_\theta}(s',a)\right) \\
&= \sum_{s'\in\mathcal{S}} l_\theta^*(s') \left(\sum_{a\in\mathcal{A}} (\pi_\theta(a|s') - \pi_{\theta*}(a|s'))Q^{\pi_\theta}(s',a)\right) \\
&= \sum_{s'\in\mathcal{S}} l_\theta^*(s') \langle \pi_\theta(\cdot|s') - \pi_{\theta*}(\cdot|s'), Q^{\pi_\theta}(s',\cdot)\rangle \\
&\leq \sum_{s'\in\mathcal{S}} l_\theta^*(s') \max_{\bar{\pi}_{s'}(\cdot|s')\in\Delta(\mathcal{A})} \langle \pi_\theta(\cdot|s') - \bar{\pi}_{s'}(\cdot|s'), Q^{\pi_\theta}(s',\cdot)\rangle,
\end{aligned} \tag{59}$$

where $\pi(\cdot|s') \in \Delta(\mathcal{A})$ and $Q^{\pi_\theta}(s',\cdot) = (Q^{\pi_\theta}(s',a_1), Q^{\pi_\theta}(s',a_2), ..., Q^{\pi_\theta}(s',a_{|\mathcal{A}|})) \in \mathbb{R}^{|\mathcal{A}|}$. Note that

$$\max_{\bar{\pi}_{s'}(\cdot|s')\in\Delta(\mathcal{A})} \langle \pi_\theta(\cdot|s') - \bar{\pi}_{s'}(\cdot|s'), Q^{\pi_\theta}(s',\cdot)\rangle \geq 0, \tag{60}$$

which is because $\max_{\bar{\pi}_{s'}(\cdot|s')\in\Delta(\mathcal{A})} \langle \pi_\theta(\cdot|s') - \bar{\pi}_{s'}(\cdot|s'), Q^{\pi_\theta}(s',\cdot)\rangle \geq \langle \pi_\theta(\cdot|s') - \pi_\theta(\cdot|s'), Q^{\pi_\theta}(s',\cdot)\rangle = 0$. Hence (59) can be further bounded as

$$\begin{aligned}
&\sum_{s'\in\mathcal{S}} \frac{l_\theta^*(s')}{l_\theta(s')} l_\theta(s') \cdot \left(\sum_{a\in\mathcal{A}} (\pi_\theta(a|s') - \pi_{\theta*}(a|s'))Q^{\pi_\theta}(s',a)\right) \\
&\leq \sum_{s'\in\mathcal{S}} \frac{l_\theta^*(s')}{l_\theta(s')} l_\theta(s') \max_{\bar{\pi}'_s(\cdot|s')\in\Delta(\mathcal{A})} \langle \pi_\theta(\cdot|s') - \bar{\pi}_{s'}(\cdot|s'), Q^{\pi_\theta}(s',\cdot)\rangle \\
&\leq C_{PL} \sum_{s'\in\mathcal{S}} l_\theta(s') \max_{\bar{\pi}_{s'}(\cdot|s')\in\Delta(\mathcal{A})} \langle \pi_\theta(\cdot|s') - \bar{\pi}_{s'}(\cdot|s'), Q^{\pi_\theta}(s',\cdot)\rangle.
\end{aligned} \tag{61}$$

If we denote $\pi_{s'}(\cdot|s') \triangleq \arg\max_{\bar{\pi}_{s'}(\cdot|s')\in\Delta(\mathcal{A})} \langle \pi_\theta(\cdot|s') - \bar{\pi}_{s'}(\cdot|s'), Q^{\pi_\theta}(s',\cdot)\rangle$, then

$$\begin{aligned}
&\sum_{s'\in\mathcal{S}} l_\theta(s') \max_{\bar{\pi}_{s'}(\cdot|s')\in\Delta(\mathcal{A})} \langle \pi_\theta(\cdot|s') - \bar{\pi}_{s'}(\cdot|s'), Q^{\pi_\theta}(s',\cdot)\rangle \\
&= \sum_{s'\in\mathcal{S}} l_\theta(s') \langle \pi_\theta(\cdot|s') - \pi_{s'}(\cdot|s'), Q^{\pi_\theta}(s',\cdot)\rangle \\
&= \sum_{s'\in\mathcal{S}} l_\theta(s') \sum_{a\in\mathcal{A}} (\pi_\theta(a|s') - \pi_{s'}(a|s'))Q^{\pi_\theta}(s',a) \\
&= \sum_{s'\in\mathcal{S}} \sum_{a\in\mathcal{A}} (\pi_\theta(a|s') - \pi_{s'}(a|s'))l_\theta(s')Q^{\pi_\theta}(s',a) \\
&= \langle \pi_\theta - \bar{\pi}, l_\theta(\cdot)Q^{\pi_\theta}(\cdot,\cdot)\rangle,
\end{aligned} \tag{62}$$

where $\pi_\theta = \left(\pi_\theta(\cdot|s_1), ..., \pi_\theta(\cdot|s_{|\mathcal{S}|})\right)^\top \in (\Delta(\mathcal{A}))^{|\mathcal{S}|}$, $\bar{\pi} = \left(\pi_{s_1}(\cdot|s_1), ..., \pi_{s_{|\mathcal{S}|}}(\cdot|s_{|\mathcal{S}|})\right)^\top \in (\Delta(\mathcal{A}))^{|\mathcal{S}|}$, $l_\theta(\cdot)Q^{\pi_\theta}(\cdot,\cdot) \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ and $(l_\theta(\cdot)Q^{\pi_\theta}(\cdot,\cdot))_{s',a} = l_\theta(s')Q^{\pi_\theta}(s',a)$. Clearly, we have that

$$\langle \pi_\theta - \bar{\pi}, l_\theta(\cdot)Q^{\pi_\theta}(\cdot,\cdot)\rangle \leq \max_{\hat{\pi}\in(\Delta(\mathcal{A}))^{|\mathcal{S}|}} \langle \pi_\theta - \hat{\pi}, l_\theta(\cdot)Q^{\pi_\theta}(\cdot,\cdot)\rangle. \tag{63}$$

Note that $\psi_\mu(\theta)_{s,b} = l_\theta(s)Q^{\pi_\theta}(s,b)$, hence combining all the inequalities above implies that

$$J_\rho(\theta) - J_\rho^* \leq C_{PL} \max_{\hat{\pi}\in(\Delta(\mathcal{A}))^{|\mathcal{S}|}} \langle \pi_\theta - \hat{\pi}, \psi_\mu(\theta)\rangle\,, \tag{64}$$

which completes the proof.

A.4. Proof of Theorem 4.2: Convergence and Global Optimality

The theorem can be proved following the standard results of stochastic differential inclusion in (Ruszczyński, 2020; Majewski et al., 2018; Borkar, 2009; Borkar & Meyn, 2000).

Lemma A.3. *(Theorem 4.1 in (Majewski et al., 2018)) If the sequence $\{\theta_t\}$ is generated by a stochastic algorithm $\theta_{t+1} \leftarrow \prod_K(\theta_t - \alpha_t Y_t)$, where $Y_t \in F(\theta_t)$ for some set-value function F, where $F(\theta) = -\partial_C f(\theta)$ for some function f, where ∂_C denotes the Clarke sub-gradient (Clarke, 1990). Denote the normal cone $N_K(\theta) \triangleq \{g \in \mathbb{R}^d : \langle g, \theta' - \theta\rangle \leq 0 \text{ for any } \theta' \in K\}$. If*

(1). the set-valued map $\theta \mapsto F(\theta)$ is upper hemicontinuous, convex-compact valued and locally bounded;

(2). the step-sizes satisfy $\alpha_t > 0$, $\sum_t^\infty \alpha_t = \infty$ and $\sum_t^\infty \alpha_t^2 < \infty$;

(3). denote the stationary points $\mathcal{X}^* \triangleq \{x : 0 \in \partial_C f(x) + N_K(x)\}$, and $f(\mathcal{X}^*)$ has no interior;

then θ_t convergence to $\mathcal{X}^*$ almost surely.

Note that according to (Clarke, 1990; Bolte et al., 2007b), $\partial J_\mu(\theta) \subseteq \partial_C J_\mu(\theta)$, hence a Fréchet sub-gradient is also a Clarke sub-gradient, and our robust policy gradient method can be viewed as a Clarke sub-gradient descent method.

It is shown in Proposition 2.1.5 and Proposition 2.1.2 in (Clarke, 1990) that the Clarke sub-gradient of a Lipschitz function is convex-compact valued, locally bounded, and lower hemicontinuous. Hence the map $-\partial J_C^\mu(\theta)$ satisfies assumption (1). The assumption (2) is easy to satisfy, e.g., $\alpha_t = \frac{1}{t+1}$.

Before verifying Assumption (3), we first show a lemma.

Lemma A.4. *For any $\hat{\theta} \in \{\theta : 0 \in \partial J_C^\mu(\theta)\}$, $J_\rho(\hat{\theta}) = J_\rho^*$.*

Proof. From Rademacher's theorem (Federer, 2014), we know that $J_\mu(\theta)$ is differentiable almost everywhere. And from (Kruger, 2003), we have that if $J_\mu(\theta)$ is differentiable at θ, then $\partial J_\mu(\theta) = \{\nabla J_\mu(\theta)\}$ and hence $\psi_\mu(\theta) = \nabla J_\mu(\theta)$.

Now consider any $\hat{\theta} \in \{\theta : 0 \in \partial J_C^\mu(\theta)\}$. From (Clarke, 1990; Zhang et al., 2020b), the Clarke sub-gradient can be viewed as a convex hull of limit gradients:

$$\partial_C J_\mu(\hat{\theta}) = \text{conv}\left(\left\{g : \exists \theta_n \to \hat{\theta},\ \text{s.t. } \nabla J_\mu(\theta_n) \to g\right\}\right). \tag{65}$$

Denote $S \triangleq \left\{g : \exists \theta_n \to \hat{\theta},\ \text{s.t. } \nabla J_\mu(\theta_n) \to g\right\}$. Note that if $0 \in \text{conv}(S)$, then $\exists \lambda_i \in [0,1]$ and $g_i \in S, i = 1, ..., k$, such that $\sum_i^k \lambda_i = 1$ and $0 = \sum_i^k \lambda_i g_i$. Note that $g_i \in S$ is a limit of a sequence of gradient $\{\psi_\mu(\theta_i^n)\}_n$ where $\theta_i^n \to \hat{\theta}$. However, note that from (53), every entry of $\psi_\mu(\theta)$ is non-negative for any θ, and hence g_i also has non-negative entries.

This implies that $\exists j \in \{1, ..., k\}$, such that $g_j = 0$. This further implies that there exists a sequence $\theta_j^t \to \hat{\theta}$, with $\psi_\mu(\theta_j^t) = \nabla J_\mu(\theta_j^t) \to 0$, i.e., $\|\nabla J_\mu(\theta_j^t)\| \to 0$.

From the PL-condition in Theorem 4.1, we know that $J_\rho(\theta_j^t) - J_\rho^* \leq C_{PL}|\mathcal{S}||\mathcal{A}|\|\nabla J_\mu(\theta_j^t)\| \to 0$ and hence $J_\rho(\theta_j^t) \to J_\rho^*$. As $J_\rho(\theta)$ is a continuous function, $J_\rho(\theta_j^t) \to J_\rho(\hat{\theta})$ and hence $J_\rho(\hat{\theta}) = J_\rho^*$. This means any $\hat{\theta} \in \{\theta : 0 \in \partial_C J_\mu(\theta)\}$ is a global optimal point, i.e., $J_\rho(\hat{\theta}) = J_\rho^*$. □

We then verify the last assumption (3). First note that from the definition of Clarke sub-differential, it can be verified that $\mathcal{X}^* \subseteq \{\theta : 0 \in \partial_C f(\theta)\}$. Hence from Lemma A.4, $\mathcal{X}^*$ is a subset of global optimal points of J_μ (set $\rho = \mu$ in Lemma A.4), hence $J_\mu(\mathcal{X}^*) = \left\{J_\mu^*\right\}$ is a singleton and contains no interior.

Hence the lemma A.3 holds for our robust policy gradient algorithm. Then, by the lemma, any sequence $\{\theta_t\}$ generated by Algorithm 1 converges to a stationary point a.s., i.e., $\lim_t \theta_t \in \left\{\theta : 0 \in \partial J_C^\mu(\theta) + N_{(\Delta(\mathcal{A}))^{|\mathcal{S}|}}(\theta)\right\} \subseteq \{\theta : 0 \in \partial J_C^\mu(\theta)\}$ almost surely.

Note that Lemma A.4 shows that any $\theta \in \{\theta : 0 \in \partial J_C^\mu(\theta)\}$ is a global optimal point of J_ρ, and this implies that the sequence $\{\theta_t\}$ generated by the Algorithm 1 converges to the global optimum of J_ρ, i.e., $J_\rho(\theta_t) \to J_\rho^*$ almost surely. And hence this completes the proof of Theorem 4.2.

B. Smoothed Robust Policy Gradient

In the remaining parts, if not specified, we omit σ in LSE function, i.e., denote $\text{LSE}(V) \triangleq \text{LSE}(\sigma, V)$.

B.1. Proof of Theorem 5.1: Gradient of J_σ

In this section we derive the gradient of $J_\sigma(\theta) \triangleq \sum_{s' \in \mathcal{S}} \rho(s') V_\sigma^{\pi_\theta}(s')$, and prove Theorem 5.1.

From the fact that $V_\sigma^{\pi_\theta}(s) = \sum_{a \in \mathcal{A}} \pi_\theta(a|s) Q_\sigma^{\pi_\theta}(s,a)$, we have that

$$
\begin{aligned}
\nabla V_\sigma^{\pi_\theta}(s) &= \nabla \left(\sum_{a \in \mathcal{A}} \pi_\theta(a|s) Q_\sigma^{\pi_\theta}(s,a) \right) \\
&= \sum_{a \in \mathcal{A}} \nabla \pi_\theta(a|s) Q_\sigma^{\pi_\theta}(s,a) + \sum_{a \in \mathcal{A}} \pi_\theta(a|s) \nabla Q_\sigma^{\pi_\theta}(s,a) \\
&\overset{(a)}{=} \sum_{a \in \mathcal{A}} \nabla \pi_\theta(a|s) Q_\sigma^{\pi_\theta}(s,a) + \sum_{a \in \mathcal{A}} \pi_\theta(a|s) \nabla \left(c(s,a) + \gamma(1-R) \sum_{s' \in \mathcal{S}} p_{s,s'}^a V_\sigma^{\pi_\theta}(s') + \gamma R \cdot \text{LSE}(V_\sigma^{\pi_\theta}) \right) \\
&= \sum_{a \in \mathcal{A}} \nabla \pi_\theta(a|s) Q_\sigma^{\pi_\theta}(s,a) + \sum_{a \in \mathcal{A}} \pi_\theta(a|s) \left(\gamma(1-R) \sum_{s' \in \mathcal{S}} p_{s,s'}^a \nabla V_\sigma^{\pi_\theta}(s') + \gamma R \cdot \nabla \text{LSE}(V_\sigma^{\pi_\theta}) \right) \\
&= \sum_{a \in \mathcal{A}} \nabla \pi_\theta(a|s) Q_\sigma^{\pi_\theta}(s,a) + \gamma R \cdot \nabla \text{LSE}(V_\sigma^{\pi_\theta}) + \sum_{a \in \mathcal{A}} \pi_\theta(a|s) \left(\gamma(1-R) \sum_{s' \in \mathcal{S}} p_{s,s'}^a \nabla V_\sigma^{\pi_\theta}(s') \right) \\
&= \sum_{a \in \mathcal{A}} \nabla \pi_\theta(a|s) Q_\sigma^{\pi_\theta}(s,a) + \gamma R \cdot \nabla \text{LSE}(V_\sigma^{\pi_\theta}) \\
&\quad + \sum_{a \in \mathcal{A}} \pi_\theta(a|s) \left(\gamma(1-R) \sum_{s' \in \mathcal{S}} p_{s,s'}^a \nabla \left(\sum_{a' \in \mathcal{A}} \pi_\theta(a'|s') Q_\sigma^{\pi_\theta}(s',a') \right) \right) \\
&= \sum_{a \in \mathcal{A}} \nabla \pi_\theta(a|s) Q_\sigma^{\pi_\theta}(s,a) + \gamma R \cdot \nabla \text{LSE}(V_\sigma^{\pi_\theta}) + \gamma(1-R) \sum_{a \in \mathcal{A}} \pi_\theta(a|s) \sum_{s' \in \mathcal{S}} p_{s,s'}^a \sum_{a'} \nabla \pi_\theta(a'|s') Q_\sigma^{\pi_\theta}(s',a') \\
&\quad + \gamma(1-R) \sum_{a \in \mathcal{A}} \pi_\theta(a|s) \sum_{s' \in \mathcal{S}} p_{s,s'}^a \sum_{a'} \pi_\theta(a'|s') \nabla Q_\sigma^{\pi_\theta}(s',a') \\
&= \sum_{a \in \mathcal{A}} \nabla \pi_\theta(a|s) Q_\sigma^{\pi_\theta}(s,a) \\
&\quad + \gamma R \cdot \nabla \text{LSE}(V_\sigma^{\pi_\theta}) + \gamma(1-R) \sum_{s' \in \mathcal{S}} \mathbb{P}(S_1 = s' | S_0 = s, \pi_\theta) \sum_{a'} \nabla \pi_\theta(a'|s') Q_\sigma^{\pi_\theta}(s',a') \\
&\quad + \gamma(1-R) \sum_{s' \in \mathcal{S}} \mathbb{P}(S_1 = s' | S_0 = s, \pi_\theta) \sum_{a'} \pi_\theta(a'|s') \nabla Q_\sigma^{\pi_\theta}(s',a') \\
&= \dots \\
&\overset{(b)}{=} \sum_{k=0}^{\infty} \gamma^k (1-R)^k \sum_{s' \in \mathcal{S}} \mathbb{P}(S_k = s' | S_0 = s, \pi_\theta) \sum_{a \in \mathcal{A}} \nabla \pi_\theta(a|s') Q_\sigma^{\pi_\theta}(s',a) + \frac{\gamma R}{1-\gamma+\gamma R} \nabla \text{LSE}(V_\sigma^{\pi_\theta}) \\
&= \frac{1}{1-\gamma+\gamma R} \sum_{s' \in \mathcal{S}} d_s^\pi(s') \sum_{a \in \mathcal{A}} \nabla \pi_\theta(a|s') Q_\sigma^{\pi_\theta}(s',a) + \frac{\gamma R}{1-\gamma+\gamma R} \nabla \text{LSE}(V_\sigma^{\pi_\theta}),
\end{aligned}
\tag{66}
$$

where (a) is from the definition of $Q_\sigma^{\pi_\theta}$, and (b) is to recursively apply the previous steps.

To simplify the notations, we denote $B(s,\theta) \triangleq \frac{1}{1-\gamma+\gamma R}\sum_{s'\in\mathcal{S}} d_s^{\pi_\theta}(s')\sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s')Q_\sigma^{\pi_\theta}(s',a)$. We then study the second term $\nabla\text{LSE}(V_\sigma^{\pi_\theta})$ in (66). Note that

$$\begin{aligned}\nabla\text{LSE}(V_\sigma^{\pi_\theta}) &= \nabla\left(\frac{1}{\sigma}\log\left(\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}\right)\right)\\ &= \frac{1}{\sigma}\frac{\nabla\left(\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}\right)}{\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}}\\ &= \frac{1}{\sigma}\frac{\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}\sigma\nabla V_\sigma^{\pi_\theta}(s)}{\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}}\\ &= \frac{\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}\nabla V_\sigma^{\pi_\theta}(s)}{\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}}.\end{aligned} \tag{67}$$

Then we plug (66) in (67), we have that

$$\begin{aligned}\nabla\text{LSE}(V_\sigma^{\pi_\theta}) &= \frac{1}{\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}}\sum_{s\in\mathcal{S}}\left(e^{\sigma V_\sigma^{\pi_\theta}(s)}\left(B(s,\theta)+\frac{\gamma R}{1-\gamma+\gamma R}\nabla\text{LSE}(V_\sigma^{\pi_\theta})\right)\right)\\ &= \frac{\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}B(s,\theta)}{\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}}+\frac{\gamma R}{1-\gamma+\gamma R}\nabla\text{LSE}(V_\sigma^{\pi_\theta}).\end{aligned} \tag{68}$$

This implies that

$$\nabla\text{LSE}(V_\sigma^{\pi_\theta}) = \frac{1-\gamma+\gamma R}{1-\gamma}\frac{\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}B(s,\theta)}{\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}}. \tag{69}$$

Hence plugging it in (66) implies that

$$\nabla V_\sigma^{\pi_\theta}(s) = B(s,\theta)+\frac{\gamma R}{1-\gamma}\frac{\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}B(s,\theta)}{\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}}. \tag{70}$$

And finally it is easy to see that

$$\nabla J_\sigma(\theta) = B(\rho,\theta)+\frac{\gamma R}{1-\gamma}\frac{\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}B(s,\theta)}{\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}}, \tag{71}$$

where $B(\rho,\theta) \triangleq \frac{1}{1-\gamma+\gamma R}\sum_{s'\in\mathcal{S}} d_\rho^{\pi_\theta}(s')\sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s')Q_\sigma^{\pi_\theta}(s',a)$.

B.2. Proof of Lemma 5.4: Smoothness of J_σ

In this section we prove Lemma 5.4 that the smoothed robust value function $J_\sigma(\theta)$ is L_σ-smooth in θ. It has been shown in (Wang & Zou, 2021) that $\|V_\sigma^\pi - V^\pi\|_\infty \le \frac{\gamma R}{1-\gamma}\frac{\log|\mathcal{S}|}{\sigma}$. Moreover we have that

$$\begin{aligned}\|Q_\sigma^\pi - Q^\pi\|_\infty &= \max_{s,a}\Big|\sum_{s'\in\mathcal{S}}\gamma(1-R)p_{s,s'}^a(V_\sigma^\pi(s')-V^\pi(s'))+\gamma R(\text{LSE}(V_\sigma^\pi)-\max V^\pi)\Big|\\ &\le \gamma(1-R)\|V_\sigma^\pi - V^\pi\|_\infty + |\gamma R(\text{LSE}(V_\sigma^\pi)-\max V^\pi)|\\ &\le \gamma(1-R)\frac{\gamma R}{1-\gamma}\frac{\log|\mathcal{S}|}{\sigma}+\gamma R|\text{LSE}(V_\sigma^\pi)-\text{LSE}(V^\pi)|+\gamma R|\text{LSE}(V^\pi)-\max V^\pi|\\ &\overset{(a)}{\le} \gamma(1-R)\frac{\gamma R}{1-\gamma}\frac{\log|\mathcal{S}|}{\sigma}+\gamma R\|V_\sigma^\pi - V^\pi\|_\infty+\gamma R|\text{LSE}(V^\pi)-\max V^\pi|\end{aligned}$$

$$\overset{(b)}{\leq} \frac{\gamma^2 R}{1-\gamma}\frac{\log|\mathcal{S}|}{\sigma} + \gamma R\frac{\log|\mathcal{S}|}{\sigma}$$
$$= \frac{\gamma R}{1-\gamma}\frac{\log|\mathcal{S}|}{\sigma}, \tag{72}$$

where (a) is from the fact that $|\text{LSE}(V_1) - \text{LSE}(V_2)| \leq \|V_1 - V_2\|_\infty$ ((59) in (Wang & Zou, 2021)), and (b) is because $\text{LSE}(V) - \max V \leq \frac{\log|S|}{\sigma}$ ((61) in (Wang & Zou, 2021)).

Note that if we define a new cost function $c'(s,a) \triangleq c(s,a) + \gamma R\frac{\log|\mathcal{S}|}{\sigma}$, the robust action-value function w.r.t. new cost function c' is $Q^\pi_{c'}(s,a) = \max_\kappa \mathbb{E}_\kappa[\sum_{t=0}^\infty \gamma^t c'(S_t,A_t)^t | S_0 = s, A_0 = a, \pi] \geq \frac{\gamma R}{1-\gamma}\frac{\log|\mathcal{S}|}{\sigma}$ for any $s \in \mathcal{S}, a \in \mathcal{A}, \pi$. Hence the smoothed robust value function w.r.t. c' is non-negative: $Q^\pi_{c',\sigma}(s,a) \geq Q^\pi_{c'}(s,a) - \frac{\gamma R}{1-\gamma}\frac{\log|\mathcal{S}|}{\sigma} \geq 0$ for any $s \in \mathcal{S}, a \in \mathcal{A}, \pi$.

This means by define a new cost function c', $Q^\pi_{c',\sigma}$ and $V^\pi_{c',\sigma}$ are non-negative. In the remaining parts, we omit the subscript c' and denote the smoothed robust value functions w.r.t. c' by Q^π_σ and V^π_σ.

On the other had, the upper bounds on them can be easily derived:

$$Q^\pi_\sigma(s,a) \leq Q^\pi_{c'}(s,a) + \frac{\gamma R}{1-\gamma}\frac{\log|\mathcal{S}|}{\sigma} \leq \frac{1}{1-\gamma}(1 + \gamma R\frac{\log|\mathcal{S}|}{\sigma}) + \frac{\gamma R}{1-\gamma}\frac{\log|\mathcal{S}|}{\sigma} \leq \frac{1}{1-\gamma}(1 + 2\gamma R\frac{\log|\mathcal{S}|}{\sigma}), \tag{73}$$

$$V^\pi_\sigma(s) = \sum_{a\in\mathcal{A}} \pi(a|s) Q^\pi_\sigma(s,a) \leq \frac{1}{1-\gamma}(1 + 2\gamma R\frac{\log|\mathcal{S}|}{\sigma}). \tag{74}$$

We denote these upper bounds by C_σ. Hence we have showed that $0 \leq Q^\pi_\sigma(s,a) \leq C_\sigma$ and $0 \leq V^\pi_\sigma(s) \leq C_\sigma$ for any $s \in \mathcal{S}, a \in \mathcal{A}, \pi$.

We then prove Lemma 5.4. We first show that $B(s,\theta)$ is Lipschitz. From (70), we know that

$$\|\nabla V^{\pi_\theta}_\sigma(s)\| \leq \max_{s\in\mathcal{S}} \|B(s,\theta)\| \left(1 + \frac{\gamma R}{1-\gamma}\right) \leq \left(\frac{1-\gamma+\gamma R}{1-\gamma}\right)\frac{1}{1-\gamma+\gamma R}|\mathcal{A}|k_\pi \sup_{s,a}|Q^{\pi_\theta}_\sigma(s,a)|, \tag{75}$$

which is from $\|B(s,\theta)\| \leq \frac{1}{1-\gamma+\gamma R}|\mathcal{A}|k_\pi \sup_{s,a}|Q^{\pi_\theta}_\sigma(s,a)|$ for any $\theta \in \Theta, s \in \mathcal{S}$.

As we showed in (73), $\|Q^{\pi_\theta}_\sigma\|_\infty \leq C_\sigma$. Hence the gradient of $V^{\pi_\theta}_\sigma$ can be bounded as

$$\|\nabla V^{\pi_\theta}_\sigma(s)\| \leq \frac{1}{1-\gamma}|\mathcal{A}|k_\pi C_\sigma \triangleq C^V_\sigma, \tag{76}$$

and this means $V^{\pi_\theta}_\sigma(s)$ is C^V_σ-Lipschitz for any $s \in \mathcal{S}$. Moreover we have that

$$|Q^{\pi_{\theta_1}}_\sigma(s,a) - Q^{\pi_{\theta_2}}_\sigma(s,a)| = |\sum_{s'\in\mathcal{S}} p^a_{s,s'}\gamma(1-R)(V^{\pi_{\theta_1}}_\sigma(s') - V^{\pi_{\theta_2}}_\sigma(s')) + \gamma R(\text{LSE}(V^{\pi_{\theta_1}}_\sigma) - \text{LSE}(V^{\pi_{\theta_2}}_\sigma))|$$
$$\overset{(a)}{\leq} \gamma(1-R)C^V_\sigma\|\theta_1 - \theta_2\| + \gamma R C^V_\sigma\|\theta_1 - \theta_2\|$$
$$= C^V_\sigma\|\theta_1 - \theta_2\|, \tag{77}$$

where inequality (a) is from $|\text{LSE}(V_1) - \text{LSE}(V_2)| \leq \|V_1 - V_2\|_\infty$ and $V^{\pi_\theta}_\sigma(s)$ is C^V_σ-Lipschitz.

It was showed in (Achiam et al., 2017; Touati et al., 2020) that

$$d_{\text{TV}}(d^{\pi_1}_s, d^{\pi_2}_s) \leq \frac{2\gamma(1-R)}{1-\gamma+\gamma R}\mathbb{E}_{S\sim d^{\pi_2}_s}\left[d_{\text{TV}}(\pi_1(\cdot|S), \pi_2(\cdot|S))\right], \tag{78}$$

hence we have that for any function $f(s,\theta)$ defined on $\mathcal{S}\times\Theta$,

$$\|\mathbb{E}_{S\sim d^{\pi_{\theta_1}}_s}[f(S,\theta)] - \mathbb{E}_{S\sim d^{\pi_{\theta_2}}_s}[f(S,\theta)]\| \leq 2d_{\text{TV}}(d^{\pi_{\theta_1}}_s, d^{\pi_{\theta_1}}_s) \sup_{s\in\mathcal{S},\theta\in\Theta}\|f(s,\theta)\|$$
$$\leq \frac{4\gamma(1-R)}{1-\gamma+\gamma R}\frac{|\mathcal{A}|}{2}k_\pi\|\theta_1 - \theta_2\| \sup_{s\in\mathcal{S},\theta\in\Theta}\|f(s,\theta)\|, \tag{79}$$

where the last inequality is from

$$d_{\mathrm{TV}}(\pi_{\theta_1}(\cdot|s), \pi_{\theta_2}(\cdot|s)) = \frac{1}{2}\sum_{a\in\mathcal{A}} |\pi_{\theta_1}(a|s) - \pi_{\theta_2}(a|s)| \leq \frac{|\mathcal{A}|}{2} k_\pi \|\theta_1 - \theta_2\|. \tag{80}$$

This implies that

$$\begin{aligned}
\|B(s,\theta_1) - B(s,\theta_2)\| &= \frac{1}{1-\gamma+\gamma R}\|\mathbb{E}_{S\sim d_s^{\pi_{\theta_1}}}[f(S,\theta_1)] - \mathbb{E}_{S\sim d_s^{\pi_{\theta_2}}}[f(S,\theta_2)]\| \\
&\leq \frac{1}{1-\gamma+\gamma R} \\
&\quad \cdot \|\mathbb{E}_{S\sim d_s^{\pi_{\theta_1}}}[f(S,\theta_1)] - \mathbb{E}_{S\sim d_s^{\pi_{\theta_1}}}[f(S,\theta_2)] + \mathbb{E}_{S\sim d_s^{\pi_{\theta_1}}}[f(S,\theta_2)] - \mathbb{E}_{S\sim d_s^{\pi_{\theta_2}}}[f(S,\theta_2)]\| \\
&\leq \frac{1}{1-\gamma+\gamma R} \\
&\quad \cdot \left(\mathbb{E}_{S\sim d_s^{\theta_1}} \|f(S,\theta_1) - f(S,\theta_2)\| + \frac{2k_\pi|\mathcal{A}|\gamma(1-R)}{1-\gamma+\gamma R} \sup_{x\in\mathcal{S},\theta\in\Theta} \|f(x,\theta)\| \|\theta_1 - \theta_2\| \right),
\end{aligned} \tag{81}$$

where $f(x,\theta) = \sum_{a\in\mathcal{A}} \nabla\pi_\theta(a|x) Q_\sigma^{\pi_\theta}(x,a)$, and the last inequality is from (79).

Note that for any $s \in \mathcal{S}$ and $\theta \in \Theta$,

$$\|f(s,\theta)\| = \left\| \sum_{a\in\mathcal{A}} \nabla\pi_\theta(a|x) Q_\sigma^{\pi_\theta}(x,a) \right\| \leq |\mathcal{A}| k_\pi C_\sigma, \tag{82}$$

and

$$\begin{aligned}
\|f(x,\theta_1) - f(x,\theta_2)\| &= \left\| \sum_{a\in\mathcal{A}} \left(\nabla\pi_{\theta_1}(a|x) Q_\sigma^{\pi_{\theta_1}}(x,a) - \nabla\pi_{\theta_2}(a|x) Q_\sigma^{\pi_{\theta_2}}(x,a) \right) \right\| \\
&\leq \left\| \sum_{a\in\mathcal{A}} \left(\nabla\pi_{\theta_1}(a|x) Q_\sigma^{\pi_{\theta_1}}(x,a) - \nabla\pi_{\theta_2}(a|x) Q_\sigma^{\pi_{\theta_1}}(x,a) \right) \right\| \\
&\quad + \left\| \sum_{a\in\mathcal{A}} \left(\nabla\pi_{\theta_2}(a|x) Q_\sigma^{\pi_{\theta_1}}(x,a) - \nabla\pi_{\theta_2}(a|x) Q_\sigma^{\pi_{\theta_2}}(x,a) \right) \right\| \\
&\leq |\mathcal{A}| C_\sigma l_\pi \|\theta_1 - \theta_2\| + |\mathcal{A}| k_\pi C_\sigma^V \|\theta_1 - \theta_2\| \\
&= \left(|\mathcal{A}| C_\sigma l_\pi + |\mathcal{A}| k_\pi C_\sigma^V \right) \|\theta_1 - \theta_2\|,
\end{aligned} \tag{83}$$

where the last inequality is because $\nabla\pi_\theta(a|x)$ is l_π-Lipschitz, and $Q_\sigma^{\pi_\theta}(x,a)$ is C_σ^V-Lipschitz.

Hence plugging (82) and (83) in (81) implies that

$$\begin{aligned}
\|B(s,\theta_1) - B(s,\theta_2)\| &\leq \frac{1}{1-\gamma+\gamma R} \left(|\mathcal{A}| C_\sigma l_\pi + |\mathcal{A}| k_\pi C_\sigma^V \right) \|\theta_1 - \theta_2\| \\
&\quad + \frac{2|\mathcal{A}|^2\gamma(1-R)}{(1-\gamma+\gamma R)^2} k_\pi^2 C_\sigma \|\theta_1 - \theta_2\| \\
&\triangleq k_B \|\theta_1 - \theta_2\|,
\end{aligned} \tag{84}$$

Hence we showed that $B(s,\theta)$ and $B(\rho,\theta)$ are both k_B-Lipschitz.

Next we show that $\frac{\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)} B(s,\theta)}{\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}}$ is also Lipschitz. Note that it is the inner production of $B(\theta) \triangleq (B(s_1,\theta), ..., B(s_{|\mathcal{S}|},\theta))$ and $p_\sigma^\theta \triangleq \left(\frac{e^{\sigma V_\sigma^{\pi_\theta}(s_1)}}{\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}}, ..., \frac{e^{\sigma V_\sigma^{\pi_\theta}(s_{|\mathcal{S}|})}}{\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}} \right)$. Hence

$$\langle B(\theta_1), p_\sigma^{\theta_1} \rangle - \langle B(\theta_2), p_\sigma^{\theta_2} \rangle = \langle B(\theta_1), p_\sigma^{\theta_1} \rangle - \langle B(\theta_2), p_\sigma^{\theta_1} \rangle + \langle B(\theta_2), p_\sigma^{\theta_1} \rangle - \langle B(\theta_2), p_\sigma^{\theta_2} \rangle$$

$$
\begin{aligned}
&= \langle B(\theta_1) - B(\theta_2), p_\sigma^{\theta_1} \rangle + \langle B(\theta_2), p_\sigma^{\theta_1} - p_\sigma^{\theta_2} \rangle \\
&\leq \|B(\theta_1) - B(\theta_2)\| + \|B(\theta_2)\| \|p_\sigma^{\theta_1} - p_\sigma^{\theta_2}\|, \qquad (85)
\end{aligned}
$$

where the last inequality is because $\|p_\sigma^\theta\| \leq 1$ for any $\theta \in \Theta$.

The Lipschitz smoothness of p_σ^θ can be showed as follows. Note that

$$
\begin{aligned}
\nabla p_\sigma^\theta(j) &= \nabla \left(\frac{e^{\sigma V_\sigma^{\pi_\theta}(j)}}{\sum_{s \in \mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}} \right) \\
&= \frac{e^{\sigma V_\sigma^{\pi_\theta}(j)} \sigma \nabla V_\sigma^{\pi_\theta}(j) \left(\sum_{s \in \mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)} \right) - \left(\sum_{s \in \mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)} \sigma \nabla V_\sigma^{\pi_\theta}(s) \right) e^{V_\sigma^{\pi_\theta}(j)}}{\left(\sum_{s \in \mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)} \right)^2} \\
&= \frac{\sigma e^{V_\sigma^{\pi_\theta}(j)} \sum_{s \in \mathcal{S}} \left(e^{\sigma V_\sigma^{\pi_\theta}(s)} \left(\nabla V_\sigma^{\pi_\theta}(j) - \nabla V_\sigma^{\pi_\theta}(s) \right) \right)}{\left(\sum_{s \in \mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)} \right)^2}, \qquad (86)
\end{aligned}
$$

and hence

$$
\begin{aligned}
\|\nabla p_\sigma^\theta(j)\| &\leq \frac{\sigma e^{V_\sigma^{\pi_\theta}(j)} \cdot 2 \max_{s \in \mathcal{S}} \|\nabla V_\sigma^{\pi_\theta}(s)\|}{\sum_{s \in \mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}} \\
&\leq 2\sigma C_\sigma^V, \qquad (87)
\end{aligned}
$$

which is from (76).

Thus we have $\|p_\sigma^{\theta_1} - p_\sigma^{\theta_2}\| = \sqrt{\sum_{s \in \mathcal{S}} \left(p_\sigma^{\theta_1}(s) - p_\sigma^{\theta_2}(s) \right)^2} \leq 2\sigma \sqrt{|\mathcal{S}|} C_\sigma^V \|\theta_1 - \theta_2\|$. Moreover, note that

$$
\|B(\theta_1) - B(\theta_2)\| \leq \sqrt{|\mathcal{S}|} k_B \|\theta_1 - \theta_2\|, \qquad (88)
$$

combining both inequality implies that

$$
\left\| \frac{\sum_{s \in \mathcal{S}} e^{\sigma V_\sigma^{\theta_1}(s)} B(s, \theta_1)}{\sum_{s \in \mathcal{S}} e^{\sigma V_\sigma^{\theta_1}(s)}} - \frac{\sum_{s \in \mathcal{S}} e^{\sigma V_\sigma^{\theta_2}(s)} B(s, \theta_2)}{\sum_{s \in \mathcal{S}} e^{\sigma V_\sigma^{\theta_2}(s)}} \right\| \leq \left(\sqrt{|\mathcal{S}|} k_B + 2\sigma |\mathcal{S}| C_\sigma^V \frac{1}{1 - \gamma + \gamma R} k_\pi |\mathcal{A}| C_\sigma \right) \|\theta_1 - \theta_2\|, \quad (89)
$$

where we use the fact that $\|B(s, \theta)\| \leq \frac{\sqrt{|\mathcal{S}|}}{1 - \gamma + \gamma R} k_\pi |\mathcal{A}| C_\sigma$ for any $s \in \mathcal{S}$ and $\theta \in \Theta$.

Hence recall the expression of $\nabla J_\sigma(\theta)$ in (11), $\nabla J_\sigma(\theta)$ is Lipschitz with constant $k_B + \frac{\gamma R}{1 - \gamma} \left(\sqrt{|\mathcal{S}|} k_B + 2\sigma |\mathcal{S}| C_\sigma^V \frac{1}{1 - \gamma + \gamma R} k_\pi |\mathcal{A}| C_\sigma \right) \triangleq L_\sigma$. And this completes the proof.

B.3. Proof of Theorem 5.2: PL-Condition of Smoothed Robust Policy Gradient

In the following, we will show that the smoothed objective function $J_\sigma(\theta)$ satisfies the PL-condition.

It can be shown that

$$
\begin{aligned}
V_\sigma^{\pi_{\theta_1}}(s) - V_\sigma^{\pi_{\theta_2}}(s) &= \sum_{a \in \mathcal{A}} \pi_{\theta_1}(a|s) Q_\sigma^{\pi_{\theta_1}}(s, a) - \sum_{a \in \mathcal{A}} \pi_{\theta_2}(a|s) Q_\sigma^{\pi_{\theta_2}}(s, a) \\
&= \sum_{a \in \mathcal{A}} \left(\pi_{\theta_1}(a|s) - \pi_{\theta_2}(a|s) \right) Q_\sigma^{\pi_{\theta_1}}(s, a) + \sum_{a \in \mathcal{A}} \pi_{\theta_2}(a|s) \left(Q_\sigma^{\pi_{\theta_1}}(s, a) - Q_\sigma^{\pi_{\theta_2}}(s, a) \right) \\
&= \sum_{a \in \mathcal{A}} \left(\pi_{\theta_1}(a|s) - \pi_{\theta_2}(a|s) \right) Q_\sigma^{\pi_{\theta_1}}(s, a) \\
&\quad + \sum_{a \in \mathcal{A}} \pi_{\theta_2}(a|s) \left(\gamma (1 - R) \sum_{s' \in \mathcal{S}} p_{s,s'}^a \left(V_\sigma^{\pi_{\theta_1}}(s') - V_\sigma^{\pi_{\theta_2}}(s') \right) + \gamma R \left(\mathrm{LSE}_1 - \mathrm{LSE}_2 \right) \right)
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(a)}{\leq} \sum_{a\in\mathcal{A}} \left(\pi_{\theta_1}(a|s) - \pi_{\theta_2}(a|s)\right) Q_\sigma^{\pi_{\theta_1}}(s,a) \\
&\quad + \gamma(1-R) \sum_{s'\in\mathcal{S}} \mathbb{P}(S_1 = s'|S_0 = s, \pi_{\theta_2}) \left(V_\sigma^{\pi_{\theta_1}}(s') - V_\sigma^{\pi_{\theta_2}}(s')\right) + \gamma R\left(\mathrm{LSE}_1 - \mathrm{LSE}_2\right) \\
&\overset{(b)}{\leq} \frac{1}{1-\gamma+\gamma R} \sum_{s'\in\mathcal{S}} d_s^{\pi_{\theta_2}}(s') \sum_{a\in\mathcal{A}} \left(\pi_{\theta_1}(a|s') - \pi_{\theta_2}(a|s')\right) Q_\sigma^{\pi_{\theta_1}}(s',a) \\
&\quad + \frac{\gamma R}{1-\gamma+\gamma R}\left(\mathrm{LSE}_1 - \mathrm{LSE}_2\right),
\end{aligned} \tag{90}
$$

where (a) follows from the definition of Q_σ^π and we denote $\mathrm{LSE}(V_\sigma^{\pi_{\theta_i}})$ by LSE_i; and (b) follows by recursively apply the above steps.

From $|\mathrm{LSE}(V) - \max V| \leq \frac{\log|\mathcal{S}|}{\sigma}$, we further have

$$
\mathrm{LSE}_1 - \mathrm{LSE}_2 \leq \max V_\sigma^{\pi_{\theta_1}} - \max V_\sigma^{\pi_{\theta_2}} + \frac{2\log|\mathcal{S}|}{\sigma}. \tag{91}
$$

Hence (90) can be further bounded as

$$
\begin{aligned}
V_\sigma^{\pi_{\theta_1}}(s) - V_\sigma^{\pi_{\theta_2}}(s) \leq{}& \frac{1}{1-\gamma+\gamma R} \sum_{s'\in\mathcal{S}} d_s^{\pi_{\theta_2}}(s') \sum_{a\in\mathcal{A}} \left(\pi_{\theta_1}(a|s') - \pi_{\theta_2}(a|s')\right) Q_\sigma^{\pi_{\theta_1}}(s',a) \\
&+ \frac{\gamma R}{1-\gamma+\gamma R}\left(\frac{2\log|\mathcal{S}|}{\sigma} + V_\sigma^{\pi_{\theta_1}}\left(s_{\theta_1}^\sigma\right) - V_\sigma^{\pi_{\theta_2}}\left(s_{\theta_2}^\sigma\right)\right),
\end{aligned} \tag{92}
$$

where $s_{\theta_j}^\sigma \in \arg\max_s V_\sigma^{\pi_{\theta_j}}(s), j = 1,2$. Note that $V_\sigma^{\pi_{\theta_1}}\left(s_{\theta_1}^\sigma\right) - V_\sigma^{\pi_{\theta_2}}\left(s_{\theta_2}^\sigma\right) = V_\sigma^{\pi_{\theta_1}}\left(s_{\theta_1}^\sigma\right) - V_\sigma^{\pi_{\theta_2}}\left(s_{\theta_1}^\sigma\right) + V_\sigma^{\pi_{\theta_2}}\left(s_{\theta_1}^\sigma\right) - V_\sigma^{\pi_{\theta_2}}\left(s_{\theta_2}^\sigma\right) \leq V_\sigma^{\pi_{\theta_1}}\left(s_{\theta_1}^\sigma\right) - V_\sigma^{\pi_{\theta_2}}\left(s_{\theta_1}^\sigma\right)$ and hence (90) can be bounded as

$$
\begin{aligned}
V_\sigma^{\pi_{\theta_1}}(s) - V_\sigma^{\pi_{\theta_2}}(s) \leq{}& \frac{1}{1-\gamma+\gamma R} \sum_{s'\in\mathcal{S}} d_s^{\pi_{\theta_2}}(s') \sum_{a\in\mathcal{A}} \left(\pi_{\theta_1}(a|s') - \pi_{\theta_2}(a|s')\right) Q_\sigma^{\pi_{\theta_1}}(s',a) \\
&+ \frac{\gamma R}{1-\gamma+\gamma R}\left(\frac{2\log|\mathcal{S}|}{\sigma} + V_\sigma^{\pi_{\theta_1}}\left(s_{\theta_1}^\sigma\right) - V_\sigma^{\pi_{\theta_2}}\left(s_{\theta_1}^\sigma\right)\right).
\end{aligned} \tag{93}
$$

If we set $s = s_{\theta_1}^\sigma$ in this inequality, we have that

$$
\begin{aligned}
V_\sigma^{\pi_{\theta_1}}\left(s_{\theta_1}^\sigma\right) - V_\sigma^{\pi_{\theta_2}}\left(s_{\theta_1}^\sigma\right) \leq{}& \frac{1}{1-\gamma+\gamma R} \sum_{s'\in\mathcal{S}} d_{s_{\theta_1}^\sigma}^{\pi_{\theta_2}}(s') \sum_{a\in\mathcal{A}} \left(\pi_{\theta_1}(a|s') - \pi_{\theta_2}(a|s')\right) Q_\sigma^{\pi_{\theta_1}}(s',a) \\
&+ \frac{\gamma R}{1-\gamma+\gamma R}\left(\frac{2\log|\mathcal{S}|}{\sigma} + V_\sigma^{\pi_{\theta_1}}\left(s_{\theta_1}^\sigma\right) - V_\sigma^{\pi_{\theta_2}}\left(s_{\theta_1}^\sigma\right)\right),
\end{aligned} \tag{94}
$$

and hence

$$
\begin{aligned}
&V_\sigma^{\pi_{\theta_1}}\left(s_{\theta_1}^\sigma\right) - V_\sigma^{\pi_{\theta_2}}\left(s_{\theta_1}^\sigma\right) \\
&\leq \left(\frac{1-\gamma+\gamma R}{1-\gamma}\right) \\
&\quad \cdot \left(\frac{\gamma R}{1-\gamma+\gamma R}\frac{2\log|\mathcal{S}|}{\sigma} + \frac{1}{1-\gamma+\gamma R} \sum_{s'\in\mathcal{S}} d_{s_{\theta_1}^\sigma}^{\pi_{\theta_2}}(s') \sum_{a\in\mathcal{A}} \left(\pi_{\theta_1}(a|s') - \pi_{\theta_2}(a|s')\right) Q_\sigma^{\pi_{\theta_1}}(s',a)\right).
\end{aligned} \tag{95}
$$

Now plugging (95) in (93) implies that for any $s \in \mathcal{S}$

$$
\begin{aligned}
V_\sigma^{\pi_{\theta_1}}(s) - V_\sigma^{\pi_{\theta_2}}(s) \leq{}& \frac{1}{1-\gamma+\gamma R} \sum_{s'\in\mathcal{S}} d_s^{\pi_{\theta_2}}(s') \sum_{a\in\mathcal{A}} \left(\pi_{\theta_1}(a|s') - \pi_{\theta_2}(a|s')\right) Q_\sigma^{\pi_{\theta_1}}(s',a) \\
&+ \frac{\gamma R}{1-\gamma+\gamma R}\left(\frac{2\log|\mathcal{S}|}{\sigma} + \left(\frac{1-\gamma+\gamma R}{1-\gamma}\right) \cdot \left(\frac{\gamma R}{1-\gamma+\gamma R}\frac{2\log|\mathcal{S}|}{\sigma}\right.\right.
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{1-\gamma+\gamma R}\sum_{s'\in\mathcal{S}}d_{s_{\theta_1}^{\sigma}}^{\pi_{\theta_2}}\left(s'\right)\sum_{a\in\mathcal{A}}\left(\pi_{\theta_1}\left(a|s'\right)-\pi_{\theta_2}\left(a|s'\right)\right)Q_{\sigma}^{\pi_{\theta_1}}\left(s',a\right)\Bigg)\Bigg)\\
&=\frac{1}{1-\gamma+\gamma R}\sum_{s'\in\mathcal{S}}d_{s}^{\pi_{\theta_2}}\left(s'\right)\sum_{a\in\mathcal{A}}\left(\pi_{\theta_1}\left(a|s'\right)-\pi_{\theta_2}\left(a|s'\right)\right)Q_{\sigma}^{\pi_{\theta_1}}\left(s',a\right)\\
&+\frac{\gamma R}{\left(1-\gamma\right)\left(1-\gamma+\gamma R\right)}\left(\sum_{s'\in\mathcal{S}}d_{s_{\theta_1}^{\sigma}}^{\pi_{\theta_2}}\left(s'\right)\sum_{a\in\mathcal{A}}\left(\pi_{\theta_1}\left(a|s'\right)-\pi_{\theta_2}\left(a|s'\right)\right)Q_{\sigma}^{\pi_{\theta_1}}\left(s',a\right)\right)\\
&+\left(\frac{\gamma R}{1-\gamma}\right)\frac{2\log|\mathcal{S}|}{\sigma}. \qquad (96)
\end{aligned}
$$

Set $\theta_1=\theta$ and $\theta_2=\theta_{\sigma}^{*}\in\arg\max_{\theta\in(\Delta(\mathcal{A}))^{|\mathcal{S}|}}J_{\sigma}(\theta)$, we have

$$
\begin{aligned}
V_{\sigma}^{\pi_\theta}\left(s\right)-V_{\sigma}^{\pi_{\theta_{\sigma}^{*}}}\left(s\right)\leq&\frac{1}{1-\gamma+\gamma R}\sum_{s'\in\mathcal{S}}d_{s}^{\pi_{\theta_{\sigma}^{*}}}\left(s'\right)\sum_{a\in\mathcal{A}}\left(\pi_{\theta}\left(a|s'\right)-\pi_{\theta_{\sigma}^{*}}\left(a|s'\right)\right)Q_{\sigma}^{\pi_\theta}\left(s',a\right)\\
&+\frac{\gamma R}{\left(1-\gamma\right)\left(1-\gamma+\gamma R\right)}\left(\sum_{s'\in\mathcal{S}}d_{s_{\theta}^{\sigma}}^{\pi_{\theta_{\sigma}^{*}}}\left(s'\right)\sum_{a\in\mathcal{A}}\left(\pi_{\theta}\left(a|s'\right)-\pi_{\theta_{\sigma}^{*}}\left(a|s'\right)\right)Q_{\sigma}^{\pi_\theta}\left(s',a\right)\right)\\
&+\left(\frac{\gamma R}{1-\gamma}\right)\frac{2\log|\mathcal{S}|}{\sigma}. \qquad (97)
\end{aligned}
$$

Recall the definition of J_{σ}, we can show that

$$
\begin{aligned}
J_{\sigma}\left(\theta\right)-J_{\sigma}^{*}\leq&\frac{1}{1-\gamma+\gamma R}\sum_{s'\in\mathcal{S}}d_{\rho}^{\pi_{\theta_{\sigma}^{*}}}\left(s'\right)\sum_{a\in\mathcal{A}}\left(\pi_{\theta}\left(a|s'\right)-\pi_{\theta_{\sigma}^{*}}\left(a|s'\right)\right)Q_{\sigma}^{\pi_\theta}\left(s',a\right)\\
&+\frac{\gamma R}{\left(1-\gamma\right)\left(1-\gamma+\gamma R\right)}\left(\sum_{s'\in\mathcal{S}}d_{s_{\theta}^{\sigma}}^{\pi_{\theta_{\sigma}^{*}}}\left(s'\right)\sum_{a\in\mathcal{A}}\left(\pi_{\theta}\left(a|s'\right)-\pi_{\theta_{\sigma}^{*}}\left(a|s'\right)\right)Q_{\sigma}^{\pi_\theta}\left(s',a\right)\right)\\
&+\left(\frac{\gamma R}{1-\gamma}\right)\frac{2\log|\mathcal{S}|}{\sigma}\\
=&\sum_{s'\in\mathcal{S}}\left(\frac{1}{1-\gamma+\gamma R}d_{\rho}^{\pi_{\theta_{\sigma}^{*}}}\left(s'\right)+\frac{\gamma R}{\left(1-\gamma\right)\left(1-\gamma+\gamma R\right)}d_{s_{\theta}^{\sigma}}^{\pi_{\theta_{\sigma}^{*}}}\left(s'\right)\right)\sum_{a\in\mathcal{A}}\left(\pi_{\theta}\left(a|s'\right)-\pi_{\theta_{\sigma}^{*}}\left(a|s'\right)\right)Q_{\sigma}^{\pi_\theta}\left(s',a\right)\\
&+\left(\frac{\gamma R}{1-\gamma}\right)\frac{2\log|\mathcal{S}|}{\sigma}\\
\triangleq&\sum_{s'\in\mathcal{S}}l_{\theta}^{*}\left(s'\right)\sum_{a\in\mathcal{A}}\left(\pi_{\theta}\left(a|s'\right)-\pi_{\theta_{\sigma}^{*}}\left(a|s'\right)\right)Q_{\sigma}^{\pi_\theta}\left(s',a\right)+\left(\frac{\gamma R}{1-\gamma}\right)\frac{2\log|\mathcal{S}|}{\sigma}, \qquad (98)
\end{aligned}
$$

where $l_{\theta}^{*}\left(s'\right)=\frac{1}{1-\gamma+\gamma R}d_{\rho}^{\pi_{\theta_{\sigma}^{*}}}\left(s'\right)+\frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)}d_{s_{\theta}^{\sigma}}^{\pi_{\theta_{\sigma}^{*}}}\left(s'\right)$. Recall that

$$
\begin{aligned}
\left(\nabla J_{\sigma}\left(\theta\right)\right)_{s,a}=&B\left(\rho,\theta\right)_{s,a}+\frac{\gamma R}{1-\gamma}\frac{\sum_{s'\in\mathcal{S}}e^{\sigma V_{\sigma}^{\pi_\theta}\left(s'\right)}B\left(s,\theta\right)_{s',a}}{\sum_{s\in\mathcal{S}}e^{\sigma V_{\sigma}^{\pi_\theta}\left(s\right)}}\\
=&\frac{1}{1-\gamma+\gamma R}d_{\rho}^{\pi_\theta}\left(s\right)Q_{\sigma}^{\pi_\theta}\left(s,a\right)+\frac{\gamma R}{1-\gamma}\frac{\sum_{s'\in\mathcal{S}}e^{\sigma V_{\sigma}^{\pi_\theta}\left(s'\right)}\frac{1}{1-\gamma+\gamma R}d_{s'}^{\pi}\left(s\right)Q_{\sigma}^{\pi_\theta}\left(s,a\right)}{\sum_{s\in\mathcal{S}}e^{\sigma V_{\sigma}^{\pi_\theta}\left(s\right)}}\\
=&\left(\frac{1}{1-\gamma+\gamma R}d_{\rho}^{\pi_\theta}\left(s\right)+\frac{\gamma R}{1-\gamma}\frac{\sum_{s'\in\mathcal{S}}e^{\sigma V_{\sigma}^{\pi_\theta}\left(s'\right)}\frac{1}{1-\gamma+\gamma R}d_{s'}^{\pi}\left(s\right)}{\sum_{s\in\mathcal{S}}e^{\sigma V_{\sigma}^{\pi_\theta}\left(s\right)}}\right)Q_{\sigma}^{\pi_\theta}\left(s,a\right)\\
\triangleq&\, l_{\theta}\left(s\right)Q_{\sigma}^{\pi_\theta}\left(s,a\right), \qquad (99)
\end{aligned}
$$

where $l_\theta(s) = \frac{1}{1-\gamma+\gamma R} d_\rho^{\pi_\theta}(s) + \frac{\gamma R}{1-\gamma} \frac{\sum_{s'\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s')} \frac{1}{1-\gamma+\gamma R} d_{s'}^{\pi}(s)}{\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}}$. It can be verified that

$$\sup_{s\in\mathcal{S}} \frac{l_\theta^*(s)}{l_\theta(s)} = \sup_{s\in\mathcal{S}} \frac{\frac{1}{1-\gamma+\gamma R} d_\rho^{\pi_{\theta_\sigma^*}}(s) + \frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)} d_{s_\theta^{\sigma}}^{\pi_{\theta_\sigma^*}}(s)}{\frac{1}{1-\gamma+\gamma R} d_\rho^{\pi_\theta}(s) + \frac{\gamma R}{1-\gamma} \frac{\sum_{s'\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s')} \frac{1}{1-\gamma+\gamma R} d_{s'}^{\pi}(s)}{\sum_{s'\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s')}}} \leq \frac{1}{(1-\gamma)\rho_{\min}} = C_{PL}, \tag{100}$$

where the last inequality is from $d_\rho^{\pi_\theta} \geq (1-\gamma+\gamma R)\rho_{\min}$, and $\frac{\sum_{s'\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s')} \frac{1}{1-\gamma+\gamma R} d_{s'}^{\pi}(s)}{\sum_{s'\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s')}} \geq 0$.

Hence similar to Theorem 4.1, we show that

$$J_\sigma(\theta) - J_\sigma^* \leq C_{PL} \max_{\hat{\pi}\in(\Delta(|\mathcal{A}|))^{|\mathcal{S}|}} \langle \pi_\theta - \hat{\pi}, \nabla J_\sigma(\theta)\rangle + \left(\frac{\gamma R}{1-\gamma}\right) \frac{2\log|\mathcal{S}|}{\sigma}, \tag{101}$$

which completes the proof.

B.4. Proof of Theorem 5.5: Convergence Rate of Smoothed Robust Policy Gradient

In this section we prove Theorem 5.5 and show the convergence rate of our smoothed robust policy gradient algorithm.

The following lemma can be derived directly using existing classic results:

Lemma B.1. *(Theorem 10.15 in (Beck, 2017)) Set the step sizes $\alpha_t = \alpha = \frac{1}{L_\sigma}$, and define the gradient mapping as $G^\alpha(\theta) \triangleq \frac{1}{\alpha}\left(\theta - \prod_{(\Delta(\mathcal{A})^{|\mathcal{S}|}}(\theta - \alpha\nabla J_\sigma(\theta))\right)$, then*

$$\min_{t=0,\ldots,T-1} \|G^\alpha(\theta_t)\| \leq \sqrt{\frac{2L_\sigma(J_\sigma(\theta_0) - J_\sigma^*)}{T}}. \tag{102}$$

We then prove Theorem 5.5.

It has been shown in Lemma 3 in (Ghadimi & Lan, 2016) that if $\|G^\alpha(\theta)\| \leq \epsilon$, then

$$-\nabla J_\sigma(\theta^+) \in N_{\Delta(\mathcal{A})^{|\mathcal{S}|}}(\theta^+) + 2\epsilon B_2, \tag{103}$$

where $N_{\Delta(\mathcal{A})^{|\mathcal{S}|}}(x) \triangleq \left\{g \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} : \langle g, y - x\rangle \leq 0 \text{ for any } y \in \Delta(\mathcal{A})^{|\mathcal{S}|}\right\}$ is the normal cone, B_2 is the unit l_2 ball and $\theta^+ = \theta - \alpha G^\alpha(\theta)$. From PL-condition in Theorem 5.2, it can be shown that

$$\begin{aligned}
\min_{t\leq T-1} J_\sigma(\theta_t) - J_\sigma^* &\leq C_{PL} \min_{t\leq T-1} \max_{\hat{\pi}\in(\Delta(|\mathcal{A}|))^{|\mathcal{S}|}} \langle \pi_{\theta_t} - \hat{\pi}, \nabla J_\sigma(\theta_t)\rangle + \left(\frac{\gamma R}{1-\gamma}\right)\frac{2\log|\mathcal{S}|}{\sigma} \\
&\leq C_{PL} \max_{\hat{\pi}\in(\Delta(|\mathcal{A}|))^{|\mathcal{S}|}} \langle \pi_{\theta_W} - \hat{\pi}, \nabla J_\sigma(\theta_W)\rangle + \left(\frac{\gamma R}{1-\gamma}\right)\frac{2\log|\mathcal{S}|}{\sigma}
\end{aligned} \tag{104}$$

where $W \triangleq 1 + \arg\min_{t\leq T-1} \|G^\alpha(\theta_t)\|$. Recall that in Lemma B.1, we showed that

$$\|G^\alpha(\theta_{W-1})\| \leq \sqrt{\frac{2L_\sigma(J_\sigma(\theta_0) - J_\sigma^*)}{T}}. \tag{105}$$

Hence if we set

$$T = \frac{64|\mathcal{S}|C_{PL}^2 L_\sigma C_\sigma}{\epsilon^2}, \tag{106}$$

then

$$\|G^\alpha(\theta_{W-1})\| \leq \sqrt{\frac{2L_\sigma(J_\sigma(\theta_0) - J_\sigma^*)}{T}} \leq \frac{\epsilon}{4\sqrt{|\mathcal{S}|}C_{PL}}, \tag{107}$$

which is from $J_\sigma(\theta_0) - J_\sigma^* \leq 2\sup_{\theta\in\Delta(\mathcal{A})^{|\mathcal{S}|}} J_\sigma(\theta) \leq 2C_\sigma$.

Hence from (103), we know that $-\nabla J_\sigma(\theta_W) = g_1 + g_2$, where $g_1 \in N_{\Delta(\mathcal{A})^{|\mathcal{S}|}}(\theta_W)$, and $g_2 \in \frac{\epsilon}{2\sqrt{|\mathcal{S}|}C_{PL}} B_2$. Thus

$$\begin{aligned}\max_{\hat{\pi}\in(\Delta(|\mathcal{A}|))^{|\mathcal{S}|}} \langle \pi_{\theta_W} - \hat{\pi}, \nabla J_\sigma(\theta_W)\rangle &= \max_{\hat{\pi}\in(\Delta(|\mathcal{A}|))^{|\mathcal{S}|}} \langle \hat{\pi} - \pi_{\theta_W}, -\nabla J_\sigma(\theta_W)\rangle \\ &= \max_{\hat{\pi}\in(\Delta(|\mathcal{A}|))^{|\mathcal{S}|}} \langle \hat{\pi} - \pi_{\theta_W}, g_1 + g_2\rangle \\ &\leq \sup_{\hat{\pi}\in(\Delta(|\mathcal{A}|))^{|\mathcal{S}|}} \|\hat{\pi} - \pi_{\theta_W}\| \frac{\epsilon}{2\sqrt{|\mathcal{S}|}C_{PL}}, \end{aligned} \tag{108}$$

where the last inequality is from $g_1 \in N_{\Delta(\mathcal{A})^{|\mathcal{S}|}}(\theta_W)$ and $\langle \pi - \pi_{\theta_W}, g_1\rangle \leq 0$ for any $\pi \in (\Delta(|\mathcal{A}|))^{|\mathcal{S}|}$.

Note that $\|\pi_{\theta_1} - \pi_{\theta_2}\| \leq 2\sqrt{|\mathcal{S}|}$, hence from (108),

we have that

$$\max_{\hat{\pi}\in(\Delta(|\mathcal{A}|))^{|\mathcal{S}|}} \langle \pi_{\theta_W} - \hat{\pi}, \nabla J_\sigma(\theta_W)\rangle \leq \frac{\epsilon}{C_{PL}}, \tag{109}$$

which further implies that

$$\min_{t\leq T-1} J_\sigma(\theta_t) - J_\sigma^* \leq \epsilon + \left(\frac{\gamma R}{1-\gamma}\right)\frac{2\log|\mathcal{S}|}{\sigma}, \tag{110}$$

Hence if we set

$$\sigma = \frac{2\log|\mathcal{S}|\left(\frac{\gamma R}{1-\gamma}\right)}{\epsilon} = \mathcal{O}(\epsilon^{-1}), \tag{111}$$

$$T = \frac{64|\mathcal{S}|C_{PL}^2 L_\sigma C_\sigma}{\epsilon^2} = \mathcal{O}(\epsilon^{-3}), \tag{112}$$

then

$$\min_{t\leq T-1} J_\sigma(\theta_t) - J_\sigma^* \leq 2\epsilon, \tag{113}$$

which means Algorithm 2 finds a global ϵ-optimum of J_σ in $\mathcal{O}(\epsilon^{-3})$ steps.

Further it can be shown that

$$\min_{t<T} J(\theta_t) - J^* \leq \min_{t<T} J_\sigma(\theta_t) - J_\sigma^* + \frac{\gamma R}{1-\gamma}\frac{2\log|\mathcal{S}|}{\sigma} \leq 3\epsilon. \tag{114}$$

This hence completes the proof.

C. Robust Actor-Critic under Tabular Setting

C.1. Convergence of Robust TD under Tabular Setting

In this section we prove that robust TD and smoothed robust TD converge asymptotically under the tabular setting. Note that if we set $Q_\zeta = \zeta$ with ζ being the Q-table in Algorithm 3, it reduces to the robust TD algorithm. For completeness, we also present the smoothed robust TD algorithm under the tabular setting in Algorithm 5.

Theorem C.1. *If step-sizes α_t satisfy $\sum_{t=0}^\infty \alpha_t = \infty$, $\sum_{t=0}^\infty \alpha_t^2 < \infty$, then Algorithm 5 (Algorithm 3) converges to the smoothed robust action-value function Q_σ^π (robust action action-value function Q^π) almost surely.*

Proof. We present the proof for smoothed robust TD here. The proof of non-smoothed one can be similarly derived.

Define the smoothed robust Bellman operator for the robust Q-function as follows:

$$\mathbf{T}_\sigma^\pi Q(s,a) = c(s,a) + \gamma(1-R)\sum_{s'\in\mathcal{S}} p_{s,s'}^a \left(\sum_b \pi(b|s')Q(s',b)\right) + \gamma R \cdot \text{LSE}\left(\sum_b \pi(b|s')Q(s',b)\right). \tag{115}$$

Algorithm 5 Smoothed Robust TD (Tabular Setting)

Input: T_c, π, σ
Initialization: Q_0, s_0
  for $t = 0, 1, ..., T_c - 1$ **do**
    Choose $a_t \sim \pi(\cdot|s_t)$ and observe c_t, s_{t+1}
    $V_t(s) \leftarrow \sum_{a\in\mathcal{A}} \pi(a|s)Q_t(s,a)$ for all $s \in \mathcal{S}$
    $Q_{t+1}(s_t, a_t) \leftarrow Q_t(s_t, a_t) + \alpha_t(c_t + \gamma(1-R)\cdot V_t(s_{t+1}) + \gamma R \cdot \text{LSE}(V_t) - Q_t(s_t, a_t))$
  end for
Output: Q_{T_c}

Note that

$$\begin{aligned}
\|\mathbf{T}_\sigma^\pi Q_1 - \mathbf{T}_\sigma^\pi Q_2\|_\infty = \max_{s,a} \Bigg| & \gamma(1-R)\sum_{s'\in\mathcal{S}} p_{s,s'}^a \left(\sum_b \pi(b|s')(Q_1(s',b) - Q_2(s',b))\right) \\
& + \gamma R \cdot \text{LSE}\left(\sum_b \pi(b|s')Q_1(s',b)\right) - \gamma R \cdot \text{LSE}\left(\sum_b \pi(b|s')Q_2(s',b)\right)\Bigg| \\
\overset{(a)}{\leq} & \gamma(1-R)\sum_{s'\in\mathcal{S}} p_{s,s'}^a \|Q_1 - Q_2\|_\infty + \left|\gamma R \cdot \left(\sum_b \pi(b|s')(Q_1(s',b) - Q_2(s',b))\right)\right| \\
\leq & \gamma\|Q_1 - Q_2\|_\infty, \qquad (116)
\end{aligned}$$

where (a) is from the fact that LSE is 1-Lipschitz for any $\sigma > 0$, i.e., $|\text{LSE}(V_1) - \text{LSE}(V_2)| \leq \|V_1 - V_2\|_\infty$. Therefore, $\mathbf{T}_\sigma^\pi$ is a contraction and Q_σ^π is its fixed point. Hence following (Borkar & Meyn, 2000), smoothed robust TD converges to its fixed point Q_σ^π almost surely.

The proof for the convergence of the non-smoothed robust TD (Algorithm 5) follows similarly, and is omitted here. □

C.2. Robust Actor-Critic under Tabular Setting

Algorithm 6 Robust Smoothed Actor-Critic under Tabular Setting

Input: T, T_c, σ, M
Initialization: θ_0
  for $t = 0, 1, ..., T - 1$ **do**
    Run Algorithm 5 for T_c times until $\|Q_{T_c} - Q_\sigma^{\pi_\theta}\| \leq \epsilon_{\text{est}}$
    $Q_t \leftarrow Q_{T_c}$
    $V_t(s) \leftarrow \sum_{a\in\mathcal{A}} \pi_\theta(a|s)Q_t(s,a)$ for all $s \in \mathcal{S}$
    for $i = 1, ..., M$ **do**
      Sample $T^i \sim$ **Geom**$(1 - \gamma + \gamma R)$
      Sample $s_0^i \sim \rho$
      Sample trajectory starting from s_0^i: $(s_0^i, a_0^i, ..., s_{T^i}^i)$
      $B_t^i \leftarrow \frac{1}{1-\gamma+\gamma R}\sum_{a\in\mathcal{A}} \nabla\pi_\theta(a|s_{T^i}^i)Q_t(s_{T^i}^i, a)$
      Sample $x_0^i \sim \text{softmax}(\sigma, V^t)$
      Sample trajectory starting from x_0^i: $(x_0^i, b_0^i, ..., x_{T^i}^i)$
      $D_t^i \leftarrow \frac{1}{1-\gamma+\gamma R}\sum_{a\in\mathcal{A}} \nabla\pi_\theta(a|x_{T^i}^i)Q_t(x_{T^i}^i, a)$
      $g_t^i \leftarrow B_t^i + \frac{\gamma R}{1-\gamma} D_t^i$
    end for
    $g_t \leftarrow \frac{\sum_{i=1}^M g_t^i}{M}$
    $\theta_{t+1} \leftarrow \prod_\Theta(\theta_t - \alpha_t g_t)$
  end for
Output: θ_T

We then show the convergence of Algorithm 6. First we show that the update g_t is an unbiased estimate of the gradient ∇J_σ

if $Q_t = Q_\sigma^\pi$.

Lemma C.2. *If $Q_t = Q_\sigma^{\theta_t}$, then g_t is an unbiased estimate of ∇J_σ, i.e.,*

$$\mathbb{E}\left[g_t|\mathcal{F}_t\right] = \nabla J_\sigma(\theta_t), \tag{117}$$

where $\mathcal{F}_t$ denotes the σ-field generated by all the randomness until the t-th iteration.

Proof. First note that T^i is generated following the geometry distribution $\mathbf{Geom}(1-\gamma+\gamma R)$, thus for any $s, s' \in \mathcal{S}$

$$\begin{aligned}
\mathbb{P}(S_{T^i} = s'|S_0 = s) &= \sum_{k=0}^{\infty} \mathbb{P}(T^i = k)\mathbb{P}(S_k = s'|S_0 = s, \pi_\theta) \\
&= \sum_{k=0}^{\infty} (1-\gamma+\gamma R)(\gamma-\gamma R)^k \mathbb{P}(S_k = s'|S_0 = s, \pi_\theta) \\
&= (1-\gamma+\gamma R)\sum_{k=0}^{\infty} (\gamma-\gamma R)^k \mathbb{P}(S_k = s'|S_0 = s, \pi_\theta) \\
&= d_s^{\pi_\theta}(s').
\end{aligned} \tag{118}$$

Hence,

$$\begin{aligned}
\mathbb{E}[B_t^i|\mathcal{F}_t] &= \mathbb{E}\left[\frac{1}{1-\gamma+\gamma R}\sum_{a\in\mathcal{A}} \nabla\pi_{\theta_t}(a|S_{T^i}^i)Q_t(S_{T^i}^i, a)\Big|\mathcal{F}_t\right] \\
&= \frac{1}{1-\gamma+\gamma R}\sum_{s\in\mathcal{S}} \rho(s)\sum_{s'\in\mathcal{S}} \mathbb{P}(S_{T^i}^i = s'|S_0^i = s)\sum_{a\in\mathcal{A}} \nabla\pi_{\theta_t}(a|s')Q_t(s', a) \\
&= \frac{1}{1-\gamma+\gamma R}\sum_{s'\in\mathcal{S}} d_\rho^{\pi_{\theta_t}}(s')\sum_{a\in\mathcal{A}} \nabla\pi_{\theta_t}(a|s')Q_t(s', a) \\
&= B(\rho, \theta_t).
\end{aligned} \tag{119}$$

According to the algorithm,

$$\mathbb{P}(x_0^i = s') = \frac{e^{\sigma V_t(s')}}{\sum_{s\in\mathcal{S}} e^{\sigma V^t(s)}}, \tag{120}$$

hence we further have that

$$\begin{aligned}
\mathbb{E}[D_t^i|\mathcal{F}_t] &= \frac{1}{1-\gamma+\gamma R}\sum_{s'\in\mathcal{S}} \mathbb{P}(x_0^i = s')\sum_{s''\in\mathcal{S}} \mathbb{P}(x_{T^i}^i = s''|x_0^i = s')\left(\sum_{a\in\mathcal{A}} \nabla\pi_{\theta_t}(a|s'')Q_t(s'', a)\right) \\
&= \sum_{s'\in\mathcal{S}} \frac{e^{\sigma V_t(s')}}{\sum_{s\in\mathcal{S}} e^{\sigma V^t(s)}}\sum_{s''\in\mathcal{S}} d_{s'}^{\pi_{\theta_t}}(s'')\left(\sum_{a\in\mathcal{A}} \nabla\pi_{\theta_t}(a|s'')Q_t(s'', a)\right) \\
&= \sum_{s'\in\mathcal{S}} \frac{e^{\sigma V_t(s')}}{\sum_{s\in\mathcal{S}} e^{\sigma V^t(s)}} B(s', \theta) \\
&= \sum_{s'\in\mathcal{S}} \frac{e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}}{\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}} B(s', \theta).
\end{aligned} \tag{121}$$

Combining (119) and (121) thus implies that

$$\begin{aligned}
\mathbb{E}\left[g_t^i|\mathcal{F}_t\right] &= B(\rho, \theta_t) + \frac{\gamma R}{1-\gamma}\sum_{s'\in\mathcal{S}} \frac{e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}}{\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}} B(s', \theta) \\
&= \nabla J_\sigma(\theta_t),
\end{aligned} \tag{122}$$

which completes the proof. □

However, in Algorithm 6 we use an estimate Q_t instead of $Q_\sigma^{\theta_t}$, and therefore the estimate g_t may be biased. The next lemma develops an upper bound on this bias.

Lemma C.3. *Consider Algorithm 6. If T_c is chosen such that $\|Q_{T_c} - Q_\sigma^{\theta_t}\|_\infty \le \epsilon_{est}$, then we have that for any i, t,*

$$\begin{aligned}\|\mathbb{E}[g_t^i|\mathcal{F}_t] - \nabla J_\sigma(\theta_t)\| &\le 2\sigma\epsilon_{est}e^{\sigma\epsilon_{est}}\frac{\gamma R}{1-\gamma}\frac{|\mathcal{A}|(\epsilon_{est}+C_\sigma)}{1-\gamma+\gamma R} + \frac{\gamma R}{1-\gamma}\frac{|\mathcal{A}|\epsilon_{est}}{1-\gamma+\gamma R} + \frac{|\mathcal{A}|\epsilon_{est}}{1-\gamma+\gamma R}\\ &= \mathcal{O}(\epsilon_{est} + \sigma\epsilon_{est}e^{\sigma\epsilon_{est}}).\end{aligned} \tag{123}$$

Proof. We first have that

$$\begin{aligned}&\mathbb{E}[B_t^i|\mathcal{F}_t]\\ &= \mathbb{E}\left[\frac{1}{1-\gamma+\gamma R}\sum_{a\in\mathcal{A}}\nabla\pi_{\theta_t}(a|s_{T^i}^i)Q_t(s_{T^i}^i,a)\Big|\mathcal{F}_t\right]\\ &= \mathbb{E}\left[\frac{1}{1-\gamma+\gamma R}\sum_{a\in\mathcal{A}}\nabla\pi_{\theta_t}(a|s_{T^i}^i)Q_\sigma^{\theta_t}(s_{T^i}^i,a)\Big|\mathcal{F}_t\right]\\ &\quad + \mathbb{E}\left[\frac{1}{1-\gamma+\gamma R}\sum_{a\in\mathcal{A}}\nabla\pi_{\theta_t}(a|s_{T^i}^i)(Q_t(s_{T^i}^i,a) - Q_\sigma^{\theta_t}(s_{T^i}^i,a))\Big|\mathcal{F}_t\right]\\ &= B(\rho,\theta_t) + \mathbb{E}\left[\frac{1}{1-\gamma+\gamma R}\sum_{a\in\mathcal{A}}\nabla\pi_{\theta_t}(a|s_{T^i}^i)(Q_t(s_{T^i}^i,a) - Q_\sigma^{\theta_t}(s_{T^i}^i,a))\Big|\mathcal{F}_t\right].\end{aligned} \tag{124}$$

We can show that

$$\left\|\mathbb{E}\left[\frac{1}{1-\gamma+\gamma R}\sum_{a\in\mathcal{A}}\nabla\pi_{\theta_t}(a|s_{T^i}^i)(Q_t(s_{T^i}^i,a) - Q_\sigma^{\theta_t}(s_{T^i}^i,a))\Big|\mathcal{F}_t\right]\right\| \le \frac{|\mathcal{A}|\epsilon_{\text{est}}}{1-\gamma+\gamma R}, \tag{125}$$

which is from $\|\nabla\pi_{\theta_t}(a|s)\| = 1$ and $\mathbb{E}[|Q_t(s_{T^i}^i,a) - Q_\sigma^{\theta_t}(s_{T^i}^i,a)|] \le \epsilon_{\text{est}}$. Hence

$$\left\|\mathbb{E}[B_t^i|\mathcal{F}_t] - B(s,\theta_t)\right\| \le \frac{|\mathcal{A}|\epsilon_{\text{est}}}{1-\gamma+\gamma R}, \tag{126}$$

which means the bias of B_t^i is bounded by $\mathcal{O}(\epsilon_{\text{est}})$.

We then bound the bias of D_t^i. We first have that

$$\begin{aligned}&\frac{\sum_{s'\in\mathcal{S}}B(s',\theta_t)e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}}{\sum_{s'\in\mathcal{S}}e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}} - \mathbb{E}[D_t^i|\mathcal{F}_t]\\ &= \frac{\sum_{s'\in\mathcal{S}}B(s',\theta_t)e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}}{\sum_{s'\in\mathcal{S}}e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}} - \frac{1}{1-\gamma+\gamma R}\sum_{s'\in\mathcal{S}}\frac{e^{\sigma V_t(s')}}{\sum_{s'\in\mathcal{S}}e^{\sigma V_t(s')}}\sum_{s\in\mathcal{S}}d_{s'}^{\pi_{\theta_t}}(s)\sum_{a\in\mathcal{A}}\nabla\pi_{\theta_t}(a|s)Q_t(s,a).\end{aligned} \tag{127}$$

Define $B_t(s') \triangleq \frac{1}{1-\gamma+\gamma R}\sum_{s\in\mathcal{S}}d_{s'}^{\pi_{\theta_t}}(s)\sum_{a\in\mathcal{A}}\nabla\pi_{\theta_t}(a|s)Q_t(s,a)$, then the bias of D_t^i can be further written as

$$\begin{aligned}&\frac{\sum_{s'\in\mathcal{S}}B(s',\theta_t)e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}}{\sum_{s'\in\mathcal{S}}e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}} - \mathbb{E}[D_t^i|\mathcal{F}_t]\\ &= \frac{\sum_{s'\in\mathcal{S}}B(s',\theta_t)e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}}{\sum_{s'\in\mathcal{S}}e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}} - \frac{\sum_{s'\in\mathcal{S}}e^{\sigma V_t(s')}B_t(s')}{\sum_{s'\in\mathcal{S}}e^{\sigma V_t(s')}}\\ &= \frac{\sum_{s,s'\in\mathcal{S}}B(s',\theta_t)e^{\sigma V_t(s)}e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')} - \sum_{s,s'\in\mathcal{S}}B_t(s')e^{\sigma V_\sigma^{\pi_{\theta_t}}(s)}e^{\sigma V_t(s')}}{\left(\sum_{s'\in\mathcal{S}}e^{\sigma V_t(s')}\right)\left(\sum_{s'\in\mathcal{S}}e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}\right)}\end{aligned}$$

$$= \frac{\sum_{s,s'\in\mathcal{S}} B(s',\theta_t)e^{\sigma V_t(s)}e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')} - \sum_{s,s'\in\mathcal{S}} B_t(s')e^{\sigma V_t(s)}e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}}{\left(\sum_{s'\in\mathcal{S}} e^{\sigma V_t(s')}\right)\left(\sum_{s'\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}\right)}$$
$$+ \frac{\sum_{s,s'\in\mathcal{S}} B_t(s')e^{\sigma V_t(s)}e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')} - \sum_{s,s'\in\mathcal{S}} B_t(s')e^{\sigma V_\sigma^{\pi_{\theta_t}}(s)}e^{\sigma V_t(s')}}{\left(\sum_{s'\in\mathcal{S}} e^{\sigma V_t(s')}\right)\left(\sum_{s'\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}\right)}. \tag{128}$$

The first term can be bounded as

$$\frac{\sum_{s,s'\in\mathcal{S}} B(s',\theta_t)e^{\sigma V_t(s)}e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')} - \sum_{s,s'\in\mathcal{S}} B_t(s')e^{\sigma V_t(s)}e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}}{\left(\sum_{s'\in\mathcal{S}} e^{\sigma V_t(s')}\right)\left(\sum_{s'\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}\right)}$$
$$= \frac{\sum_{s,s'\in\mathcal{S}} e^{\sigma V_t(s)}e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}(B(s',\theta_t) - B_t(s'))}{\left(\sum_{s'\in\mathcal{S}} e^{\sigma V_t(s')}\right)\left(\sum_{s'\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}\right)}$$
$$\leq \frac{|\mathcal{A}|\epsilon_{\text{est}}}{1-\gamma+\gamma R}, \tag{129}$$

where the last inequality is from (126).

We then bound the second term in (128). Consider the numerator,

$$\sum_{s,s'\in\mathcal{S}} B_t(s')e^{\sigma V_t(s)}e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')} - \sum_{s,s'\in\mathcal{S}} B_t(s')e^{\sigma V_\sigma^{\pi_{\theta_t}}(s)}e^{\sigma V_t(s')}$$
$$= \left(\sum_{s\in\mathcal{S}} e^{\sigma V_t(s)}\right)\left(\sum_{s'\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}B_t(s')\right) - \left(\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s)}\right)\left(\sum_{s'\in\mathcal{S}} e^{\sigma V_t(s')}B_t(s')\right)$$
$$= \left(\sum_{s\in\mathcal{S}} e^{\sigma V_t(s)}\right)\left(\sum_{s'\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}B_t(s')\right) - \left(\sum_{s\in\mathcal{S}} e^{\sigma V_t(s)}\right)\left(\sum_{s'\in\mathcal{S}} e^{\sigma V_t(s')}B_t(s')\right)$$
$$+ \left(\sum_{s\in\mathcal{S}} e^{\sigma V_t(s)}\right)\left(\sum_{s'\in\mathcal{S}} e^{\sigma V_t(s')}B_t(s')\right) - \left(\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s)}\right)\left(\sum_{s'\in\mathcal{S}} e^{\sigma V_t(s')}B_t(s')\right)$$
$$= \underbrace{\left(\sum_{s\in\mathcal{S}} e^{\sigma V_t(s)}\right)\left(\sum_{s'\in\mathcal{S}} \left(e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')} - e^{\sigma V_t(s')}\right)B_t(s'))\right)}_{(a)}$$
$$+ \underbrace{\left(\sum_{s\in\mathcal{S}} e^{\sigma V_t(s)} - \sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s)}\right)\left(\sum_{s'\in\mathcal{S}} e^{\sigma V_t(s')}B_t(s')\right)}_{(b)}. \tag{130}$$

To bound term (a), we first have that

$$e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')} - e^{\sigma V_t(s')} = (V_\sigma^{\pi_{\theta_t}}(s') - V_t(s'))\sigma e^{\sigma(V_\sigma^{\pi_{\theta_t}}(s') + \lambda(V_t(s') - V_\sigma^{\pi_{\theta_t}}(s')))} \leq \sigma\epsilon_{\text{est}}e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}e^{\sigma\epsilon_{\text{est}}}, \tag{131}$$

where the first equation is from the mean-value theorem for some $0 \leq \lambda \leq 1$, and the last inequality is because $V_\sigma^{\pi_{\theta_t}}(s') - V_t(s') \leq \epsilon_{\text{est}}$ and $1 - \lambda \leq 1$. Hence term (a) can be bounded as follows:

$$\left\|\left(\sum_{s\in\mathcal{S}} e^{\sigma V_t(s)}\right)\left(\sum_{s'\in\mathcal{S}} \left(e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')} - e^{\sigma V_t(s')}\right)B_t(s'))\right)\right\|$$
$$\leq \left\|\left(\sum_{s\in\mathcal{S}} e^{\sigma V_t(s)}\right)\left(\sum_{s'\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}\right)\sigma\epsilon_{\text{est}}e^{\sigma\epsilon_{\text{est}}}\right\| \sup_s \|B_t(s)\|. \tag{132}$$

To bound term (b), for some $\lambda \in [0,1]$ and by the mean value theorem, we have that

$$\begin{aligned}
&\left\| \left(\sum_{s\in\mathcal{S}} e^{\sigma V_t(s)} - \sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s)} \right) \left(\sum_{s'\in\mathcal{S}} e^{\sigma V_t(s')} B_t(s') \right) \right\| \\
&= \left\| \left(\sum_{s'\in\mathcal{S}} e^{\sigma V_t(s')} B_t(s') \right) \right\| \left| \left(\sum_{s\in\mathcal{S}} \left(\sigma e^{\sigma V_\sigma^{\pi_{\theta_t}}(s) + \sigma(1-\lambda)(V_t(s) - V_\sigma^{\pi_{\theta_t}}(s))} (V_t(s) - V_\sigma^{\pi_{\theta_t}}(s)) \right) \right) \right| \\
&\leq \sigma \epsilon_{\text{est}} e^{\sigma\epsilon_{\text{est}}} \sup_s \|B_t(s)\| \left(\sum_{s\in\mathcal{S}} e^{\sigma V_t(s)} \right) \left(\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s)} \right). \qquad (133)
\end{aligned}$$

Hence combine the bounds on (a) and (b), we obtain a bound of the second term in (128):

$$\left\| \frac{\sum_{s,s'\in\mathcal{S}} B(s',\theta_t) e^{\sigma V_t(s)} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')} - \sum_{s,s'\in\mathcal{S}} B_t(s') e^{\sigma V_t(s)} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}}{\left(\sum_{s'\in\mathcal{S}} e^{\sigma V_t(s')}\right)\left(\sum_{s'\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}\right)} \right\| \leq 2\sigma\epsilon_{\text{est}} e^{\sigma\epsilon_{\text{est}}} \sup_s \|B_t(s)\|. \qquad (134)$$

Note that $B_t(s') = \frac{1}{1-\gamma+\gamma R} \sum_{s\in\mathcal{S}} d_{s'}^{\pi_{\theta_t}}(s) \sum_{a\in\mathcal{A}} \nabla\pi_{\theta_t}(a|s) Q_t(s,a)$, and $\|Q_t\|_\infty \leq \epsilon_{\text{est}} + \|Q_\sigma^{\theta_t}\|_\infty \leq \epsilon_{\text{est}} + C_\sigma$, hence

$$\sup_s \|B_t(s)\| \leq \frac{|\mathcal{A}|(\epsilon_{\text{est}} + C_\sigma)}{1-\gamma+\gamma R}, \qquad (135)$$

and thus,

$$\left\| \frac{\sum_{s,s'\in\mathcal{S}} B(s',\theta_t) e^{\sigma V_t(s)} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')} - \sum_{s,s'\in\mathcal{S}} B_t(s') e^{\sigma V_t(s)} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}}{\left(\sum_{s'\in\mathcal{S}} e^{\sigma V_t(s')}\right)\left(\sum_{s'\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}\right)} \right\| \leq 2\sigma\epsilon_{\text{est}} e^{\sigma\epsilon_{\text{est}}} \frac{|\mathcal{A}|(\epsilon_{\text{est}} + C_\sigma)}{1-\gamma+\gamma R}. \qquad (136)$$

Finally combining the bounds in (129) and (136) and plugging them in (128) implies that

$$\begin{aligned}
&\left\| \mathbb{E}\left[\frac{\sum_{s'\in\mathcal{S}} B(s',\theta_t) e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}}{\sum_{s'\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}} - D_t^i \middle| \mathcal{F}_t \right] \right\| \\
&\leq 2\sigma\epsilon_{\text{est}} e^{\sigma\epsilon_{\text{est}}} \frac{|\mathcal{A}|(\epsilon_{\text{est}} + C_\sigma)}{1-\gamma+\gamma R} + \frac{|\mathcal{A}|\epsilon_{\text{est}}}{1-\gamma+\gamma R}. \qquad (137)
\end{aligned}$$

Hence the bias of g_t^i can be bounded as follows

$$\begin{aligned}
&\|\mathbb{E}[g_t^i|\mathcal{F}_t] - \nabla J_\sigma(\theta_t)\| \\
&\leq \|\mathbb{E}[B_t^i|\mathcal{F}_t] - B(\rho,\theta_t)\| + \frac{\gamma R}{1-\gamma} \left\| \mathbb{E}\left[\frac{\sum_{s'\in\mathcal{S}} B(s',\theta_t) e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}}{\sum_{s'\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_{\theta_t}}(s')}} - D_t^i \middle| \mathcal{F}_t \right] \right\| \\
&\leq 2\sigma\epsilon_{\text{est}} e^{\sigma\epsilon_{\text{est}}} \frac{\gamma R}{1-\gamma} \frac{|\mathcal{A}|(\epsilon_{\text{est}} + C_\sigma)}{1-\gamma+\gamma R} + \frac{\gamma R}{1-\gamma} \frac{|\mathcal{A}|\epsilon_{\text{est}}}{1-\gamma+\gamma R} + \frac{|\mathcal{A}|\epsilon_{\text{est}}}{1-\gamma+\gamma R} \\
&\triangleq b_g = \mathcal{O}(\epsilon_{\text{est}} + \sigma\epsilon_{\text{est}} e^{\sigma\epsilon_{\text{est}}}). \qquad (138)
\end{aligned}$$

□

This theorem implies that the Algorithm 4 is actually a projected stochastic gradient descent with bias $b_g = \mathcal{O}(\epsilon_{\text{est}} + \sigma\epsilon_{\text{est}} e^{\sigma\epsilon_{\text{est}}})$.

C.3. Proof of Theorem 6.1: Global Convergence of Robust Actor-Critic under Tabular Setting

Denote $\Omega_t = g_t - \nabla J_\sigma(\theta_t)$, and define the stochastic gradient map by $H_t = \frac{1}{\alpha_t}(\theta_t - \prod_{(\Delta(\mathcal{A})^{|\mathcal{S}|}}(\theta_t - \alpha_t g_t))$. Note that J_σ is L_σ-smooth, hence similar to the proof of Theorem 1 in (Ghadimi et al., 2016), we can show that

$$J_\sigma(\theta_{t+1}) \leq J_\sigma(\theta_t) - \left(\alpha_t - \frac{L_\sigma}{2}\alpha_t^2 \right) \|H_t\|^2 + \alpha_t \langle \Omega_t, G_t \rangle + \alpha_t \|\Omega_t\| \|H_t - G_t\|$$

$$\leq J_\sigma(\theta_t) - \left(\alpha_t - \frac{L_\sigma}{2}\alpha_t^2\right)\|H_t\|^2 + \alpha_t\langle\Omega_t, G_t\rangle + \alpha_t\|\Omega_t\|^2, \tag{139}$$

where $G_t \triangleq \frac{1}{\alpha_t}(\theta_t - \prod_{(\Delta(\mathcal{A})^{|\mathcal{S}|}}(\theta_t - \alpha_t\nabla J_\sigma(\theta_t)))$, and the last inequality is from Proposition 1 in (Ghadimi et al., 2016). Summing up from $t = 0$ to $T-1$, we have

$$\begin{aligned}\sum_{t=0}^{T-1}(\alpha_t - L_\sigma\alpha_t^2)\|H_t\|^2 &\leq J_\sigma(\theta_0) - J_\sigma(\theta_{T+1}) + \sum_{t=0}^{T-1}\left(\alpha_t\langle\Omega_t, G_t\rangle + \alpha_t\|\Omega_t\|^2\right)\\ &\leq J_\sigma(\theta_0) - J_\sigma^* + \sum_{t=0}^{T-1}\left(\alpha_t\langle\Omega_t, G_t\rangle + \alpha_t\|\Omega_t\|^2\right).\end{aligned} \tag{140}$$

Note that G_t is deterministic given θ_t, i.e., $\mathcal{F}_t$. Hence we have

$$\mathbb{E}[\langle\Omega_t, G_t\rangle|\mathcal{F}_t] = \langle\mathbb{E}[\Omega_t|\mathcal{F}_t], G_t\rangle \leq \|G_t\|b_g. \tag{141}$$

Define $\Omega_t^j = g_t^j - \nabla J_\sigma(\theta_t)$. Then we have that

$$\begin{aligned}\mathbb{E}[\|\Omega_t\|^2|\mathcal{F}_t] &= \mathbb{E}\left[\left\|\frac{\sum_{j=1}^M \Omega_t^j}{M}\right\|^2\middle|\mathcal{F}\right]\\ &= \mathbb{E}\left[\sum_{i,j}\left\langle\frac{\Omega_t^j}{M}, \frac{\Omega_t^i}{M}\right\rangle\middle|\mathcal{F}_t\right]\\ &\overset{(a)}{=} \sum_{i=1}^M\mathbb{E}\left[\left\|\frac{\Omega_t^i}{M}\right\|^2 + \sum_{i\neq j}\left\langle\frac{\Omega_t^j}{M}, \frac{\Omega_t^i}{M}\right\rangle\middle|\mathcal{F}_t\right]\\ &\leq \frac{\sup_i\mathbb{E}[\|\Omega_t^i\|^2]}{M} + \|\mathbb{E}[\Omega_t^i|\mathcal{F}_t]\|^2\\ &\leq \frac{\sup_i\mathbb{E}[\|\Omega_t^i\|^2]}{M} + b_g^2,\end{aligned} \tag{142}$$

where (a) is from the fact that Ω_t^i and Ω_t^j are independent for $i \neq j$.

Note that for any j,

$$\mathbb{E}[\|\Omega_t^j\|^2] = \mathbb{E}[\|g_t^j - \nabla J_\sigma(\theta_t)^2\|] \leq 2\left(\sup_j\|g_t^j\|^2 + \sup_\theta\|\nabla J_\sigma(\theta)\|^2\right). \tag{143}$$

And we have shown that

$$\sup_j\|g_t^j\|^2 \leq \left(\frac{\gamma R}{1-\gamma} + 1\right)^2\frac{|\mathcal{A}|^2}{(1-\gamma+\gamma R)^2}(\sup_t\|Q_t\|)^2 \leq \left(\frac{\gamma R}{1-\gamma} + 1\right)^2\frac{|\mathcal{A}|^2}{(1-\gamma+\gamma R)^2}(C_\sigma + \epsilon_{\text{est}})^2 \triangleq C_g^2, \tag{144}$$

and hence

$$\mathbb{E}[\|\Omega_t^j\|^2] \leq 2\left(C_g^2 + (C_\sigma^V)^2\right), \tag{145}$$

where we use the fact $\|\nabla J_\sigma\| \leq C_\sigma^V$.

Thus

$$\mathbb{E}[\|\Omega_t\|^2] \leq b_g^2 + \frac{2\left(C_g^2 + {C_\sigma^V}^2\right)}{M} \triangleq C_\Omega. \tag{146}$$

Thus plugging all the inequalities (141) and (146) in (140), we have that

$$\sum_{t=0}^{T-1}(\alpha_t - L_\sigma\alpha_t^2)\mathbb{E}\left[\|H_t\|^2|\mathcal{F}_t\right] \leq J_\sigma(\theta_0) - J_\sigma^* + \sum_{t=0}^{T-1}\alpha_t\|G_t\|b_g + \sum_{t=0}^{T-1}\alpha_t C_\Omega. \tag{147}$$

Note that $\|G_t\| \leq \|\nabla J_\sigma(\theta_t)\| \leq C_\sigma^V$, hence the last inequality becomes

$$\sum_{t=0}^{T-1}(\alpha_t - L_\sigma \alpha_t^2)\mathbb{E}\left[\|H_t\|^2|\mathcal{F}_t\right] \leq J_\sigma(\theta_0) - J_\sigma^* + \sum_{t=0}^{T-1}\alpha_t C_\sigma^V b_g + \sum_{t=0}^{T-1}\alpha_t C_\Omega. \tag{148}$$

Set $\alpha_t = \frac{1}{2L_\sigma}$, it then follows that

$$\frac{1}{4L_\sigma}\sum_{t=0}^{T-1}\mathbb{E}\left[\|H_t\|^2|\mathcal{F}_t\right] \leq J_\sigma(\theta_0) - J_\sigma^* + \frac{TC_\sigma^V b_g}{2L_\sigma} + \frac{T}{2L_\sigma}C_\Omega, \tag{149}$$

and

$$\mathbb{E}[\|H_U\|^2] \leq \frac{4L_\sigma(J_\sigma(\theta_0) - J_\sigma^*)}{T} + 2C_\sigma^V b_g + 2C_\Omega, \tag{150}$$

where $U \sim \text{Uniform}(0, ..., T-1)$. Similar to Corollary 3 in (Ghadimi et al., 2016), we have that

$$\begin{aligned}\mathbb{E}\left[\|G_U\|^2\right] &\leq 2\mathbb{E}\left[\|H_U\|^2\right] + 2\mathbb{E}\left[\|H_U - \nabla J_\sigma(\theta_U)\|^2\right] \\ &= 2\mathbb{E}\left[\|H_U\|^2\right] + 2\mathbb{E}\left[\|\Omega_U\|^2\right] \\ &\leq \frac{8L_\sigma(J_\sigma(\theta_0) - J_\sigma^*)}{T} + 4C_\sigma^V b_g + 6C_\Omega \\ &\triangleq \epsilon_G.\end{aligned} \tag{151}$$

Note that G_t is fully determined by θ_t, hence we have that

$$\mathbb{E}\left[\|G_t\|^2\right] = \sum_{\theta \in (\Delta(\mathcal{A}))^{|\mathcal{S}|}} \mathbb{P}(\theta_t = \theta)\|G(\theta)\|^2, \tag{152}$$

where $G(\theta) = \frac{1}{\alpha}\left(\theta - \prod_{(\Delta(\mathcal{A})^{|\mathcal{S}|}}(\theta - \alpha\nabla J_\sigma(\theta))\right)$, and we denote $\theta^+ = \theta - \alpha G(\theta)$.

If we denote $\|G(\theta)\|^2 \triangleq \epsilon_\theta$, then $\int_{\theta \in (\Delta(\mathcal{A}))^{|\mathcal{S}|}} \epsilon_\theta d\mathbb{P}(\theta_t = \theta) = \mathbb{E}[\|G_t\|^2]$. Following the proof of Theorem 5.5, and because of $\|G(\theta)\| = \sqrt{\epsilon_\theta}$, we have

$$-\nabla J_\sigma(\theta^+) \in N_{(\Delta(\mathcal{A}))^{|\mathcal{S}|}}(\theta^+) + \sqrt{\epsilon_\theta}B_2. \tag{153}$$

From the PL-condition in Theorem 5.2, we further have that

$$\begin{aligned}J_\sigma(\theta^+) - J_\sigma^* &\leq C_{PL}\max_{\hat{\pi} \in (\Delta(\mathcal{A}))^{|\mathcal{S}|}}\langle \pi_{\theta^+} - \hat{\pi}, \nabla J_\sigma(\theta^+)\rangle + \left(\frac{\gamma R}{1-\gamma}\right)\frac{2\log|\mathcal{S}|}{\sigma} \\ &\leq 2C_{PL}\sqrt{|\mathcal{S}|\epsilon_\theta} + \left(\frac{\gamma R}{1-\gamma}\right)\frac{2\log|\mathcal{S}|}{\sigma}.\end{aligned} \tag{154}$$

Hence we have that for any t:

$$\begin{aligned}\mathbb{E}[J_\sigma(\theta_t^+) - J_\sigma^*] &= \int_{\theta \in \Theta}(J_\sigma(\theta^+) - J_\sigma^*)d\mathbb{P}(\theta_t = \theta) \\ &\leq \int_{\theta \in \Theta} 2C_{PL}\sqrt{|\mathcal{S}|\epsilon_\theta}\mathbb{P}(\theta_t = \theta) + \left(\frac{\gamma R}{1-\gamma}\right)\frac{2\log|\mathcal{S}|}{\sigma} \\ &\leq 2C_{PL}\sqrt{|\mathcal{S}|}\sqrt{\int_{\theta \in \Theta}\epsilon_\theta d\mathbb{P}(\theta_t = \theta)} + \left(\frac{\gamma R}{1-\gamma}\right)\frac{2\log|\mathcal{S}|}{\sigma} \\ &= 2C_{PL}\sqrt{|\mathcal{S}|}\sqrt{\mathbb{E}[\|G_t\|^2]} + \left(\frac{\gamma R}{1-\gamma}\right)\frac{2\log|\mathcal{S}|}{\sigma}.\end{aligned} \tag{155}$$

And further we have that

$$\begin{aligned}\mathbb{E}[J_\sigma(\theta_U^+) - J_\sigma^*] &= \frac{\sum_{t=0}^{T-1}\mathbb{E}[J_\sigma(\theta_t^+) - J_\sigma^*]}{T} \\ &= \frac{\sum_{t=0}^{T-1} 2C_{PL}\sqrt{|\mathcal{S}|}\sqrt{\mathbb{E}[\|G_t\|^2]}}{T} + \left(\frac{\gamma R}{1-\gamma}\right)\frac{2\log|\mathcal{S}|}{\sigma} \\ &\leq 2C_{PL}\sqrt{|\mathcal{S}|}\sqrt{\frac{\sum_{t=0}^{T-1}\mathbb{E}[\|G_t\|^2]}{T}} + \left(\frac{\gamma R}{1-\gamma}\right)\frac{2\log|\mathcal{S}|}{\sigma} \\ &\leq 2C_{PL}\sqrt{|\mathcal{S}|\epsilon_G} + \left(\frac{\gamma R}{1-\gamma}\right)\frac{2\log|\mathcal{S}|}{\sigma}. \end{aligned} \tag{156}$$

Now consider $J_\sigma(\theta_{U+1})$. Note that

$$\begin{aligned}\mathbb{E}[J_\sigma(\theta_{U+1}) - J_\sigma(\theta_U^+)] &\leq \mathbb{E}[C_\sigma^V\|\theta_U^+ - \theta_{U+1}\|] \\ &\overset{(a)}{\leq} C_\sigma^V\alpha\mathbb{E}[\|g_U - \nabla J_\sigma(\theta_U)\|] \\ &= C_\sigma^V\alpha\mathbb{E}[\|\Omega_U\|] \\ &\leq C_\sigma^V\alpha\sqrt{C_\Omega}, \end{aligned} \tag{157}$$

where (a) is from Lemma 2 in (Ghadimi et al., 2016), and the last inequality is from (146) and the fact that $\mathbb{E}[\|\Omega_t\|] \leq \sqrt{\mathbb{E}[\|\Omega_t\|^2]} \leq \sqrt{C_\Omega}$ for any t.

Hence we have

$$\begin{aligned}\mathbb{E}[J_\sigma(\theta_{U+1}) - J_\sigma^*] &= \mathbb{E}[J_\sigma(\theta_{U+1}) - J_\sigma(\theta_U^+) + J_\sigma(\theta_U^+) - J_\sigma^*] \\ &\leq C_\sigma^V\alpha\sqrt{C_\Omega} + 2C_{PL}\sqrt{|\mathcal{S}|\epsilon_G} + \left(\frac{\gamma R}{1-\gamma}\right)\frac{2\log|\mathcal{S}|}{\sigma} \\ &\leq 2C_\sigma^V\frac{\epsilon}{L_\sigma} + 2C_{PL}\sqrt{|\mathcal{S}|\epsilon_G} + \epsilon. \end{aligned} \tag{158}$$

Moreover, we have that

$$\begin{aligned}|J_\sigma(\theta) - J(\theta)| &\leq \epsilon, \\ |J_\sigma^* - J^*| &\leq \epsilon, \end{aligned} \tag{159}$$

hence

$$\begin{aligned}\mathbb{E}[J(\theta_{U+1}) - J^*] &\leq 2\epsilon + 2C_\sigma^V\frac{\epsilon}{L_\sigma} + 2C_{PL}\sqrt{|\mathcal{S}|\epsilon_G} + \epsilon \\ &= 2C_{PL}\sqrt{|\mathcal{S}|}\sqrt{\frac{8L_\sigma(J_\sigma(\theta_0) - J_\sigma^*)}{T} + 4C_\sigma^V b_g + 6C_\Omega} + 3\epsilon + 2C_\sigma^V\frac{\epsilon}{L_\sigma}. \end{aligned} \tag{160}$$

Plug in the definition of C_Ω, and we further have that

$$\begin{aligned}\mathbb{E}[J(\theta_{U+1}) - J^*] &\leq 2\epsilon + 2C_\sigma^V\frac{\epsilon}{L_\sigma} + 2C_{PL}\sqrt{|\mathcal{S}|\epsilon_G} + \epsilon \\ &= 2C_{PL}\sqrt{|\mathcal{S}|}\sqrt{\frac{8L_\sigma(J_\sigma(\theta_0) - J_\sigma^*)}{T} + 4C_\sigma^V b_g + 6\left(b_g^2 + \frac{2\left(C_g^2 + {C_\sigma^V}^2\right)}{M}\right)} + 3\epsilon + 2C_\sigma^V\frac{\epsilon}{L_\sigma} \\ &\leq 2C_{PL}\sqrt{|\mathcal{S}|}\left(\sqrt{\frac{16L_\sigma C_\sigma}{T}} + \sqrt{4C_\sigma^V b_g + 6b_g^2} + \sqrt{\frac{12\left(C_g^2 + {C_\sigma^V}^2\right)}{M}}\right) \end{aligned} \tag{161}$$

Hence if we set

$$b_g \leq \frac{\epsilon}{2\sqrt{6|\mathcal{S}|}C_{PL}}, \tag{162}$$

$$M = \frac{48(C_g^2 + {C_\sigma^V}^2)|\mathcal{S}|C_{PL}^2}{\epsilon^2}, \tag{163}$$

$$T = \frac{64L_\sigma C_\sigma |\mathcal{S}|C_{PL}^2}{\epsilon^2}, \tag{164}$$

we have that

$$\min_{1\leq t\leq T} \mathbb{E}[J(\theta_t)] - J^* \leq \mathbb{E}[J(\theta_{U+1}) - J^*] \leq 6\epsilon + 2C_\sigma^V \frac{\epsilon}{L_\sigma} \leq 7\epsilon. \tag{165}$$

D. Useful Lemmas

Lemma D.1. *Let* $F(x) = \max\{f_1(x), ..., f_n(x)\}$, *and for any* x, *denote* $I(x) \triangleq \arg\max_i \{f_i(x)\}$. *Then* $\{\partial f_i(x) : i \in I_x\} \subseteq \partial F(x)$.

Proof. From the definition, we know that for any $i \in I_x$ and $g \in \partial f_i(x)$, we have

$$\liminf_{y\to x} \frac{f_i(y) - f_i(x) - \langle g, y - x\rangle}{\|y - x\|} \geq 0. \tag{166}$$

It can be showed that $g \in \partial F(x)$ as follows:

$$\liminf_{y\to x} \frac{F(y) - F(x) - \langle g, y - x\rangle}{\|y - x\|} \geq \liminf_{y\to x} \frac{f_i(y) - f_i(x) - \langle g, y - x\rangle}{\|y - x\|} \geq 0, \tag{167}$$

which is from $F(y) \geq f_i(y)$ and $F(x) = f_i(x)$. And this completes the proof. □

E. Constants

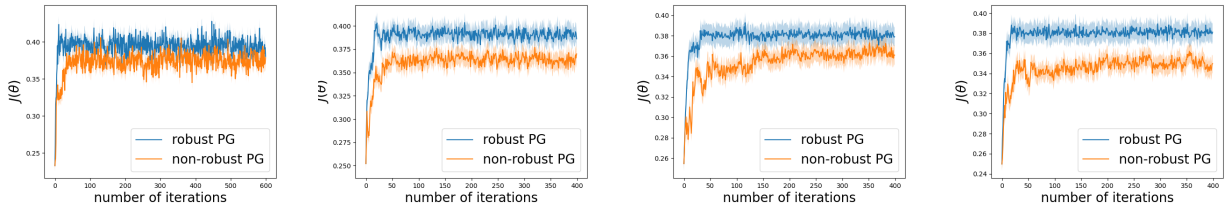
In this section we list the definition of all the constants in this paper.

$$\begin{aligned}
L_V &= \frac{k_\pi|\mathcal{A}|}{(1-\gamma)^2},\\
C_{PL} &= \frac{1}{(1-\gamma)\mu_{\min}},\\
C_\sigma &= \frac{1}{1-\gamma}(1 + 2\gamma R\frac{\log|\mathcal{S}|}{\sigma}),\\
C_\sigma^V &= \frac{1}{1-\gamma}|\mathcal{A}|k_\pi C_\sigma,\\
k_B &= \frac{1}{1-\gamma+\gamma R}\left(|\mathcal{A}|C_\sigma l_\pi + |\mathcal{A}|k_\pi C_\sigma^V\right) + \frac{2|\mathcal{A}|^2\gamma(1-R)}{(1-\gamma+\gamma R)^2}k_\pi^2 C_\sigma,\\
L_\sigma &= k_B + \frac{\gamma R}{1-\gamma}\left(\sqrt{|\mathcal{S}|}k_B + 2\sigma|\mathcal{S}|C_\sigma^V\frac{1}{1-\gamma+\gamma R}k_\pi|\mathcal{A}|C_\sigma\right),\\
b_g &= 2\sigma\epsilon_{\text{est}}e^{\sigma\epsilon_{\text{est}}}\frac{\gamma R}{1-\gamma}\frac{|\mathcal{A}|(\epsilon_{\text{est}} + C_\sigma)}{1-\gamma+\gamma R} + \frac{\gamma R}{1-\gamma}\frac{|\mathcal{A}|\epsilon_{\text{est}}}{1-\gamma+\gamma R} + \frac{|\mathcal{A}|\epsilon_{\text{est}}}{1-\gamma+\gamma R},\\
C_g &= \left(\frac{\gamma R}{1-\gamma} + 1\right)\frac{|\mathcal{A}|}{(1-\gamma+\gamma R)}(C_\sigma + \epsilon_{\text{est}}),\\
C_\Omega &= b_g^2 + \frac{2\left(C_g^2 + {C_\sigma^V}^2\right)}{M},\\
\epsilon_G &= \frac{8L_\sigma(J_\sigma(\theta_0) - J_\sigma^*)}{T} + 4C_\sigma^V b_g + 6C_\Omega.
\end{aligned} \tag{168}$$

F. Additional Experiments

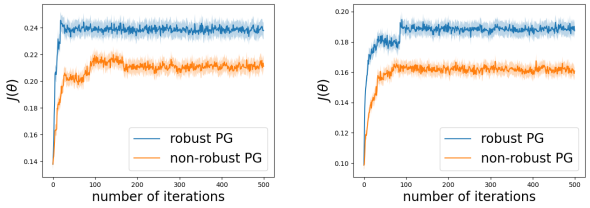
In this section we present some additional experiments to demonstrate our theoretical results.

Robust Policy Gradient. We provide more experiment results on robust policy gradient v.s. non-robust one. In Figure 8 we compare the two algorithms on $\mathcal{G}(12, 6)$, and in Figure 9, we compare them on $\mathcal{G}(20, 10)$. All the results show that our robust policy gradient can find a policy that has higher accumulated discounted reward under the worst-case transition kernel.



(a) $R = 0.12$. (b) $R = 0.15$. (c) $R = 0.2$. (d) $R = 0.25$.

Figure 8. Robust Policy Gradient v.s. Non-robust Policy Gradient on Garnet Problem $\mathcal{G}(12, 6)$.

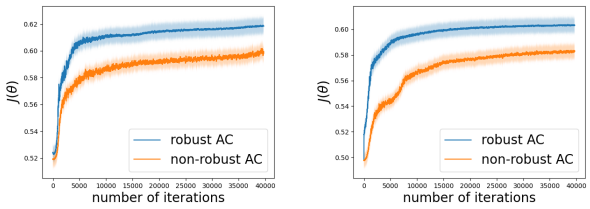


(a) $R = 0.1$. (b) $R = 0.2$.

Figure 9. Robust Policy Gradient v.s. Non-robust Policy Gradient on Garnet Problem $\mathcal{G}(20, 10)$.

Robust Actor-Critic.

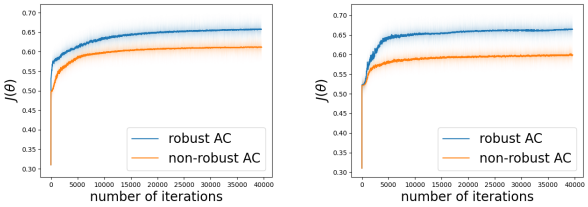
In Figure 10, we compare robust actor-critic and vanilla one on Garnet problem $\mathcal{G}(40, 15)$ using neural policy and neural network to approximate robust value functions. We plot the discounted accumulative reward under the worst case v.s. number of iterations. Our results suggest that robust actor-critic is more robust than vanilla actor-critic under the model mismatch.



(a) $R = 0.15$. (b) $R = 0.2$.

Figure 10. Robust Actor-Critic v.s. Non-robust Actor-Critic on Garnet Problem $\mathcal{G}(40, 15)$.

In Figure 11, we consider Garnet problem $\mathcal{G}(30, 10)$ using direct policy parameterization, and we use a two-layer neural network (with 20 neurons in the hidden layer) in the critic to approximate the robust value function. As the results show, our robust actor-critic algorithm finds a policy that achieves a higher accumulated discounted reward under the worst-case transition kernel than the vanilla actor-critic algorithm.

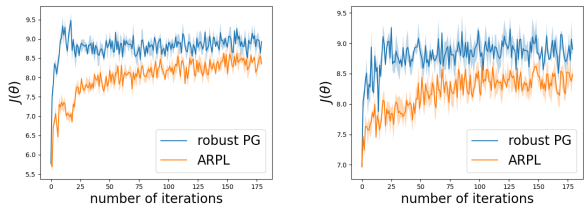


(a) $R = 0.15$. (b) $R = 0.2$.

Figure 11. Robust Actor-Critic v.s. Non-robust Actor-Critic on Garnet Problem $\mathcal{G}(30, 10)$.

Comparison with ARPL

In Figure 12, we compare robust PG with ARPL (Mandlekar et al., 2017) on Frozen-Lake problem. We plot the discounted accumulative reward under the worst case v.s. number of iterations. The results also show that our method is more robust than ARPL.



(a) $R = 0.15$. (b) $R = 0.2$.

Figure 12. Robust PG v.s. ARPL on Frozen-Lake Problem.