Robust Constrained Reinforcement Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Constrained reinforcement learning is to maximize the expected reward subject to constraints on utilities/costs. However, the training environment may not be the same as the test one, due to, e.g., modeling error, adversarial attack, non-stationarity, resulting in severe performance degradation and more importantly constraint violation. We propose a framework of robust constrained reinforcement learning under model uncertainty, where the MDP is not fixed but lies in some uncertainty set, the goal is to guarantee that constraints on utilities/costs are satisfied for all MDPs in the uncertainty set, and to maximize the worst-case reward performance over the uncertainty set. We design a robust primal-dual approach, and further theoretically develop guarantee on its convergence, complexity and robust feasibility. We then investigate a concrete example of δ-contamination uncertainty set, design an online and model-free algorithm and theoretically characterize its finite-sample error bound. We demonstrate the robustness and the worst-case feasibility of our approach numerically.

1 Introduction

In many practical reinforcement learning (RL) applications, it is critical for an agent to meet certain constraints on utilities and costs while maximizing the reward, e.g., safety constraint in autonomous driving [23] and robotics [51]. However, in practice, it is often the case that the environment on which a learned policy will be deployed deviates from the one that was used to generate the policy, due to, e.g., modeling error of the simulator, adversarial attack, and non-stationarity. This could lead to a significant performance degradation in reward, and more importantly, constraints may not be satisfied anymore, which is severe in safety-critical applications. For example, a drone may run out of battery due to model deviation between the training and test environments, resulting in a crash. To solve these issues, we propose a framework of robust constrained RL under model uncertainty. Specifically, the Markov decision process (MDP) is not fixed and lies in an uncertainty set [48, 31, 8], and the goal is to maximize the worst-case accumulative discounted reward over the uncertainty set while guaranteeing that constraints are satisfied for all MDPs in the uncertainty set at the same time.

Despite of its practical importance, studies on the problem of robust constrained RL are limited in the literature. Two closely related topics are robust RL [8, 48, 31] and constrained RL [5]. The problem of constrained RL [5] aims to find a policy that optimizes an objective reward while satisfying certain constraints on costs/utilities. For the problem of robust RL [8, 48, 31], the MDP is not fixed but lies in some uncertainty set, and the goal is to find a policy that optimizes the robust value function, which measures the worst-case accumulative reward over the uncertainty set. The problem of robust constrained RL was investigated in [64, 44], where two heuristic approaches were proposed. The basic idea in [64, 44] is to first evaluate the worst-case performance of the policy over the uncertainty set, and then use that together with classical policy improvement methods, e.g., policy gradient [68], to update the policy. However, as will be discussed in more details later, these approaches may not necessarily lead to an improved policy, and thus may not perform well in practice.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

In this paper, we design the robust primal-dual algorithm for the problem of robust constrained RL. Our approach employs the true gradient of the Lagrangian function, which is the weighted sum of two robust value functions, instead of approximating the gradient heuristically as in [64]. We theoretically characterize the convergence and complexity of our robust primal-dual method, and prove the robust feasibility of our solution for all MDPs in the uncertainty set. We further present a concrete example of δ-contamination uncertainty set [30, 21, 29, 49, 50, 58, 59, 74, 75], for which we extend our algorithm to the online and model-free setting, and theoretically characterize its finite-time error bound. In particular, the challenges and our major contributions are summarized as follows.

- In our primal-dual method, the Lagrangian function is the sum of two robust value functions. In the non-robust setting, the sum of two value functions is actually a value function of the combined reward. However, this does not hold in the robust setting, since the worst-case transition kernels for the two robust value functions are not necessarily the same, and therefore, the sum of two robust value functions cannot be written as a robust value function for the combined reward as being done in the non-robust setting. In the non-robust setting, although the value function is non-convex, it satisfies the Polyak-Łojasiewicz (PL) condition [57, 42], and thus is convex-like [3, 12]. Though it was shown in [75] that robust value function (for the case with δ-contamination uncertainty set) also satisfies the PL condition, the sum of two robust value functions may not do so. Therefore, the geometry of our Lagrangian function is much more complicated than the geometry of the non-robust constrained RL problem and the robust RL problem without constraints. In this paper, we formulate the dual problem of the robust constrained RL problem as a minimax linear-nonconcave optimization problem, and show that the optimal dual variable is bounded. We then construct a robust primal-dual algorithm by alternatively updating the primal and dual variables. We theoretically prove the convergence to stationary points, and characterize its complexity. More importantly, we prove that the solution we obtain is feasible for all MDPs in the uncertainty set.
- Despite being a constrained optimization problem with non-convex objective and constraints, existing studies [5, 54] show that non-robust constrained RL has zero duality gap, based on which global optimality for various primal-dual methods can be established [19, 18, 34, 40, 81]. The zero duality gap result for non-robust constrained RL relies on the fact that the set of all visitation distributions is convex. To show zero duality gap for robust constrained RL, we will need the set of all robust visitation distributions induced by the policy and its corresponding worst-case transition kernel being convex. However, this actually does not hold. We construct a counter example, and show that such set is actually non-convex.
- We provide a concrete example with δ-contamination uncertainty set. For this example, the robust value function is not differentiable [75]. We then propose a smoothed approximation of the robust value function towards a better geometry. We theoretically characterize the approximation error of doing so, and show that the smoothness condition required for convergence can be satisfied. We further investigate the practical online and model-free setting where only samples can be taken from the centroid of the uncertainty set. We design an online and model-free robust primal-dual algorithm, and further develop its finite-time error bound for the tabular case.

1.1 Related Works

We discuss works related to robust constrained RL. We focus on "soft" constraints that the trained policy is guaranteed to satisfy the constraints. There are also works on "hard" constraints that during the training, constraints are also satisfied, e.g., [71, 10, 61, 14, 4], which is not the focus here.

Robust constrained RL. In [64], the robust constrained RL problem was studied, and a heuristic approach was developed. The basic idea is to estimate the robust value functions, and then to use the vanilla policy gradient method [68] with the vanilla value function replaced by the robust value function. However, this approach did not take into consideration the fact that the worst-case transition kernel is also a function of the policy (see Section 3.1 in [64]), and therefore the "gradient" therein is not actually the gradient of the robust value function. Thus, its performance and convergence cannot be theoretically guaranteed. The other work [44] studied the same robust constrained RL problem under the continuous control setting, and proposed a similar heuristic algorithm. They first proposed a robust Bellman operator and used it to estimate the robust value function, which is further combined with some non-robust continuous control algorithm to update the policy. Both approaches in [64] and [44] inherit the heuristic structure of "robust policy evaluation" + "non-robust vanilla policy improvement", which may not necessarily guarantee an improved policy in general. In this paper, we employ a "robust policy evaluation" + "***robust*** policy improvement" approach, which guarantees an improvement in the policy, and more importantly, we provide theoretical convergence guarantee, robust feasibility guarantee, and complexity analysis for our algorithms.

Constrained RL. The most commonly used method for constrained RL is the primal-dual method [5, 54, 53, 35, 67, 70, 82, 85, 22, 7], which augments the objective with a sum of constraints weighted by their corresponding Lagrange multipliers, and then alternatively updates the primal and dual variables. It was shown that the strong duality holds for constrained RL, and hence the primal-dual method has zero duality gap [54, 5]. The convergence rate of the primal-dual method was investigated in [19, 18, 34, 40, 81]. Another class of method is the primal method, which is to enforce the constraints without resorting to the Lagrangian formulation [2, 41, 15, 17, 78]. The above studies, when directly applied to *robust* constrained RL, cannot guarantee the constraints when there is model deviation. Moreover, the objective and constraints in this paper take min over the uncertainty set (see (2)), and therefore have much more complicated geometry than the non-robust case.

Robust RL under model uncertainty. Model-based robust RL was firstly introduced and studied in [31, 48, 8, 66, 76, 36, 77, 83, 37, 69], where the uncertainty set is assumed to be known, and the problem can be solved using robust dynamic programming. It was then extended to the model-free setting, where the uncertainty set is unknown, and only samples from its centroid can be collected [63, 74, 75, 86, 80, 52, 25, 26]. There are also empirical studies on robust RL, e.g., [72, 56, 1, 27, 62, 28, 32, 39, 55, 43]. These works focus on robust RL without constraints, whereas in this paper we investigate robust RL with constraints, which is more challenging.

2 Preliminaries

Constrained MDP. A constrained MDP (CMDP) can be specified by a tuple $(\mathcal{S}, \mathcal{A}, \mathsf{P}, r, c_1, ..., c_m, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, $\mathsf{P} = \{p_s^a \in \Delta_{\mathcal{S}}, a \in \mathcal{A}, s \in \mathcal{S}\}$ is the transition kernel[1], $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, $c_i : \mathcal{S} \times \mathcal{A} \to [0, 1], i = 1, ..., m$ are utility functions in the constraint, and $\gamma \in [0, 1)$ is the discount factor. A stationary policy π is a mapping $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$, where $\pi(a|s)$ denotes the probability of taking action a when the agent is at state s. The set of all the stationary policies is denoted by Π.

The non-robust value function of reward r and a policy π is defined as the expected accumulative discounted reward if the agent follows policy π: $\mathbb{E}_{\pi,\mathsf{P}}[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t)|S_0 = s]$, where $\mathbb{E}_{\pi,\mathsf{P}}$ denotes the expectation when the policy is π and the transition kernel is P. Similarly, the non-robust value function of c is defined as $\mathbb{E}_{\pi,\mathsf{P}}[\sum_{t=0}^{\infty} \gamma^t c_i(S_t, A_t)|S_0 = s]$. The goal of CMDP is to find a policy that maximizes the expected reward subject to constraints on the expected utility:

$$\max_{\pi \in \Pi} \mathbb{E}_{\pi,\mathsf{P}} \left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t)|S_0 \sim \rho \right], \text{ s.t. } \mathbb{E}_{\pi,\mathsf{P}} \left[\sum_{t=0}^{\infty} \gamma^t c_i(S_t, A_t)|S_0 \sim \rho \right] \geq b_i, 1 \leq i \leq m, \quad (1)$$

where b_i's are some positive thresholds and ρ is the initial state distribution.

Define the visitation distribution induced by policy π and transition kernel P: $d_{\rho,\mathsf{P}}^{\pi}(s, a) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(S_t = s, A_t = a|S_0 \sim \rho, \pi, \mathsf{P})$. It can be shown that the set of the visitation distributions of all policies $\{d_{\rho,\mathsf{P}}^{\pi} \in \Delta_{\mathcal{S}\times\mathcal{A}} : \pi \in \Pi\}$ is convex [53, 5]. A standard assumption in the literature is the Slater's condition [11, 18]: There exists a constant $\zeta > 0$ and a policy $\pi \in \Pi$ s.t. $\forall i$, $\mathbb{E}_{\pi,\mathsf{P}} \left[\sum_{t=0}^{\infty} \gamma^t c_i(S_t, A_t)|S_0 \sim \rho\right] - b_i \geq \zeta$. Based on the convexity of the set of all visitation distributions and Slater's condition, strong duality can be established [5, 54].

Robust MDP. A robust MDP can be specified by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$. In this paper, we focus on the (s, a)-rectangular uncertainty set [48, 31], i.e., $\mathcal{P} = \bigotimes_{s,a} \mathcal{P}_s^a$, where $\mathcal{P}_s^a \subseteq \Delta_{\mathcal{S}}$. Denote the transition kernel at time t by P_t, and let $\kappa = (\mathsf{P}_0, \mathsf{P}_1, ...)$ be the dynamic model, where $\mathsf{P}_t \in \mathcal{P}, \forall t \geq 0$. We then define the robust value function of a policy π as the worst-case expected accumulative discounted reward following policy π over all MDPs in the uncertainty set [48, 31]:

$$V_r^{\pi}(s) = \min_{\kappa \in \bigotimes_{t \geq 0} \mathcal{P}} \mathbb{E}_{\kappa} \left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t)|S_0 = s, \pi \right], \quad (2)$$

[1] $\Delta_{\mathcal{S}}$ denotes the probability simplex supported on $\mathcal{S}$.

where $\mathbb{E}_\kappa$ denotes the expectation when the state transits according to κ. It has been shown that the robust value function is the fixed point of the robust Bellman operator [48, 31, 60]: $\mathbf{T}_\pi V(s) \triangleq \sum_{a\in\mathcal{A}} \pi(a|s)\left(r(s,a) + \gamma\sigma_{\mathcal{P}_s^a}(V)\right)$, where $\sigma_{\mathcal{P}_s^a}(V) \triangleq \min_{p\in\mathcal{P}_s^a} p^\top V$ is the support function of V on $\mathcal{P}_s^a$. Similarly, we can define the robust action-value function for a policy π: $Q_r^\pi(s,a) = \min_\kappa \mathbb{E}_\kappa \left[\sum_{t=0}^\infty \gamma^t r(S_t, A_t)|S_0 = s, A_0 = a, \pi\right]$.

Note that the minimizer of (2), κ^*, is stationary in time [31], which we denote by $\kappa^* = \{\mathsf{P}^\pi, \mathsf{P}^\pi, ...\}$, and refer to P^π as the worst-case transition kernel. Then the robust value function V_r^π is actually the value function under policy π and transition kernel P^π.

The goal of robust RL is to find the optimal robust policy π^* that maximizes the worst-case accumulative discounted reward: $\pi^* = \arg\max_\pi V_r^\pi(s), \forall s \in \mathcal{S}$. We also denote the optimal robust value function $V_r^{\pi^*}$ and the optimal robust action-value function $Q_r^{\pi^*}$ by V_r^* and Q_r^*, respectively, For robust MDP, the following robust analogue of the Bellman recursion was provided in [48, 31]: $V_r^*(s) = \max_{a\in\mathcal{A}}(r(s,a) + \gamma\sigma_{\mathcal{P}_s^a}(V_r^*))$, and $Q_r^*(s,a) = r(s,a) + \gamma\sigma_{\mathcal{P}_s^a}(V_r^*)$.

3 Robust Constrained RL

In this section, we introduce the framework of robust constrained RL, propose our robust primal-dual algorithm, characterize its convergence, complexity and robust feasibility theoretically.

We focus on a general parameterized policy class, i.e., $\pi_\theta \in \Pi_\Theta$, where $\Theta \subseteq \mathbb{R}^d$ is a parameter set and Π_Θ is a class of parameterized policies, e.g., direct parameterized policy, softmax or neural network policy. Robust constrained RL is to solve the following constrained optimization problem:

$$\max_{\theta\in\Theta} V_r^{\pi_\theta}(\rho), \text{ s.t. } V_{c_i}^{\pi_\theta}(\rho) \geq b_i, 1 \leq i \leq m, \tag{3}$$

where $V_{c_i}^{\pi_\theta}(\rho)$ is the robust value function of c_i and π_θ. We assume that the policy class Π_Θ is Lipschitz and smooth. This assumption can be easily satisfied by many policy classes, e.g., direct parameterization [3], soft-max [45, 33, 84, 73], or neural network with Lipschitz and smooth activation functions [20, 47, 46].

Assumption 1. *The policy class Π_Θ is k-Lipschitz and l-smooth, i.e., for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$ and for any $\theta \in \Theta$, there exist universal constants k, l, such that $\|\nabla\pi_\theta(a|s)\| \leq k$, and $\|\nabla^2\pi_\theta(a|s)\| \leq l$.*

Similar to the non-robust case, the problem (3) is equivalent to the following max-min problem:

$$\max_{\theta\in\Theta}\min_{\lambda_i\geq 0} V_r^{\pi_\theta}(\rho) + \sum_{i=1}^m \lambda_i(V_{c_i}^{\pi_\theta}(\rho) - b_i). \tag{4}$$

Unlike non-robust CMDP, strong duality for robust constrained RL may not hold. For robust RL, the robust value function can be viewed as the value function for policy π under its worst-case transition kernel P^π, and therefore can be written as the inner product between the reward (utility) function and the visitation distribution induced by π and P^π (referred to as robust visitation distribution of π). The following lemma shows that the set of robust visitation distributions may not be convex, and therefore, the approach used in [5, 54] to show strong duality cannot be applied here.

Lemma 1. *There exists a robust MDP, such that the set of robust visitation distributions is non-convex.*

In the following, we focus on the dual problem of (4). For simplicity, we investigate the case with one constraint, and extension to the case with multiple constraints is straightforward:

$$\min_{\lambda\geq 0}\max_{\theta\in\Theta} V_r^{\pi_\theta}(\rho) + \lambda(V_c^{\pi_\theta}(\rho) - b). \tag{5}$$

We make an assumption of Slater's condition, assuming there exists a strictly feasible policy [11, 18].

Assumption 2. *There exists $\zeta > 0$ and a policy $\pi \in \Pi_\Theta$, s.t. $V_c^\pi(\rho) - b \geq \zeta$.*

Under Assumption 2, we show that the optimal dual variable of (5) is bounded.

Lemma 2. *Denote the optimal solution of* (5) *by (λ^*, π^*). Then, $\lambda^* \in \left[0, \frac{2}{\zeta(1-\gamma)}\right]$.*

Lemma 2 suggests that the dual problem (5) is equivalent to a bounded min-max problem:

$$\min_{\lambda\in\left[0,\frac{2}{\zeta(1-\gamma)}\right]}\max_{\theta\in\Theta} V_r^{\pi_\theta}(\rho)+\lambda(V_c^{\pi_\theta}(\rho)-b). \tag{6}$$

The problem (6) is a bounded linear-nonconcave optimization problem. We then propose our robust primal-dual algorithm for robust constrained RL in Algorithm 1.

Algorithm 1 Robust Primal-Dual algorithm (RPD)

Input: T, α_t, β_t, b_t
Initialization: λ_0, θ_0
for $t = 0, 1, ..., T-1$ **do**
$\lambda_{t+1} \leftarrow \prod_{[0,\Lambda^*]}\left(\lambda_t - \frac{1}{\beta_t}\left(V_c^{\pi_{\theta_t}}(\rho)-b\right)-\frac{b_t}{\beta_t}\lambda_t\right)$
$\theta_{t+1} \leftarrow \prod_{\Theta}\left(\theta_t + \frac{1}{\alpha_t}\left(\nabla_\theta V_r^{\pi_{\theta_t}}(\rho)+\lambda_{t+1}\nabla_\theta V_c^{\pi_{\theta_t}}(\rho)\right)\right)$
end for
Output: θ_T

The basic idea of Algorithm 1 is to perform gradient descent-ascent w.r.t. λ and θ alternatively. When the policy π violates the constraint, the dual variable λ increases such that λV_c^π dominates V_r^π. Then the gradient ascent will update θ until the policy satisfies the constraint. Therefore, this approach is expected to find a feasible policy (as will be shown in Lemma 5).

Here, $\prod_{\mathcal{X}}(x)$ denotes the projection of x to the set $\mathcal{X}$, and $\{b_t\}$ is a non-negative monotone decreasing sequence, which will be specified later. Algorithm 1 reduces to the vanilla gradient descent-ascent algorithm in [38] if $b_t = 0$. However, b_t is critical to the convergence of Algorithm 1 [79]. The outer problem of (6) is actually linear, and after introducing b_t, the update of λ_t can be viewed as a gradient descent of a strongly-convex function $\lambda(V_c - b) + \frac{b_t}{2}\lambda^2$, which converges more stable and faster.

Denote that Lagrangian function by $V^L(\theta,\lambda) \triangleq V_r^{\pi_\theta}(\rho)+\lambda(V_c^{\pi_\theta}(\rho)-b)$, and further denote the gradient mapping of Algorithm 1 by

$$G_t \triangleq \begin{bmatrix} \beta_t\left(\lambda_t - \prod_{[0,\Lambda^*]}\left(\lambda_t - \frac{1}{\beta_t}\left(\nabla_\lambda V^L(\theta_t,\lambda_t)\right)\right)\right) \\ \alpha_t\left(\theta_t - \prod_{\Theta}\left(\theta_t + \frac{1}{\alpha_t}\left(\nabla_\theta V^L(\theta_t,\lambda_t)\right)\right)\right)\end{bmatrix}. \tag{7}$$

The gradient mapping is a standard measure of convergence for projected optimization approaches [9]. Intuitively, it reduces to the gradient $(\nabla_\lambda V^L, \nabla_\theta V^L)$, when $\Lambda^* = \infty$ and $\Theta = \mathbb{R}^d$, and it measures the updates of θ and λ at time step t. If $\|G_t\| \to 0$, the updates of both variables are small, and hence the algorithm converges to a stationary solution.

To show the convergence of Algorithm 1, we make the following Lipschitz smoothness assumption.

Assumption 3. *The gradients of the Lagrangian function are Lipschitz:*

$$\|\nabla_\lambda V^L(\theta,\lambda)|_{\theta_1} - \nabla_\lambda V^L(\theta,\lambda)|_{\theta_2}\| \le L_{11}\|\theta_1-\theta_2\|, \tag{8}$$
$$\|\nabla_\lambda V^L(\theta,\lambda)|_{\lambda_1} - \nabla_\lambda V^L(\theta,\lambda)|_{\lambda_2}\| \le L_{12}|\lambda_1-\lambda_2|, \tag{9}$$
$$\|\nabla_\theta V^L(\theta,\lambda)|_{\theta_1} - \nabla_\theta V^L(\theta,\lambda)|_{\theta_2}\| \le L_{21}\|\theta_1-\theta_2\|, \tag{10}$$
$$\|\nabla_\theta V^L(\theta,\lambda)|_{\lambda_1} - \nabla_\theta V^L(\theta,\lambda)|_{\lambda_2}\| \le L_{22}|\lambda_1-\lambda_2|. \tag{11}$$

In general, Assumption 3 may or may not hold depending on the uncertainty set model. As will be shown in Section 4, even if Assumption 3 does not hold, we can design a smoothed approximation of the robust value function, so that the assumption holds for the smoothed problem.

In the following theorem, we show that our robust primal-dual algorithm converges to a stationary point of the min-max problem (14), with a complexity of $\mathcal{O}(\epsilon^{-4})$.

Theorem 1. *Under Assumption 3, if we set step sizes α_t, β_t, and b_t as in Section I and $T = \mathcal{O}(\epsilon^{-4})$, then* $\min_{1\le t\le T}\|G_t\| \le 2\epsilon$.

The next proposition characterizes the feasibility of the obtained policy.

Proposition 1. *Denote by $W \triangleq \arg\min_{1\leq t\leq T}\|G_t\|$. If $\lambda_W - \frac{1}{\beta_W}\left(\nabla_\lambda V_\sigma^L(\theta_W,\lambda_W)\right) \in [0,\Lambda^*)$, then π_W satisfies the constraint with a 2ϵ-violation.*

Intuitively, if we set Λ^* larger so that the optimal solution $\lambda^* \in [0,\Lambda^*)$, then Algorithm 1 is expected to converge to an interior point of $[0,\Lambda^*]$ and therefore, π_W is feasible. On the other hand, Λ^* can't be set too large. Note that the complexity in Theorem 1 depends on Λ^* (see (52) in the appendix), and a larger Λ^* means a higher complexity.

4 δ-Contamination Uncertainty Set

In this section, we investigate a concrete example of robust constrained RL with δ-contamination uncertainty set. The δ-contamination uncertainty set models the scenario where the state transition of the MDP could be arbitrarily perturbed with a small probability δ. This model is widely used to model distributional uncertainty in the literature of robust learning and optimization, e.g., [30, 21, 29, 49, 50, 58, 59, 74, 75]. Specifically, let $\mathsf{P} = \{p_s^a | s\in\mathcal{S}, a\in\mathcal{A}\}$ be the centroid transition kernel, then the δ-contamination uncertainty set centered at P is defined as $\mathcal{P} \triangleq \bigotimes_{s\in\mathcal{S},a\in\mathcal{A}}\mathcal{P}_s^a$, where $\mathcal{P}_s^a \triangleq \{(1-\delta)p_s^a + \delta q | q\in\Delta_{\mathcal{S}}\}, s\in\mathcal{S}, a\in\mathcal{A}$.

Under the δ-contamination setting, the robust Bellman operator can be explicitly computed: $\mathbf{T}_\pi V(s) = \sum_{a\in\mathcal{A}}\pi(a|s)\left(r(s,a) + \gamma\left(\delta\min_{s'}V(s') + (1-\delta)\sum_{s'\in\mathcal{S}}p_{s,s'}^a V(s')\right)\right)$. In this case, the robust value function is not differentiable due to the min term, and hence Assumption 3 does not hold. One possible approach is to use sub-gradient [16, 75], which, however, is less stable, and its convergence is difficult to characterize. In the following, we design a differentiable and smooth approximation of the robust value function. Specifically, consider a smoothed robust Bellman operator $\mathbf{T}_\sigma^\pi$ using the LSE function [74, 75]:

$$\mathbf{T}_\sigma^\pi V(s) = \mathbb{E}_{A\sim\pi(\cdot|s)}\left[r(s,A) + \gamma(1-\delta)\sum_{s'\in\mathcal{S}}p_{s,s'}^A V(s') + \gamma\delta\mathrm{LSE}(\sigma,V)\right], \tag{12}$$

where $\mathrm{LSE}(\sigma,V) = \frac{\log(\sum_{i=1}^d e^{\sigma V(i)})}{\sigma}$ for $V\in\mathbb{R}^d$ and some $\sigma<0$. The approximation error $|\mathrm{LSE}(\sigma,V) - \min V| \to 0$ as $\sigma\to-\infty$, and hence the fixed point of $\mathbf{T}_\sigma^\pi$, denoted by V_σ^π, is an approximation of the robust value function V^π [75]. We refer to V_σ^π as the smoothed robust value function and define the smoothed robust action-value function as $Q_\sigma^\pi(s,a) \triangleq r(s,a) + \gamma(1-\delta)\sum_{s'\in\mathcal{S}}p_{s,s'}^a V_\sigma^\pi(s') + \gamma\delta\mathrm{LSE}(\sigma,V_\sigma^\pi)$. It can be shown that for any π, as $\sigma\to-\infty$, $\|V_r^\pi - V_{\sigma,r}^\pi\|\to 0$ and $\|V_c^\pi - V_{\sigma,c}^\pi\|\to 0$ [74].

The gradient of $V_\sigma^{\pi_\theta}$ can be computed explicitly [75]: $\nabla V_\sigma^{\pi_\theta}(s) = B(s,\theta) + \frac{\gamma\delta\sum_{s\in\mathcal{S}}e^{\sigma V_\sigma^{\pi_\theta}(s)}B(s,\theta)}{(1-\gamma)\sum_{s\in\mathcal{S}}e^{\sigma V_\sigma^{\pi_\theta}(s)}}$, where $B(s,\theta) \triangleq \frac{1}{1-\gamma+\gamma\delta}\sum_{s'\in\mathcal{S}}d_{s,\mathsf{P}}^{\pi_\theta}(s')\sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s')Q_\sigma^{\pi_\theta}(s',a)$, and $d_{s,\mathsf{P}}^{\pi_\theta}(\cdot)$ is the visitation distribution of π_θ under the centroid kernel P starting from s. Denote the smoothed Lagrangian function by $V_\sigma^L(\theta,\lambda) \triangleq V_{\sigma,r}^{\pi_\theta}(\rho) + \lambda(V_{\sigma,c}^{\pi_\theta}(\rho) - b)$. The following lemma shows that ∇V_σ^L is Lipschitz.

Lemma 3. *∇V_σ^L is Lipschitz in θ and λ.*

Hence Assumption 3 holds for V_σ^L. A natural idea is to use the smoothed robust value functions to replace the ones in the original problem (5):

$$\min_{\lambda\geq 0}\max_{\pi\in\Pi_\Theta} V_{\sigma,r}^\pi(\rho) + \lambda(V_{\sigma,c}^\pi(\rho) - b). \tag{13}$$

As will be shown below in Lemma 6, this approximation can be arbitrarily close to the original problem in (5) as $\sigma\to-\infty$. We first show that under Assumption 2, the following Slater's condition holds for the smoothed problem in (13).

Lemma 4. *Let σ be sufficiently small such that $\|V_{\sigma,c}^\pi - V_c^\pi\| < \zeta$ for any π, then there exists $\zeta'>0$ and a policy $\pi'\in\Pi_\Theta$ s.t. $V_{\sigma,c}^{\pi'}(\rho) - b \geq \zeta'$.*

The following lemma shows that the optimal dual variable for (13) is also bounded.

Lemma 5. *Denote the optimal solution of (13) by (λ^*,π^*). Then $\lambda^*\in\left[0,\frac{2C_\sigma}{\zeta'}\right]$, where C_σ is the upper bound of smoothed robust value functions $V_{\sigma,c}^\pi$.*

Denote by $\Lambda^* = \max\left\{\frac{2C_\sigma}{\zeta'}, \frac{2}{\zeta(1-\gamma)}\right\}$, then problems (6) and (13) are equivalent to the following bounded ones: $\min_{\lambda\in[0,\Lambda^*]}\max_{\pi\in\Pi_\Theta} V_r^\pi(\rho) + \lambda(V_c^\pi(\rho) - b)$, and

$$\min_{\lambda\in[0,\Lambda^*]}\max_{\pi\in\Pi_\Theta} V_{\sigma,r}^\pi(\rho) + \lambda(V_{\sigma,c}^\pi(\rho) - b). \tag{14}$$

The following lemma shows that the two problems are within a gap of $\mathcal{O}(\epsilon)$.

Lemma 6. *Choose a small enough σ such that $\|V_r^\pi - V_{\sigma,r}^\pi\| \le \epsilon$ and $\|V_c^\pi - V_{\sigma,c}^\pi\| \le \epsilon$. Then*

$$\left|\min_{\lambda\in[0,\Lambda^*]}\max_{\pi\in\Pi_\Theta} V_{\sigma,r}^\pi(\rho) + \lambda(V_{\sigma,c}^\pi(\rho) - b) - \min_{\lambda\in[0,\Lambda^*]}\max_{\pi\in\Pi_\Theta} V_r^\pi(\rho) + \lambda(V_c^\pi(\rho) - b)\right| \le (1+\Lambda^*)\,\epsilon.$$

In the following, we hence focus on the smoothed dual problem in (14), which is an accurate approximation of the original problem (6). Denote the gradient mapping of the smoothed Lagrangian function V_σ^L by

$$G_t \triangleq \begin{bmatrix} \beta_t\left(\lambda_t - \prod_{[0,\Lambda^*]}\left(\lambda_t - \frac{1}{\beta_t}\left(\nabla_\lambda V_\sigma^L(\theta_t,\lambda_t)\right)\right)\right) \\ \alpha_t\left(\theta_t - \prod_\Theta\left(\theta_t + \frac{1}{\alpha_t}\left(\nabla_\theta V_\sigma^L(\theta_t,\lambda_t)\right)\right)\right) \end{bmatrix}. \tag{15}$$

Applying our RPD algorithm in (14), we have the following convergence guarantee.

Corollary 1. *If we set step sizes α_t, β_t, and b_t as in Section I and set $T = \mathcal{O}(\epsilon^{-4})$, then $\min_{1\le t\le T}\|G_t\| \le 2\epsilon$.*

This corollary implies that our robust primal-dual algorithm converges to a stationary point of the min-max problem (14) under the δ-contamination model, with a complexity of $\mathcal{O}(\epsilon^{-4})$.

Algorithm 2 Smoothed Robust TD [75]

Input: T_c, π, σ, ω_t
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 $s \in \mathcal{S}$
    $Q_{t+1}(s_t,a_t) \leftarrow Q_t(s_t,a_t) + \alpha_t\big(c_t + \gamma(1-\delta)\cdot V_t(s_{t+1}) + \gamma\delta\cdot \mathrm{LSE}(V_t) - Q_t(s_t,a_t)\big)$
  end for
Output: Q_{T_c}

Note that Algorithm 1 assumes knowledge of the smoothed robust value functions which may not be available in practice. Different from the non-robust value function which can be estimated using Monte Carlo, robust value functions are the value function corresponding to the worst-case transition kernel from which no samples are directly taken. To solve this issue, we adopt the smoothed robust TD algorithm (Algorithm 2) from [75] to estimate the smoothed robust value functions.

It was shown that the smoothed robust TD algorithm converges to the smoothed robust value function with a sample complexity of $\mathcal{O}(\epsilon^{-2})$ [75] under the tabular case. We then construct our online and model-free RPD algorithm as in Algorithm 3. We note that Algorithm 3 is for the tabular setting with finite $\mathcal{S}$ and $\mathcal{A}$. It can be easily extended to the case with large/continuous $\mathcal{S}$ and $\mathcal{A}$ using function approximation.

Algorithm 3 Online Robust Primal-Dual algorithm

Input: T, σ, ϵ_{est}, β_t, α_t, b_t
Initialization: λ_0, θ_0
  for $t = 0, 1, ..., T - 1$ **do**
    Set $T_c = \mathcal{O}\left(\frac{(t+1)^{1.5}}{\epsilon_{\text{est}}^2}\right)$ and run Algorithm 2 for r and c, output $Q_{T_c,r}$, $Q_{T_c,c}$
    $\hat{V}_{\sigma,r}^{\pi_{\theta_t}}(s) \leftarrow \sum_a \pi_{\theta_t}(a|s)Q_{T_c,r}(s,a)$, $\hat{V}_{\sigma,c}^{\pi_{\theta_t}}(s) \leftarrow \sum_a \pi_{\theta_t}(a|s)Q_{T_c,c}(s,a)$
    $\hat{V}_{\sigma,r}^{\pi_{\theta_t}}(\rho) \leftarrow \sum_s \rho(s)\hat{V}_{\sigma,r}^{\pi_{\theta_t}}(s)$, $\hat{V}_{\sigma,c}^{\pi_{\theta_t}}(\rho) \leftarrow \sum_s \rho(s)\hat{V}_{\sigma,c}^{\pi_{\theta_t}}(s)$
    $\lambda_{t+1} \leftarrow \prod_{[0,\Lambda^*]}\left(\lambda_t - \frac{1}{\beta_t}\left(\hat{V}_{\sigma,c}^{\pi_{\theta_t}}(\rho) - b\right) - \frac{b_t}{\beta_t}\lambda_t\right)$
    $\theta_{t+1} \leftarrow \prod_\Theta\left(\theta_t + \frac{1}{\alpha_t}\left(\nabla_\theta \hat{V}_{\sigma,r}^{\pi_{\theta_t}}(\rho) + \lambda_{t+1}\nabla_\theta \hat{V}_{\sigma,c}^{\pi_{\theta_t}}(\rho)\right)\right)$
  end for
Output: θ_T

Algorithm 3 can be viewed as a biased stochastic gradient descent-ascent algorithm. It is a sample-based algorithm without assuming any knowledge of robust value functions, and can be performed in an online fashion. We further extend the convergence results in Theorem 1 to the model-free setting, and characterize the following finite-time error bound of Algorithm 3. Similarly, Algorithm 3 can be shown to achieve a 2ϵ-feasible policy almost surely.

Theorem 2. *Consider the same conditions as in Theorem 1. Let $\epsilon_{est} = \mathcal{O}(\epsilon^2)$ and $T = \mathcal{O}(\epsilon^{-4})$, then* $\min_{1\leq t\leq T} \|G_t\| \leq (1+\sqrt{2})\epsilon$.

5 Numerical Results

In this section, we numerically demonstrate the robustness of our algorithm in terms of both maximizing robust reward value function and satisfying constraints under model uncertainty. We compare our RPD algorithm with the heuristic algorithms in [65, 44] and the vanilla non-robust primal-dual method. Based on the idea of "robust policy evaluation" + "non-robust policy improvement" in [65, 44], we combine the robust TD algorithm 2 with non-robust vanilla policy gradient method [68], which we refer to as the heuristic primal-dual algorithm. Several environments, including Garnet [6], 8×8 Frozen-Lake, Taxi and N-chain environments from OpenAI [13], are investigated.

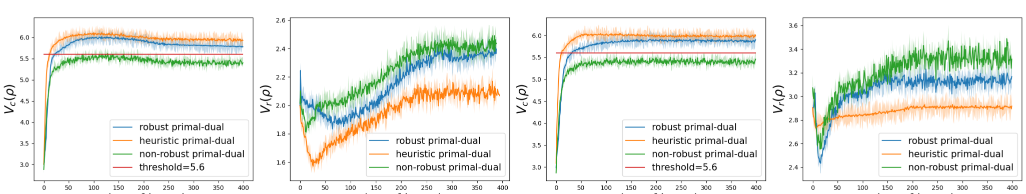
We first run the algorithm and store the obtained policies π_t at each time step. At each time step, we run robust TD with a sample size 200 for 30 times to estimate the objective $V_r(\rho)$ and the constraint $V_c(\rho)$. We then plot them v.s. the number of iterations t. The upper and lower envelopes of the curves correspond to the 95 and 5 percentiles of the 30 curves, respectively. We repeat the experiment for two different values of $\delta = 0.2, 0.3$.

Garnet problem. A 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, and receives a randomly generated reward/utility signal. The transition kernels are also randomly generated. The comparison results are shown in Fig.1.

8×8 **Frozen-Lake problem.** We then compare the three algorithms under the 8×8 Frozen-lake problem setting in Fig.2. The Frozen-Lake problem involves a frozen lake of size 8×8 which contains several "holes". The agent aims to cross the lake from the start point to the end point without falling into any holes. The agent receives $r = -10$ when falling in a hole, receives $r = 20$ when arrive at the end point, and receives $r = 0$ at other times.

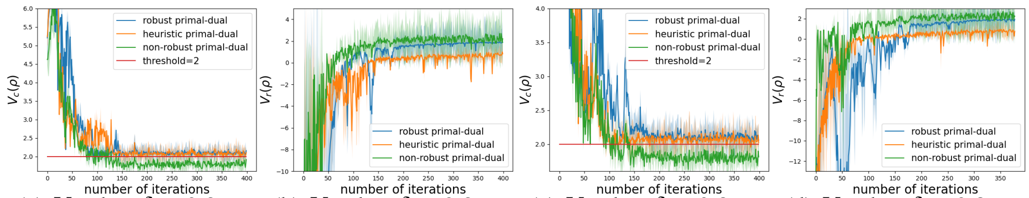
Taxi problem. We then compare the three algorithms under the Taxi problem environment. The taxi problem simulates a taxi driver in a 5×5 map. There are four designated locations in the grid world and a passenger occurs at a random location of the designated four locations at the start of each episode. The goal of the driver is to first pick up the passenger and then to drop off at another specific location. The driver receives $r = 20$ for each successful drop-off, and always receives $r = -1$ at other times. We randomly generate utility signal for each state-action pair. The results are shown in Fig.3.

N**-Chain problem.** We then compare three algorithms under the N-Chain problem environment. The N-chain problem involves a chain contains N nodes. The agent can either move to its left or right node. When it goes to left, it receives a reward-utility signal $(1, 0)$; When it goes right, it receives a reward-utility signal $(0, 2)$, and if the agent arrives the N-th node, it receives a bonus reward of 40. There is also a small probability that the agent slips to the different direction of its action. In this experiment, we set $N = 40$. The results are shown in Fig.4.



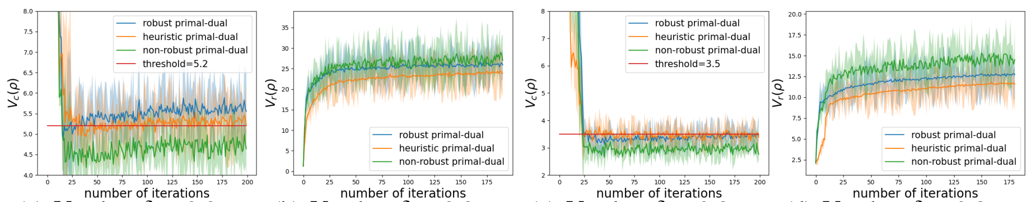
(a) V_c when $\delta = 0.2$. (b) V_r when $\delta = 0.2$. (c) V_c when $\delta = 0.3$. (d) V_r when $\delta = 0.3$.

Figure 1: Comparison on Garnet Problem $\mathcal{G}(20, 10)$.



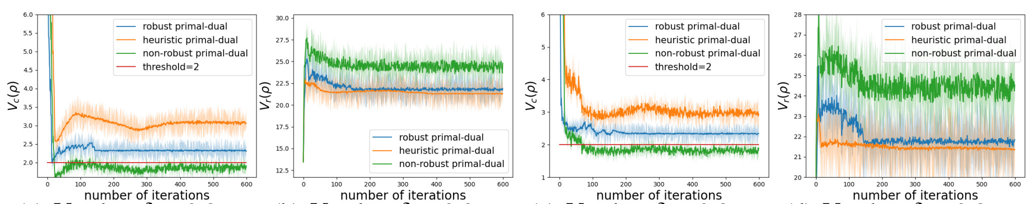
(a) V_c when $\delta = 0.2$. (b) V_r when $\delta = 0.2$. (c) V_c when $\delta = 0.3$. (d) V_r when $\delta = 0.3$.

Figure 2: Comparison on 8×8 Frozen-Lake Problem.



(a) V_c when $\delta = 0.2$. (b) V_r when $\delta = 0.2$. (c) V_c when $\delta = 0.3$. (d) V_r when $\delta = 0.3$.

Figure 3: Comparison on Taxi Problem.



(a) V_c when $\delta = 0.2$. (b) V_r when $\delta = 0.2$. (c) V_c when $\delta = 0.3$. (d) V_r when $\delta = 0.3$.

Figure 4: Comparison on N-Chain Problem.

It can be seen that both our RPD algorithm and the heuristic primal-dual approach find feasible policies satisfying the constraint under the worst-case scenario, i.e., $V_c^\pi \geq b$. However, the non-robust primal-dual method fails to find a feasible solution that satisfy the constraint under the worst-case scenario. In terms of reward, it can be seen that the non-robust primal-dual achieves the largest accumulative reward. This is because the policy it finds violates the robust constraint. Although both our RPD algorithm and the heuristic primal-dual algorithm find feasible solutions, our RPD algorithm achieves a higher reward than the heuristic primal-dual algorithm. Thus the experiments verify that among the three algorithms, our RPD algorithm is the best that it optimizes the worst-case reward performance while satisfying the robust constraint on the utility.

6 Conclusion

In this paper, we formulate the problem of robust constrained reinforcement learning under model uncertainty, where the goal is to guarantee that constraints are satisfied for all MDPs in the uncertainty set, and to maximize the worst-case reward performance over the uncertainty set. We propose a robust primal-dual algorithm, and theoretically characterize its convergence, complexity and robust feasibility. Our algorithm guarantees convergence to a feasible solution, and outperforms the other two heuristic algorithms. We further investigate a concrete example with δ-contamination uncertainty set, and construct online and model-free robust primal-dual algorithm. **Limitations:** It is of future interest to generalize our results to other types of uncertainty sets, e.g., ones defined by KL divergence, total variation, Wasserstein distance. **Negative societal impact:** This work is a theoretical study. To the best of the authors' knowledge, it does not have any potential negative impact on the society.

References

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

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

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

[4] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu. Safe reinforcement learning via shielding. In *Proc. Conference on Artificial Intelligence (AAAI)*, volume 32, 2018.

[5] E. Altman. *Constrained Markov decision processes: stochastic modeling*. Routledge, 1999.

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

[7] P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, volume 21, 2008.

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

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

[10] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause. Safe model-based reinforcement learning with stability guarantees. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, volume 30, 2017.

[11] D. P. Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

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

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

[14] R. Cheng, G. Orosz, R. M. Murray, and J. W. Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proc. Conference on Artificial Intelligence (AAAI)*, volume 33, pages 3387–3395, 2019.

[15] Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

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

[17] G. Dalal, K. Dvijotham, M. Vecerik, T. Hester, C. Paduraru, and Y. Tassa. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*, 2018.

[18] D. Ding, X. Wei, Z. Yang, Z. Wang, and M. Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *Proc. International Conference on Artifical Intelligence and Statistics (AISTATS)*, pages 3304–3312. PMLR, 2021.

[19] D. Ding, K. Zhang, T. Basar, and M. Jovanovic. Natural policy gradient primal-dual method for constrained Markov decision processes. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 8378–8390, 2020.

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

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

[22] Y. Efroni, S. Mannor, and M. Pirotta. Exploration-exploitation in constrained MDPs. *arXiv preprint arXiv:2003.02189*, 2020.

[23] J. F. Fisac, A. K. Akametalu, M. N. Zeilinger, S. Kaynama, J. Gillula, and C. J. Tomlin. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 64(7):2737–2752, 2018.

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

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

[26] C. P. Ho, M. Petrik, and W. Wiesemann. Partial policy iteration for l1-robust Markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46, 2021.

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

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

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

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

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

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

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

[34] T. Li, Z. Guan, S. Zou, T. Xu, Y. Liang, and G. Lan. Faster algorithm and sharper analysis for constrained Markov decision process. *arXiv preprint arXiv:2110.10351*, 2021.

[35] Q. Liang, F. Que, and E. Modiano. Accelerated primal-dual policy optimization for safe reinforcement learning. *arXiv preprint arXiv:1802.06480*, 2018.

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

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

[38] T. Lin, C. Jin, and M. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *Proc. International Conference on Machine Learning (ICML)*, pages 6083–6093. PMLR, 2020.

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

[40] T. Liu, R. Zhou, D. Kalathil, P. Kumar, and C. Tian. Fast global convergence of policy optimization for constrained MDPs. *arXiv preprint arXiv:2111.00552*, 2021.

[41] Y. Liu, J. Ding, and X. Liu. Ipo: Interior-point policy optimization under constraints. In *Proc. Conference on Artificial Intelligence (AAAI)*, volume 34, pages 4940–4947, 2020.

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

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

[44] D. J. Mankowitz, D. A. Calian, R. Jeong, C. Paduraru, N. Heess, S. Dathathri, M. Riedmiller, and T. Mann. Robust constrained reinforcement learning for continuous control with model misspecification. *arXiv preprint arXiv:2010.10644*, 2020.

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

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

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

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

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

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

[51] M. Ono, M. Pavone, Y. Kuwata, and J. Balaram. Chance-constrained dynamic programming with application to risk-aware robotic space exploration. *Autonomous Robots*, 39(4):555–571, 2015.

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

[53] S. Paternain, M. Calvo-Fullana, L. F. Chamon, and A. Ribeiro. Safe policies for reinforcement learning via primal-dual methods. *IEEE Transactions on Automatic Control*, 2022.

[54] S. Paternain, L. Chamon, M. Calvo-Fullana, and A. Ribeiro. Constrained reinforcement learning has zero duality gap. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

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

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

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

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

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

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

[61] Z. Qin, K. Zhang, Y. Chen, J. Chen, and C. Fan. Learning safe multi-agent control with decentralized neural barrier certificates. *arXiv preprint arXiv:2101.05436*, 2021.

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

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

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

[65] R. H. Russel, M. Benosman, J. Van Baar, and R. Corcodel. Lyapunov robust constrained-MDPs: Soft-constrained robustly stable policy optimization under model uncertainty. *arXiv preprint arXiv:2108.02701*, 2021.

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

[67] A. Stooke, J. Achiam, and P. Abbeel. Responsive safety in reinforcement learning by pid lagrangian methods. In *Proc. International Conference on Machine Learning (ICML)*, pages 9133–9143. PMLR, 2020.

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

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

[70] C. Tessler, D. J. Mankowitz, and S. Mannor. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*, 2018.

[71] G. Thomas, Y. Luo, and T. Ma. Safe reinforcement learning by imagining the near future. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

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

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

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

[75] Y. Wang and S. Zou. Policy gradient method for robust reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, 2022.

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

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

[78] T. Xu, Y. Liang, and G. Lan. Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *Proc. International Conference on Machine Learning (ICML)*, pages 11480–11491. PMLR, 2021.

[79] Z. Xu, H. Zhang, Y. Xu, and G. Lan. A unified single-loop alternating gradient projection algorithm for nonconvex-concave and convex-nonconcave minimax problems. *arXiv preprint arXiv:2006.02032*, 2020.

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

[81] D. Ying, Y. Ding, and J. Lavaei. A dual approach to constrained Markov decision processes with entropy regularization. *arXiv preprint arXiv:2110.08923*, 2021.

[82] M. Yu, Z. Yang, M. Kolar, and Z. Wang. Convergent policy optimization for safe reinforcement learning. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

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

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

[85] L. Zheng and L. Ratliff. Constrained upper confidence reinforcement learning. In *Learning for Dynamics and Control*, pages 620–629. PMLR, 2020.

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

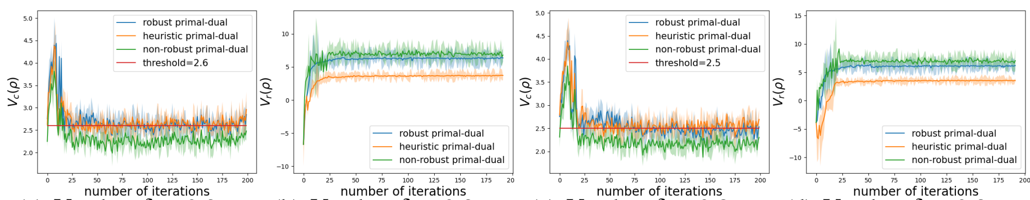
Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

A Additional Experiments on 4×4 Frozen Lake

The 4×4 frozen lake is similar to the 8×8 one but with a smaller map. Similarly, we randomly generate the utility signal for each state-action pair. The results are shown in Fig.5.



(a) V_c when $\delta = 0.2$. (b) V_r when $\delta = 0.2$. (c) V_c when $\delta = 0.3$. (d) V_r when $\delta = 0.3$.

Figure 5: Comparison on 4×4 Frozen-Lake Problem.

B Proof of Lemma 1

Denote by $\mathsf{P}^\pi = \{(p^\pi)^a_s \in \Delta_{\mathcal{S}} : s \in \mathcal{S}, a \in \mathcal{A}\}$ the worst-case transition kernel corresponding to the policy π. We consider the δ-contamination uncertainty set defined in Section 4. We then show that under δ-contamination model, the set of visitation distributions is non-convex. The robust visitation distribution set can be written as follows:

$$\left\{ d \in \Delta_{\mathcal{S}\times\mathcal{A}} : \exists \pi \in \Pi, \text{ s.t. } \forall (s,a), \begin{cases} d(s,a) = \pi(a|s) \sum_b d(s,b), \\ \gamma \sum_{s',a'} (p^\pi)^{a'}_{s',s} d(s',a') + (1-\gamma)\rho(s) = \sum_a d(s,a). \end{cases} \right\}. \tag{16}$$

Under the δ-contamination model, P^π can be explicated as $(p^\pi)^a_{s,s'} = (1-\delta)p^a_{s,s'} + \delta \mathbb{1}_{\{s' = \arg\min V^\pi\}}$. Hence the set in (16) can be rewritten as

$$\left\{ d \in \Delta_{\mathcal{S}\times\mathcal{A}} : \exists \pi, \text{ s.t. } \forall (s,a), \begin{cases} d(s,a) = \pi(a|s) \left(\sum_b d(s,b)\right), \\ \gamma(1-\delta) \sum_{s',a'} p^{a'}_{s',s} d(s',a') + \gamma\delta \mathbb{1}_{\{s = \arg\min V^\pi\}} \\ \qquad + (1-\gamma)\rho(s) = \sum_a d(s,a). \end{cases} \right\}. \tag{17}$$

Now consider any two pairs $(\pi_1, d_1), (\pi_2, d_2)$ of policy and their worst-case visitation distribution, to show that the set is convex, we need to find a pair (π', d') such that $\forall \lambda \in [0,1]$ and $\forall s, a$,

$$\lambda d_1(s,a) + (1-\lambda) d_2(s,a) = d'(s,a), \tag{18}$$

$$d'(s,a) = \pi'(a|s) \left(\sum_b d'(s,b)\right), \tag{19}$$

$$\sum_{a'} d'(s,a') = \gamma(1-\delta) \sum_{s',a'} p^{a'}_{s',s} d'(s',a') + \gamma\delta \mathbb{1}_{\{s = \arg\min V^{\pi'}\}} + (1-\gamma)\rho(s). \tag{20}$$

(20) firstly implies that $\forall s$,

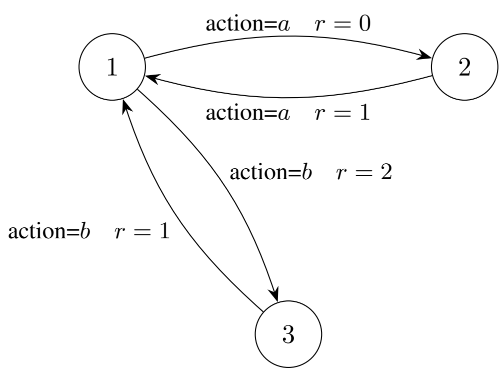
$$\lambda \mathbb{1}_{\{s = \arg\min V^{\pi_1}\}} + (1-\lambda) \mathbb{1}_{\{s = \arg\min V^{\pi_2}\}} = \mathbb{1}_{\{s = \arg\min V^{\pi'}\}}, \tag{21}$$

where from (18) and (19), π' should be

$$\pi'(a|s) = \frac{d'(s,a)}{\sum_b d'(s,b)} = \frac{\lambda d_1(s,a) + (1-\lambda) d_2(s,a)}{\sum_b (\lambda d_1(s,b) + (1-\lambda) d_2(s,b))}. \tag{22}$$

We then construct the following counterexample, which shows that there exists a robust MDP, two policy-distribution pairs $(\pi_1, d_1), (\pi_2, d_2)$, and $\lambda \in (0,1)$, such that $\lambda \mathbb{1}_{\{s=\arg\min V^{\pi_1}\}} + (1-\lambda)\mathbb{1}_{\{s=\arg\min V^{\pi_2}\}} \neq \mathbb{1}_{\{s=\arg\min V^{\pi'}\}}$, and therefore the set of robust visitation distribution is non-convex.

Consider the following Robust MDP. It has three states $1, 2, 3$ and two actions a, b. When the agent is at state 1, if it takes action a, the system will transit to state 2 and receive reward $r = 0$; if it takes action b, the system will transit to state 3 and receive reward $r = 2$. When the agent is at state $2/3$, it can only take action a/b, the system can only transits back to state 1 and the agent will receive reward $r = 1$. The initial distribution is $\mathbb{1}_{s=1}$.



Clearly all policy can be written as $\pi = (p, 1-p)$, where p is the probability of taking action a at state 1. We consider two policies, $\pi_1 = (1, 0)$ and $\pi_2 = (0, 1)$.

It can be verified that $\arg\min V^{\pi_1} = 1$, and its robust visitation distribution, denoted by d_1, is

$$d_1(1, a) = \frac{1-\gamma}{1-\gamma^2}, \tag{23}$$
$$d_1(1, b) = 0, \tag{24}$$
$$d_1(2, a) = \frac{\gamma(1-\gamma)}{1-\gamma^2}, \tag{25}$$
$$d_1(2, b) = 0, \tag{26}$$
$$d_1(3, a) = 0, \tag{27}$$
$$d_1(3, b) = 0. \tag{28}$$

Similarly, $\arg\min V^{\pi_2} = 2$, and and its robust visitation distribution, denoted by d_2, is

$$d_2(1, a) = 0, \tag{29}$$
$$d_2(1, b) = \frac{1-\gamma}{1-\gamma^2}, \tag{30}$$
$$d_2(2, a) = 0, \tag{31}$$
$$d_2(2, b) = 0, \tag{32}$$
$$d_2(3, a) = 0, \tag{33}$$
$$d_2(3, b) = \frac{\gamma(1-\gamma)}{1-\gamma^2}. \tag{34}$$

Hence according to (22), π' should be as follows:

$$\pi'(a|1) = \lambda, \pi'(b|1) = 1-\lambda, \pi'(a|2) = 1, \pi'(b|3) = 1. \tag{35}$$

We then show that there exists $\lambda \in [0, 1]$, such that $\lambda \mathbb{1}_{\{s=1\}} + (1-\lambda)\mathbb{1}_{\{s=2\}} \neq \mathbb{1}_{\{\arg\min V^{\pi'}\}}$.

Clearly (21) holds only if $V^{\pi'}(1) = V^{\pi'}(2) = \min_s V^{\pi'}(s)$. However, according to the Bellman equations for π', we have that

$$V^{\pi'}(1) = \lambda(\gamma(1-\delta)V^{\pi'}(2) + \gamma\delta \min V^{\pi'}) + (1-\lambda)(2 + \gamma(1-\delta)V^{\pi'}(3) + \gamma\delta \min V^{\pi'}), \tag{36}$$

$$V^{\pi'}(2) = 1 + \gamma(1-\delta)V^{\pi'}(1) + \gamma\delta \min V^{\pi'}, \tag{37}$$
$$V^{\pi'}(3) = 1 + \gamma(1-\delta)V^{\pi'}(1) + \gamma\delta \min V^{\pi'}. \tag{38}$$

If we set $\lambda = \frac{1}{3}$,

$$V^{\pi'}(1) = \frac{4}{3} + \gamma\delta \min V^{\pi'} + \gamma(1-\delta)V^{\pi'}(2), \tag{39}$$
$$V^{\pi'}(2) = 1 + \gamma\delta \min V^{\pi'} + \gamma(1-\delta)V^{\pi'}(1). \tag{40}$$

Clearly, $V^{\pi'}(1) \neq V^{\pi'}(2)$, and hence $\lambda \mathbb{1}_{\{\arg\min V^1\}} + (1-\lambda)\mathbb{1}_{\{\arg\min V^2\}} \neq \mathbb{1}_{\left\{\arg\min V^{\pi'}\right\}}$.

C Proof of Lemmas 2 and 5

Proof of Lemma 2

Proof. We first set $C = V_r^{\pi^*}(\rho) + \lambda^*(V_c^{\pi^*}(\rho) - b)$, clearly $\max_{\pi\in\Pi} V_r^\pi(\rho) + \lambda^*(V_c^\pi(\rho) - b) = C$, and hence

$$C = \max_{\pi\in\Pi} V_r^\pi(\rho) + \lambda^*(V_c^\pi(\rho) - b) \geq V_r^{\pi^\zeta}(\rho) + \lambda^*(V_c^{\pi^\zeta}(\rho) - b) \geq V_r^{\pi^\zeta}(\rho) + \lambda^*\zeta. \tag{41}$$

Thus we have that

$$\lambda^* \leq \frac{C - V_r^{\pi^\zeta}(\rho)}{\zeta}. \tag{42}$$

Note that

$$C = \min_{\lambda\geq 0}\max_{\pi\in\Pi} V_r^\pi(\rho) + \lambda(V_c^\pi(\rho) - b) \overset{(a)}{\leq} \max_{\pi\in\Pi} V_r^\pi(\rho) \leq \frac{1}{1-\gamma}, \tag{43}$$

where (a) is because $\min_{\lambda\geq 0}\max_{\pi\in\Pi} V_r^\pi(\rho) + \lambda(V_c^\pi(\rho) - b)$ is less than the optimal value of inner problem when $\lambda = 0$, i.e., $\max_{\pi\in\Pi} V_r^\pi(\rho)$, and $\frac{1}{1-\gamma}$ is the upper bound of robust value functions. Hence we have that

$$\lambda^* \leq \frac{1}{(1-\gamma)\zeta}, \tag{44}$$

which completes the proof. □

Proof of Lemma 5

Proof. Set $C = V_{\sigma,r}^{\pi^*}(\rho) + \lambda^*(V_{\sigma,c}^{\pi^*}(\rho) - b)$, then

$$C = \max_{\pi\in\Pi} V_{\sigma,r}^\pi(\rho) + \lambda^*(V_{\sigma,c}^\pi(\rho) - b) \geq V_{\sigma,r}^{\pi^{\zeta'}}(\rho) + \lambda^*(V_{\sigma,c}^{\pi^{\zeta'}}(\rho) - b) \geq V_{\sigma,r}^{\pi^{\zeta'}}(\rho) + \lambda^*\zeta'. \tag{45}$$

Thus we have that

$$C \geq V_{\sigma,r}^{\pi^\zeta}(\rho) + \lambda^*\zeta', \tag{46}$$

hence

$$\lambda^* \leq \frac{C - V_{\sigma,r}^{\pi^\zeta}(\rho)}{\zeta'}. \tag{47}$$

Note that

$$C = \min_{\lambda\geq 0}\max_{\pi\in\Pi} V_{\sigma,r}^\pi(\rho) + \lambda(V_{\sigma,c}^\pi(\rho) - b) \leq \max_{\pi\in\Pi} V_{\sigma,r}^\pi(\rho) \leq C_\sigma, \tag{48}$$

where C_σ is the upper bound of smoothed robust value functions [75]: $C_\sigma = \frac{1}{1-\gamma}(1 + 2\gamma R\frac{\log|\mathcal{S}|}{\sigma})$. Hence we have that

$$\lambda^* \leq \frac{C_\sigma}{\zeta'}, \tag{49}$$

which completes the proof. □

D Proof of Lemma 6

Proof. For any λ, denote the optimal value of the inner problems $\max_{\pi\in\Pi_\Theta} V^{\pi}_{\sigma,r}(\rho)+\lambda(V^{\pi}_{\sigma,c}(\rho)-b)$ and $\max_{\pi\in\Pi_\Theta} V^{\pi}_{r}(\rho)+\lambda(V^{\pi}_{c}(\rho)-b)$ by $V^D(\lambda)$ and $V^D_\sigma(\lambda)$. It is then easy to verify that

$$|V^D(\lambda)-V^D_\sigma(\lambda)|\leq(1+\lambda)\epsilon\leq(1+\Lambda^*)\epsilon. \tag{50}$$

Denote the optimal solutions of $\min_{\lambda\in[0,\Lambda^*]} V^D(\lambda)$ and $\min_{\lambda\in[0,\Lambda^*]} V^D_\sigma(\lambda)$ by λ^D and λ^D_σ. We thus conclude that $|V^D_\sigma(\lambda^D_\sigma)-V^D(\lambda^D)|\leq(1+\Lambda^*)\,\epsilon$, and this thus completes the proof. □

E Proof of Theorem 1

We restate Theorem 1 with all the specific step sizes as follows.

Set $b_t=\frac{19}{20\xi t^{0.25}}$, $\mu_t=\xi(C^V_\sigma)^2+\frac{16\tau(C^V_\sigma)^2}{\xi(b_{t+1})^2}-2\nu$, $\beta_t=\frac{1}{\xi}$, $\alpha_t=\nu+\mu_t$, where $\xi>\frac{2\nu+(1+\Lambda^*)L_\sigma}{(C^V_\sigma)^2}$, ν is any positive number and τ is any number greater than 2, then

$$\min_{1\leq t\leq T}\|G_t\|^2\leq 2\epsilon, \tag{51}$$

when

$$T=\max\left\{\frac{7(\Lambda^*)^4}{\xi^4\epsilon^4},\left(2+\frac{9\xi(\tau-2)(C^V_\sigma)^2uK}{\epsilon^2}\right)^2\right\}=\mathcal{O}(\epsilon^{-4}). \tag{52}$$

The definitions of u, K can be found in Section I.

Theorem 1 can be proved similarly as Theorem 2, and hence the proof is omitted here.

F Proof of Lemma 3

Proof. Recall that $V^L_\sigma(\theta,\lambda)=V^{\pi_\theta}_{\sigma,r}(\rho)+\lambda(V^{\pi_\theta}_{\sigma,c}(\rho)-b)$, hence we have that

$$\nabla_\lambda V^L_\sigma(\theta,\lambda)=V^{\pi_\theta}_{\sigma,c}(\rho)-b, \tag{53}$$

$$\nabla_\theta V^L_\sigma(\theta,\lambda)=\nabla_\theta V^{\pi_\theta}_{\sigma,r}(\rho)+\lambda\nabla_\theta V^{\pi_\theta}_{\sigma,c}(\rho). \tag{54}$$

Note that in [75], it has been shown that

$$\|V^{\pi_{\theta_1}}_{\sigma,r}-V^{\pi_{\theta_2}}_{\sigma,r}\|\leq C^V_\sigma\|\theta_1-\theta_2\|, \tag{55}$$

$$\|\nabla_\theta V^{\pi_{\theta_1}}_{\sigma,r}-\nabla_\theta V^{\pi_{\theta_2}}_{\sigma,r}\|\leq L_\sigma\|\theta_1-\theta_2\|, \tag{56}$$

where the definition of constants C^V_σ and L_σ can be found in Section I. Hence

$$\|\nabla_\lambda V^L_\sigma(\theta,\lambda)|_{\theta_1}-\nabla_\lambda V^L_\sigma(\theta,\lambda)|_{\theta_2}\|=\|V^{\pi_{\theta_1}}_{\sigma,c}(\rho)-V^{\pi_{\theta_2}}_{\sigma,c}(\rho)\|\leq C^V_\sigma\|\theta_1-\theta_2\|, \tag{57}$$

$$\|\nabla_\lambda V^L_\sigma(\theta,\lambda)|_{\lambda_1}-\nabla_\lambda V^L_\sigma(\theta,\lambda)|_{\lambda_2}\|=0. \tag{58}$$

Similarly, we have that

$$\|\nabla_\theta V^L_\sigma(\theta,\lambda)|_{\theta_1}-\nabla_\theta V^L_\sigma(\theta,\lambda)|_{\theta_2}\|\leq(1+\lambda)L_\sigma\|\theta_1-\theta_2\|\leq(1+\Lambda^*)L_\sigma\|\theta_1-\theta_2\|, \tag{59}$$

$$\|\nabla_\theta V^L_\sigma(\theta,\lambda)|_{\lambda_1}-\nabla_\theta V^L_\sigma(\theta,\lambda)|_{\lambda_2}\|\leq|(\lambda_1-\lambda_2)|\max_{\theta\in\Theta}\|\nabla_\theta V^{\pi_\theta}_{\sigma,c}(\rho)\|\leq C^V_\sigma|\lambda_1-\lambda_2|. \tag{60}$$

This completes the proof. □

G Proof of Proposition 1

Proof. The λ-entry of G_W is smaller than 2ϵ, i.e.,

$$|(G_W)_\lambda|=\left|\beta_W\left(\lambda_W-\prod\nolimits_{[0,\Lambda^*]}\left(\lambda_W-\frac{1}{\beta_W}\left(\nabla_\lambda V^L_\sigma(\theta_W,\lambda_W)\right)\right)\right)\right|<2\epsilon. \tag{61}$$

Denote $\lambda^+ \triangleq \prod_{[0,\Lambda^*]}\left(\lambda_W - \frac{1}{\beta_W}\left(\nabla_\lambda V_\sigma^L(\theta_W, \lambda_W)\right)\right)$. From Lemma 3 in [24], $-\nabla_\lambda V_\sigma^L(\theta_W, \lambda^+)$ can be rewritten as the sum of two parts: $-\nabla_\lambda V_\sigma^L(\theta_W, \lambda^+) \in N_{[0,\Lambda^*]}(\lambda^+) + 4\epsilon B$, where $N_K(x) \triangleq \left\{g \in \mathbb{R}^d : \langle g, y - x\rangle \leq 0 : \forall y \in K\right\}$ is the normal cone, and B is the unit ball.

This hence implies that for any $\lambda \in [0, \Lambda^*]$, $(\lambda - \lambda^+)(V_c^W - b) \geq -4(\lambda - \lambda^+)\epsilon$. By setting $\lambda = \Lambda^*$, we have $V_c^W + 4\epsilon \geq b$, which means π_W is feasible with a 4ϵ-violation. □

H Proof of Theorem 2

We then prove Theorem 2. Our proof extends the one in [79] to the biased setting.

To simplify notations, we denote the updates in Algorithm 3 by $\hat{f}(\theta_t) \triangleq \hat{V}_{\sigma,c}^{\pi_{\theta_t}}(\rho) - b$, and $\hat{g}(\theta_t, \lambda_{t+1}) \triangleq \nabla_\theta \hat{V}_{\sigma,r}^{\pi_{\theta_t}}(\rho) + \lambda_{t+1}\nabla_\theta \hat{V}_{\sigma,c}^{\pi_{\theta_t}}(\rho)$, and denote the update functions in Algorithm 1 by $f(\theta_t) \triangleq V_{\sigma,c}^{\pi_{\theta_t}}(\rho) - b$, and $g(\theta_t, \lambda_{t+1}) \triangleq \nabla_\theta V_{\sigma,r}^{\pi_{\theta_t}}(\rho) + \lambda_{t+1}\nabla_\theta V_{\sigma,c}^{\pi_{\theta_t}}(\rho)$. Here $\hat{f}$ and $\hat{g}$ can be viewed as biased estimations of f and g.

In the following, we will first show several technical lemmas that will be useful in the proof of Theorem 2.

Lemma 7. *Recall that the step size $\alpha_t = \nu + \mu_t$. If $\mu_t > (1 + \Lambda^*)L_\sigma$, $\forall t \geq 0$, then*

$$\begin{aligned} V_\sigma^L(\theta_{t+1}, \lambda_{t+1}) - V_\sigma^L(\theta_t, \lambda_{t+1}) \geq & \langle \theta_{t+1} - \theta_t, -\hat{g}(\theta_t, \lambda_{t+1}) + g(\theta_t, \lambda_{t+1})\rangle \\ & + \left(\frac{\mu_t}{2} + \nu\right)\|\theta_{t+1} - \theta_t\|^2. \end{aligned} \tag{62}$$

Proof. Note that from the update of θ_t and proposition of projection, it implies that

$$\left\langle \theta_t + \frac{1}{\alpha_t}\hat{g}(\theta_t, \lambda_{t+1}) - \theta_{t+1}, \theta_t - \theta_{t+1} \right\rangle \leq 0. \tag{63}$$

Hence

$$\langle \hat{g}(\theta_t, \lambda_{t+1}) - \alpha_t(\theta_{t+1} - \theta_t), \theta_t - \theta_{t+1}\rangle \leq 0. \tag{64}$$

From Lemma 3, we have that

$$V_\sigma^L(\theta_{t+1}, \lambda_{t+1}) - V_\sigma^L(\theta_t, \lambda_{t+1}) \geq \langle \theta_{t+1} - \theta_t, g(\theta_t, \lambda_{t+1})\rangle - \frac{(1 + \Lambda^*)L_\sigma}{2}\|\theta_{t+1} - \theta_t\|^2. \tag{65}$$

Summing up the two inequalities implies

$$\begin{aligned} & V_\sigma^L(\theta_{t+1}, \lambda_{t+1}) - V_\sigma^L(\theta_t, \lambda_{t+1}) \\ & \geq \langle \theta_{t+1} - \theta_t, -\hat{g}(\theta_t, \lambda_{t+1}) + g(\theta_t, \lambda_{t+1}) + \alpha_t(\theta_{t+1} - \theta_t)\rangle - \frac{(1 + \Lambda^*)L_\sigma}{2}\|\theta_{t+1} - \theta_t\|^2 \\ & \geq \langle \theta_{t+1} - \theta_t, -\hat{g}(\theta_t, \lambda_{t+1}) + g(\theta_t, \lambda_{t+1})\rangle + \left(\alpha_t - \frac{L_\sigma(1 + \Lambda^*)}{2}\right)\|\theta_{t+1} - \theta_t\|^2 \\ & \geq \langle \theta_{t+1} - \theta_t, -\hat{g}(\theta_t, \lambda_{t+1}) + g(\theta_t, \lambda_{t+1})\rangle + \left(\frac{\mu_t}{2} + \nu\right)\|\theta_{t+1} - \theta_t\|^2, \end{aligned} \tag{66}$$

and hence completes the proof. □

Lemma 8. *Recall that the step size $\beta_t = \frac{1}{\xi}$, and set $\xi \leq \frac{1}{b_0}$, then*

$$\begin{aligned} & V_\sigma^L(\theta_{t+1}, \lambda_{t+1}) - V_\sigma^L(\theta_t, \lambda_t) \\ & \geq (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}))(\lambda_{t+1} - \lambda_t) + \langle \theta_{t+1} - \theta_t, -\hat{g}(\theta_t, \lambda_{t+1}) + g(\theta_t, \lambda_{t+1})\rangle - \frac{\xi(C_\sigma^V)^2}{2}\|\theta_t - \theta_{t-1}\|^2 \\ & \quad + \left(\frac{\mu_t}{2} + \nu\right)\|\theta_{t+1} - \theta_t\|^2 + \frac{b_{t-1}}{2}(\lambda_t^2 - \lambda_{t+1}^2) - \frac{1}{\xi}(\lambda_{t+1} - \lambda_t)^2 - \frac{1}{2\xi}(\lambda_t - \lambda_{t-1})^2. \end{aligned} \tag{67}$$

Proof. For any $t > 1$, define $\tilde{V}_t(\theta, \lambda) \triangleq V_\sigma^L(\theta, \lambda) + \frac{b_{t-1}}{2}\lambda^2$. Thus we have

$$|\nabla_\lambda \tilde{V}_t(\theta_t, \lambda_{t+1}) - \nabla_\lambda \tilde{V}_t(\theta_t, \lambda_t)| = b_{t-1}|\lambda_{t+1} - \lambda_t| \leq b_0|\lambda_{t+1} - \lambda_t|, \tag{68}$$

where that last inequality is due to $b_{t-1} \leq b_0$. Note that $\tilde{V}_t(\theta, \lambda)$ is b_{t-1}-strongly convex in λ, hence we have

$$
\begin{aligned}
&(\nabla_\lambda \tilde{V}_t(\theta, \lambda_{t+1}) - \nabla_\lambda \tilde{V}_t(\theta, \lambda_t))(\lambda_{t+1} - \lambda_t) \\
&\geq b_{t-1}(\lambda_{t+1} - \lambda_t)^2 \\
&\geq b_{t-1} \left(\frac{b_{t-1} + b_0}{b_{t-1} + b_0} \right) (\lambda_{t+1} - \lambda_t)^2 \\
&= \frac{b_{t-1} b_0}{b_{t-1} + b_0} (\lambda_{t+1} - \lambda_t)^2 + \frac{b_{t-1}^2}{b_{t-1} + b_0} (\lambda_{t+1} - \lambda_t)^2 \\
&\geq \frac{b_{t-1} b_0}{b_{t-1} + b_0} (\lambda_{t+1} - \lambda_t)^2 + \frac{1}{b_{t-1} + b_0} (\nabla_\lambda \tilde{V}_t(\theta_t, \lambda_{t+1}) - \nabla_\lambda \tilde{V}_t(\theta_t, \lambda_t))^2,
\end{aligned} \tag{69}
$$

where the last inequality is from (68).

Recall the update of λ_t in Algorithm 3 which can be rewritten as

$$
\lambda_{t+1} = \prod\nolimits_{[0, \Lambda^*]} \left(\lambda_t - \frac{1}{\beta_t} \nabla_\lambda \tilde{V}_{t+1}(\theta_t, \lambda_t) + \frac{1}{\beta_t} (f(\theta_t) - \hat{f}(\theta_t)) \right), \tag{70}
$$

This further implies that $\forall \lambda \in [0, \Lambda^*]$:

$$
(\beta_t(\lambda_{t+1} - \lambda_t) + \nabla_\lambda \tilde{V}_{t+1}(\theta_t, \lambda_t) - f(\theta_t) + \hat{f}(\theta_t))(\lambda - \lambda_{t+1}) \geq 0. \tag{71}
$$

Hence setting $\lambda = \lambda_k$ implies that

$$
(\beta_t(\lambda_{t+1} - \lambda_t) + \nabla_\lambda \tilde{V}_{t+1}(\theta_t, \lambda_t) - f(\theta_t) + \hat{f}(\theta_t))(\lambda_t - \lambda_{t+1}) \geq 0. \tag{72}
$$

Similarly, we have that

$$
(\beta_t(\lambda_t - \lambda_{t-1}) + \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}) - f(\theta_{t-1}) + \hat{f}(\theta_{t-1}))(\lambda_{t+1} - \lambda_t) \geq 0. \tag{73}
$$

Note that $\tilde{V}_t$ is convex, we hence have that

$$
\begin{aligned}
&\tilde{V}_t(\theta_t, \lambda_{t+1}) - \tilde{V}_t(\theta_t, \lambda_t) \\
&\geq (\nabla_\lambda \tilde{V}_t(\theta_t, \lambda_t))(\lambda_{t+1} - \lambda_t) \\
&= (\nabla_\lambda \tilde{V}_t(\theta_t, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))(\lambda_{t+1} - \lambda_t) + (\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))(\lambda_{t+1} - \lambda_t) \\
&\overset{(a)}{\geq} (\nabla_\lambda \tilde{V}_t(\theta_t, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1})(\lambda_{t-1}), \lambda_{t+1} - \lambda_t) \\
&\quad + (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - \beta_t(\lambda_t - \lambda_{t-1}))(\lambda_{t+1} - \lambda_t),
\end{aligned} \tag{74}
$$

where (a) is from (73). The first term in the RHS of (74) can be further bounded as follows.

$$
\begin{aligned}
&(\nabla_\lambda \tilde{V}_t(\theta_t, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))(\lambda_{t+1} - \lambda_t) \\
&= (\nabla_\lambda \tilde{V}_t(\theta_t, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t))(\lambda_{t+1} - \lambda_t) \\
&\quad + (\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))(\lambda_{t+1} - \lambda_t) \\
&= (\nabla_\lambda \tilde{V}_t(\theta_t, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t))(\lambda_{t+1} - \lambda_t) \\
&\quad + (\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))(\lambda_t - \lambda_{t-1}) \\
&\quad + m_{t+1}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1})),
\end{aligned} \tag{75}
$$

where $m_{t+1} \triangleq (\lambda_{t+1} - \lambda_t) - (\lambda_t - \lambda_{t-1})$. Plug it in (74) and we have that

$$
\begin{aligned}
&\tilde{V}_t(\theta_t, \lambda_{t+1}) - \tilde{V}_t(\theta_t, \lambda_t) \\
&\geq (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - \beta_t(\lambda_t - \lambda_{t-1}))(\lambda_{t+1} - \lambda_t) \\
&\quad + \underbrace{(\nabla_\lambda \tilde{V}_t(\theta_t, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t))(\lambda_{t+1} - \lambda_t)}_{(a)} \\
&\quad + \underbrace{(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))(\lambda_t - \lambda_{t-1})}_{(b)}
\end{aligned}
$$

$$+ \underbrace{(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))m_{t+1}}_{(c)}. \tag{76}$$

We then provide bounds for each term in (76) as follows.

Term (a) can be bounded as follows:

$$\begin{aligned}
&(\nabla_\lambda \tilde{V}_t(\theta_t, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t))(\lambda_{t+1} - \lambda_t) \\
&= (\nabla_\lambda V_\sigma^L(\theta_t, \lambda_t) - \nabla_\lambda V_\sigma^L(\theta_{t-1}, \lambda_t))(\lambda_{t+1} - \lambda_t) \\
&\geq \frac{-(\lambda_{t+1} - \lambda_t)^2}{2\xi} - \frac{\xi}{2}(\nabla_\lambda V_\sigma^L(\theta_t, \lambda_t) - \nabla_\lambda V_\sigma^L(\theta_{t-1}, \lambda_t))^2 \\
&\geq \frac{-(\lambda_{t+1} - \lambda_t)^2}{2\xi} - \frac{\xi(C_\sigma^V)^2}{2}\|\theta_t - \theta_{t-1}\|^2,
\end{aligned} \tag{77}$$

which is from Cauchy–Schwarz inequality and C_σ^V-smoothness of $V_\sigma^L(\theta, \lambda)$.

Term (b) can be bounded as follows:

$$\begin{aligned}
&(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))(\lambda_t - \lambda_{t-1}) \\
&\geq \frac{1}{b_{t-1} + b_0}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))^2,
\end{aligned} \tag{78}$$

which is from (69).

Term (c) can be bounded as follows by Cauchy–Schwarz inequality:

$$\begin{aligned}
&m_{t+1}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1})) \\
&\geq -\frac{\xi}{2}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))^2 - \frac{1}{2\xi}m_{t+1}^2
\end{aligned} \tag{79}$$

Moreover, it can be shown that

$$\frac{1}{\xi}(\lambda_{t+1} - \lambda_t)(\lambda_t - \lambda_{t-1}) = \frac{1}{2\xi}(\lambda_{t+1} - \lambda_t)^2 + \frac{1}{2\xi}(\lambda_t - \lambda_{t-1})^2 - \frac{1}{2\xi}m_{t+1}^2. \tag{80}$$

Plug (77) to (80) in 76, and we have that

$$\begin{aligned}
&\tilde{V}_t(\theta_t, \lambda_{t+1}) - \tilde{V}_t(\theta_t, \lambda_t) \\
&\geq (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}))(\lambda_{t+1} - \lambda_t) - \beta_t(\lambda_t - \lambda_{t-1})(\lambda_{t+1} - \lambda_t) \\
&\quad + (\nabla_\lambda \tilde{V}_t(\theta_t, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t))(\lambda_{t+1} - \lambda_t) + (\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))(\lambda_t - \lambda_{t-1}) \\
&\quad + m_{t+1}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1})) \\
&\geq (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}))(\lambda_{t+1} - \lambda_t) - \frac{1}{2\xi}(\lambda_{t+1} - \lambda_t)^2 - \frac{1}{2\xi}(\lambda_t - \lambda_{t-1})^2 + \frac{1}{2\xi}m_{t+1}^2 \\
&\quad - \frac{(\lambda_{t+1} - \lambda_t)^2}{2\xi} - \frac{\xi(C_\sigma^V)^2}{2}\|\theta_t - \theta_{t-1}\|^2 + \frac{1}{b_{t-1} + b_0}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))^2 \\
&\quad - \frac{\xi}{2}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))^2 - \frac{1}{2\xi}m_{t+1}^2 \\
&\geq (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}))(\lambda_{t+1} - \lambda_t) - \frac{1}{\xi}(\lambda_{t+1} - \lambda_t)^2 - \frac{1}{2\xi}(\lambda_t - \lambda_{t-1})^2 - \frac{\xi(C_\sigma^V)^2}{2}\|\theta_t - \theta_{t-1}\|^2.
\end{aligned} \tag{81}$$

From the definition of $\tilde{V}_t$, we have that

$$\begin{aligned}
&\tilde{V}_t(\theta_t, \lambda_{t+1}) - \tilde{V}_t(\theta_t, \lambda_t) \\
&= V_\sigma^L(\theta_t, \lambda_{t+1}) + \frac{b_{t-1}}{2}\lambda_{t+1}^2 - V_\sigma^L(\theta_t, \lambda_t) - \frac{b_{t-1}}{2}\lambda_t^2.
\end{aligned} \tag{82}$$

Then we have that

$$V_\sigma^L(\theta_t, \lambda_{t+1}) - V_\sigma^L(\theta_t, \lambda_t)$$

$$\geq \frac{b_{t-1}}{2}(\lambda_t^2 - \lambda_{t+1}^2) + (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}))(\lambda_{t+1} - \lambda_t)$$
$$- \frac{1}{\xi}(\lambda_{t+1} - \lambda_t)^2 - \frac{1}{2\xi}(\lambda_t - \lambda_{t-1})^2 - \frac{\xi (C_\sigma^V)^2}{2}\|\theta_t - \theta_{t-1}\|^2. \tag{83}$$

Combining with Lemma 7, if $\forall t$, $\mu_t > (1 + \Lambda^*)L_\sigma$, we then have that

$$\begin{aligned}&V_\sigma^L(\theta_{t+1}, \lambda_{t+1}) - V_\sigma^L(\theta_t, \lambda_t)\\ &\geq (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}))(\lambda_{t+1} - \lambda_t) + \langle \theta_{t+1} - \theta_t, -\hat{g}(\theta_t, \lambda_{t+1}) + g(\theta_t, \lambda_{t+1})\rangle - \frac{\xi (C_\sigma^V)^2}{2}\|\theta_t - \theta_{t-1}\|^2\\ &\quad + \left(\frac{\mu_t}{2} + \nu\right)\|\theta_{t+1} - \theta_t\|^2 + \frac{b_{t-1}}{2}(\lambda_t^2 - \lambda_{t+1}^2) - \frac{1}{\xi}(\lambda_{t+1} - \lambda_t)^2 - \frac{1}{2\xi}(\lambda_t - \lambda_{t-1})^2. \end{aligned}\tag{84}$$

□

Lemma 9. *Define*

$$\begin{aligned}F_{t+1} \triangleq& -\frac{8}{\xi^2 b_{t+1}}(\lambda_t - \lambda_{t+1})^2 - \frac{8}{\xi}\left(1 - \frac{b_t}{b_{t+1}}\right)\lambda_{t+1}^2 + V_\sigma^L(\theta_{t+1}, \lambda_{t+1}) + \frac{b_t}{2}\lambda_{t+1}^2\\ &+ \left(-\frac{16 (C_\sigma^V)^2}{\xi b_{t+1}^2} - \frac{\xi (C_\sigma^V)^2}{2}\right)\|\theta_{t+1} - \theta_t\|^2 + \left(\frac{8}{\xi} - \frac{1}{2\xi}\right)(\lambda_{t+1} - \lambda_t)^2,\end{aligned}\tag{85}$$

and if $\frac{1}{b_{t+1}} - \frac{1}{b_t} \leq \frac{\xi}{5}$, *then*

$$\begin{aligned}&F_{t+1} - F_t\\ &\geq S_t + \left(\frac{\mu_t}{2} + \nu - \frac{16 (C_\sigma^V)^2}{\xi b_{t+1}^2} - \frac{\xi (C_\sigma^V)^2}{2}\right)\|\theta_{t+1} - \theta_t\|^2 + \frac{b_t - b_{t-1}}{2}\lambda_{t+1}^2\\ &\quad + \frac{9}{10\xi}(\lambda_{t+1} - \lambda_t)^2 + \frac{8}{\xi}\left(\frac{b_t}{b_{t+1}} - \frac{b_{t-1}}{b_t}\right)\lambda_{t+1}^2,\end{aligned}\tag{86}$$

where $S_t \triangleq \frac{16}{b_t \xi}(f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - f(\theta_t) + \hat{f}(\theta_t))(-\lambda_t + \lambda_{t+1}) + (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}))(\lambda_{t+1} - \lambda_t) + \langle \theta_{t+1} - \theta_t, -\hat{g}(\theta_t, \lambda_{t+1}) + g(\theta_t, \lambda_{t+1})\rangle$.

Proof. From (72) and (73), we have that

$$\begin{aligned}\beta_t m_{t+1}(\lambda_t - \lambda_{t+1}) \geq& (\nabla_\lambda \tilde{V}_{t+1}(\theta_t, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))(-\lambda_t + \lambda_{t+1})\\ &+ (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - f(\theta_t) + \hat{f}(\theta_t))(-\lambda_t + \lambda_{t+1}).\end{aligned}\tag{87}$$

The first term can be rewritten as

$$\begin{aligned}&(\nabla_\lambda \tilde{V}_{t+1}(\theta_t, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))(\lambda_{t+1} - \lambda_t)\\ &= (\nabla_\lambda \tilde{V}_{t+1}(\theta_t, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t))(\lambda_{t+1} - \lambda_t) + (\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))(\lambda_{t+1} - \lambda_t)\\ &= (\nabla_\lambda \tilde{V}_{t+1}(\theta_t, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t))(\lambda_{t+1} - \lambda_t) + (\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))(\lambda_t - \lambda_{t-1})\\ &\quad + m_{t+1}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1})).\end{aligned}\tag{88}$$

The first term in (88) can be bounded as

$$\begin{aligned}&(\nabla_\lambda \tilde{V}_{t+1}(\theta_t, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t))(\lambda_{t+1} - \lambda_t)\\ &= (\nabla_\lambda V_\sigma^L(\theta_t, \lambda_t) - \nabla_\lambda V_\sigma^L(\theta_{t-1}, \lambda_t))(\lambda_{t+1} - \lambda_t) + (b_t \lambda_t - b_{t-1}\lambda_t)(\lambda_{t+1} - \lambda_t)\\ &\overset{(a)}{\geq} -\frac{1}{2h}(\nabla_\lambda V_\sigma^L(\theta_t, \lambda_t) - \nabla_\lambda V_\sigma^L(\theta_{t-1}, \lambda_t))^2 - \frac{h}{2}(\lambda_{t+1} - \lambda_t)^2\\ &\quad + \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1}^2 - \lambda_t^2) - \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1} - \lambda_t)^2\\ &\overset{(b)}{\geq} -\frac{(C_\sigma^V)^2}{2h}\|\theta_t - \theta_{t-1}\|^2 - \frac{h}{2}(\lambda_{t+1} - \lambda_t)^2\\ &\quad + \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1}^2 - \lambda_t^2) - \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1} - \lambda_t)^2,\end{aligned}\tag{89}$$

where (a) is from the Cauchy–Schwarz inequality and (b) is from the C_σ^V-smoothness of V_σ^L, for any $h > 0$.

Similar to (69), the second term in (88) can be bounded as

$$\begin{aligned}
&(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))(\lambda_t - \lambda_{t-1}) \\
&\geq \frac{b_{t-1} b_0}{b_{t-1} + b_0}(\lambda_t - \lambda_{t-1})^2 + \frac{1}{b_{t-1} + b_0}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))^2. \qquad (90)
\end{aligned}$$

The third term in (88) can be bounded as

$$\begin{aligned}
&m_{t+1}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1})) \\
&\geq -\frac{\xi}{2}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))^2 - \frac{1}{2\xi} m_{t+1}^2. \qquad (91)
\end{aligned}$$

Hence combine (89) to (90) and plug in (88), we have that

$$\begin{aligned}
&(\nabla_\lambda \tilde{V}_{t+1}(\theta_t, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))(\lambda_{t+1} - \lambda_t) \\
&\geq -\frac{(C_\sigma^V)^2}{2h}\|\theta_t - \theta_{t-1}\|^2 - \frac{h}{2}(\lambda_{t+1} - \lambda_t)^2 \\
&\quad + \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1}^2 - \lambda_t^2) - \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1} - \lambda_t)^2 \\
&\quad + \frac{b_{t-1} b_0}{b_{t-1} + b_0}(\lambda_t - \lambda_{t-1})^2 + \frac{1}{b_{t-1} + b_0}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))^2 \\
&\quad - \frac{\xi}{2}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))^2 - \frac{1}{2\xi} m_{t+1}^2. \qquad (92)
\end{aligned}$$

Hence (87) can be further bounded as

$$\begin{aligned}
&(\beta_t m_{t+1})(\lambda_t - \lambda_{t+1}) \\
&\geq (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - f(\theta_t) + \hat{f}(\theta_t))(-\lambda_t + \lambda_{t+1}) \\
&\quad - \frac{(C_\sigma^V)^2}{2h}\|\theta_t - \theta_{t-1}\|^2 - \frac{h}{2}(\lambda_{t+1} - \lambda_t)^2 \\
&\quad + \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1}^2 - \lambda_t^2) - \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1} - \lambda_t)^2 \\
&\quad + \frac{b_{t-1} b_0}{b_{t-1} + b_0}(\lambda_t - \lambda_{t-1})^2 + \frac{1}{b_{t-1} + b_0}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))^2 \\
&\quad - \frac{\xi}{2}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))^2 - \frac{1}{2\xi} m_{t+1}^2. \qquad (93)
\end{aligned}$$

It can be directly verified that

$$m_{t+1}(\lambda_t - \lambda_{t+1}) = \frac{1}{2}(\lambda_t - \lambda_{t-1})^2 - \frac{1}{2}(\lambda_t - \lambda_{t+1})^2 - \frac{m_{t+1}^2}{2}. \qquad (94)$$

Recall that $\beta_t = \frac{1}{\xi}$, hence

$$\begin{aligned}
&\frac{1}{2\xi}(\lambda_t - \lambda_{t-1})^2 - \frac{1}{2\xi}(\lambda_t - \lambda_{t+1})^2 - \frac{m_{t+1}^2}{2\xi} \\
&\geq (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - f(\theta_t) + \hat{f}(\theta_t))(-\lambda_t + \lambda_{t+1}) \\
&\quad - \frac{(C_\sigma^V)^2}{2h}\|\theta_t - \theta_{t-1}\|^2 - \frac{h}{2}(\lambda_{t+1} - \lambda_t)^2 \\
&\quad + \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1}^2 - \lambda_t^2) - \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1} - \lambda_t)^2 \\
&\quad + \frac{b_{t-1} b_0}{b_{t-1} + b_0}(\lambda_t - \lambda_{t-1})^2 + \frac{1}{b_{t-1} + b_0}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))^2
\end{aligned}$$

$$- \frac{\xi}{2}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))^2 - \frac{1}{2\xi} m_{t+1}^2. \tag{95}$$

From $\xi \leq \frac{1}{b_0} \leq \frac{2}{b_0+b_{t-1}}$, we have $\frac{1}{b_{t-1}+b_0}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))^2 - \frac{\xi}{2}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))^2 \geq 0$. Also, it can be shown that $\frac{b_{t-1}b_0}{b_{t-1}+b_0} \geq \frac{b_{t-1}b_0}{2b_0} = \frac{b_{t-1}}{2}$. Thus, it follows that

$$\begin{aligned}
&\frac{1}{2\xi}(\lambda_t - \lambda_{t-1})^2 - \frac{1}{2\xi}(\lambda_t - \lambda_{t+1})^2 - \frac{m_{t+1}^2}{2\xi} \\
&\geq (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - f(\theta_t) + \hat{f}(\theta_t))(-\lambda_t + \lambda_{t+1}) \\
&\quad - \frac{(C_\sigma^V)^2}{2h}\|\theta_t - \theta_{t-1}\|^2 - \frac{h}{2}(\lambda_{t+1} - \lambda_t)^2 \\
&\quad + \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1}^2 - \lambda_t^2) - \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1} - \lambda_t)^2 \\
&\quad + \frac{b_{t-1}}{2}(\lambda_t - \lambda_{t-1})^2 - \frac{1}{2\xi}m_{t+1}^2.
\end{aligned} \tag{96}$$

Re-arrange the terms, it follows that

$$\begin{aligned}
&- \frac{1}{2\xi}(\lambda_t - \lambda_{t+1})^2 - \frac{b_t - b_{t-1}}{2}\lambda_{t+1}^2 \\
&\geq (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - f(\theta_t) + \hat{f}(\theta_t))(-\lambda_t + \lambda_{t+1}) - \frac{(C_\sigma^V)^2}{2h}\|\theta_t - \theta_{t-1}\|^2 - \frac{h}{2}(\lambda_{t+1} - \lambda_t)^2 \\
&\quad - \frac{(b_t - b_{t-1})}{2}\lambda_t^2 - \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1} - \lambda_t)^2 + \frac{b_{t-1}}{2}(\lambda_t - \lambda_{t-1})^2 - \frac{1}{2\xi}(\lambda_t - \lambda_{t-1})^2 \\
&\geq -\frac{1}{2\xi}(\lambda_t - \lambda_{t-1})^2 - \frac{(b_t - b_{t-1})}{2}\lambda_t^2 + (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - f(\theta_t) + \hat{f}(\theta_t))(-\lambda_t + \lambda_{t+1}) \\
&\quad - \frac{(C_\sigma^V)^2}{2h}\|\theta_t - \theta_{t-1}\|^2 - \frac{h}{2}(\lambda_{t+1} - \lambda_t)^2 + \frac{b_{t-1}}{2}(\lambda_t - \lambda_{t-1})^2,
\end{aligned} \tag{97}$$

where the last inequality is from the fact that b_t is decreasing.

Now multiply $\frac{2}{\xi b_t}$ on both sides, we further have that

$$\begin{aligned}
&- \frac{1}{\xi^2 b_t}(\lambda_t - \lambda_{t+1})^2 - \frac{1}{\xi}\left(1 - \frac{b_{t-1}}{b_t}\right)\lambda_{t+1}^2 \\
&\geq -\frac{1}{\xi^2 b_t}(\lambda_t - \lambda_{t-1})^2 - \frac{1}{\xi}\left(1 - \frac{b_{t-1}}{b_t}\right)\lambda_t^2 + \frac{2}{\xi b_t}(f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - f(\theta_t) + \hat{f}(\theta_t))(-\lambda_t + \lambda_{t+1}) \\
&\quad - \frac{(C_\sigma^V)^2}{h\xi b_t}\|\theta_t - \theta_{t-1}\|^2 - \frac{h}{\xi b_t}(\lambda_{t+1} - \lambda_t)^2 + \frac{1}{\xi}(\lambda_t - \lambda_{t-1})^2.
\end{aligned} \tag{98}$$

If we set $h = \frac{b_t}{2}$, (98) can be rewritten as

$$\begin{aligned}
&- \frac{1}{\xi^2 b_t}(\lambda_t - \lambda_{t+1})^2 - \frac{1}{\xi}\left(1 - \frac{b_{t-1}}{b_t}\right)\lambda_{t+1}^2 \\
&\geq -\frac{1}{\xi^2 b_t}(\lambda_t - \lambda_{t-1})^2 - \frac{1}{\xi}\left(1 - \frac{b_{t-1}}{b_t}\right)\lambda_t^2 + \frac{2}{\xi b_t}(f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - f(\theta_t) + \hat{f}(\theta_t))(-\lambda_t + \lambda_{t+1}) \\
&\quad - \frac{2(C_\sigma^V)^2}{\xi b_t^2}\|\theta_t - \theta_{t-1}\|^2 - \frac{1}{2\xi}(\lambda_{t+1} - \lambda_t)^2 + \frac{1}{\xi}(\lambda_t - \lambda_{t-1})^2.
\end{aligned} \tag{99}$$

Further we have that

$$\begin{aligned}
&- \frac{1}{\xi^2 b_{t+1}}(\lambda_t - \lambda_{t+1})^2 + \left(\frac{1}{\xi^2 b_{t+1}} - \frac{1}{\xi^2 b_t}\right)(\lambda_t - \lambda_{t+1})^2 - \frac{1}{\xi}\left(1 - \frac{b_t}{b_{t+1}}\right)\lambda_{t+1}^2 + \frac{1}{\xi}\left(\frac{b_{t-1}}{b_t} - \frac{b_t}{b_{t+1}}\right)\lambda_{t+1}^2 \\
&\geq -\frac{1}{\xi^2 b_t}(\lambda_t - \lambda_{t-1})^2 - \frac{1}{\xi}\left(1 - \frac{b_{t-1}}{b_t}\right)\lambda_t^2 + \frac{2}{\xi b_t}(f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - f(\theta_t) + \hat{f}(\theta_t))(-\lambda_t + \lambda_{t+1})
\end{aligned}$$

$$-\frac{2(C_\sigma^V)^2}{\xi b_t^2}\|\theta_t-\theta_{t-1}\|^2-\frac{1}{2\xi}(\lambda_{t+1}-\lambda_t)^2+\frac{1}{\xi}(\lambda_t-\lambda_{t-1})^2. \tag{100}$$

Re-arranging the terms in (100) implies that

$$\begin{aligned}
&-\frac{1}{\xi^2 b_{t+1}}(\lambda_t-\lambda_{t+1})^2-\frac{1}{\xi}\left(1-\frac{b_t}{b_{t+1}}\right)\lambda_{t+1}^2-\left(-\frac{1}{\xi^2 b_t}(\lambda_t-\lambda_{t-1})^2-\frac{1}{\xi}\left(1-\frac{b_{t-1}}{b_t}\right)\lambda_t^2\right)\\
&\geq-\left(\frac{1}{\xi^2 b_{t+1}}-\frac{1}{\xi^2 b_t}\right)(\lambda_t-\lambda_{t+1})^2-\frac{1}{\xi}\left(\frac{b_{t-1}}{b_t}-\frac{b_t}{b_{t+1}}\right)\lambda_{t+1}^2\\
&\quad-\frac{2(C_\sigma^V)^2}{\xi b_t^2}\|\theta_t-\theta_{t-1}\|^2-\frac{1}{2\xi}(\lambda_{t+1}-\lambda_t)^2+\frac{1}{\xi}(\lambda_t-\lambda_{t-1})^2\\
&\quad+\frac{2}{\xi b_t}(f(\theta_{t-1})-\hat{f}(\theta_{t-1})-f(\theta_t)+\hat{f}(\theta_t))(-\lambda_t+\lambda_{t+1})\\
&\geq-\frac{7}{10\xi}(-\lambda_t+\lambda_{t+1})^2-\frac{2(C_\sigma^V)^2}{\xi b_t^2}\|\theta_t-\theta_{t-1}\|^2+\frac{1}{\xi}(\lambda_t-\lambda_{t-1})^2+\frac{1}{\xi}\left(\frac{b_t}{b_{t+1}}-\frac{b_{t-1}}{b_t}\right)\lambda_{t+1}^2\\
&\quad+\frac{2}{\xi b_t}(f(\theta_{t-1})-\hat{f}(\theta_{t-1})-f(\theta_t)+\hat{f}(\theta_t))(-\lambda_t+\lambda_{t+1}),
\end{aligned} \tag{101}$$

where the last inequality is from $\frac{1}{b_{t+1}}-\frac{1}{b_t}\leq\frac{\xi}{5}$. Recall in Lemma 8, we showed that

$$\begin{aligned}
&V_\sigma^L(\theta_{t+1},\lambda_{t+1})-V_\sigma^L(\theta_t,\lambda_t)\\
&\geq(f(\theta_{t-1})-\hat{f}(\theta_{t-1}))(\lambda_{t+1}-\lambda_t)+\langle\theta_{t+1}-\theta_t,-\hat{g}(\theta_t,\lambda_{t+1})+g(\theta_t,\lambda_{t+1})\rangle-\frac{\xi(C_\sigma^V)^2}{2}\|\theta_t-\theta_{t-1}\|^2\\
&\quad+\left(\frac{\mu_t}{2}+\nu\right)\|\theta_{t+1}-\theta_t\|^2+\frac{b_{t-1}}{2}(\lambda_t^2-\lambda_{t+1}^2)-\frac{1}{\xi}(\lambda_{t+1}-\lambda_t)^2-\frac{1}{2\xi}(\lambda_t-\lambda_{t-1})^2.
\end{aligned} \tag{102}$$

Combine both inequality together, and we further have that

$$\begin{aligned}
&-\frac{8}{\xi^2 b_{t+1}}(\lambda_t-\lambda_{t+1})^2-\frac{8}{\xi}\left(1-\frac{b_t}{b_{t+1}}\right)\lambda_{t+1}^2-\left(-\frac{8}{\xi^2 b_t}(\lambda_t-\lambda_{t-1})^2-\frac{8}{\xi}\left(1-\frac{b_{t-1}}{b_t}\right)\lambda_t^2\right)\\
&\quad+V_\sigma^L(\theta_{t+1},\lambda_{t+1})-V_\sigma^L(\theta_t,\lambda_t)\\
&\geq-\frac{28}{5\xi}(-\lambda_t+\lambda_{t+1})^2-\frac{16(C_\sigma^V)^2}{\xi b_t^2}\|\theta_t-\theta_{t-1}\|^2+\frac{8}{\xi}(\lambda_t-\lambda_{t-1})^2+\frac{8}{\xi}\left(\frac{b_t}{b_{t+1}}-\frac{b_{t-1}}{b_t}\right)\lambda_{t+1}^2\\
&\quad+\frac{16}{b_t\xi}(f(\theta_{t-1})-\hat{f}(\theta_{t-1})-f(\theta_t)+\hat{f}(\theta_t))(-\lambda_t+\lambda_{t+1})\\
&\quad+(f(\theta_{t-1})-\hat{f}(\theta_{t-1}))(\lambda_{t+1}-\lambda_t)+\langle\theta_{t+1}-\theta_t,-\hat{g}(\theta_t,\lambda_{t+1})+g(\theta_t,\lambda_{t+1})\rangle-\frac{\xi(C_\sigma^V)^2}{2}\|\theta_t-\theta_{t-1}\|^2\\
&\quad+\left(\frac{\mu_t}{2}+\nu\right)\|\theta_{t+1}-\theta_t\|^2+\frac{b_{t-1}}{2}(\lambda_t^2-\lambda_{t+1}^2)-\frac{1}{\xi}(\lambda_{t+1}-\lambda_t)^2-\frac{1}{2\xi}(\lambda_t-\lambda_{t-1})^2\\
&=S_t+\left(-\frac{16(C_\sigma^V)^2}{\xi b_t^2}-\frac{\xi(C_\sigma^V)^2}{2}\right)\|\theta_t-\theta_{t-1}\|^2+\left(-\frac{28}{5\xi}-\frac{1}{\xi}\right)(-\lambda_t+\lambda_{t+1})^2+\frac{b_{t-1}}{2}(\lambda_t^2-\lambda_{t+1}^2)\\
&\quad+\left(\frac{8}{\xi}-\frac{1}{2\xi}\right)(\lambda_t-\lambda_{t-1})^2+\frac{8}{\xi}\left(\frac{b_t}{b_{t+1}}-\frac{b_{t-1}}{b_t}\right)\lambda_{t+1}^2+\left(\frac{\mu_t}{2}+\nu\right)\|\theta_{t+1}-\theta_t\|^2,
\end{aligned} \tag{103}$$

where $S_t\triangleq\frac{16}{b_t\xi}(f(\theta_{t-1})-\hat{f}(\theta_{t-1})-f(\theta_t)+\hat{f}(\theta_t))(-\lambda_t+\lambda_{t+1})+(f(\theta_{t-1})-\hat{f}(\theta_{t-1}))(\lambda_{t+1}-\lambda_t)+\langle\theta_{t+1}-\theta_t,-\hat{g}(\theta_t,\lambda_{t+1})+g(\theta_t,\lambda_{t+1})\rangle$. Now

$$\begin{aligned}
&-\frac{8}{\xi^2 b_{t+1}}(\lambda_t-\lambda_{t+1})^2-\frac{8}{\xi}\left(1-\frac{b_t}{b_{t+1}}\right)\lambda_{t+1}^2-\left(-\frac{8}{\xi^2 b_t}(\lambda_t-\lambda_{t-1})^2-\frac{8}{\xi}\left(1-\frac{b_{t-1}}{b_t}\right)\lambda_t^2\right)\\
&\quad+V_\sigma^L(\theta_{t+1},\lambda_{t+1})-V_\sigma^L(\theta_t,\lambda_t)+\frac{b_t}{2}\lambda_{t+1}^2-\frac{b_{t-1}}{2}\lambda_t^2\\
&\quad+\left(-\frac{16(C_\sigma^V)^2}{\xi b_{t+1}^2}-\frac{\xi(C_\sigma^V)^2}{2}\right)\|\theta_{t+1}-\theta_t\|^2-\left(-\frac{16(C_\sigma^V)^2}{\xi b_t^2}-\frac{\xi(C_\sigma^V)^2}{2}\right)\|\theta_t-\theta_{t-1}\|^2
\end{aligned}$$

$$
\begin{aligned}
&+\left(\frac{8}{\xi}-\frac{1}{2\xi}\right)(\lambda_{t+1}-\lambda_t)^2-\left(\frac{8}{\xi}-\frac{1}{2\xi}\right)(\lambda_t-\lambda_{t-1})^2\\
\geq\ & S_t+\left(\frac{\mu_t}{2}+\nu-\frac{16(C_\sigma^V)^2}{\xi b_{t+1}^2}-\frac{\xi(C_\sigma^V)^2}{2}\right)\|\theta_{t+1}-\theta_t\|^2+\frac{b_t-b_{t-1}}{2}\lambda_{t+1}^2\\
&+\left(\frac{8}{\xi}-\frac{1}{2\xi}-\frac{28}{5\xi}-\frac{1}{\xi}\right)(\lambda_{t+1}-\lambda_t)^2+\frac{8}{\xi}\left(\frac{b_t}{b_{t+1}}-\frac{b_{t-1}}{b_t}\right)\lambda_{t+1}^2\\
=\ & S_t+\left(\frac{\mu_t}{2}+\nu-\frac{16(C_\sigma^V)^2}{\xi b_{t+1}^2}-\frac{\xi(C_\sigma^V)^2}{2}\right)\|\theta_{t+1}-\theta_t\|^2+\frac{b_t-b_{t-1}}{2}\lambda_{t+1}^2\\
&+\frac{9}{10\xi}(\lambda_{t+1}-\lambda_t)^2+\frac{8}{\xi}\left(\frac{b_t}{b_{t+1}}-\frac{b_{t-1}}{b_t}\right)\lambda_{t+1}^2,
\end{aligned}
\tag{104}
$$

which then completes the proof. □

We now restate Theorem 2 with all the specific step sizes. The definitions of these constants can also be found in Section I.

Theorem 3. *(Restatement of Theorem 2) Set* $b_t = \frac{19}{20\xi t^{0.25}}$, $\mu_t = \xi(C_\sigma^V)^2 + \frac{16\tau(C_\sigma^V)^2}{\xi(b_{t+1})^2} - 2\nu$, $\beta_t = \frac{1}{\xi}, \alpha_t = \nu + \mu_t$, *where* $\xi > \frac{2\nu+(1+\Lambda^*)L_\sigma}{(C_\sigma^V)^2}$, ν *is any positive number and* τ *is any number greater than* 2. *Moreover, set* $\epsilon_{est} = \frac{1}{t^{0.5}L_\Omega}\frac{1}{32t^{0.25}\Lambda^*+2\Lambda^*+\frac{1}{\alpha_1}(1+\Lambda^*)C_\sigma^V}\frac{19^2\epsilon^2}{3200\xi(\tau-2)(C_\sigma^V)^2uL_\Omega} = \mathcal{O}(\frac{\epsilon^2}{t^{0.75}})$, *then*

$$
\min_{1\leq t\leq T}\|G_t\|^2 \leq (1+\sqrt{2})\epsilon,
\tag{105}
$$

when $T = \mathcal{O}(\epsilon^{-4})$.

Proof. Denote by $p_t \triangleq \frac{8(\tau-2)(C_\sigma^V)^2}{\xi b_{t+1}^2}$ and $M_1 \triangleq \frac{16\tau^2}{(\tau-2)^2} + \frac{(\xi(C_\sigma^V)^2-\nu)^2}{64(\tau-2)^2(C_\sigma^V)^2\xi^2}$. Then it can be verified that $\nu + \frac{\mu_t}{2} - \frac{\xi(C_\sigma^V)^2}{2} - \frac{16(C_\sigma^V)^2}{\xi b_{t+1}^2} = p_t$. Then (104) can be rewritten as

$$
\begin{aligned}
F_{t+1}-F_t \geq S_t + p_t\|\theta_{t+1}-\theta_t\|^2 + \frac{b_t-b_{t-1}}{2}\lambda_{t+1}^2&\\
+\frac{9}{10\xi}(\lambda_{t+1}-\lambda_t)^2+\frac{8}{\xi}\left(\frac{b_t}{b_{t+1}}-\frac{b_{t-1}}{b_t}\right)\lambda_{t+1}^2&.
\end{aligned}
\tag{106}
$$

From the definition, we have that

$$
G_t = \begin{bmatrix} \beta_t\left(\lambda_t - \prod_{[0,\Lambda^*]}\left(\lambda_t - \frac{1}{\beta_t}\left(\nabla_\lambda V_\sigma^L(\theta_t,\lambda_t)\right)\right)\right) \\ \alpha_t\left(\theta_t - \prod_\Theta\left(\theta_t + \frac{1}{\alpha_t}\left(\nabla_\theta V_\sigma^L(\theta_t,\lambda_t)\right)\right)\right) \end{bmatrix},
\tag{107}
$$

and denote by

$$
\tilde{G}_t \triangleq \begin{bmatrix} \beta_t\left(\lambda_t - \prod_{[0,\Lambda^*]}\left(\lambda_t - \frac{1}{\beta_t}\left(\nabla_\lambda \tilde{V}_t(\theta_t,\lambda_t)\right)\right)\right) \\ \alpha_t\left(\theta_t - \prod_\Theta\left(\theta_t + \frac{1}{\alpha_t}\left(\nabla_\theta \tilde{V}_t(\theta_t,\lambda_t)\right)\right)\right) \end{bmatrix}.
\tag{108}
$$

It can be verified that

$$
\|G_t\| - \|\tilde{G}_t\| \leq b_{t-1}|\lambda_t|.
\tag{109}
$$

From Theorem 4.2 in [79], it can be shown that

$$
\|\tilde{G}_t\|^2 \leq 2(\mu_t+\nu)^2\|\theta_{t+1}-\theta_t\|^2 + \left(2(C_\sigma^V)^2 + \frac{1}{\xi^2}\right)(\lambda_{t+1}-\lambda_t)^2,
\tag{110}
$$

and

$$
M_1 \geq \frac{2(\nu+\mu_t)^2}{p_t^2}.
\tag{111}
$$

Hence

$$\|\tilde{G}_t\|^2 \leq M_1 p_t^2 \|\theta_{t+1} - \theta_t\|^2 + \left(2(C_\sigma^V)^2 + \frac{1}{\xi^2}\right)(\lambda_{t+1} - \lambda_t)^2. \tag{112}$$

Set $u_t \triangleq \frac{1}{\max\left\{M_1 p_t, \frac{10+20\xi^2(C_\sigma^V)^2}{9\xi}\right\}}$, then from (106), we have that

$$u_t\|\tilde{G}_t\|^2 \leq F_{t+1} - F_t - S_t - \frac{b_t - b_{t-1}}{2}\lambda_{t+1}^2 - \frac{8}{\xi}\left(\frac{b_t}{b_{t+1}} - \frac{b_{t-1}}{b_t}\right)\lambda_{t+1}^2. \tag{113}$$

Summing the inequality above from $t = 1$ to T, then

$$\begin{aligned}\sum_{t=1}^{T} u_t\|\tilde{G}_t\|^2 &\leq F_{T+1} - F_1 - \sum_{t=1}^{T} S_t + \frac{8}{\xi}\left(\frac{b_0}{b_1}\lambda_2^2 - \frac{b_T}{b_{T+1}}\lambda_{T+1}^2\right) + \left(\frac{b_0 - b_T}{2}(\Lambda^*)^2\right)\\ &\leq F_{T+1} - F_1 - \sum_{t=1}^{T} S_t + \frac{8}{\xi}\frac{b_0}{b_1}(\Lambda^*)^2 + \left(\frac{b_0 - b_T}{2}(\Lambda^*)^2\right),\end{aligned} \tag{114}$$

which is from b_t is decreasing and $\lambda_t < \Lambda^*$. Note that

$$\begin{aligned}\max_{t\geq 1}\max_{\theta\in\Theta,\lambda\in[0,\Lambda^*]} F_t &= \max\Bigg\{ -\frac{8}{\xi^2 b_{t+1}}(\lambda_t - \lambda_{t+1})^2 - \frac{8}{\xi}\left(1 - \frac{b_t}{b_{t+1}}\right)\lambda_{t+1}^2 + V_\sigma^L(\theta_{t+1}, \lambda_{t+1}) + \frac{b_t}{2}\lambda_{t+1}^2\\ &\quad + \left(-\frac{16(C_\sigma^V)^2}{\xi b_{t+1}^2} - \frac{\xi(C_\sigma^V)^2}{2}\right)\|\theta_{t+1} - \theta_t\|^2 + \left(\frac{8}{\xi} - \frac{1}{2\xi}\right)(\lambda_{t+1} - \lambda_t)^2\Bigg\}\\ &\leq \frac{1.6}{\xi}(\Lambda^*)^2 + (1 + \Lambda^*)(2C_\sigma) + \frac{b_1}{2}(\Lambda^*)^2 + \frac{15}{2\xi}(\Lambda^*)^2\\ &\triangleq F^*,\end{aligned} \tag{115}$$

which is from the definition of b_t, and $8(\frac{b_t}{b_{t+1}} - 1) \leq 8(\frac{(t+1)^{0.25}}{t^{0.25}} - 1) \leq 8(\frac{2^{0.25}}{1} - 1) < 1.6$. Then plugging in the definition of b_t implies that

$$\sum_{t=1}^{T} u_t\|\tilde{G}_t\|^2 \leq F^* - F_1 - \sum_{t=1}^{T} S_t + \frac{8}{\xi}(\Lambda^*)^2 + \left(\frac{b_0}{2}(\Lambda^*)^2\right). \tag{116}$$

If moreover set $u \triangleq \max\left\{M_1, \frac{10+20\xi^2(C_\sigma^V)^2}{9\xi p_2}\right\}$, then $u_t \geq \frac{1}{up_t}$, and hence

$$\frac{\sum_{t=1}^{T}\frac{1}{p_t}\|\tilde{G}_t\|^2}{\sum_{t=1}^{T}\frac{1}{p_t}} \leq \frac{u}{\sum_{t=1}^{T}\frac{1}{p_t}}\left(F^* - F_1 - \sum_{t=1}^{T} S_t + \frac{8}{\xi}(\Lambda^*)^2 + \left(\frac{b_0}{2}(\Lambda^*)^2\right)\right). \tag{117}$$

Plug in the definition of p_t then we have that

$$\frac{\sum_{t=1}^{T}\frac{1}{p_t}\|\tilde{G}_t\|^2}{\sum_{t=1}^{T}\frac{1}{p_t}} \leq \frac{3200\xi(\tau - 2)(C_\sigma^V)^2 d}{19^2(\sqrt{T} - 2)}\left(F^* - F_1 - \sum_{t=1}^{T} S_t + \frac{8}{\xi}(\Lambda^*)^2 + \left(\frac{b_0}{2}(\Lambda^*)^2\right)\right). \tag{118}$$

We moreover have that

$$\begin{aligned}|S_t| &= \Bigg|\frac{16}{b_t\xi}(f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - f(\theta_t) + \hat{f}(\theta_t))(-\lambda_t + \lambda_{t+1}) + (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}))(\lambda_{t+1} - \lambda_t)\\ &\quad + \langle\theta_{t+1} - \theta_t, -\hat{g}(\theta_t, \lambda_{t+1}) + g(\theta_t, \lambda_{t+1})\rangle\Bigg|\\ &\leq 32t^{0.25}\Lambda^*(\Omega_{t-1} + \Omega_t) + 2\Lambda^*\Omega_{t-1} + \frac{1}{\alpha_t}(1 + \Lambda^*)C_\sigma^V\Omega_t,\end{aligned} \tag{119}$$

where $\Omega_t \triangleq \max\left\{\|g(\theta_t, \lambda_{t+1}) - \hat{g}(\theta_t, \lambda_{t+1})\|, |f(\theta_t) - \hat{f}(\theta_t)|\right\}$. Note that it has been shown in [75] that $\Omega_t \leq L_\Omega \max\left\{\|Q_{\sigma,r} - \hat{Q}_{\sigma,r}\|, \|Q_{\sigma,c} - \hat{Q}_{\sigma,c}\|\right\} = L_\Omega\epsilon_{\text{est}}$, and hence Ω_t can be controlled by setting ϵ_{est}.

Note that $\alpha_t = \nu + \mu_t$ is increasing, hence $\frac{1}{\alpha_t} \leq \frac{1}{\alpha_1}$. Hence if we set $\epsilon_{\text{est}} = \frac{1}{t^{0.5} L_\Omega} \frac{1}{32 t^{0.25} \Lambda^* + 2\Lambda^* + \frac{1}{\alpha_1}(1+\Lambda^*) C_\sigma^V} \frac{19^2 \epsilon^2}{3200\xi(\tau-2)(C_\sigma^V)^2 u L_\Omega} = \mathcal{O}(\frac{\epsilon^2}{t^{0.75}})$, then

$$|S_t| \leq \frac{1}{t^{0.5}} \frac{19^2 \epsilon^2}{3200\xi(\tau-2)(C_\sigma^V)^2 u L_\Omega}, \tag{120}$$

and hence

$$\left| \sum_{t=1}^{T} S_t \right| \leq \sqrt{T} \frac{19^2 \epsilon^2}{3200\xi(\tau-2)(C_\sigma^V)^2 u L_\Omega}. \tag{121}$$

Thus plug in (118) and we have that

$$\frac{\sum_{t=1}^{T} \frac{1}{p_t} \|\tilde{G}_t\|^2}{\sum_{t=1}^{T} \frac{1}{p_t}} \leq \frac{3200\xi(\tau-2)(C_\sigma^V)^2 u}{19^2(\sqrt{T}-2)} K + \epsilon^2, \tag{122}$$

where $K = F^* - F_1 + \frac{8}{\xi}(\Lambda^*)^2 + \left(\frac{b_1}{2}(\Lambda^*)^2\right)$. When $T = (2 + \frac{3200\xi(\tau-2)(C_\sigma^V)^2 u K}{19^2 \epsilon^2})^2$, we have that

$$\frac{\sum_{t=1}^{T} \frac{1}{p_t} \|\tilde{G}_t\|^2}{\sum_{t=1}^{T} \frac{1}{p_t}} \leq 2\epsilon^2. \tag{123}$$

Similarly to Theorem 4.2 in [79], if $t > \frac{19^4 (\Lambda^*)^4}{2 \cdot 10^4 \xi^4 \epsilon^4}$, then $b_{t-1} < \frac{\epsilon}{\Lambda^*}$ and $b_{t-1}\lambda_t < \epsilon$. Hence combine with (109) we finally have that

$$\min_{1 \leq t \leq T} \|G_t\| \leq (1+\sqrt{2})\epsilon, \tag{124}$$

when $T = \max\left\{ \frac{7(\Lambda^*)^4}{\xi^4 \epsilon^4}, \left(2 + \frac{9\xi(\tau-2)(C_\sigma^V)^2 u K}{\epsilon^2}\right)^2 \right\} = \mathcal{O}(\epsilon^{-4})$. $\square$

Remark 1. *Note that the sample complexity of robust TD algorithm to achieve an ϵ_{est}-error bound is $\mathcal{O}(\epsilon_{est}^{-2})$, hence the sample complexity at the time step t is $\mathcal{O}(\epsilon_{est}^{-2}) = \mathcal{O}(\frac{t^{1.5}}{\epsilon^4})$. Thus the total sample complexity to find an ϵ-stationary solution is $\sum_{t=1}^{T} \frac{t^{1.5}}{\epsilon^4} = \mathcal{O}(\epsilon^{-14})$. This great increasing of complexity is due to the estimation of robust value functions.*

I Constants

In this section, we summarize the definitions of all the constants we used in this paper.

$$\begin{aligned}
L_V &= \frac{k|\mathcal{A}|}{(1-\gamma)^2}, \\
C_\sigma &= \frac{1}{1-\gamma}(1 + 2\gamma\delta \frac{\log|\mathcal{S}|}{\sigma}), \\
C_\sigma^V &= \frac{1}{1-\gamma}|\mathcal{A}| k C_\sigma, \\
k_B &= \frac{1}{1-\gamma+\gamma\delta}\left(|\mathcal{A}| C_\sigma l + |\mathcal{A}| k C_\sigma^V\right) + \frac{2|\mathcal{A}|^2\gamma(1-\delta)}{(1-\gamma+\gamma\delta)^2} k^2 C_\sigma, \\
L_\sigma &= k_B + \frac{\gamma\delta}{1-\gamma}\left(\sqrt{|\mathcal{S}|} k_B + 2\sigma|\mathcal{S}| C_\sigma^V \frac{1}{1-\gamma+\gamma\delta} k|\mathcal{A}| C_\sigma\right), \\
b_t &= \frac{19}{20\xi t^{0.25}}, \\
M_1 &= \frac{16\tau^2}{(\tau-2)^2} + \frac{(\xi(C_\sigma^V)^2 - \nu)^2}{64(\tau-2)^2(C_\sigma^V)^2\xi^2}, \\
u &= \max\left\{ M_1, \frac{10 + 20\xi^2(C_\sigma^V)^2}{9\xi p_2} \right\},
\end{aligned}$$

$$
\begin{aligned}
F^* &= \frac{1.6}{\xi}(\Lambda^*)^2 + (1+\Lambda^*)(2C_\sigma) + \frac{b_1}{2}(\Lambda^*)^2 + \frac{15}{2\xi}(\Lambda^*)^2,\\
K &= F^* - F_1 + \frac{8}{\xi}(\Lambda^*)^2 + \left(\frac{b_1}{2}(\Lambda^*)^2\right),\\
\mu_t &= \xi(C_\sigma^V)^2 + \frac{16\tau(C_\sigma^V)^2}{\xi(b_{t+1})^2} - 2\nu,\\
\beta_t &= \frac{1}{\xi},\\
\alpha_t &= \nu + \mu_t.
\end{aligned}
\tag{125}
$$