# Robust Constrained Reinforcement Learning

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Constrained reinforcement learning is to maximize the reward subject to constraints on utilities/costs. However, in practice it is often the case that the training environment is not 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 in the test environment. To address this challenge, we formulate the framework of robust constrained reinforcement learning under model uncertainty, where the MDP is not fixed but lies in some uncertainty set. The goal is two fold: 1) to guarantee that constraints on utilities/costs are satisfied for all MDPs in the uncertainty set, and 2) to maximize the worst-case reward performance over the uncertainty set. We design a robust primal-dual approach, and further develop theoretical guarantee on its convergence, complexity and robust feasibility. We then investigate a concrete example of $\delta$-contamination uncertainty set, design an online and model-free algorithm and theoretically characterize its sample complexity.

## 1 Introduction

In many practical reinforcement learning (RL) applications, it is critical for an agent to meet certain constraints on utilities/costs while maximizing the reward. This problem is usually modeled as the constrained Markov decision processes (CMDPs) Altman (1999). Consider a CMDP with state space $\mathcal{S}$, action space $\mathcal{A}$, transition kernel $\mathsf{P} = \{p_s^a \in \Delta_{\mathcal{S}}{}^1 : s \in \mathcal{S}, a \in \mathcal{A}\}$, reward and utility functions: $r, c_i : \mathcal{S} \times \mathcal{A} \to [0, 1]$, $1 \leq i \leq m$, and discount factor $\gamma$. The goal of CMDP is to find a stationary policy $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$ that maximizes the expected reward subject to constraints on the utility:

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

where $\rho$ is the initial state distribution, $b_i$ 's are some thresholds and $\mathbb{E}_{\pi,\mathsf{P}}$ denotes the expectation when the agent follows policy $\pi$ and the environment transits following $\mathsf{P}$.

In practice, it is often the case that the environment on which the learned policy will deploy (the test environment) possibly deviates from the training one, 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 and crash due to mismatch between training and test environments. This hence motivates the study of robust constrained RL in this paper. In this paper, we take a pessimistic approach in face of uncertainty. Specifically, consider a set of transition kernels $\mathcal{P}$, which is usually constructed in a way to include the test environment with high probability Iyengar (2005); Nilim & El Ghaoui (2004); Bagnell et al. (2001). The learned policy should satisfy the constraints under all these environments in $\mathcal{P}$, i.e., $\forall \mathsf{P} \in \mathcal{P}$,

$$\mathbb{E}_{\pi,\mathsf{P}} \left[ \sum_{t=0}^{\infty} \gamma^t c_i(S_t, A_t) | S_0 \sim \rho \right] \geq b_i, \tag{2}$$

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

which is equivalent to $\min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}_{\pi,\mathsf{P}}\left[\sum_{t=0}^{\infty}\gamma^t c(S_t,A_t)|S_0\sim\rho\right]\geq b_i$. At the same time, we aim to optimize the worst-case reward performance over $\mathcal{P}$:

$$\begin{aligned}&\max_{\pi\in\Pi}\min_{\mathsf{P}\in\mathcal{P}}\mathbb{E}_{\pi,\mathsf{P}}\left[\sum_{t=0}^{\infty}\gamma^t r(S_t,A_t)|S_0\sim\rho\right],\\ &\text{s.t.}\ \min_{\mathsf{P}\in\mathcal{P}}\mathbb{E}_{\pi,\mathsf{P}}\left[\sum_{t=0}^{\infty}\gamma^t c_i(S_t,A_t)|S_0\sim\rho\right]\geq b_i, 1\leq i\leq m.\end{aligned}\tag{3}$$

On one hand, a feasible solution to eq. (3) always satisfies eq. (2), and on the other hand, the solution to eq. (3) provides a performance guarantee for any $\mathsf{P}\in\mathcal{P}$. We note that our approach and analysis can also be applied to the optimistic approach in face of uncertainty.

In this paper, we design and analyze a robust primal-dual algorithm for the problem of robust constrained RL. In particular, the technical challenges and our major contributions are as follows.

- We take the Lagrange multiplier method to solve the constrained policy optimization problem. A first question is that whether the primal problem is equivalent to the dual problem, i.e., whether the duality gap is zero. For non-robust constrained RL, the Lagrange function has a zero duality gap Paternain et al. (2019); Altman (1999). However, we show that this is not necessarily true in the robust constrained setting. Note that the set of visitation distribution being convex is one key property to show zero duality gap of constrained MDP Altman (1999); Paternain et al. (2019). In this paper, we constructed a novel counter example showing that the set of robust visitation distributions for our robust problem is non-convex.
- In the dual problem of non-robust CMDPs, 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. Therefore, the geometry of our Lagrangian function is much more complicated. 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.
- In general, convergence to stationary points of the Lagrangian function does not necessarily imply that the solution is feasible Lin et al. (2020); **?**. We design a novel proof to show that the gradient belongs to the normal cone of the feasible set, based on which we further prove the robust feasibility of the obtained policy.
- We apply and extend our results on an important uncertainty set referred to as $\delta$-contamination model Huber (1965). Under this model, the robust value functions are not differentiable and we hence propose a smoothed approximation of the robust value function towards a better geometry. We further investigate the practical online and model-free setting and design an actor-critic type algorithm. We also establish its convergence, sample complexity, and robust feasibility.

We then discuss works related to robust constrained RL.

**Robust constrained RL.** In Russel et al. (2020), 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 Sutton et al. (1999) 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 Russel et al. (2020)), 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 Mankowitz et al. (2020) 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 Russel et al.

(2020) and Mankowitz et al. (2020) 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 Altman (1999); Paternain et al. (2019; 2022); Auer et al. (2008), which augments the objective with a sum of constraints weighted by their corresponding Lagrange multipliers, and then alternatively updates the primal and dual variables. Under the non-robust constrained MDP problems, it was shown that the strong duality holds, and hence the primal-dual method has zero duality gap Paternain et al. (2019); Altman (1999). The convergence rate of the primal-dual method was moreover investigated in Ding et al. (2020; 2021); Li et al. (2021b); Liu et al. (2021); Ying et al. (2021).

Another class of method is the primal method, which does not introduce any dual variables or Lagrange multipliers; Instead, it checks if the constraints are satisfied after each update, and enforces the constraints if not Achiam et al. (2017); Liu et al. (2020); Chow et al. (2018); Dalal et al. (2018); Xu et al. (2021); Yang et al. (2020).

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 eq. (4)), and therefore have much more complicated geometry than the non-robust case.

**Non-Constrained Robust RL under model uncertainty.** The non-constrained robust RL was first introduced and studied in Iyengar (2005); Nilim & El Ghaoui (2004), where the uncertainty set is assumed to be known, and the problem can be solved using robust dynamic programming. It was further extended to the model-free setting, where the learner assumes no knowledge of the environment, and can only get samples from its centroid transition kernel Roy et al. (2017); Wang & Zou (2021; 2022); Zhou et al. (2021); Yang et al. (2021); Panaganti & Kalathil (2021); Ho et al. (2018; 2021).

These works, however, mainly focus on robust RL without constraints, whereas in this paper we investigate robust RL with constraints, which is more challenging. There is a related line of works on (robust) imitation learning Ho & Ermon (2016); Fu et al. (2017); Torabi et al. (2018); Viano et al. (2022), which can be formulated as a constrained problem. But their problem settings and approaches are fundamentally different from ours.

## 2 Preliminaries

**Constrained MDP.** Consider the CMDP problem in eq. (1). Define the visitation distribution induced by policy $\pi$ and transition kernel $\mathsf{P}$: $d^{\pi}_{\rho,\mathsf{P}}(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}(S_t = s, A_t = a | S_0 \sim \rho, \pi, \mathsf{P})$. It can be shown that the set of the visitation distributions of all policies $\{d^{\pi}_{\rho,\mathsf{P}} \in \Delta_{\mathcal{S}\times\mathcal{A}} : \pi \in \Pi\}$ is convex Paternain et al. (2022); Altman (1999). Based on this convexity, the strong duality of CMDP can be established Altman (1999); Paternain et al. (2019) under a standard assumption referred as Slater's condition: Bertsekas (2014); Ding et al. (2021): there exists a constant $\zeta > 0$ and a policy $\pi \in \Pi$ s.t. $\forall i$, $V^{\pi}_{c_i,\mathsf{P}} - b_i \geq \zeta$.

**Robust MDP.** In this paper, we focus on the $(s,a)$-rectangular uncertainty set Nilim & El Ghaoui (2004); Iyengar (2005), i.e., $\mathcal{P} = \bigotimes_{s,a}\mathcal{P}^a_s$, where $\mathcal{P}^a_s \subseteq \Delta_{\mathcal{S}}$. At each time step, the environment transits following a transition kernel belonging to the uncertainty set $\mathsf{P}_t \in \mathcal{P}$. The robust value function of a policy $\pi$ is then defined as the worst-case expected accumulative discounted reward following policy $\pi$ over all MDPs in the uncertainty set Nilim & El Ghaoui (2004); Iyengar (2005):

$$V^{\pi}_{r,\mathcal{P}}(s) \triangleq \min_{\kappa=(\mathsf{P}_0,\mathsf{P}_1,\ldots)\in\bigotimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\kappa}\left[\sum_{t=0}^{\infty}\gamma^t r(S_t,A_t)|S_0 = s,\pi\right], \tag{4}$$

where $\mathbb{E}_{\kappa}$ denotes the expectation when the state transits according to $\kappa$. It was shown that the robust value function is the fixed point of the robust Bellman operator Nilim & El Ghaoui (2004); Iyengar (2005); Puterman (2014): $\mathbf{T}_{\pi}V(s) \triangleq \sum_{a\in\mathcal{A}}\pi(a|s)\left(r(s,a) + \gamma\sigma_{\mathcal{P}^a_s}(V)\right)$, where $\sigma_{\mathcal{P}^a_s}(V) \triangleq \min_{p\in\mathcal{P}^a_s} p^{\top}V$ is the support function of $V$ on $\mathcal{P}^a_s$.

Note that the minimizer of eq. (4), $\kappa^*$, is stationary in time Iyengar (2005), which we denote by $\kappa^* = \{\mathsf{P}^\pi, \mathsf{P}^\pi, ...\}$, and refer to $\mathsf{P}^\pi$ as the worst-case transition kernel. Then the robust value function $V^\pi_{r,\mathcal{P}}$ is actually the value function under policy $\pi$ and transition kernel $\mathsf{P}^\pi$. The goal of robust RL is to find the optimal robust policy $\pi^*$ that maximizes the worst-case accumulative discounted reward: $\pi^* = \arg\max_\pi V^\pi_{r,\mathcal{P}}(s), \forall s \in \mathcal{S}$.

## 3 Robust Constrained RL

Recall the robust constrained RL formulated in eq. (3):

$$\max_{\theta\in\Theta} V_r^{\pi_\theta}(\rho), \text{ s.t. } V_{c_i}^{\pi_\theta}(\rho) \geq b_i, 1 \leq i \leq m, \tag{5}$$

where for simplicity we omit the subscript $\mathcal{P}$ in $V^{\pi_\theta}_{\diamond,\mathcal{P}}$ and denote by $V_{c_i}^{\pi_\theta}(\rho)$ and $V_r^{\pi_\theta}(\rho)$ the robust value function for $c_i$ and $r$ under $\pi_\theta$. The goal of eq. (5) is to find a policy that maximizes the robust reward value function among those feasible solutions. Here, any feasible solution to eq. (5) can guarantee that under any MDP in the uncertainty set, its accumulative discounted utility is always no less than $b_i$, which guarantees robustness to constraint violation under model uncertainty. Furthermore, the optimal solution to eq. (5) achieves the best "worst-case reward performance" among all feasible solutions. If we use the optimal solution to eq. (5), then under any MDP in the uncertainty set, we have a guaranteed reward no less than the value of eq. (5).

In this paper, we focus on the parameterized policy class, i.e., $\pi_\theta \in \Pi_\Theta$, where $\Theta \subseteq \mathbb{R}^d$ is a parameter set and $\Pi_\Theta$ is a class of parameterized policies, e.g., direct parameterized policy, softmax or neural network policy. For technical convenience, we adopt a standard assumption on the policy class.

**Assumption 1.** *The policy class $\Pi_\Theta$ 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 $\|\nabla^2\pi_\theta(a|s)\| \leq l$.*

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

The problem eq. (5) is equivalent to the following max-min problem:

$$\max_{\theta\in\Theta}\min_{\lambda_i\geq 0} V_r^{\pi_\theta}(\rho) + \sum_{i=1}^m \lambda_i(V_{c_i}^{\pi_\theta}(\rho) - b_i). \tag{6}$$

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 $\pi$ under its worst-case transition kernel $\mathsf{P}^\pi$, and therefore can be written as the inner product between the reward (utility) function and the visitation distribution induced by $\pi$ and $\mathsf{P}^\pi$ (referred to as robust visitation distribution of $\pi$). The following lemma shows that the set of robust visitation distributions may not be convex, and therefore, the approach used in Altman (1999); Paternain et al. (2019) 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 eq. (6). For simplicity, we investigate the case with one constraint, and extension to the case with multiple constraints is straightforward:

$$\min_{\lambda\geq 0}\max_{\theta\in\Theta} V_r^{\pi_\theta}(\rho) + \lambda(V_c^{\pi_\theta}(\rho) - b). \tag{7}$$

We make an assumption of Slater's condition, assuming there exists at least one strictly feasible policy Bertsekas (2014); Ding et al. (2021), under which, we further show that the optimal dual variable of eq. (7) is bounded. This assumption is

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

**Lemma 2.** *Denote the optimal solution of eq. (7) by $(\lambda^*, \pi_{\theta^*})$. Then, $\lambda^* \in \left[0, \frac{2}{\zeta(1-\gamma)}\right]$.*

Lemma 2 suggests that the dual problem eq. (7) is equivalent to a bounded min-max problem:

$$\min_{\lambda \in \left[0, \frac{2}{\zeta(1-\gamma)}\right]} \max_{\theta \in \Theta} V_r^{\pi_\theta}(\rho) + \lambda(V_c^{\pi_\theta}(\rho) - b). \tag{8}$$

The problem in eq. (8) is a bounded linear-nonconcave optimization problem. We then propose our robust primal-dual algorithm for robust constrained RL in Algorithm 1. The basic idea of Algorithm 1 is to perform

**Algorithm 1** Robust Primal-Dual algorithm (RPD)

**Input**: $T$, $\alpha_t$, $\beta_t$, $b_t$
**Initialization**: $\lambda_0$, $\theta_0$
  **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**: $\theta_T$

gradient descent-ascent w.r.t. $\lambda$ and $\theta$ alternatively. When the policy $\pi$ violates the constraint, the dual variable $\lambda$ increases such that $\lambda V_c^\pi$ dominates $V_r^\pi$. Then the gradient ascent will update $\theta$ 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 Lin et al. (2020) if $b_t = 0$. However, $b_t$ is critical to the convergence of Algorithm 1 **?**. The outer problem of eq. (8) is actually linear, and after introducing $b_t$, the update of $\lambda_t$ 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

$$G_t \triangleq \begin{bmatrix} \beta_t\left(\lambda_t - \prod_{[0,\Lambda^*]}\left(\lambda_t - \frac{1}{\beta_t}\left(\nabla_\lambda V^L(\theta_t, \lambda_t)\right)\right)\right) \\ \alpha_t\left(\theta_t - \prod_{\Theta}\left(\theta_t + \frac{1}{\alpha_t}\left(\nabla_\theta V^L(\theta_t, \lambda_t)\right)\right)\right) \end{bmatrix}. \tag{9}$$

The gradient mapping is a standard measure of convergence for projected optimization approaches Beck (2017). 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 $\theta$ and $\lambda$ 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:*

$$\|\nabla_\lambda V^L(\theta, \lambda)|_{\theta_1} - \nabla_\lambda V^L(\theta, \lambda)|_{\theta_2}\| \le L_{11}\|\theta_1 - \theta_2\|, \tag{10}$$

$$\|\nabla_\lambda V^L(\theta, \lambda)|_{\lambda_1} - \nabla_\lambda V^L(\theta, \lambda)|_{\lambda_2}\| \le L_{12}|\lambda_1 - \lambda_2|, \tag{11}$$

$$\|\nabla_\theta V^L(\theta, \lambda)|_{\theta_1} - \nabla_\theta V^L(\theta, \lambda)|_{\theta_2}\| \le L_{21}\|\theta_1 - \theta_2\|, \tag{12}$$

$$\|\nabla_\theta V^L(\theta, \lambda)|_{\lambda_1} - \nabla_\theta V^L(\theta, \lambda)|_{\lambda_2}\| \le L_{22}|\lambda_1 - \lambda_2|. \tag{13}$$

As will be shown in Section 4, Assumption 3 can be satisfied with a smoothed approximation of the robust value function.

In the following theorem, we show that our robust primal-dual algorithm converges to a stationary point of the min-max problem eq. (16), with a complexity of $\mathcal{O}(\epsilon^{-4})$. The accurate statement can be found in Appendix E.

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

We note that this convergence rate matches the rate of general convex-nonconcave optimization problems in Xu et al. (2023), which is State-of-the-Art according to our best knowledge.

The next proposition characterizes the feasibility of the obtained policy.

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

In general, convergence to stationary points of the Lagrangian function does not necessarily imply that the solution is feasible. Proposition 1 shows that Algorithm 1 always return a policy that is robust feasible, i.e., satisfying the constraints in eq. (5). Intuitively, if we set $\Lambda^*$ 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, $\pi_W$ is feasible. On the other hand, $\Lambda^*$ can't be set too large. Note that the complexity in Theorem 1 depends on $\Lambda^*$ (see eq. (54) in the appendix), and a larger $\Lambda^*$ means a higher complexity.

## 4 $\delta$-Contamination Uncertainty Set

In this section, we investigate a concrete example of robust constrained RL with $\delta$-contamination uncertainty set. The method we developed here can be similarly extended to other types of uncertainty sets like KL-divergence or total variation. The $\delta$-contamination uncertainty set models the scenario where the state transition of the MDP could be arbitrarily perturbed with a small probability $\delta$. This model is widely used to model distributional uncertainty in the literature of robust learning and optimization, e.g., Huber (1965); Du et al. (2018); Huber & Ronchetti (2009); Nishimura & Ozaki (2004; 2006); Prasad et al. (2020a;b); Wang & Zou (2021; 2022). Specifically, let $\mathsf{P} = \{p_s^a | s \in \mathcal{S}, a \in \mathcal{A}\}$ be the centroid transition kernel, then the $\delta$-contamination uncertainty set centered at $\mathsf{P}$ 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 $\delta$-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 non-differentiable due to the min term, and hence Assumption 3 does not hold. One possible approach is to use sub-gradient, 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:

$$\mathbf{T}_\sigma^\pi V(s) = \mathbb{E}_{A\sim\pi(\cdot|s)}\left[r(s,A) + \gamma(1-\delta)\sum_{s'\in\mathcal{S}} p_{s,s'}^A V(s') + \gamma\delta \mathrm{LSE}(\sigma, V)\right], \tag{14}$$

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 $V_\sigma^\pi$, is an approximation of the robust value function $V^\pi$ Wang & Zou (2022). We refer to $V_\sigma^\pi$ 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 $\pi$, as $\sigma \to -\infty$, $\|V_r^\pi - V_{\sigma,r}^\pi\| \to 0$ and $\|V_c^\pi - V_{\sigma,c}^\pi\| \to 0$.

The gradient of $V_\sigma^{\pi_\theta}$ can be computed explicitly Wang & Zou (2022): $\nabla V_\sigma^{\pi_\theta}(s) = B(s,\theta) + \frac{\gamma\delta\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)} B(s,\theta)}{(1-\gamma)\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}}$, 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 $\pi_\theta$ under $\mathsf{P}$ 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 $\nabla V_\sigma^L$ is Lipschitz.

**Lemma 3.** $\nabla V_\sigma^L$ *is Lipschitz in* $\theta$ *and* $\lambda$. *And hence Assumption 3 holds for* $V_\sigma^L$.

A natural idea is to use the smoothed robust value functions to replace the ones in eq. (7):

$$\min_{\lambda\ge 0}\max_{\pi\in\Pi_\Theta} V_{\sigma,r}^\pi(\rho) + \lambda(V_{\sigma,c}^\pi(\rho) - b). \tag{15}$$

As will be shown below in Lemma 6, this approximation can be arbitrarily close to the original problem in eq. (7) as $\sigma \to -\infty$. We first show that under Assumption 2, the following Slater's condition holds for the smoothed problem in eq. (15).

**Lemma 4.** *Let $\sigma = \mathcal{O}(\epsilon^{-1})$ be sufficiently large such that $\|V^{\pi}_{\sigma,c} - V^{\pi}_{c}\| < \zeta$ for any $\pi$, then the Slater's condition also holds for $V^{\pi}_{\sigma,c}$, i.e., there exists $\zeta' > 0$ and a policy $\pi' \in \Pi_\Theta$ s.t. $V^{\pi'}_{\sigma,c}(\rho) - b \geq \zeta'$.*

The following lemma shows that the optimal dual variable for eq. (15) is also bounded.

**Lemma 5.** *Denote the optimal solution of eq. (15) by $(\lambda^*, \pi_{\theta^*})$. Then $\lambda^* \in \left[0, \frac{2C_\sigma}{\zeta'}\right]$, where $C_\sigma$ is the upper bound of smoothed robust value functions $V^{\pi}_{\sigma,c}$.*

Denote by $\Lambda^* = \max\left\{\frac{2C_\sigma}{\zeta'}, \frac{2}{\zeta(1-\gamma)}\right\}$, then problems eq. (8) and eq. (15) are equivalent to the following bounded ones: $\min_{\lambda\in[0,\Lambda^*]} \max_{\pi\in\Pi_\Theta} V^{\pi}_{r}(\rho) + \lambda(V^{\pi}_{c}(\rho) - b)$, and

$$\min_{\lambda\in[0,\Lambda^*]} \max_{\pi\in\Pi_\Theta} V^{\pi}_{\sigma,r}(\rho) + \lambda(V^{\pi}_{\sigma,c}(\rho) - b). \tag{16}$$

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

**Lemma 6.** *Choose a large enough $\sigma = \mathcal{O}(\epsilon^{-1})$ such that $\|V^{\pi}_{r} - V^{\pi}_{\sigma,r}\| \leq \epsilon$ and $\|V^{\pi}_{c} - V^{\pi}_{\sigma,c}\| \leq \epsilon$. Then*

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

In the following, we hence focus on the smoothed dual problem in eq. (16), which is an accurate approximation of the original problem eq. (8). Denote the gradient mapping of the smoothed Lagrangian function $V^{L}_{\sigma}$ by

$$G_t \triangleq \begin{bmatrix} \beta_t\left(\lambda_t - \prod_{[0,\Lambda^*]}\left(\lambda_t - \frac{1}{\beta_t}\left(\nabla_\lambda V^{L}_{\sigma}(\theta_t,\lambda_t)\right)\right)\right) \\ \alpha_t\left(\theta_t - \prod_{\Theta}\left(\theta_t + \frac{1}{\alpha_t}\left(\nabla_\theta V^{L}_{\sigma}(\theta_t,\lambda_t)\right)\right)\right) \end{bmatrix}. \tag{17}$$

Applying our RPD algorithm in eq. (16), we have the following convergence guarantee.

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

This corollary implies that our robust primal-dual algorithm converges to a stationary point of the min-max problem eq. (16) under the $\delta$-contamination model, with a complexity of $\mathcal{O}((\Lambda^*)^4\epsilon^{-4})$. Also, we note that the convergence rate does not depend on $\sigma$.

## 5 Robust Primal-Dual Method for Model-Free Setting

In the previous sections, we assumed full knowledge of the environment, the smoothed robust value functions, and their gradients. However, such information may not be readily available in real-world applications. To address this challenge, we consider a more practical model-free setting, where we have no prior knowledge of the environment or robust value functions and can only obtain samples from the nominal transition kernel. This setting presents new challenges in terms of designing effective algorithms for constrained robust reinforcement learning. By extending our approach to the model-free setting, we aim to provide a more practical and applicable solution to robust constrained RL problems.

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 Wang & Zou (2022) to estimate the smoothed robust value functions.

**Algorithm 2** Smoothed Robust TD Wang & Zou (2022)

**Input**: $T_{\text{inner}}$, $\pi$, $\sigma$, $c$
**Initialization**: $Q_0$, $s_0$
**for** $t = 0, 1, ..., T_{\text{inner}} - 1$ **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 $s \in \mathcal{S}$
$Q_{t+1}(s_t, a_t) \leftarrow Q_t(s_t, a_t) + \alpha_t\big(c_t + \gamma(1-\delta)\cdot V_t(s_{t+1}) + \gamma\delta \cdot \text{LSE}(\sigma, V_t) - Q_t(s_t, a_t)\big)$
**end for**
**Output**: $Q_{T_{\text{inner}},c} \triangleq Q_{T_{\text{inner}}}$

It was shown that the smoothed robust TD algorithm converges to the smoothed robust value function with a sample complexity of $\mathcal{O}(\epsilon^{-2})$ Wang & Zou (2022) 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**: $T$, $\sigma$, $\epsilon_{\text{est}}$, $\beta_t$, $\alpha_t$, $b_t$,$r$,$c$
**Initialization**: $\lambda_0$, $\theta_0$
**for** $t = 0, 1, ..., T-1$ **do**
Set $T_{\text{inner}} = \mathcal{O}\left(\frac{1}{\epsilon_{\text{est}}^2}\right)$ and run Algorithm 2 for $r$ and $c$, output $Q_{T_{\text{inner}},r}$, $Q_{T_{\text{inner}},c}$
$\hat{V}_{\sigma,r}^{\pi_{\theta_t}}(s) \leftarrow \sum_a \pi_{\theta_t}(a|s)Q_{T_{\text{inner}},r}(s,a)$, $\hat{V}_{\sigma,c}^{\pi_{\theta_t}}(s) \leftarrow \sum_a \pi_{\theta_t}(a|s)Q_{T_{\text{inner}},c}(s,a)$
$\hat{V}_{\sigma,r}^{\pi_{\theta_t}}(\rho) \leftarrow \sum_s \rho(s)\hat{V}_{\sigma,r}^{\pi_{\theta_t}}(s)$, $\hat{V}_{\sigma,c}^{\pi_{\theta_t}}(\rho) \leftarrow \sum_s \rho(s)\hat{V}_{\sigma,c}^{\pi_{\theta_t}}(s)$
$\lambda_{t+1} \leftarrow \prod_{[0,\Lambda^*]}\left(\lambda_t - \frac{1}{\beta_t}\left(\hat{V}_{\sigma,c}^{\pi_{\theta_t}}(\rho) - b\right) - \frac{b_t}{\beta_t}\lambda_t\right)$
$\theta_{t+1} \leftarrow \prod_{\Theta}\left(\theta_t + \frac{1}{\alpha_t}\left(\nabla_\theta \hat{V}_{\sigma,r}^{\pi_{\theta_t}}(\rho) + \lambda_{t+1}\nabla_\theta \hat{V}_{\sigma,c}^{\pi_{\theta_t}}(\rho)\right)\right)$
**end for**
**Output**: $\theta_T$

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 $2\epsilon$-feasible policy almost surely.

Under the online model-free setting, the estimation of the robust value functions is biased. Therefore, the analysis is more challenging than the existing literature, where it is usually assumed that the gradients are exact. We develop a new method to bound the bias accumulated in every iteration of the algorithm and establish the final convergence results. The formal statement and discussion on the sample complexity can be found in Theorem 3.

**Theorem 2.** *Set step sizes $\alpha_t, \beta_t$, and $b_t$ as in Section I, and let $\epsilon_{est} = \mathcal{O}(\frac{\epsilon 2}{\sigma^2 t^{0.75}})$, $T = \mathcal{O}(\epsilon^{-4})$, then* $\min_{1\le t\le T} \|G_t\| \le (1+\sqrt{2})\epsilon$.

## 6 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 Russel et al. (2021); Mankowitz et al. (2020) and the vanilla non-robust primal-dual method. Based on the idea of "robust policy evaluation" + "non-robust policy improvement" in Russel et al. (2021); Mankowitz et al. (2020), we combine the robust TD algorithm 2 with non-robust vanilla policy gradient method Sutton et al. (1999), which we refer to as the heuristic primal-dual

algorithm. Several environments, including Garnet Archibald et al. (1995), $8 \times 8$ Frozen-Lake and Taxi environments from OpenAI Brockman et al. (2016), are investigated.

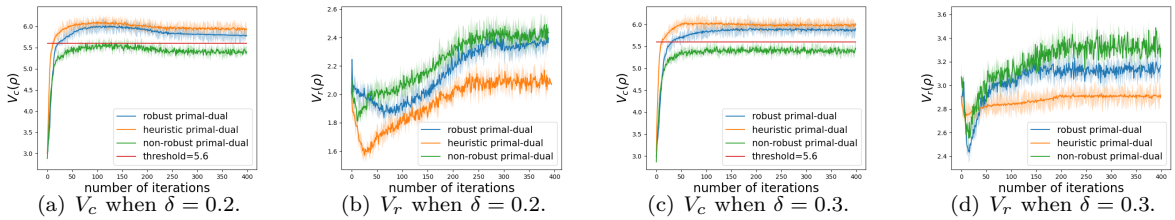
We first run the algorithm and store the obtained policies $\pi_t$ at each time step. At each time step, we run robust TD with a sample size 200 for 30 times to estimate the objective $V_r(\rho)$ and the constraint $V_c(\rho)$. We then plot them v.s. the number of iterations $t$. The upper and lower envelopes of the curves correspond to the 95 and 5 percentiles of the 30 curves, respectively. We deploy our experiments on the $\delta$-contamination model, and repeat 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 generated from the uniform distribution on [0,1]. The transition kernels are also randomly generated. The comparison results are shown in Fig.1.

$8 \times 8$ **Frozen-Lake problem.** We then compare the three algorithms under the $8 \times 8$ Frozen-lake problem setting in Fig.2. The Frozen-Lake problem involves a frozen lake of size $8 \times 8$ 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 $r = -10$ and $c = 0$ when falling in a hole, receives $r = 20$ and $c = 1$ when arrive at the end point; At other times, the agent receives $r = 0$ and a randomly generated utility $c$ according to the uniform distribution on [0,1].

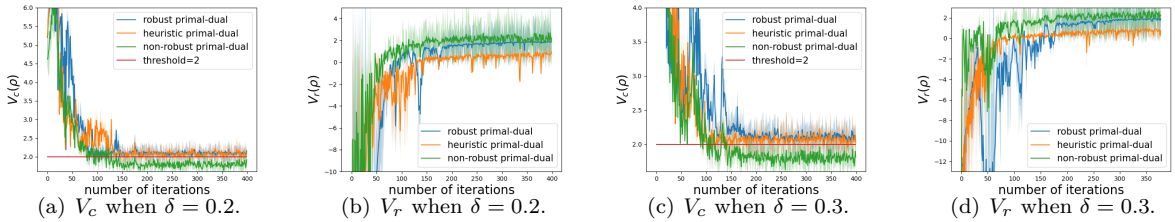
**Taxi problem.** We then compare the three algorithms under the Taxi problem environment. The taxi problem simulates a taxi driver in a $5 \times 5$ map. There are four designated locations in the grid world and a passenger occurs at a random location of the designated four locations at the start of each episode. The goal of the driver is to first pick up the passengers and then drop them 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 the utility according to the uniform distribution on [0,1] for each state-action pair. The results are shown in Fig.3.

From the experiment results above, it can be seen that: (1) 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. (2) Compared to the heuristic PD method, our RPD method can obtain more reward and can find a more robust policy while satisfying the robust constraint. Note that the non-robust PD method obtain more reward, but this is because the policy it finds violates the robust constraint. Our experiments demonstrate that among the three algorithms, our RPD algorithm is the best one which optimizes the worst-case reward performance while satisfying the robust constraints on the utility.



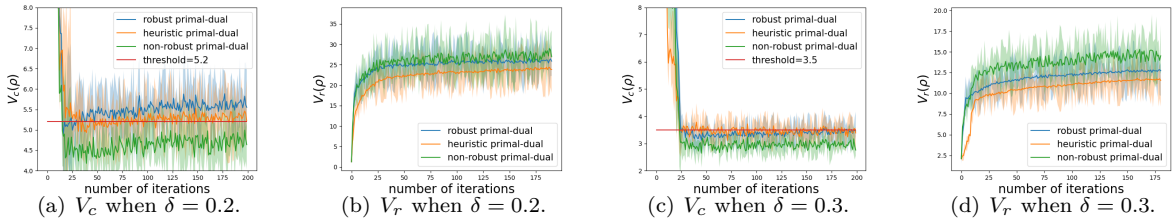

(a) $V_c$ when $\delta = 0.2$. (b) $V_r$ when $\delta = 0.2$. (c) $V_c$ when $\delta = 0.3$. (d) $V_r$ when $\delta = 0.3$.

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 $8 \times 8$ 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.

## 7 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 $\delta$-contamination uncertainty set, and construct online and model-free robust primal-dual algorithm. Our methodology can also be readily extended to problems with other uncertainty sets like KL-divergence, total variation and Wasserstein distance. The major challenge lies in deriving the robust policy gradient, and further designing model-free algorithm to estimate the robust value function.

## References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proc. International Conference on Machine Learning (ICML)*, pp. 22–31. PMLR, 2017.

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

Eitan Altman. *Constrained Markov decision processes: stochastic modeling*. Routledge, 1999.

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

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, volume 21, 2008.

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

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

Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016.

Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

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

Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained Markov decision processes. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 8378–8390, 2020.

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *Proc. International Conference on Artifical Intelligence and Statistics (AISTATS)*, pp. 3304–3312. PMLR, 2021.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1675–1685. PMLR, 2019.

Simon S Du, Yining Wang, Sivaraman Balakrishnan, Pradeep Ravikumar, and Aarti Singh. Robust nonparametric regression under Huber's $\epsilon$-contamination model. *arXiv preprint arXiv:1805.10406*, 2018.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

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

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Fast Bellman updates for robust MDPs. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1979–1988. PMLR, 2018.

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

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

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

PJ Huber and EM Ronchetti. *Robust Statistics*. John Wiley & Sons, Inc, 2009.

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

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Softmax policy gradient methods can take exponential time to converge. *arXiv preprint arXiv:2102.11270*, 2021a.

Tianjiao Li, Ziwei Guan, Shaofeng Zou, Tengyu Xu, Yingbin Liang, and Guanghui Lan. Faster algorithm and sharper analysis for constrained Markov decision process. *arXiv preprint arXiv:2110.10351*, 2021b.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *Proc. International Conference on Machine Learning (ICML)*, pp. 6083–6093. PMLR, 2020.

Tao Liu, Ruida Zhou, Dileep Kalathil, PR Kumar, and Chao Tian. Fast global convergence of policy optimization for constrained MDPs. *arXiv preprint arXiv:2111.00552*, 2021.

Yongshuai Liu, Jiaxin Ding, and Xin Liu. Ipo: Interior-point policy optimization under constraints. In *Proc. Conference on Artificial Intelligence (AAAI)*, volume 34, pp. 4940–4947, 2020.

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

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

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *Proc. International Conference on Learning Representations (ICLR)*, 2018.

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

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

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

Kiyohiko G. Nishimura and Hiroyuki Ozaki. An axiomatic approach to $\epsilon$-contamination. *Economic Theory*, 27(2):333–340, 2006.

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

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

Santiago Paternain, Miguel Calvo-Fullana, Luiz FO Chamon, and Alejandro Ribeiro. Safe policies for reinforcement learning via primal-dual methods. *IEEE Transactions on Automatic Control*, 2022.

Adarsh Prasad, Vishwak Srinivasan, Sivaraman Balakrishnan, and Pradeep Ravikumar. On learning ising models under Huber's contamination model. *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020a.

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

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

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

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

Reazul Hasan Russel, Mouhacine Benosman, Jeroen Van Baar, and Radu Corcodel. Lyapunov robust constrained-MDPs: Soft-constrained robustly stable policy optimization under model uncertainty. *arXiv preprint arXiv:2108.02701*, 2021.

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

Faraz Torabi, Garrett Warnell, and Peter Stone. Generative adversarial imitation from observation. *arXiv preprint arXiv:1807.06158*, 2018.

Luca Viano, Yu-Ting Huang, Parameswaran Kamalaruban, Craig Innes, Subramanian Ramamoorthy, and Adrian Weller. Robust learning from observation with model misspecification. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 1337–1345, 2022.

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

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

Yue Wang and Shaofeng Zou. Policy gradient method for robust reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, 2022.

Tengyu Xu, Yingbin Liang, and Guanghui Lan. Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *Proc. International Conference on Machine Learning (ICML)*, pp. 11480–11491. PMLR, 2021.

Zi Xu, Huiling Zhang, Yang Xu, and Guanghui Lan. A unified single-loop alternating gradient projection algorithm for nonconvex-concave and convex-nonconcave minimax problems. *arXiv preprint arXiv:2006.02032*, 2023.

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

Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Towards theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics. *arXiv preprint arXiv:2105.03863*, 2021.

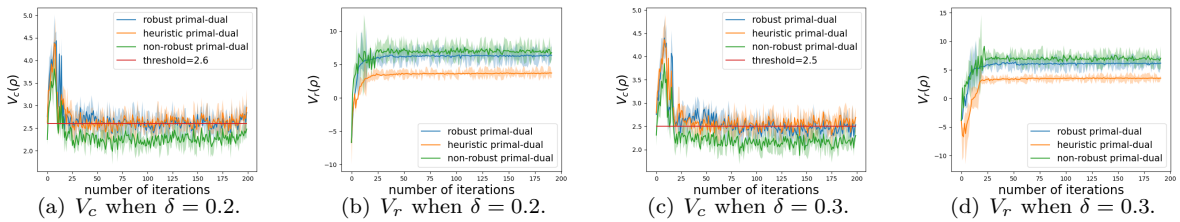
Donghao Ying, Yuhao Ding, and Javad Lavaei. A dual approach to constrained Markov decision processes with entropy regularization. *arXiv preprint arXiv:2110.08923*, 2021.

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

# Appendix

## A Additional Experiments

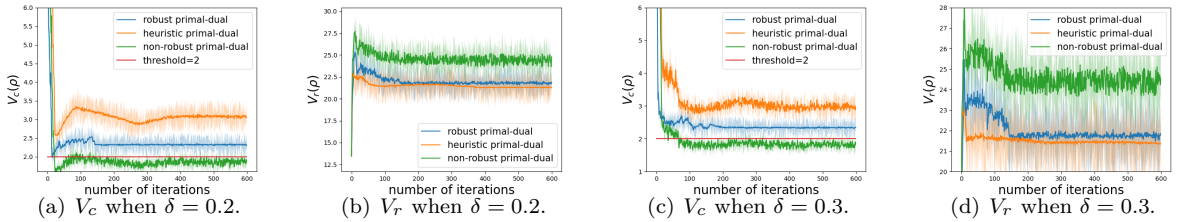
$4 \times 4$ **Frozen Lake problem.** The $4 \times 4$ frozen lake is similar to the $8 \times 8$ one but with a smaller map. Similarly, we randomly generate the utility signal for each state-action pair. The results are shown in Fig.4.




(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 4: Comparison on $4 \times 4$ Frozen-Lake problem.

$N$**-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.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 $N$-Chain problem.

## B Proof of Lemma 1

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

$$\left\{ d \in \Delta_{\mathcal{S}\times\mathcal{A}} : \exists \pi \in \Pi, \text{ s.t. } \forall (s,a), \begin{cases} d(s,a) = \pi(a|s) \sum_b d(s,b), \\ \gamma \sum_{s',a'} (p^\pi)_{s',s}^{a'} d(s',a') + (1-\gamma)\rho(s) = \sum_a d(s,a). \end{cases} \right\}. \tag{18}$$

Under the $\delta$-contamination model, $\mathsf{P}^\pi$ 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 eq. (18) can be rewritten as

$$\left\{ d \in \Delta_{\mathcal{S}\times\mathcal{A}} : \exists \pi, \text{ s.t. } \forall (s,a), \left\{ \begin{array}{l} d(s,a) = \pi(a|s)\left(\sum_b d(s,b)\right), \\ \gamma(1-\delta)\sum_{s',a'} p^{a'}_{s',s} d(s',a') + \gamma\delta \mathbb{1}_{\{s=\arg\min V^\pi\}} \\ \qquad + (1-\gamma)\rho(s) = \sum_a d(s,a). \end{array} \right\} \right\}. \tag{19}$$

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 $(\pi', d')$ such that $\forall \lambda \in [0,1]$ and $\forall s, a$,

$$\lambda d_1(s,a) + (1-\lambda) d_2(s,a) = d'(s,a), \tag{20}$$

$$d'(s,a) = \pi'(a|s)\left(\sum_b d'(s,b)\right), \tag{21}$$

$$\sum_{a'} d'(s,a') = \gamma(1-\delta)\sum_{s',a'} p^{a'}_{s',s} d'(s',a') + \gamma\delta \mathbb{1}_{\left\{s=\arg\min V^{\pi'}\right\}} + (1-\gamma)\rho(s). \tag{22}$$

eq. (22) firstly implies that $\forall s$,

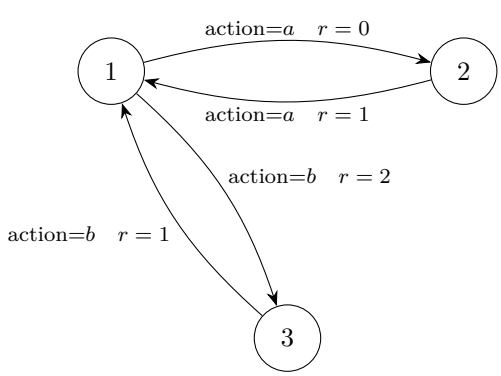
$$\lambda \mathbb{1}_{\{s=\arg\min V^{\pi_1}\}} + (1-\lambda)\mathbb{1}_{\{s=\arg\min V^{\pi_2}\}} = \mathbb{1}_{\left\{s=\arg\min V^{\pi'}\right\}}, \tag{23}$$

where from eq. (20) and eq. (21), $\pi'$ should be

$$\pi'(a|s) = \frac{d'(s,a)}{\sum_b d'(s,b)} = \frac{\lambda d_1(s,a) + (1-\lambda) d_2(s,a)}{\sum_b (\lambda d_1(s,b) + (1-\lambda) d_2(s,b))}. \tag{24}$$

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

$$d_1(1,a) = \frac{1-\gamma}{1-\gamma^2}, \tag{25}$$
$$d_1(1,b) = 0, \tag{26}$$
$$d_1(2,a) = \frac{\gamma(1-\gamma)}{1-\gamma^2}, \tag{27}$$
$$d_1(2,b) = 0, \tag{28}$$
$$d_1(3,a) = 0, \tag{29}$$
$$d_1(3,b) = 0. \tag{30}$$

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

$$d_2(1,a) = 0, \tag{31}$$
$$d_2(1,b) = \frac{1-\gamma}{1-\gamma^2}, \tag{32}$$
$$d_2(2,a) = 0, \tag{33}$$
$$d_2(2,b) = 0, \tag{34}$$
$$d_2(3,a) = 0, \tag{35}$$
$$d_2(3,b) = \frac{\gamma(1-\gamma)}{1-\gamma^2}. \tag{36}$$

Hence according to eq. (24), $\pi'$ should be as follows:

$$\pi'(a|1) = \lambda, \pi'(b|1) = 1-\lambda, \pi'(a|2) = 1, \pi'(b|3) = 1. \tag{37}$$

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

Clearly eq. (23) holds only if $V^{\pi'}(1) = V^{\pi'}(2) = \min_s V^{\pi'}(s)$. However, according to the Bellman equations for $\pi'$, we have that

$$V^{\pi'}(1) = \lambda(\gamma(1-\delta)V^{\pi'}(2) + \gamma\delta \min V^{\pi'}) + (1-\lambda)(2 + \gamma(1-\delta)V^{\pi'}(3) + \gamma\delta \min V^{\pi'}), \tag{38}$$
$$V^{\pi'}(2) = 1 + \gamma(1-\delta)V^{\pi'}(1) + \gamma\delta \min V^{\pi'}, \tag{39}$$
$$V^{\pi'}(3) = 1 + \gamma(1-\delta)V^{\pi'}(1) + \gamma\delta \min V^{\pi'}. \tag{40}$$

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

$$V^{\pi'}(1) = \frac{4}{3} + \gamma\delta \min V^{\pi'} + \gamma(1-\delta)V^{\pi'}(2), \tag{41}$$
$$V^{\pi'}(2) = 1 + \gamma\delta \min V^{\pi'} + \gamma(1-\delta)V^{\pi'}(1). \tag{42}$$

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 $C = V_r^{\pi_{\theta^*}}(\rho) + \lambda^*(V_c^{\pi_{\theta^*}}(\rho) - b)$, clearly $\max_{\pi\in\Pi} V_r^{\pi}(\rho) + \lambda^*(V_c^{\pi}(\rho) - b) = C$, and hence

$$C = \max_{\pi\in\Pi} V_r^{\pi}(\rho) + \lambda^*(V_c^{\pi}(\rho) - b) \geq V_r^{\pi^\zeta}(\rho) + \lambda^*(V_c^{\pi^\zeta}(\rho) - b) \geq V_r^{\pi^\zeta}(\rho) + \lambda^*\zeta. \tag{43}$$

Thus we have that

$$\lambda^* \leq \frac{C - V_r^{\pi^\zeta}(\rho)}{\zeta}. \tag{44}$$

Note that

$$C = \min_{\lambda \geq 0} \max_{\pi \in \Pi} V_r^\pi(\rho) + \lambda(V_c^\pi(\rho) - b) \overset{(a)}{\leq} \max_{\pi \in \Pi} V_r^\pi(\rho) \leq \frac{1}{1-\gamma}, \tag{45}$$

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

$$\lambda^* \leq \frac{1}{(1-\gamma)\zeta}, \tag{46}$$

which completes the proof. □

**Proof of Lemma 5**

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

$$C = \max_{\pi \in \Pi} V_{\sigma,r}^\pi(\rho) + \lambda^*(V_{\sigma,c}^\pi(\rho) - b) \geq V_{\sigma,r}^{\pi^{\zeta'}}(\rho) + \lambda^*(V_{\sigma,c}^{\pi^{\zeta'}}(\rho) - b) \geq V_{\sigma,r}^{\pi^{\zeta'}}(\rho) + \lambda^*\zeta'. \tag{47}$$

Thus we have that

$$C \geq V_{\sigma,r}^{\pi^\zeta}(\rho) + \lambda^*\zeta', \tag{48}$$

hence

$$\lambda^* \leq \frac{C - V_{\sigma,r}^{\pi^\zeta}(\rho)}{\zeta'}. \tag{49}$$

Note that

$$C = \min_{\lambda \geq 0} \max_{\pi \in \Pi} V_{\sigma,r}^\pi(\rho) + \lambda(V_{\sigma,c}^\pi(\rho) - b) \leq \max_{\pi \in \Pi} V_{\sigma,r}^\pi(\rho) \leq C_\sigma, \tag{50}$$

where $C_\sigma$ is the upper bound of smoothed robust value functions Wang & Zou (2022): $C_\sigma = \frac{1}{1-\gamma}(1 + 2\gamma R \frac{\log|\mathcal{S}|}{\sigma})$. Hence we have that

$$\lambda^* \leq \frac{C_\sigma}{\zeta'}, \tag{51}$$

which completes the proof. □

## D Proof of Lemma 6

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

$$|V^D(\lambda) - V_\sigma^D(\lambda)| \leq (1+\lambda)\epsilon \leq (1+\Lambda^*)\epsilon. \tag{52}$$

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

## E Proof of Theorem 1

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

Set $b_t = \frac{19}{20\xi t^{0.25}}$, $\mu_t = \xi L_{21}^2 + \frac{16\tau L_{21}^2}{\xi(b_{t+1})^2} - 2\nu$, $\beta_t = \frac{1}{\xi}, \alpha_t = \nu + \mu_t$, where $\xi < \frac{1}{10L_{11}}$, $\nu$ is any positive number and $\tau > \max\left\{2, K_3 \frac{L_{22}+2\mu-\xi(L_{21})^2}{\xi L_{21}^2}\right\}$, then

$$\min_{1\le t\le T} \|G_t\|^2 \le 2\epsilon, \tag{53}$$

when

$$T = \max\left\{\frac{7(\Lambda^*)^4}{\xi^4\epsilon^4}, \left(2 + \frac{9\xi(\tau-2)(L_{21})^2 D_2 D_3}{\epsilon^2}\right)^2\right\} = \mathcal{O}(\epsilon^{-4}). \tag{54}$$

The definitions of the constants 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 Proposition 1

*Proof.* The $\lambda$-entry of $G_W$ is smaller than $2\epsilon$, i.e.,

$$|(G_W)_\lambda| = \left|\beta_W\left(\lambda_W - \prod\nolimits_{[0,\Lambda^*]}\left(\lambda_W - \frac{1}{\beta_W}\left(\nabla_\lambda V^L(\theta_W,\lambda_W)\right)\right)\right)\right| < 2\epsilon. \tag{55}$$

Denote $\lambda^+ \triangleq \prod_{[0,\Lambda^*]}\left(\lambda_W - \frac{1}{\beta_W}\left(\nabla_\lambda V^L(\theta_W,\lambda_W)\right)\right)$. From Lemma 3 in Ghadimi & Lan (2016), $-\nabla_\lambda V^L(\theta_W,\lambda^+)$ can be rewritten as the sum of two parts: $-\nabla_\lambda V^L(\theta_W,\lambda^+) \in N_{[0,\Lambda^*]}(\lambda^+) + 4\epsilon B$, where $N_K(x) \triangleq \left\{g \in \mathbb{R}^d : \langle g, y-x\rangle \le 0 : \forall y \in K\right\}$ is the normal cone, and $B$ is the unit ball.

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

## G Proof of Lemma 3

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

$$\nabla_\lambda V_\sigma^L(\theta,\lambda) = V_{\sigma,c}^{\pi_\theta}(\rho) - b, \tag{56}$$

$$\nabla_\theta V_\sigma^L(\theta,\lambda) = \nabla_\theta V_{\sigma,r}^{\pi_\theta}(\rho) + \lambda\nabla_\theta V_{\sigma,c}^{\pi_\theta}(\rho). \tag{57}$$

Note that in Wang & Zou (2022), it has been shown that

$$\|V_{\sigma,r}^{\pi_{\theta_1}} - V_{\sigma,r}^{\pi_{\theta_2}}\| \le C_\sigma^V\|\theta_1-\theta_2\|, \tag{58}$$

$$\|\nabla_\theta V_{\sigma,r}^{\pi_{\theta_1}} - \nabla_\theta V_{\sigma,r}^{\pi_{\theta_2}}\| \le L_\sigma\|\theta_1-\theta_2\|, \tag{59}$$

where the definition of constants $C_\sigma^V$ and $L_\sigma$ can be found in Section I. Hence

$$\|\nabla_\lambda V_\sigma^L(\theta,\lambda)|_{\theta_1} - \nabla_\lambda V_\sigma^L(\theta,\lambda)|_{\theta_2}\| = \|V_{\sigma,c}^{\pi_{\theta_1}}(\rho) - V_{\sigma,c}^{\pi_{\theta_2}}(\rho)\| \le C_\sigma^V\|\theta_1-\theta_2\|, \tag{60}$$

$$\|\nabla_\lambda V_\sigma^L(\theta,\lambda)|_{\lambda_1} - \nabla_\lambda V_\sigma^L(\theta,\lambda)|_{\lambda_2}\| = 0. \tag{61}$$

Similarly, we have that

$$\|\nabla_\theta V_\sigma^L(\theta,\lambda)|_{\theta_1} - \nabla_\theta V_\sigma^L(\theta,\lambda)|_{\theta_2}\| \le (1+\lambda)L_\sigma\|\theta_1-\theta_2\| \le (1+\Lambda^*)L_\sigma\|\theta_1-\theta_2\|, \tag{62}$$

$$\|\nabla_\theta V_\sigma^L(\theta,\lambda)|_{\lambda_1} - \nabla_\theta V_\sigma^L(\theta,\lambda)|_{\lambda_2}\| \le |(\lambda_1-\lambda_2)|\max_{\theta\in\Theta}\|\nabla_\theta V_{\sigma,c}^{\pi_\theta}(\rho)\| \le C_\sigma^V|\lambda_1-\lambda_2|. \tag{63}$$

This completes the proof. □

## H Proof of Theorem 2

We then prove Theorem 2. Our proof extends the one in **?** 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*

$$V_\sigma^L(\theta_{t+1}, \lambda_{t+1}) - V_\sigma^L(\theta_t, \lambda_{t+1}) \geq \langle \theta_{t+1} - \theta_t, -\hat{g}(\theta_t, \lambda_{t+1}) + g(\theta_t, \lambda_{t+1})\rangle + \left(\frac{\mu_t}{2} + \nu\right)\|\theta_{t+1} - \theta_t\|^2. \tag{64}$$

*Proof.* Note that from the update of $\theta_t$ and proposition of projection, it implies that

$$\left\langle \theta_t + \frac{1}{\alpha_t}\hat{g}(\theta_t, \lambda_{t+1}) - \theta_{t+1}, \theta_t - \theta_{t+1}\right\rangle \leq 0. \tag{65}$$

Hence

$$\langle \hat{g}(\theta_t, \lambda_{t+1}) - \alpha_t(\theta_{t+1} - \theta_t), \theta_t - \theta_{t+1}\rangle \leq 0. \tag{66}$$

From Lemma 3, we have that

$$V_\sigma^L(\theta_{t+1}, \lambda_{t+1}) - V_\sigma^L(\theta_t, \lambda_{t+1}) \geq \langle \theta_{t+1} - \theta_t, g(\theta_t, \lambda_{t+1})\rangle - \frac{(1+\Lambda^*)L_\sigma}{2}\|\theta_{t+1} - \theta_t\|^2. \tag{67}$$

Summing up the two inequalities implies

$$\begin{aligned}
&V_\sigma^L(\theta_{t+1}, \lambda_{t+1}) - V_\sigma^L(\theta_t, \lambda_{t+1}) \\
&\geq \langle \theta_{t+1} - \theta_t, -\hat{g}(\theta_t, \lambda_{t+1}) + g(\theta_t, \lambda_{t+1}) + \alpha_t(\theta_{t+1} - \theta_t)\rangle - \frac{(1+\Lambda^*)L_\sigma}{2}\|\theta_{t+1} - \theta_t\|^2 \\
&\geq \langle \theta_{t+1} - \theta_t, -\hat{g}(\theta_t, \lambda_{t+1}) + g(\theta_t, \lambda_{t+1})\rangle + \left(\alpha_t - \frac{L_\sigma(1+\Lambda^*)}{2}\right)\|\theta_{t+1} - \theta_t\|^2 \\
&\geq \langle \theta_{t+1} - \theta_t, -\hat{g}(\theta_t, \lambda_{t+1}) + g(\theta_t, \lambda_{t+1})\rangle + \left(\frac{\mu_t}{2} + \nu\right)\|\theta_{t+1} - \theta_t\|^2,
\end{aligned} \tag{68}$$

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*

$$\begin{aligned}
&V_\sigma^L(\theta_{t+1}, \lambda_{t+1}) - V_\sigma^L(\theta_t, \lambda_t) \\
&\geq (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}))(\lambda_{t+1} - \lambda_t) + \langle \theta_{t+1} - \theta_t, -\hat{g}(\theta_t, \lambda_{t+1}) + g(\theta_t, \lambda_{t+1})\rangle - \frac{\xi(C_\sigma^V)^2}{2}\|\theta_t - \theta_{t-1}\|^2 \\
&\quad + \left(\frac{\mu_t}{2} + \nu\right)\|\theta_{t+1} - \theta_t\|^2 + \frac{b_{t-1}}{2}(\lambda_t^2 - \lambda_{t+1}^2) - \frac{1}{\xi}(\lambda_{t+1} - \lambda_t)^2 - \frac{1}{2\xi}(\lambda_t - \lambda_{t-1})^2.
\end{aligned} \tag{69}$$

*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

$$|\nabla_\lambda \tilde{V}_t(\theta_t, \lambda_{t+1}) - \nabla_\lambda \tilde{V}_t(\theta_t, \lambda_t)| = b_{t-1}|\lambda_{t+1} - \lambda_t| \leq b_0|\lambda_{t+1} - \lambda_t|, \tag{70}$$

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 $\lambda$, hence we have

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

$$
\begin{aligned}
&\geq b_{t-1}\left(\frac{b_{t-1}+b_0}{b_{t-1}+b_0}\right)(\lambda_{t+1}-\lambda_t)^2 \\
&= \frac{b_{t-1}b_0}{b_{t-1}+b_0}(\lambda_{t+1}-\lambda_t)^2 + \frac{b_{t-1}^2}{b_{t-1}+b_0}(\lambda_{t+1}-\lambda_t)^2 \\
&\geq \frac{b_{t-1}b_0}{b_{t-1}+b_0}(\lambda_{t+1}-\lambda_t)^2 + \frac{1}{b_{t-1}+b_0}(\nabla_\lambda \tilde{V}_t(\theta_t,\lambda_{t+1}) - \nabla_\lambda \tilde{V}_t(\theta_t,\lambda_t))^2,
\end{aligned} \tag{71}
$$

where the last inequality is from eq. (70).

Recall the update of $\lambda_t$ in Algorithm 3 which can be rewritten as

$$
\lambda_{t+1} = \prod\nolimits_{[0,\Lambda^*]}\left(\lambda_t - \frac{1}{\beta_t}\nabla_\lambda \tilde{V}_{t+1}(\theta_t,\lambda_t) + \frac{1}{\beta_t}(f(\theta_t) - \hat{f}(\theta_t))\right), \tag{72}
$$

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

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

Hence setting $\lambda = \lambda_k$ implies that

$$
(\beta_t(\lambda_{t+1}-\lambda_t) + \nabla_\lambda \tilde{V}_{t+1}(\theta_t,\lambda_t) - f(\theta_t) + \hat{f}(\theta_t))(\lambda_t - \lambda_{t+1}) \geq 0. \tag{74}
$$

Similarly, we have that

$$
(\beta_t(\lambda_t-\lambda_{t-1}) + \nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}) - f(\theta_{t-1}) + \hat{f}(\theta_{t-1}))(\lambda_{t+1} - \lambda_t) \geq 0. \tag{75}
$$

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

$$
\begin{aligned}
&\tilde{V}_t(\theta_t,\lambda_{t+1}) - \tilde{V}_t(\theta_t,\lambda_t) \\
&\geq (\nabla_\lambda \tilde{V}_t(\theta_t,\lambda_t))(\lambda_{t+1}-\lambda_t) \\
&= (\nabla_\lambda \tilde{V}_t(\theta_t,\lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))(\lambda_{t+1}-\lambda_t) + (\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))(\lambda_{t+1}-\lambda_t) \\
&\overset{(a)}{\geq} (\nabla_\lambda \tilde{V}_t(\theta_t,\lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1})(\lambda_{t-1}), \lambda_{t+1}-\lambda_t) \\
&\quad + (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - \beta_t(\lambda_t - \lambda_{t-1}))(\lambda_{t+1}-\lambda_t),
\end{aligned} \tag{76}
$$

where $(a)$ is from eq. (75). The first term in the RHS of eq. (76) can be further bounded as follows.

$$
\begin{aligned}
&(\nabla_\lambda \tilde{V}_t(\theta_t,\lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))(\lambda_{t+1}-\lambda_t) \\
&= (\nabla_\lambda \tilde{V}_t(\theta_t,\lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t))(\lambda_{t+1}-\lambda_t) \\
&\quad + (\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))(\lambda_{t+1}-\lambda_t) \\
&= (\nabla_\lambda \tilde{V}_t(\theta_t,\lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t))(\lambda_{t+1}-\lambda_t) \\
&\quad + (\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))(\lambda_t-\lambda_{t-1}) \\
&\quad + m_{t+1}(\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1})),
\end{aligned} \tag{77}
$$

where $m_{t+1} \triangleq (\lambda_{t+1}-\lambda_t) - (\lambda_t - \lambda_{t-1})$. Plug it in eq. (76) and we have that

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

$$+\underbrace{(\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t)-\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))m_{t+1}}_{(c)}. \tag{78}$$

We then provide bounds for each term in eq. (78) as follows.

Term $(a)$ can be bounded as follows:

$$\begin{aligned}
&(\nabla_\lambda \tilde{V}_t(\theta_t,\lambda_t)-\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t))(\lambda_{t+1}-\lambda_t)\\
&=(\nabla_\lambda V_\sigma^L(\theta_t,\lambda_t)-\nabla_\lambda V_\sigma^L(\theta_{t-1},\lambda_t))(\lambda_{t+1}-\lambda_t)\\
&\geq \frac{-(\lambda_{t+1}-\lambda_t)^2}{2\xi}-\frac{\xi}{2}(\nabla_\lambda V_\sigma^L(\theta_t,\lambda_t)-\nabla_\lambda V_\sigma^L(\theta_{t-1},\lambda_t))^2\\
&\geq \frac{-(\lambda_{t+1}-\lambda_t)^2}{2\xi}-\frac{\xi(C_\sigma^V)^2}{2}\|\theta_t-\theta_{t-1}\|^2,
\end{aligned} \tag{79}$$

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

Term $(b)$ can be bounded as follows:

$$\begin{aligned}
&(\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t)-\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))(\lambda_t-\lambda_{t-1})\\
&\geq \frac{1}{b_{t-1}+b_0}(\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t)-\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))^2,
\end{aligned} \tag{80}$$

which is from eq. (71).

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

$$\begin{aligned}
&m_{t+1}(\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t)-\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))\\
&\geq -\frac{\xi}{2}(\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t)-\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))^2-\frac{1}{2\xi}m_{t+1}^2
\end{aligned} \tag{81}$$

Moreover, it can be shown that

$$\frac{1}{\xi}(\lambda_{t+1}-\lambda_t)(\lambda_t-\lambda_{t-1})=\frac{1}{2\xi}(\lambda_{t+1}-\lambda_t)^2+\frac{1}{2\xi}(\lambda_t-\lambda_{t-1})^2-\frac{1}{2\xi}m_{t+1}^2. \tag{82}$$

Plug eq. (79) to eq. (82) in 78, and we have that

$$\begin{aligned}
&\tilde{V}_t(\theta_t,\lambda_{t+1})-\tilde{V}_t(\theta_t,\lambda_t)\\
&\geq (f(\theta_{t-1})-\hat{f}(\theta_{t-1}))(\lambda_{t+1}-\lambda_t)-\beta_t(\lambda_t-\lambda_{t-1})(\lambda_{t+1}-\lambda_t)\\
&\quad+(\nabla_\lambda \tilde{V}_t(\theta_t,\lambda_t)-\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t))(\lambda_{t+1}-\lambda_t)+(\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t)-\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))(\lambda_t-\lambda_{t-1})\\
&\quad+m_{t+1}(\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t)-\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))\\
&\geq (f(\theta_{t-1})-\hat{f}(\theta_{t-1}))(\lambda_{t+1}-\lambda_t)-\frac{1}{2\xi}(\lambda_{t+1}-\lambda_t)^2-\frac{1}{2\xi}(\lambda_t-\lambda_{t-1})^2+\frac{1}{2\xi}m_{t+1}^2\\
&\quad-\frac{(\lambda_{t+1}-\lambda_t)^2}{2\xi}-\frac{\xi(C_\sigma^V)^2}{2}\|\theta_t-\theta_{t-1}\|^2+\frac{1}{b_{t-1}+b_0}(\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t)-\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))^2\\
&\quad-\frac{\xi}{2}(\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t)-\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))^2-\frac{1}{2\xi}m_{t+1}^2\\
&\geq (f(\theta_{t-1})-\hat{f}(\theta_{t-1}))(\lambda_{t+1}-\lambda_t)-\frac{1}{\xi}(\lambda_{t+1}-\lambda_t)^2-\frac{1}{2\xi}(\lambda_t-\lambda_{t-1})^2-\frac{\xi(C_\sigma^V)^2}{2}\|\theta_t-\theta_{t-1}\|^2.
\end{aligned} \tag{83}$$

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

$$\begin{aligned}
&\tilde{V}_t(\theta_t,\lambda_{t+1})-\tilde{V}_t(\theta_t,\lambda_t)\\
&=V_\sigma^L(\theta_t,\lambda_{t+1})+\frac{b_{t-1}}{2}\lambda_{t+1}^2-V_\sigma^L(\theta_t,\lambda_t)-\frac{b_{t-1}}{2}\lambda_t^2.
\end{aligned} \tag{84}$$

Then we have that

$$\begin{aligned}
&V_\sigma^L(\theta_t, \lambda_{t+1}) - V_\sigma^L(\theta_t, \lambda_t) \\
&\geq \frac{b_{t-1}}{2}(\lambda_t^2 - \lambda_{t+1}^2) + (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}))(\lambda_{t+1} - \lambda_t) \\
&\quad - \frac{1}{\xi}(\lambda_{t+1} - \lambda_t)^2 - \frac{1}{2\xi}(\lambda_t - \lambda_{t-1})^2 - \frac{\xi(C_\sigma^V)^2}{2}\|\theta_t - \theta_{t-1}\|^2. \qquad (85)
\end{aligned}$$

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

$$\begin{aligned}
&V_\sigma^L(\theta_{t+1}, \lambda_{t+1}) - V_\sigma^L(\theta_t, \lambda_t) \\
&\geq (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}))(\lambda_{t+1} - \lambda_t) + \langle \theta_{t+1} - \theta_t, -\hat{g}(\theta_t, \lambda_{t+1}) + g(\theta_t, \lambda_{t+1})\rangle - \frac{\xi(C_\sigma^V)^2}{2}\|\theta_t - \theta_{t-1}\|^2 \\
&\quad + \left(\frac{\mu_t}{2} + \nu\right)\|\theta_{t+1} - \theta_t\|^2 + \frac{b_{t-1}}{2}(\lambda_t^2 - \lambda_{t+1}^2) - \frac{1}{\xi}(\lambda_{t+1} - \lambda_t)^2 - \frac{1}{2\xi}(\lambda_t - \lambda_{t-1})^2. \qquad (86)
\end{aligned}$$

□

**Lemma 9.** *Define*

$$\begin{aligned}
F_{t+1} \triangleq &-\frac{8}{\xi^2 b_{t+1}}(\lambda_t - \lambda_{t+1})^2 - \frac{8}{\xi}\left(1 - \frac{b_t}{b_{t+1}}\right)\lambda_{t+1}^2 + V_\sigma^L(\theta_{t+1}, \lambda_{t+1}) + \frac{b_t}{2}\lambda_{t+1}^2 \\
&+ \left(-\frac{16(C_\sigma^V)^2}{\xi b_{t+1}^2} - \frac{\xi(C_\sigma^V)^2}{2}\right)\|\theta_{t+1} - \theta_t\|^2 + \left(\frac{8}{\xi} - \frac{1}{2\xi}\right)(\lambda_{t+1} - \lambda_t)^2, \qquad (87)
\end{aligned}$$

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

$$\begin{aligned}
&F_{t+1} - F_t \\
&\geq S_t + \left(\frac{\mu_t}{2} + \nu - \frac{16(C_\sigma^V)^2}{\xi b_{t+1}^2} - \frac{\xi(C_\sigma^V)^2}{2}\right)\|\theta_{t+1} - \theta_t\|^2 + \frac{b_t - b_{t-1}}{2}\lambda_{t+1}^2 \\
&\quad + \frac{9}{10\xi}(\lambda_{t+1} - \lambda_t)^2 + \frac{8}{\xi}\left(\frac{b_t}{b_{t+1}} - \frac{b_{t-1}}{b_t}\right)\lambda_{t+1}^2, \qquad (88)
\end{aligned}$$

*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 eq. (74) and eq. (75), we have that

$$\begin{aligned}
\beta_t m_{t+1}(\lambda_t - \lambda_{t+1}) \geq &(\nabla_\lambda \tilde{V}_{t+1}(\theta_t, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))(-\lambda_t + \lambda_{t+1}) \\
&+ (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - f(\theta_t) + \hat{f}(\theta_t))(-\lambda_t + \lambda_{t+1}). \qquad (89)
\end{aligned}$$

The first term can be rewritten as

$$\begin{aligned}
&(\nabla_\lambda \tilde{V}_{t+1}(\theta_t, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))(\lambda_{t+1} - \lambda_t) \\
&= (\nabla_\lambda \tilde{V}_{t+1}(\theta_t, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t))(\lambda_{t+1} - \lambda_t) + (\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))(\lambda_{t+1} - \lambda_t) \\
&= (\nabla_\lambda \tilde{V}_{t+1}(\theta_t, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t))(\lambda_{t+1} - \lambda_t) + (\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))(\lambda_t - \lambda_{t-1}) \\
&\quad + m_{t+1}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1})). \qquad (90)
\end{aligned}$$

The first term in eq. (90) can be bounded as

$$\begin{aligned}
&(\nabla_\lambda \tilde{V}_{t+1}(\theta_t, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t))(\lambda_{t+1} - \lambda_t) \\
&= (\nabla_\lambda V_\sigma^L(\theta_t, \lambda_t) - \nabla_\lambda V_\sigma^L(\theta_{t-1}, \lambda_t))(\lambda_{t+1} - \lambda_t) + (b_t\lambda_t - b_{t-1}\lambda_t)(\lambda_{t+1} - \lambda_t)
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\geq} -\frac{1}{2h}(\nabla_\lambda V_\sigma^L(\theta_t,\lambda_t) - \nabla_\lambda V_\sigma^L(\theta_{t-1},\lambda_t))^2 - \frac{h}{2}(\lambda_{t+1}-\lambda_t)^2 \\
&\quad + \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1}^2 - \lambda_t^2) - \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1}-\lambda_t)^2 \\
&\overset{(b)}{\geq} -\frac{(C_\sigma^V)^2}{2h}\|\theta_t - \theta_{t-1}\|^2 - \frac{h}{2}(\lambda_{t+1}-\lambda_t)^2 \\
&\quad + \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1}^2 - \lambda_t^2) - \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1}-\lambda_t)^2,
\end{aligned} \tag{91}
$$

where $(a)$ is from the Cauchy–Schwarz inequality and $(b)$ is from the $C_\sigma^V$-smoothness of $V_\sigma^L$, for any $h > 0$. Similar to eq. (71), the second term in eq. (90) can be bounded as

$$
\begin{aligned}
&(\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))(\lambda_t - \lambda_{t-1}) \\
&\geq \frac{b_{t-1}b_0}{b_{t-1}+b_0}(\lambda_t-\lambda_{t-1})^2 + \frac{1}{b_{t-1}+b_0}(\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))^2.
\end{aligned} \tag{92}
$$

The third term in eq. (90) can be bounded as

$$
\begin{aligned}
&m_{t+1}(\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1})) \\
&\geq -\frac{\xi}{2}(\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))^2 - \frac{1}{2\xi}m_{t+1}^2.
\end{aligned} \tag{93}
$$

Hence combine eq. (91) to eq. (92) and plug in eq. (90), we have that

$$
\begin{aligned}
&(\nabla_\lambda \tilde{V}_{t+1}(\theta_t,\lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))(\lambda_{t+1} - \lambda_t) \\
&\geq -\frac{(C_\sigma^V)^2}{2h}\|\theta_t - \theta_{t-1}\|^2 - \frac{h}{2}(\lambda_{t+1}-\lambda_t)^2 \\
&\quad + \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1}^2 - \lambda_t^2) - \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1}-\lambda_t)^2 \\
&\quad + \frac{b_{t-1}b_0}{b_{t-1}+b_0}(\lambda_t-\lambda_{t-1})^2 + \frac{1}{b_{t-1}+b_0}(\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))^2 \\
&\quad - \frac{\xi}{2}(\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))^2 - \frac{1}{2\xi}m_{t+1}^2.
\end{aligned} \tag{94}
$$

Hence eq. (89) can be further bounded as

$$
\begin{aligned}
&(\beta_t m_{t+1})(\lambda_t - \lambda_{t+1}) \\
&\geq (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - f(\theta_t) + \hat{f}(\theta_t))(-\lambda_t + \lambda_{t+1}) \\
&\quad - \frac{(C_\sigma^V)^2}{2h}\|\theta_t - \theta_{t-1}\|^2 - \frac{h}{2}(\lambda_{t+1}-\lambda_t)^2 \\
&\quad + \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1}^2 - \lambda_t^2) - \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1}-\lambda_t)^2 \\
&\quad + \frac{b_{t-1}b_0}{b_{t-1}+b_0}(\lambda_t-\lambda_{t-1})^2 + \frac{1}{b_{t-1}+b_0}(\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))^2 \\
&\quad - \frac{\xi}{2}(\nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1},\lambda_{t-1}))^2 - \frac{1}{2\xi}m_{t+1}^2.
\end{aligned} \tag{95}
$$

It can be directly verified that

$$
m_{t+1}(\lambda_t - \lambda_{t+1}) = \frac{1}{2}(\lambda_t - \lambda_{t-1})^2 - \frac{1}{2}(\lambda_t - \lambda_{t+1})^2 - \frac{m_{t+1}^2}{2}. \tag{96}
$$

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

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

$$
\begin{aligned}
&\geq (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - f(\theta_t) + \hat{f}(\theta_t))(-\lambda_t + \lambda_{t+1}) \\
&\quad - \frac{(C_\sigma^V)^2}{2h}\|\theta_t - \theta_{t-1}\|^2 - \frac{h}{2}(\lambda_{t+1} - \lambda_t)^2 \\
&\quad + \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1}^2 - \lambda_t^2) - \frac{(b_t - b_{t-1})}{2}(\lambda_{t+1} - \lambda_t)^2 \\
&\quad + \frac{b_{t-1}b_0}{b_{t-1} + b_0}(\lambda_t - \lambda_{t-1})^2 + \frac{1}{b_{t-1} + b_0}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))^2 \\
&\quad - \frac{\xi}{2}(\nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_t) - \nabla_\lambda \tilde{V}_t(\theta_{t-1}, \lambda_{t-1}))^2 - \frac{1}{2\xi}m_{t+1}^2. \qquad (97)
\end{aligned}
$$

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

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

Re-arrange the terms, it follows that

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

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

$$
\begin{aligned}
&- \frac{1}{\xi^2 b_t}(\lambda_t - \lambda_{t+1})^2 - \frac{1}{\xi}\left(1 - \frac{b_{t-1}}{b_t}\right)\lambda_{t+1}^2 \\
&\geq -\frac{1}{\xi^2 b_t}(\lambda_t - \lambda_{t-1})^2 - \frac{1}{\xi}\left(1 - \frac{b_{t-1}}{b_t}\right)\lambda_t^2 + \frac{2}{\xi b_t}(f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - f(\theta_t) + \hat{f}(\theta_t))(-\lambda_t + \lambda_{t+1}) \\
&\quad - \frac{(C_\sigma^V)^2}{h\xi b_t}\|\theta_t - \theta_{t-1}\|^2 - \frac{h}{\xi b_t}(\lambda_{t+1} - \lambda_t)^2 + \frac{1}{\xi}(\lambda_t - \lambda_{t-1})^2. \qquad (100)
\end{aligned}
$$

If we set $h = \frac{b_t}{2}$, eq. (100) can be rewritten as

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

$$- \frac{2(C_\sigma^V)^2}{\xi b_t^2}\|\theta_t - \theta_{t-1}\|^2 - \frac{1}{2\xi}(\lambda_{t+1} - \lambda_t)^2 + \frac{1}{\xi}(\lambda_t - \lambda_{t-1})^2. \tag{101}$$

Further we have that

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

Re-arranging the terms in eq. (102) implies that

$$\begin{aligned}
&- \frac{1}{\xi^2 b_{t+1}}(\lambda_t - \lambda_{t+1})^2 - \frac{1}{\xi}\left(1 - \frac{b_t}{b_{t+1}}\right)\lambda_{t+1}^2 - \left(-\frac{1}{\xi^2 b_t}(\lambda_t - \lambda_{t-1})^2 - \frac{1}{\xi}\left(1 - \frac{b_{t-1}}{b_t}\right)\lambda_t^2\right) \\
&\geq -\left(\frac{1}{\xi^2 b_{t+1}} - \frac{1}{\xi^2 b_t}\right)(\lambda_t - \lambda_{t+1})^2 - \frac{1}{\xi}\left(\frac{b_{t-1}}{b_t} - \frac{b_t}{b_{t+1}}\right)\lambda_{t+1}^2 \\
&\quad - \frac{2(C_\sigma^V)^2}{\xi b_t^2}\|\theta_t - \theta_{t-1}\|^2 - \frac{1}{2\xi}(\lambda_{t+1} - \lambda_t)^2 + \frac{1}{\xi}(\lambda_t - \lambda_{t-1})^2 \\
&\quad + \frac{2}{\xi b_t}(f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - f(\theta_t) + \hat{f}(\theta_t))(-\lambda_t + \lambda_{t+1}) \\
&\geq -\frac{7}{10\xi}(-\lambda_t + \lambda_{t+1})^2 - \frac{2(C_\sigma^V)^2}{\xi b_t^2}\|\theta_t - \theta_{t-1}\|^2 + \frac{1}{\xi}(\lambda_t - \lambda_{t-1})^2 + \frac{1}{\xi}\left(\frac{b_t}{b_{t+1}} - \frac{b_{t-1}}{b_t}\right)\lambda_{t+1}^2 \\
&\quad + \frac{2}{\xi b_t}(f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - f(\theta_t) + \hat{f}(\theta_t))(-\lambda_t + \lambda_{t+1}),
\end{aligned} \tag{103}$$

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

$$\begin{aligned}
&V_\sigma^L(\theta_{t+1}, \lambda_{t+1}) - V_\sigma^L(\theta_t, \lambda_t) \\
&\geq (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}))(\lambda_{t+1} - \lambda_t) + \langle \theta_{t+1} - \theta_t, -\hat{g}(\theta_t, \lambda_{t+1}) + g(\theta_t, \lambda_{t+1})\rangle - \frac{\xi(C_\sigma^V)^2}{2}\|\theta_t - \theta_{t-1}\|^2 \\
&\quad + \left(\frac{\mu_t}{2} + \nu\right)\|\theta_{t+1} - \theta_t\|^2 + \frac{b_{t-1}}{2}(\lambda_t^2 - \lambda_{t+1}^2) - \frac{1}{\xi}(\lambda_{t+1} - \lambda_t)^2 - \frac{1}{2\xi}(\lambda_t - \lambda_{t-1})^2.
\end{aligned} \tag{104}$$

Combine both inequality together, and we further have that

$$\begin{aligned}
&- \frac{8}{\xi^2 b_{t+1}}(\lambda_t - \lambda_{t+1})^2 - \frac{8}{\xi}\left(1 - \frac{b_t}{b_{t+1}}\right)\lambda_{t+1}^2 - \left(-\frac{8}{\xi^2 b_t}(\lambda_t - \lambda_{t-1})^2 - \frac{8}{\xi}\left(1 - \frac{b_{t-1}}{b_t}\right)\lambda_t^2\right) \\
&\quad + V_\sigma^L(\theta_{t+1}, \lambda_{t+1}) - V_\sigma^L(\theta_t, \lambda_t) \\
&\geq -\frac{28}{5\xi}(-\lambda_t + \lambda_{t+1})^2 - \frac{16(C_\sigma^V)^2}{\xi b_t^2}\|\theta_t - \theta_{t-1}\|^2 + \frac{8}{\xi}(\lambda_t - \lambda_{t-1})^2 + \frac{8}{\xi}\left(\frac{b_t}{b_{t+1}} - \frac{b_{t-1}}{b_t}\right)\lambda_{t+1}^2 \\
&\quad + \frac{16}{b_t \xi}(f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - f(\theta_t) + \hat{f}(\theta_t))(-\lambda_t + \lambda_{t+1}) \\
&\quad + (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}))(\lambda_{t+1} - \lambda_t) + \langle \theta_{t+1} - \theta_t, -\hat{g}(\theta_t, \lambda_{t+1}) + g(\theta_t, \lambda_{t+1})\rangle - \frac{\xi(C_\sigma^V)^2}{2}\|\theta_t - \theta_{t-1}\|^2 \\
&\quad + \left(\frac{\mu_t}{2} + \nu\right)\|\theta_{t+1} - \theta_t\|^2 + \frac{b_{t-1}}{2}(\lambda_t^2 - \lambda_{t+1}^2) - \frac{1}{\xi}(\lambda_{t+1} - \lambda_t)^2 - \frac{1}{2\xi}(\lambda_t - \lambda_{t-1})^2 \\
&= S_t + \left(-\frac{16(C_\sigma^V)^2}{\xi b_t^2} - \frac{\xi(C_\sigma^V)^2}{2}\right)\|\theta_t - \theta_{t-1}\|^2 + \left(-\frac{28}{5\xi} - \frac{1}{\xi}\right)(-\lambda_t + \lambda_{t+1})^2 + \frac{b_{t-1}}{2}(\lambda_t^2 - \lambda_{t+1}^2) \\
&\quad + \left(\frac{8}{\xi} - \frac{1}{2\xi}\right)(\lambda_t - \lambda_{t-1})^2 + \frac{8}{\xi}\left(\frac{b_t}{b_{t+1}} - \frac{b_{t-1}}{b_t}\right)\lambda_{t+1}^2 + \left(\frac{\mu_t}{2} + \nu\right)\|\theta_{t+1} - \theta_t\|^2,
\end{aligned} \tag{105}$$

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

$$
\begin{aligned}
&-\frac{8}{\xi^2 b_{t+1}}(\lambda_t - \lambda_{t+1})^2 - \frac{8}{\xi}\left(1 - \frac{b_t}{b_{t+1}}\right)\lambda_{t+1}^2 - \left(-\frac{8}{\xi^2 b_t}(\lambda_t - \lambda_{t-1})^2 - \frac{8}{\xi}\left(1 - \frac{b_{t-1}}{b_t}\right)\lambda_t^2\right) \\
&\quad + V_\sigma^L(\theta_{t+1}, \lambda_{t+1}) - V_\sigma^L(\theta_t, \lambda_t) + \frac{b_t}{2}\lambda_{t+1}^2 - \frac{b_{t-1}}{2}\lambda_t^2 \\
&\quad + \left(-\frac{16(C_\sigma^V)^2}{\xi b_{t+1}^2} - \frac{\xi(C_\sigma^V)^2}{2}\right)\|\theta_{t+1} - \theta_t\|^2 - \left(-\frac{16(C_\sigma^V)^2}{\xi b_t^2} - \frac{\xi(C_\sigma^V)^2}{2}\right)\|\theta_t - \theta_{t-1}\|^2 \\
&\quad + \left(\frac{8}{\xi} - \frac{1}{2\xi}\right)(\lambda_{t+1} - \lambda_t)^2 - \left(\frac{8}{\xi} - \frac{1}{2\xi}\right)(\lambda_t - \lambda_{t-1})^2 \\
&\geq S_t + \left(\frac{\mu_t}{2} + \nu - \frac{16(C_\sigma^V)^2}{\xi b_{t+1}^2} - \frac{\xi(C_\sigma^V)^2}{2}\right)\|\theta_{t+1} - \theta_t\|^2 + \frac{b_t - b_{t-1}}{2}\lambda_{t+1}^2 \\
&\quad + \left(\frac{8}{\xi} - \frac{1}{2\xi} - \frac{28}{5\xi} - \frac{1}{\xi}\right)(\lambda_{t+1} - \lambda_t)^2 + \frac{8}{\xi}\left(\frac{b_t}{b_{t+1}} - \frac{b_{t-1}}{b_t}\right)\lambda_{t+1}^2 \\
&= S_t + \left(\frac{\mu_t}{2} + \nu - \frac{16(C_\sigma^V)^2}{\xi b_{t+1}^2} - \frac{\xi(C_\sigma^V)^2}{2}\right)\|\theta_{t+1} - \theta_t\|^2 + \frac{b_t - b_{t-1}}{2}\lambda_{t+1}^2 \\
&\quad + \frac{9}{10\xi}(\lambda_{t+1} - \lambda_t)^2 + \frac{8}{\xi}\left(\frac{b_t}{b_{t+1}} - \frac{b_{t-1}}{b_t}\right)\lambda_{t+1}^2, \qquad (106)
\end{aligned}
$$

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* $\xi > \frac{2\nu + (1+\Lambda^*)L_\sigma}{(C_\sigma^V)^2}$, $\nu$ *is any positive number and* $\tau$ *is any number greater than 2. Moreover, set* $\epsilon_{est} = \frac{1}{t^{0.5}L_\Omega}\frac{1}{32t^{0.25}\Lambda^* + 2\Lambda^* + \frac{1}{\alpha_1}(1+\Lambda^*)C_\sigma^V}\frac{19^2\epsilon^2}{3200\xi(\tau-2)(C_\sigma^V)^2 u L_\Omega} = \mathcal{O}(\frac{\epsilon 2}{\sigma^2 t^{0.75}})$, *then*

$$
\min_{1\leq t\leq T}\|G_t\|^2 \leq (1+\sqrt{2})\epsilon, \qquad (107)
$$

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

**Remark 1.** *Note that the sample complexity of robust TD algorithm to achieve an* $\epsilon_{est}$*-error bound is* $\mathcal{O}(\epsilon_{est}^{-2})$, *hence the sample complexity* $T_{inner}$ *at the time step* $t$ *is* $\mathcal{O}(\epsilon_{est}^{-2}) = \mathcal{O}(\frac{t^{1.5}}{\epsilon^8})$. *Thus the total sample complexity to find an* $\epsilon$*-stationary solution is* $\sum_{t=1}^{T}\frac{t^{1.5}}{\epsilon^8} = \mathcal{O}(\epsilon^{-18})$. *This increase in complexity is due to the model-free setting and the estimation of robust value functions.*

*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 eq. (106) can be rewritten as

$$
\begin{aligned}
F_{t+1} - F_t \geq S_t + p_t\|\theta_{t+1} - \theta_t\|^2 + \frac{b_t - b_{t-1}}{2}\lambda_{t+1}^2 \\
+ \frac{9}{10\xi}(\lambda_{t+1} - \lambda_t)^2 + \frac{8}{\xi}\left(\frac{b_t}{b_{t+1}} - \frac{b_{t-1}}{b_t}\right)\lambda_{t+1}^2. \qquad (108)
\end{aligned}
$$

From the definition, we have that

$$
G_t = \begin{bmatrix} \beta_t\left(\lambda_t - \prod_{[0,\Lambda^*]}\left(\lambda_t - \frac{1}{\beta_t}\left(\nabla_\lambda V_\sigma^L(\theta_t, \lambda_t)\right)\right)\right) \\ \alpha_t\left(\theta_t - \prod_\Theta\left(\theta_t + \frac{1}{\alpha_t}\left(\nabla_\theta V_\sigma^L(\theta_t, \lambda_t)\right)\right)\right) \end{bmatrix}, \qquad (109)
$$

and denote by

$$\tilde{G}_t \triangleq \begin{bmatrix} \beta_t \left(\lambda_t - \prod_{[0,\Lambda^*]} \left(\lambda_t - \frac{1}{\beta_t}\left(\nabla_\lambda \tilde{V}_t(\theta_t,\lambda_t)\right)\right)\right) \\ \alpha_t \left(\theta_t - \prod_{\Theta} \left(\theta_t + \frac{1}{\alpha_t}\left(\nabla_\theta \tilde{V}_t(\theta_t,\lambda_t)\right)\right)\right) \end{bmatrix}. \tag{110}$$

It can be verified that

$$\|G_t\| - \|\tilde{G}_t\| \leq b_{t-1}|\lambda_t|. \tag{111}$$

From Theorem 4.2 in **?**, it can be shown that

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

and

$$M_1 \geq \frac{2(\nu+\mu_t)^2}{p_t^2}. \tag{113}$$

Hence

$$\|\tilde{G}_t\|^2 \leq M_1 p_t^2\|\theta_{t+1}-\theta_t\|^2 + \left(2(C_\sigma^V)^2 + \frac{1}{\xi^2}\right)(\lambda_{t+1}-\lambda_t)^2. \tag{114}$$

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 eq. (108), we have that

$$u_t\|\tilde{G}_t\|^2 \leq F_{t+1} - F_t - S_t - \frac{b_t - b_{t-1}}{2}\lambda_{t+1}^2 - \frac{8}{\xi}\left(\frac{b_t}{b_{t+1}} - \frac{b_{t-1}}{b_t}\right)\lambda_{t+1}^2. \tag{115}$$

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

$$\begin{aligned}\sum_{t=1}^T u_t\|\tilde{G}_t\|^2 &\leq F_{T+1} - F_1 - \sum_{t=1}^T S_t + \frac{8}{\xi}\left(\frac{b_0}{b_1}\lambda_2^2 - \frac{b_T}{b_{T+1}}\lambda_{T+1}^2\right) + \left(\frac{b_0-b_T}{2}(\Lambda^*)^2\right) \\ &\leq F_{T+1} - F_1 - \sum_{t=1}^T S_t + \frac{8}{\xi}\frac{b_0}{b_1}(\Lambda^*)^2 + \left(\frac{b_0-b_T}{2}(\Lambda^*)^2\right),\end{aligned} \tag{116}$$

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

$$\begin{aligned}\max_{t\geq 1}\max_{\theta\in\Theta,\lambda\in[0,\Lambda^*]} F_t &= \max\Bigg\{ -\frac{8}{\xi^2 b_{t+1}}(\lambda_t-\lambda_{t+1})^2 - \frac{8}{\xi}\left(1-\frac{b_t}{b_{t+1}}\right)\lambda_{t+1}^2 + V_\sigma^L(\theta_{t+1},\lambda_{t+1}) + \frac{b_t}{2}\lambda_{t+1}^2 \\ &\quad + \left(-\frac{16(C_\sigma^V)^2}{\xi b_{t+1}^2} - \frac{\xi(C_\sigma^V)^2}{2}\right)\|\theta_{t+1}-\theta_t\|^2 + \left(\frac{8}{\xi} - \frac{1}{2\xi}\right)(\lambda_{t+1}-\lambda_t)^2\Bigg\} \\ &\leq \frac{1.6}{\xi}(\Lambda^*)^2 + (1+\Lambda^*)(2C_\sigma) + \frac{b_1}{2}(\Lambda^*)^2 + \frac{15}{2\xi}(\Lambda^*)^2 \\ &\triangleq F^*,\end{aligned} \tag{117}$$

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

$$\sum_{t=1}^T u_t\|\tilde{G}_t\|^2 \leq F^* - F_1 - \sum_{t=1}^T S_t + \frac{8}{\xi}(\Lambda^*)^2 + \left(\frac{b_0}{2}(\Lambda^*)^2\right). \tag{118}$$

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

$$\frac{\sum_{t=1}^T \frac{1}{p_t}\|\tilde{G}_t\|^2}{\sum_{t=1}^T \frac{1}{p_t}} \leq \frac{u}{\sum_{t=1}^T \frac{1}{p_t}}\left(F^* - F_1 - \sum_{t=1}^T S_t + \frac{8}{\xi}(\Lambda^*)^2 + \left(\frac{b_0}{2}(\Lambda^*)^2\right)\right). \tag{119}$$

Plug in the definition of $p_t$ then we have that

$$\frac{\sum_{t=1}^T \frac{1}{p_t}\|\tilde{G}_t\|^2}{\sum_{t=1}^T \frac{1}{p_t}} \leq \frac{3200\xi(\tau-2)(C_\sigma^V)^2 d}{19^2(\sqrt{T}-2)}\left(F^* - F_1 - \sum_{t=1}^T S_t + \frac{8}{\xi}(\Lambda^*)^2 + \left(\frac{b_0}{2}(\Lambda^*)^2\right)\right). \tag{120}$$

We moreover have that

$$\begin{aligned}|S_t| &= \Bigg|\frac{16}{b_t\xi}(f(\theta_{t-1}) - \hat{f}(\theta_{t-1}) - f(\theta_t) + \hat{f}(\theta_t))(-\lambda_t + \lambda_{t+1}) + (f(\theta_{t-1}) - \hat{f}(\theta_{t-1}))(\lambda_{t+1} - \lambda_t) \\ &\quad + \langle \theta_{t+1} - \theta_t, -\hat{g}(\theta_t, \lambda_{t+1}) + g(\theta_t, \lambda_{t+1})\rangle\Bigg| \\ &\leq 32t^{0.25}\Lambda^*(\Omega_{t-1} + \Omega_t) + 2\Lambda^*\Omega_{t-1} + \frac{1}{\alpha_t}(1+\Lambda^*)C_\sigma^V\Omega_t,\end{aligned} \tag{121}$$

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 Wang & Zou (2022) 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 $\Omega_t$ can be controlled by setting $\epsilon_{\text{est}}$.

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

$$|S_t| \leq \frac{1}{t^{0.5}}\frac{19^2\epsilon^2}{3200\xi(\tau-2)(C_\sigma^V)^2 uL_\Omega}, \tag{122}$$

and hence

$$\left|\sum_{t=1}^T S_t\right| \leq \sqrt{T}\frac{19^2\epsilon^2}{3200\xi(\tau-2)(C_\sigma^V)^2 uL_\Omega}. \tag{123}$$

Thus plug in eq. (120) and we have that

$$\frac{\sum_{t=1}^T \frac{1}{p_t}\|\tilde{G}_t\|^2}{\sum_{t=1}^T \frac{1}{p_t}} \leq \frac{3200\xi(\tau-2)(C_\sigma^V)^2 u}{19^2(\sqrt{T}-2)}K + \epsilon^2, \tag{124}$$

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

$$\frac{\sum_{t=1}^T \frac{1}{p_t}\|\tilde{G}_t\|^2}{\sum_{t=1}^T \frac{1}{p_t}} \leq 2\epsilon^2. \tag{125}$$

Similarly to Theorem 4.2 in **?**, 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 eq. (111) we finally have that

$$\min_{1\leq t\leq T}\|G_t\| \leq (1+\sqrt{2})\epsilon, \tag{126}$$

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

# I Constants

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

$$
\begin{aligned}
\bar{f} &= (1+\Lambda^*)\frac{1}{1-\gamma}, \\
\mathcal{F}_0 &= \bar{f} + K_0\frac{(\Lambda^*)^2}{\xi}, \\
D_1 &= \frac{16\tau^2}{(\tau-2)^2} + K_2\frac{(\xi L_{21}^2-\mu)^2}{(\tau-2)^2 L_{21}^4\xi^2}, \\
D_2 &= \max\left\{D_1, \frac{K_2+\xi^2 L_{21}^2}{\xi},\right\} \\
D_3 &= \mathcal{F}_0 + \frac{(\Lambda^*)^2}{\xi}, \\
L_V &= \frac{k|\mathcal{A}|}{(1-\gamma)^2}, \\
C_\sigma &= \frac{1}{1-\gamma}(1+2\gamma\delta\frac{\log|\mathcal{S}|}{\sigma}), \\
C_\sigma^V &= \frac{1}{1-\gamma}|\mathcal{A}|kC_\sigma, \\
k_B &= \frac{1}{1-\gamma+\gamma\delta}\left(|\mathcal{A}|C_\sigma l + |\mathcal{A}|kC_\sigma^V\right) + \frac{2|\mathcal{A}|^2\gamma(1-\delta)}{(1-\gamma+\gamma\delta)^2}k^2C_\sigma, \\
L_\sigma &= k_B + \frac{\gamma\delta}{1-\gamma}\left(\sqrt{|\mathcal{S}|}k_B + 2\sigma|\mathcal{S}|C_\sigma^V\frac{1}{1-\gamma+\gamma\delta}k|\mathcal{A}|C_\sigma\right), \\
b_t &= \frac{19}{20\xi t^{0.25}}, \\
M_1 &= \frac{16\tau^2}{(\tau-2)^2} + \frac{(\xi(C_\sigma^V)^2-\nu)^2}{64(\tau-2)^2(C_\sigma^V)^2\xi^2}, \\
u &= \max\left\{M_1, \frac{10+20\xi^2(C_\sigma^V)^2}{9\xi p_2}\right\}, \\
F^* &= \frac{1.6}{\xi}(\Lambda^*)^2 + (1+\Lambda^*)(2C_\sigma) + \frac{b_1}{2}(\Lambda^*)^2 + \frac{15}{2\xi}(\Lambda^*)^2, \\
K &= F^* - F_1 + \frac{8}{\xi}(\Lambda^*)^2 + \left(\frac{b_1}{2}(\Lambda^*)^2\right), \\
\mu_t &= \xi(C_\sigma^V)^2 + \frac{16\tau(C_\sigma^V)^2}{\xi(b_{t+1})^2} - 2\nu, \\
\beta_t &= \frac{1}{\xi}, \\
\alpha_t &= \nu + \mu_t.
\end{aligned} \tag{127}
$$