Policy Gradient in Robust MDPs with Global Convergence Guarantee

Qiuhao Wang [1] **Chin Pang Ho** [1] **Marek Petrik** [2]

Abstract

Robust Markov decision processes (RMDPs) provide a promising framework for computing reliable policies in the face of model errors. Many successful reinforcement learning algorithms build on variations of policy-gradient methods, but adapting these methods to RMDPs has been challenging. As a result, the applicability of RMDPs to large, practical domains remains limited. This paper proposes a new Double-Loop Robust Policy Gradient (DRPG), the first generic policy gradient method for RMDPs. In contrast with prior robust policy gradient algorithms, DRPG monotonically reduces approximation errors to guarantee convergence to a globally optimal policy in tabular RMDPs. We introduce a novel parametric transition kernel and solve the inner loop robust policy via a gradient-based method. Finally, our numerical results demonstrate the utility of our new algorithm and confirm its global convergence properties.

1. Introduction

Markov decision process (MDP) is a standard model in dynamic decision-making and reinforcement learning (Puterman, 2014; Sutton & Barto, 2018). However, a fundamental challenge with using MDPs in many applications is that model parameters, such as the transition function, are rarely known precisely. Robust Markov decision processes (RMDPs) have emerged as an effective and promising approach for mitigating the impact of model ambiguity. RMDPs assume that the transition function resides in a predefined *ambiguity set* and seek a policy that performs best for the worst-case transition function in the ambiguity set. Compared to MDPs, the performance of RMDPs is less sensitive to the parameter errors that arise when one estimates the transition function from empirical data, as is common in reinforcement learning (Xu & Mannor, 2009; Petrik, 2012; Petrik et al., 2016).

As is common in recent literature on RMDPs, we assume that the RMDP's ambiguity set satisfies certain rectangularity assumptions (Wiesemann et al., 2013; Ho et al., 2021; Panaganti & Kalathil, 2021). Albeit general RMDPs are NP-hard to solve (Wiesemann et al., 2013), they become tractable under rectangularity assumptions and can be solved using dynamic programming (Iyengar, 2005; Nilim & El Ghaoui, 2005; Kaufman & Schaefer, 2013; Ho et al., 2021). The simplest rectangularity assumption is known as (s, a)-rectangularity and allows the adversarial nature to choose the worst transition probability for each state and action independently. Because the (s, a)-rectangularity assumption can be too restrictive, we assume the more-general s-rectangular ambiguity set (Le Tallec, 2007; Wiesemann et al., 2013; Derman et al., 2021; Wang et al., 2022), which restricts the adversarial nature to choose a transition probability without observing the action. Our results also readily extend to other notions of rectangularity, including k-rectangular (Mannor et al., 2016), and r-rectangular RMDPs (Goyal & Grand-Clément, 2022).

Policy gradient techniques have gained considerable popularity in reinforcement learning due to their remarkable empirical performance and flexibility in large and complex domains (Silver et al., 2014; Xu et al., 2014). By parameterizing policies, policy gradient methods easily scale to large state and action spaces, and they also easily leverage generic optimization techniques (Konda & Tsitsiklis, 1999; Bhatnagar et al., 2009; Petrik & Subramanian, 2014; Pirotta et al., 2015; Schulman et al., 2015; 2017; Behzadian et al., 2021a). In addition, recent work shows that many policy gradient algorithms are guaranteed to find a globally-optimal policy in tabular MDPs even though they optimize a non-convex objective function (Agarwal et al., 2021; Bhandari & Russo, 2021).

As our first contribution, we propose a new policy gradient method for solving s-rectangular RMDPs. We call this method the *Double-Loop Robust Policy Gradient* (DRPG), because it is inspired by double-loop algorithms designed for solving saddle point problems (Jin et al., 2020; Luo et al., 2020; Razaviyayn et al., 2020; Zhang et al., 2020). In

[1]School of Data Science, City University of Hong Kong [2]Department of Computer Science, University of New Hampshire. Correspondence to: Chin Pang Ho <clint.ho@cityu.edu.hk>, Marek Petrik <mpetrik@cs.unh.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

particular, DRPG solves RMDPs using two nested loops: an outer loop updates policies, and an inner loop approximately computes the worst-case transition probabilities. While the outer loop resembles policy gradient updates in regular MDPs, the inner loop must optimize over an infinite number of transition probabilities in the ambiguity set. To effectively optimize the continuous transition probabilities, we use a projected gradient method with a finite but complete parametrization in tabular MDPs. To scale the algorithm to large problems, we propose to use a parametrization based on KL-divergence ambiguity sets.

As our second contribution, we show that DRPG is guaranteed to converge to a globally optimal policy in s-rectangular RMDPs. While this result mirrors similar known results for ordinary MDPs, the robust setting involves several additional non-trivial challenges. Unlike in ordinary MDPs, the RMDP return is not differentiable in terms of the policy (Razaviyayn et al., 2020), which precludes us from leveraging MDP results. Since the RMDP return is not convex, it also does not admit subgradients. Instead, we show that it is sufficient to approximate it by its Moreau envelope, which is differentiable. An additional challenge is that solving the inner loop optimally in every policy carries an unacceptable computational policy, but solving it approximately may cause oscillations. We address this problem by proposing a schedule of decreasing approximation errors that are sufficient to converge to the optimal solution. In fact, the policy updates are guaranteed to converge to the optimal policy as long as the inner loop can be solved with sufficient precision, even when the RMDP is non-rectangular.

Despite the recent advances in robust reinforcement learning (Roy et al., 2017; Badrinath & Kalathil, 2021; Wang & Zou, 2021; Panaganti & Kalathil, 2022), policy gradient methods for solving RMDPs have received only limited attention. A concurrent work proposes a policy gradient method for solving RMDPs with a particular R-contamination ambiguity sets (Wang & Zou, 2022). While this algorithm is compellingly simple, the R-contamination set is very limited in comparison with the general sets that we consider. In fact, we show in Proposition F.1 that RMDPs with R-contamination ambiguity sets simply equal to ordinary MDPs with a reduced discount factor; please see Appendix F for more details. Another recent work develops an extended mirror descent method for solving RMDPs (Li et al., 2022); however, their results are limited to (s, a)-rectangular MDPs only, and their algorithm requires the exact robust Q function to update the policy at every iteration. On the other hand, our proposed algorithm is compatible with any compact ambiguity set, and we do not require an exact optimal solution when solving the inner maximization problem. Moreover, by parameterizing the inner problem, the proposed algorithm is scalable to large problems.

While this paper exclusively focuses on RMDPs, it is worth mentioning that there is an active line of research studying a related model, called distributionally robust MDPs, which assumes the transition kernel is random and governed by an unknown probability distribution that lies in an ambiguity set (Ruszczyński, 2010; Xu & Mannor, 2010; Shapiro, 2016; Chen et al., 2019; Grand-Clément & Kroer, 2021a; Shapiro, 2021; Liu et al., 2022).

The remainder of the paper is organized as follows. Section 2 outlines RMDP and optimization properties that are needed for our results. Then, Section 3 describes the outer loop of DRPG, our proposed algorithm, and shows its global convergence guarantee. The algorithms for solving the inner loop are then described in Section 4. Finally, in Section 5, we present experimental results that illustrate the effective empirical performance of DRPG.

Notation: We reserve lowercase letters for scalars, lowercase bold characters for vectors, and uppercase bold characters for matrices. We denote Δ^S as the probability simplex in $\mathbb{R}^S_+$. For vectors, we use $\|\cdot\|$ to denote the l_2-norm. For a differentiable function $f(x, y)$, we use $\nabla_x f(x, y)$ to denote the partial gradient of f with respect to x. The symbol $\boldsymbol{e}$ denotes a vector of all ones of the size appropriate to the context.

2. Notations and Settings

An ordinary MDP is specified by a tuple $\langle \mathcal{S}, \mathcal{A}, \boldsymbol{p}, \boldsymbol{c}, \gamma, \boldsymbol{\rho} \rangle$, where $\mathcal{S} = \{1, 2, \cdots, S\}$ and $\mathcal{A} = \{1, 2, \cdots, A\}$ are the finite state and action sets, respectively. The discount factor is $\gamma \in (0, 1)$ and the distribution of the initial state is $\boldsymbol{\rho} \in \Delta^S$. The probability distribution of transiting from a current state s to a next state s' after taking an action a is denoted as a vector $\boldsymbol{p}_{sa} \in \Delta^S$ and in a part of the transition kernel $\boldsymbol{p} := (\boldsymbol{p}_{sa})_{s\in\mathcal{S}, a\in\mathcal{A}} \in (\Delta^S)^{S\times A}$. The cost of the aforementioned transition is denoted as $c_{sas'}$ for each $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. It is well-known that translating the costs by a constant or multiplying them by a positive scalar does not change the set of optimal policies. Therefore, we can assume without loss of generality that the cost function is bounded in $[0, 1]$.

Assumption 2.1 (Bounded cost)**.** For any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, the cost $c_{sas'} \in [0, 1]$.

Given a stationary randomized policy $\boldsymbol{\pi} := (\boldsymbol{\pi}_s)_{s\in\mathcal{S}}$ that lies in the policy space $\Pi = (\Delta^A)^S$, $\boldsymbol{\pi}$ maps from state $s \in \mathcal{S}$ to a distribution over action $a \in \mathcal{A}$, and the quality of a policy $\boldsymbol{\pi}$ is evaluated by the *value function* $\boldsymbol{v}^{\boldsymbol{\pi},\boldsymbol{p}} \in \mathbb{R}^S$, defined as

$$v_s^{\boldsymbol{\pi},\boldsymbol{p}} = \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{p}} \left[\sum_{t=0}^{\infty} \gamma^t \cdot c_{s_t a_t s_{t+1}} \mid s_0 = s \right],$$

where a_t follows the distribution $\boldsymbol{\pi}_{s_t}$, and $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{p}}$ denotes expectation with respect to the distribution induced by $\boldsymbol{\pi}$

and transition function $\boldsymbol{p}$ conditioned on the initial state event $\{s_0 = s\}$. Similarly, the value of taking action a at state s is referred as the *action value function* as below

$$q_{sa}^{\boldsymbol{\pi},\boldsymbol{p}} = \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{p}}\left[\sum_{t=0}^{\infty}\gamma^t c_{s_t a_t s_{t+1}} \mid s_0 = s, a_0 = a\right],$$

where it is known that $v_s^{\boldsymbol{\pi},\boldsymbol{p}} = \sum_{a\in\mathcal{A}} \pi_{sa} q_{sa}^{\boldsymbol{\pi},\boldsymbol{p}}$ (Puterman, 2014; Sutton & Barto, 2018). The objective of an MDP is to compute the optimal policy $\boldsymbol{\pi}^\star$ that yields the minimum expected cost, *i.e.*,

$$\boldsymbol{\pi}^\star = \arg\min_{\boldsymbol{\pi}\in\Pi} \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{p}}\left[\sum_{t=0}^{\infty}\gamma^t c_{s_t a_t s_{t+1}} | s_0 \sim \boldsymbol{\rho}\right]. \quad (1)$$

In most domains, the exact transition kernel and cost function are not known precisely and must be estimated from data. These estimation errors often result in policies that perform poorly when deployed. To compute reliable policies with model errors, RMDPs, defined as $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \boldsymbol{c}, \gamma, \boldsymbol{\rho}\rangle$, aim to optimize the worst-case performance with respect to plausible errors (Iyengar, 2005; Nilim & El Ghaoui, 2005; Wiesemann et al., 2013), *i.e.*

$$\min_{\boldsymbol{\pi}\in\Pi}\max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p}) := \boldsymbol{\rho}^\top \boldsymbol{v}^{\boldsymbol{\pi},\boldsymbol{p}} = \sum_{s\in\mathcal{S}} \rho_s v_s^{\boldsymbol{\pi},\boldsymbol{p}}, \quad (2)$$

where $\mathcal{P}$ is known as the *ambiguity set*. By carefully calibrating $\mathcal{P}$ so that it contains the unknown true transition kernel, the optimal policy in (2) can achieve reliable performance in practice (Russell & Petrik, 2019; Behzadian et al., 2021b; Panaganti et al., 2022).

Note that, at this point, there is no need to assume that the RMDP in (2) is rectangular, such as (s, a)-rectangular or s-rectangular (Iyengar, 2005; Nilim & El Ghaoui, 2005; Wiesemann et al., 2013; Ho et al., 2021). We do not need these assumptions to describe or analyze DRPG and only require $\mathcal{P}$ to be compact. Rectangularity assumptions will be helpful, however, when developing algorithms for solving the inner maximization problem.

Given a specific policy and transition kernel, the *occupancy measure* represents the frequencies of visits to states (Puterman, 2014), which is defined as follow.

Definition 2.2 (Occupancy measure)**.** The discounted state occupancy measure $d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}} \colon \mathcal{S} \to [0, 1]$ for an initial distribution $\boldsymbol{\rho}$, a policy $\boldsymbol{\pi} \in \Pi$, and a transition kernel $\boldsymbol{p}$ is defined as

$$d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}}(s') = (1-\gamma)\sum_{s\in\mathcal{S}}\sum_{t=0}^{\infty}\gamma^t \rho(s) p_{ss'}^{\boldsymbol{\pi}}(t). \quad (3)$$

Here, $p_{ss'}^{\boldsymbol{\pi}}(t)$ is the probability of arriving in a state s' after transiting t time steps from state s over the policy $\boldsymbol{\pi}$ and the transition kernel $\boldsymbol{p}$.

The non-convex minimax problem in (2) can be reformulated as an equivalent problem of minimizing the worst-case return:

$$\min_{\boldsymbol{\pi}\in\Pi}\left\{\Phi(\boldsymbol{\pi}) := \max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p})\right\}. \quad (4)$$

Then, it may seem natural to solve (4) by a gradient descent on the function Φ. This is, in general, not possible because the function Φ is not differentiable. In addition, since Φ is neither convex nor concave, its subgradient does not exist either (Nouiehed et al., 2019; Lin et al., 2020). These complications motivate the need for the double-loop iterative scheme to solve RMDPs in Section 3.

Next, we introduce two crucial definitions on smoothness and Lipschitz continuity, which we need to analyze DRPG.

Definition 2.3. A function $h : \mathcal{X} \to \mathbb{R}$ is *L-Lipschitz* if for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$, we have that $\|h(\boldsymbol{x}_1) - h(\boldsymbol{x}_2)\| \leq L\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$, and ℓ*-smooth* if for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$, we have $\|\nabla h(\boldsymbol{x}_1) - \nabla h(\boldsymbol{x}_2)\| \leq \ell\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$.

To discuss the global optimality of RMDPs, we introduce the following definition of weak convexity officially.

Definition 2.4 (Weak Convexity)**.** The function $h : \mathcal{X} \to \mathbb{R}$ is ℓ-weakly convex if for any $\boldsymbol{g} \in \partial h(\boldsymbol{x})$ and $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$,

$$h(\boldsymbol{x}') - h(\boldsymbol{x}) \geq \langle \boldsymbol{g}, \boldsymbol{x}' - \boldsymbol{x}\rangle - \frac{\ell}{2}\|\boldsymbol{x}' - \boldsymbol{x}\|^2.$$

Here, $\partial h(\boldsymbol{x})$ represents the Fréchet sub-differential (See Definition D.1 in the appendix) of $h(\cdot)$ at $\boldsymbol{x} \in \mathcal{X}$, which generalizes the notion of gradient for the non-smooth function (Vial, 1983; Davis & Drusvyatskiy, 2019; Thekumparampil et al., 2019).

3. Solving the Outer Loop

In this section, we describe a policy gradient approach that solves the minimization problem in (4). Surprisingly, we show that a form of gradient descent applied to (4) converges to a globally-optimal solution, even though the objective function is neither convex nor concave. This result is inspired by the recent analysis of policy gradient methods for ordinary MDPs (Agarwal et al., 2021; Bhandari & Russo, 2021). For now, we assume that there exists an oracle that solves the inner maximization problem. We provide the discussion and algorithms for solving the inner problem in Section 4.

The remainder of the section is organized as follows. In Section 3.1, we describe our new policy gradient scheme and then, in Section 3.2, we show that our scheme is guaranteed to converge to the global solution. To the best of our knowledge, this is the first generic robust policy gradient algorithm with global convergence guarantees.

Algorithm 1 Double-Loop Robust Policy Gradient (DRPG)

Input: initial policy $\boldsymbol{\pi}_0$, iteration time T, tolerance sequence $\{\epsilon_t\}_{t\geq 0}$ such that $\epsilon_{t+1} \leq \gamma\epsilon_t$, step size sequence $\{\alpha_t\}_{t\geq 0}$
for $t = 0, 1, \ldots, T-1$ **do**
  Find $\boldsymbol{p}_t$ so that $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_t, \boldsymbol{p}_t) \geq \max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_t, \boldsymbol{p}) - \epsilon_t$.
  Set $\boldsymbol{\pi}_{t+1} \leftarrow \mathrm{Proj}_{\Pi}(\boldsymbol{\pi}_t - \alpha_t \nabla_{\boldsymbol{\pi}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_t, \boldsymbol{p}_t))$. (Eq. (5))
end for
Output: $\boldsymbol{\pi}_{t^\star} \in \{\boldsymbol{\pi}_0, \ldots, \boldsymbol{\pi}_{T-1}\}$ s.t. $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_{t^\star}, \boldsymbol{p}_t) = \min_{t'\in\{0,\ldots,T-1\}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_{t'}, \boldsymbol{p}_t)$

3.1. Double-Loop Robust Policy Gradient Method (DRPG)

We now describe the proposed policy gradient scheme summarized in Algorithm 1, named *Double-Loop Robust Policy Gradient* (DRPG). We refer to DRPG as a "double loop" method in order to be consistent with the terminology in game theory literature (Nouiehed et al., 2019; Thekumparampil et al., 2019; Jin et al., 2020; Zhang et al., 2020).

The inner loop of DRPG updates the worst-case transition probabilities $\boldsymbol{p}_t$ while the outer loop updates the policies $\boldsymbol{\pi}_t$. Specifically, DRPG iteratively takes steps along the policy gradient to search for an optimal policy in (2). At each iteration t, we first solve the inner maximization problem to some specific precision ϵ_t; that is, for a policy $\boldsymbol{\pi}_t$ at iteration t, we seek for any transition kernel $\boldsymbol{p}_t$ such that

$$J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_t, \boldsymbol{p}_t) \geq \max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_t, \boldsymbol{p}) - \epsilon_t \ .$$

Once $\boldsymbol{p}_t$ is computed, DRPG then takes a *projected gradient step* to minimize $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}_t)$ subject to a constraint $\pi \in \Pi$.

When chosen appropriately, the sequence ϵ_t allows for quick policy updates in the initial stages of the algorithm without putting the global convergence in jeopardy. Similar algorithms studied in the context of zero-sum games do not include this tolerance ϵ_t (Nouiehed et al., 2019; Thekumparampil et al., 2019). The adaptive tolerance sequence $\{\epsilon_t\}_{t\geq 0}$ is inspired by prior work on algorithms for RMDPs (Ho et al., 2021). The convergence analysis below provides further guidance on appropriate choices of ϵ_t.

DRPG updates policies using projected gradient descent. The well-known proximal representation of projected gradient is (Bertsekas, 2016):

$$\begin{aligned}\boldsymbol{\pi}_{t+1} &\in \arg\min_{\boldsymbol{\pi}\in\Pi} \langle \nabla_{\boldsymbol{\pi}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_t, \boldsymbol{p}_t), \boldsymbol{\pi} - \boldsymbol{\pi}_t\rangle + \frac{1}{2\alpha_t}\|\boldsymbol{\pi} - \boldsymbol{\pi}_t\|^2 \\ &= \mathrm{Proj}_{\Pi}\left(\boldsymbol{\pi}_t - \alpha_t \nabla_{\boldsymbol{\pi}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_t, \boldsymbol{p}_t)\right), \end{aligned} \tag{5}$$

where Proj_{Π} is the projection operator onto Π and $\alpha_t > 0$ is the step size. This projected gradient update on $\boldsymbol{\pi}_t := (\boldsymbol{\pi}_{t,s})_{s\in\mathcal{S}} \in (\Delta^A)^S$ can be further decoupled to multiple projection updates that across states and take the form as

$$\boldsymbol{\pi}_{t+1,s} = \mathrm{Proj}_{\Delta^A}\left(\boldsymbol{\pi}_{t,s} - \alpha_t \nabla_{\boldsymbol{\pi}_s} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_t, \boldsymbol{p}_t)\right), \ \ \forall s \in \mathcal{S},$$

which can also be seen as a gradient step followed by a projection onto Δ^A for each state $s \in \mathcal{S}$. Note that the gradient $\nabla_{\boldsymbol{\pi}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_t, \boldsymbol{p}_t)$ used in DRPG is identical to the the gradient in ordinary MDPs, e.g., (Agarwal et al., 2021; Bhandari & Russo, 2021),

$$\frac{\partial J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})}{\partial \pi_{sa}} = \frac{1}{1-\gamma} \cdot d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}}(s) \cdot q_{sa}^{\boldsymbol{\pi},\boldsymbol{p}}. \tag{6}$$

Actor-critic RL algorithms are typically based on this form of the policy gradient.

An alternative to double-loop algorithms is to use single-loop algorithms. Single-loop algorithms interleave gradient updates to the inner and outer optimization problems (Mokhtari et al., 2020; Zhang et al., 2020). Interleaving gradient updates is fast but prone to instabilities and oscillations. The most-common approach to preventing such instabilities is to resort to two-scale step size updates (Heusel et al., 2017; Daskalakis et al., 2020; Russel et al., 2020). We focus in this work on double-loop algorithms because of their conceptual simplicity and good empirical behavior.

3.2. Convergence Analysis

We now turn to analyzing the convergence behavior of DRPG. First, recall that we assume that $\mathcal{P}$ is compact. Virtually all ambiguity sets considered in prior work, such as L_1-ambiguity sets, L_∞-ambiguity sets, L_2-ambiguity sets, and KL-ambiguity sets, are compact.

Then, the following lemma helps us to derive the weak convexity of this non-convex, non-differentiable (i.e., non-smooth) objective function $\Phi(\boldsymbol{\pi})$.

Lemma 3.1. *The objective function $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$ in (2) is $L_{\boldsymbol{\pi}}$-Lipschitz and $\ell_{\boldsymbol{\pi}}$-smooth in $\boldsymbol{\pi}$ with*

$$L_{\boldsymbol{\pi}} := \frac{\sqrt{A}}{(1-\gamma)^2}, \quad \ell_{\boldsymbol{\pi}} := \frac{2\gamma A}{(1-\gamma)^3}.$$

Furthermore, the objective $\Phi(\boldsymbol{\pi})$ is $\ell_{\boldsymbol{\pi}}$-weakly convex and $L_{\boldsymbol{\pi}}$-Lipschitz.

The proof of this lemma, as well as of all the remaining auxiliary results, are provided in the appendix. Lemma 3.1 establishes some general continuity properties of $\Phi(\boldsymbol{\pi})$ and serves as an important stepping stone for deriving the global convergence of Algorithm 1; however, weak convexity alone is insufficient to guarantee that gradient-based updates converge to a global optimum.

Recent work (Agarwal et al., 2021) proved the global convergence of policy gradient methods in ordinary MDP relying

on a "gradient dominance condition". Informally speaking, a function $h(\boldsymbol{x})$ is said to satisfy the gradient dominance condition if $h(\boldsymbol{x}) - h(\boldsymbol{x}^\star) = \mathcal{O}(G(\boldsymbol{x}))$, where $G(\cdot)$ is a suitable notion that measures the gradient of h. By having a gradient dominance condition, one can prevent the gradient from vanishing before reaching a globally optimal point.

Despite the non-smoothness of $\Phi(\boldsymbol{\pi})$, weakly convex problems naturally admit an implicit smooth approximation through the Moreau envelope (Davis & Drusvyatskiy, 2019; Mai & Johansson, 2020). Inspired by the idea of gradient dominance, we introduce the gradient of the Moreau envelope and show that $\Phi(\boldsymbol{\pi})$ satisfies a particular variant of the gradient dominance condition in the next theorem.

Theorem 3.2. *Denote $\boldsymbol{\pi}^\star$ as the global optimal policy for RMDPs. Then, for any policy $\boldsymbol{\pi}$, we have*

$$\Phi(\boldsymbol{\pi}) - \Phi(\boldsymbol{\pi}^\star) \leq \left(\frac{D\sqrt{SA}}{1-\gamma} + \frac{L_{\boldsymbol{\pi}}}{2\ell_{\boldsymbol{\pi}}} \right) \|\nabla \Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi})\|, \quad (7)$$

where $\Phi_\lambda(\boldsymbol{\pi})$ is the Moreau envelope function of $\Phi(\boldsymbol{\pi})$ (see Definition D.3) and $D := \sup_{\boldsymbol{\pi}\in\Pi, \boldsymbol{p}\in\mathcal{P}} \|{d_\rho^{\boldsymbol{\pi},\boldsymbol{p}}}/{\rho}\|_\infty < \infty$ for every $\boldsymbol{\rho}$ with $\min_{s\in\mathcal{S}} \rho_s > 0$.

Here, $\|{d_\rho^{\boldsymbol{\pi},\boldsymbol{p}}}/{\rho}\|_\infty$ is formally named as *distribution mismatch coefficient* which is often assumed to be bounded (Scherrer, 2014; Chen & Jiang, 2019; Mei et al., 2020; Agarwal et al., 2021; Leonardos et al., 2021).

This gradient-dominance type property implies that any first-order stationary point of the Moreau envelope results in an approximately global optimal policy. We are now ready to state our main result.

Theorem 3.3 (Global convergence for DRPG)**.** *Denote $\boldsymbol{\pi}_{t^\star}$ as the policy that Algorithm 1 outputs. Then, for a constant step size $\alpha := \frac{\delta}{\sqrt{T}}$ with any $\delta > 0$ and the initial tolerance $\epsilon_0 \leq \sqrt{T}$, we have*

$$\Phi(\boldsymbol{\pi}_{t^\star}) - \min_{\boldsymbol{\pi}\in\Pi} \Phi(\boldsymbol{\pi}) \leq \epsilon, \quad (8)$$

and T is chosen to be a large enough such that

$$T \geq \frac{\left(\frac{D\sqrt{SA}}{1-\gamma} + \frac{L_{\boldsymbol{\pi}}}{2\ell_{\boldsymbol{\pi}}}\right)^4 \left(\frac{4\ell_{\boldsymbol{\pi}} S}{\delta} + 2\delta\ell_{\boldsymbol{\pi}} L_{\boldsymbol{\pi}}^2 + \frac{4\ell_{\boldsymbol{\pi}}}{1-\gamma}\right)^2}{\epsilon^4} = \mathcal{O}(\epsilon^{-4}). \quad (9)$$

Compared to the ordinary MDPs, the convergence analysis for solving RMDPs poses additional difficulties as objective function $\Phi(\boldsymbol{\pi})$ is not only non-convex but also non-differentiable (Nouiehed et al., 2019; Lin et al., 2020). Theorem 3.3 shows that the proposed Algorithm 1 converges to the global optimal for RMDPs by the following strategy. We first show the existence of an ϵ-first order stationary point (see Definition D.4) of $\Phi(\boldsymbol{\pi})$. More concretely, we prove the gradient of the Moreau envelope is smaller than ϵ on the output policy. Then, by applying the derived gradient dominance condition (Theorem 3.2), we finally complete the proof as this stationary point is arbitrarily close to the global optimal solution.

Theorem 3.3 shows that DRPG converges to an ϵ global optimum within $\mathcal{O}(\epsilon^{-4})$ steps, which has a slower rate compared to standard policy gradient methods (Agarwal et al., 2021). The additional complexity arises from this need to control the approximation error in order to avoid looping. In particular, computational errors at the inner loops could break the convergence of the outer loop. Similar behaviors are also observed in policy iteration for robust MDPs (Condon, 1990; Ho et al., 2021). Nevertheless, our analysis matches and is consistent with the other minimax convergence results obtained in non-convex non-concave minimax optimization (Davis & Drusvyatskiy, 2019; Jin et al., 2020), and provides a conservative convergence guarantee.

DRPG relies on an oracle that outputs at least one worst-case transition kernel for any given $\boldsymbol{\pi}$. In fact, solving the inner loop problem could still be NP-hard for non-rectangular cases (Wiesemann et al., 2013). The following section proposes an algorithm for solving the inner loop problem.

4. Solving the Inner Loop

So far, we have described the outline of DRPG and proved its global convergence. In Algorithm 1, the transition kernel $\boldsymbol{p}_t$ is obtained by approximately solving the inner maximization problem with a fixed outer policy $\boldsymbol{\pi}_k \in \Pi$:

$$\max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_k, \boldsymbol{p}) = \max_{\boldsymbol{p}\in\mathcal{P}} \boldsymbol{\rho}^\top \boldsymbol{v}^{\boldsymbol{\pi}_k,\boldsymbol{p}}. \quad (10)$$

Whereas assumptions of boundness and compactness are used to ensure the inner maximum existing for the maximization problem, solving this maximization problem is still computationally challenging due to its non-convexity (Wiesemann et al., 2013). This section discusses two solution methods for solving the inner maximization problem, which we refer to as the *robust policy evaluation problem*. Note that the convergence results in Section 3 are independent of the method used to solve this robust policy evaluation problem.

We now introduce two broad classes of ambiguity sets that are considered in the rest of this section. An ambiguity set $\mathcal{P}$ is (s, a)-rectangular (Iyengar, 2005; Nilim & El Ghaoui, 2005; Le Tallec, 2007) if it is a Cartesian product of sets $\mathcal{P}_{s,a} \subseteq \Delta^S$ for each state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, *i.e.*,

$$\mathcal{P} = \{\boldsymbol{p} \in (\Delta^S)^{S\times A} \mid \boldsymbol{p}_{s,a} \in \mathcal{P}_{s,a},\ \forall s \in \mathcal{S}, a \in \mathcal{A}\},$$

whereas an ambiguity set $\mathcal{P}$ is s-rectangular (Wiesemann et al., 2013) if it is defined as a Cartesian product of sets

$\mathcal{P}_s \subseteq (\Delta^S)^A$, *i.e.*,

$$\mathcal{P} = \{\boldsymbol{p} \in (\Delta^S)^{S\times A} \mid (\boldsymbol{p}_{s,a})_{a\in\mathcal{A}} \in \mathcal{P}_s,\ \forall s \in \mathcal{S}\}.$$

4.1. Value-iteration Approach

The optimum of the inner problem (10) is attained by solving $\boldsymbol{v}^{\boldsymbol{\pi}_k} := \min_{\boldsymbol{p}\in\mathcal{P}} \boldsymbol{v}^{\boldsymbol{\pi}_k,\boldsymbol{p}}$, which is commonly defined as the robust value function (Iyengar, 2005; Nilim & El Ghaoui, 2005; Wiesemann et al., 2013). The robust value function $\boldsymbol{v}^{\boldsymbol{\pi}}$ of a rectangular RMDP for a policy $\boldsymbol{\pi} \in \Pi$ can be computed using the robust Bellman policy update $\mathcal{T}_{\boldsymbol{\pi}} : \mathbb{R}^S \to \mathbb{R}^S$ (Ho et al., 2021). Specifically, for (s,a)-rectangular RMDPs, the operator $\mathcal{T}_{\boldsymbol{\pi}}$ is defined for each state $s \in \mathcal{S}$

$$(\mathcal{T}_{\boldsymbol{\pi}}\boldsymbol{v})_s := \sum_{a\in\mathcal{A}} \left(\pi_{sa} \cdot \max_{\boldsymbol{p}_{sa}\in\mathcal{P}_{sa}} \boldsymbol{p}_{sa}^\top (\boldsymbol{c}_{sa} + \gamma \boldsymbol{v}) \right),$$

while for s-rectangular RMDPs, the the operator $\mathcal{T}_{\boldsymbol{\pi}}$ is defined as

$$(\mathcal{T}_{\boldsymbol{\pi}}\boldsymbol{v})_s := \max_{\boldsymbol{p}_s\in\mathcal{P}_s} \left\{ \sum_{a\in\mathcal{A}} \pi_{sa} \cdot \boldsymbol{p}_{sa}^\top (\boldsymbol{c}_{sa} + \gamma \boldsymbol{v}) \right\}.$$

For rectangular RMDPs, $\mathcal{T}_{\boldsymbol{\pi}}$ is a contraction and the robust value function is the unique solution to $\boldsymbol{v}^{\boldsymbol{\pi}} = \mathcal{T}_{\boldsymbol{\pi}}\boldsymbol{v}^{\boldsymbol{\pi}}$. To solve the robust value function, the state-of-the-art method is to compute the sequence $\boldsymbol{v}_{t+1}^{\boldsymbol{\pi}} = \mathcal{T}_{\boldsymbol{\pi}}\boldsymbol{v}_t^{\boldsymbol{\pi}}$ with any initial values $\boldsymbol{v}_0^{\boldsymbol{\pi}}$, which is similar to the policy evaluation for ordinary MDPs.

Note that computing the value function update $\boldsymbol{v}_t^{\boldsymbol{\pi}}$ to $\boldsymbol{v}_{t+1}^{\boldsymbol{\pi}}$ requires solving an optimization problem. For the common ambiguity sets which are constrained by the support information and one additional convex constraint (*e.g.* L_1-norm ball), one has to solve A convex optimization problems with $\mathcal{O}(S)$ variables and $\mathcal{O}(S)$ constraints for all $s \in \mathcal{S}$ at each iteration (Grand-Clément & Kroer, 2021b). Examples of common ambiguity sets are provided in Appendix A.

4.2. Gradient-based Approach

Unlike the extensive study of efficient value-based methods (Iyengar, 2005; Nilim & El Ghaoui, 2005; Wiesemann et al., 2013; Petrik & Subramanian, 2014; Ho et al., 2018; Behzadian et al., 2021a), there has been little work on designing gradient-based algorithms to compute the robust value function. In this subsection, a first gradient-based algorithm is proposed in Algorithm 2 to solve the inner-loop robust policy evaluation problem with a global convergence guarantee, under the assumptions of having rectangular and convex ambiguity set.

Note that the inner problem (10) could be regarded as a constrained non-concave maximization problem when the outer policy $\boldsymbol{\pi}_k$ is fixed. Therefore, the most intuitive approach to solve (10) is to iteratively update the variable by following its ascent direction within the feasible set.

Algorithm 2 Projected gradient descent for the inner problem

Input: Target fixed policy $\boldsymbol{\pi}_k$, initial transition kernel $\boldsymbol{p}_0$, iteration time T_k, step size sequence $\{\beta_t\}_{t\geq 0}$
for $t = 0, 1, \ldots, T_k - 1$ **do**
  Set $\boldsymbol{p}_{t+1} \leftarrow \text{Proj}_{\mathcal{P}}(\boldsymbol{p}_t + \beta_t \nabla_{\boldsymbol{p}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_k, \boldsymbol{p}_t))$.
end for
Output: $\boldsymbol{p}_{t^\star} \in \{\boldsymbol{p}_0, \ldots, \boldsymbol{p}_{T_k-1}\}$ s.t. $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_k, \boldsymbol{p}_{t^\star}) = \min_{t\in\{0,\ldots,T_k-1\}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_k, \boldsymbol{p}_t)$

To maximize $J_\rho(\boldsymbol{\pi}_k, \boldsymbol{p})$, Algorithm 2 iteratively computes the *projected gradient step* on $\boldsymbol{p}$; that is, at iteration t, we compute

$$\boldsymbol{p}_{t+1} = \text{Proj}_{\mathcal{P}}(\boldsymbol{p}_t + \beta_t \nabla_{\boldsymbol{p}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_k, \boldsymbol{p}_t)), \tag{11}$$

which depends on the explicit form of $\mathcal{P}$. Although (s,a)-rectangular ambiguity sets can be viewed as a special case of s-rectangular ambiguity sets in general (Wiesemann et al., 2013; Ho et al., 2021), the implementations of the projected gradient step for two rectangular ambiguity sets are different.

For (s,a)-rectangular RMDPs, this projected gradient update can be decoupled to multiple projection updates that across state-action pairs such as

$$\boldsymbol{p}_{t+1,sa} = \text{Proj}_{\mathcal{P}_{s,a}}(\boldsymbol{p}_{t,sa} + \beta_t \nabla_{\boldsymbol{p}_{sa}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_k, \boldsymbol{p}_t)).$$

Similarly, for s-rectangular RMDPs, the projected gradient update can be computed across states as

$$\boldsymbol{p}_{t+1,s} = \text{Proj}_{\mathcal{P}_s}(\boldsymbol{p}_{t,s} + \beta_t \nabla_{\boldsymbol{p}_s} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_k, \boldsymbol{p}_t)).$$

If the ambiguity set is convex, the projected update can be implemented by solving a convex optimization problem with a quadratic objective.

4.3. Inner Loop Global Optimality

To establish some general convergence properties of Algorithm 2, we first derive some continuity properties for the inner objective (10). Then, we prove the global optimality of Algorithm 2 by introducing a particular gradient dominance condition for the inner problem.

The next lemma derives the gradient for the inner loop.

Lemma 4.1 (Differentiability)**.** *The partial derivative of $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$ has the explicit form for any $(s, a, s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$,*

$$\frac{\partial J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})}{\partial p_{sas'}} = \frac{1}{1-\gamma} d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}}(s)\pi_{sa}\left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}}\right).$$

Moreover, $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$ is $L_{\boldsymbol{p}}$-Lipschitz in $\boldsymbol{p}$ with $L_{\boldsymbol{p}} := \frac{\sqrt{SA}}{(1-\gamma)^2}$.

If a function is smooth, then a gradient update with a sufficiently small step size is guaranteed to improve the objective value. As it turns out, inner problem is $\ell_{\boldsymbol{p}}$-smooth.

Lemma 4.2 (Smoothness)**.** *The function $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$ is $\ell_{\boldsymbol{p}}$-smooth in $\boldsymbol{p}$ with $\ell_{\boldsymbol{p}} := \frac{2\gamma S^2}{(1-\gamma)^3}$.*

Due to the non-convexity of $J_{\boldsymbol{\rho}}$, smoothness is not sufficient to establish the global convergence guarantee. We notice that the inner problem can be interpreted as having an adversarial nature to maximize the total reward (decision maker's cost) by selecting a proper transition kernel from the ambiguity set $\mathcal{P}$ (Lim et al., 2013; Goyal & Grand-Clément, 2022). Hence we leverage the idea from the convergence analysis of the classical policy gradient (Agarwal et al., 2021) and derive our global convergence guarantee by first deriving the following inner problem's gradient dominance condition.

Lemma 4.3 (Gradient dominance)**.** *For any fixed $\boldsymbol{\pi} \in \Pi$, $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$ satisfies the following condition for any $\boldsymbol{p} \in \mathcal{P}$ such that*

$$J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}^\star) - J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}) \leq \frac{D}{1-\gamma} \max_{\bar{\boldsymbol{p}} \in \mathcal{P}} \langle \bar{\boldsymbol{p}} - \boldsymbol{p}, \nabla_{\boldsymbol{p}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}) \rangle ,$$

where $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}^\star) := \max_{\boldsymbol{p} \in \mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$.

Using this notion of gradient dominance, we now give an iteration complexity bound for Algorithm 2.

Theorem 4.4. *Let $\boldsymbol{p}_{t^\star}$ be the point obtained by Algorithm 2 and $\epsilon_k > 0$ be the desired precision. Algorithm 2 with constant step size $\beta = \frac{(1-\gamma)^3}{2\gamma S^2}$ satisfies*

$$\max_{\boldsymbol{p} \in \mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_k, \boldsymbol{p}) - J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_k, \boldsymbol{p}_{t^\star}) \leq \epsilon_k, \tag{12}$$

whenever

$$T \geq \frac{32\gamma S^3 A D^2}{(1-\gamma)^6 \epsilon_k^2} = \mathcal{O}(\epsilon_k^{-2}). \tag{13}$$

4.4. Scalability of Parametric Transition

In standard policy-gradient methods, one considers a family of policies parametrized by lower-dimensional parameter vectors to limit the number of variables when scaling to large problems. The projected gradient step in Algorithm 2 needs to update each $p_{sas'}$, which is difficult with large state and action spaces. To overcome this problem, we provide a new approach to transition probability parameterization. To the best of our knowledge, comparable parameterizations for the inner problem have not been studied previously.

We parameterize transition kernel with the following form for any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,

$$p^{\boldsymbol{\xi}}_{sas'} := \frac{\bar{p}_{sas'} \cdot \exp(\frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(s')}{\lambda_{sa}})}{\sum_k \bar{p}_{sak} \cdot \exp(\frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(k)}{\lambda_{sa}})}, \tag{14}$$

where $\boldsymbol{\phi}(s) := [\phi_1(s), \cdots, \phi_m(s)]$ is a m-dimensional feature vector corresponding to the state $s \in \mathcal{S}$, $\boldsymbol{\xi} := (\boldsymbol{\theta}, \boldsymbol{\lambda})$ is the collection of parameters, consisting of the strictly positive parameter $\boldsymbol{\lambda} := \{\lambda_{sa} > 0 \mid \forall (s, a) \in \mathcal{S} \times \mathcal{A}\}$ and the unconstrained parameter $\boldsymbol{\theta} := [\theta_1, \cdots, \theta_m]$. The symbol $\bar{\boldsymbol{p}}$ represents the nominal transition kernel, which is typically estimated from the empirical sample of state transitions.

The parameterization in (14) is motivated by the form of the worst-case transition probabilities in RMDPs with KL-divergence constrained (s, a)-rectangular ambiguity sets (Nilim & El Ghaoui, 2005). In fact, the worst-case transitions has an identical form to (14) when linear approximation $\boldsymbol{\theta}^\top \boldsymbol{\phi}(s)$ is applied.

Then, the RMDPs problem then becomes,

$$\min_{\boldsymbol{\pi} \in \Pi} \max_{\boldsymbol{\xi} \in \Xi} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{\xi}),$$

where Ξ is the ambiguity set for the parameter $\boldsymbol{\xi}$. In practice, Ξ could be constructed via distance-type constraint; that is, we consider

$$\Xi := \{\boldsymbol{\xi} \mid D(\boldsymbol{\xi} \| \boldsymbol{\xi}_c) \leq \kappa\},$$

where $D(\cdot \| \cdot)$ represents a distance function, such as L_1-norm and L_∞-norm, $\boldsymbol{\xi}_c$ is the user-specified empirical estimation of $\boldsymbol{\xi}$, and $\kappa \in \mathbb{R}_{++}$ is a given radius.

To apply the gradient-based update on parameterized transition, we introduce the following lemma to derive the gradient of the inner problem, which is similar to the classical policy gradient theorem (Sutton et al., 1999)

Lemma 4.5. *Consider a map $\boldsymbol{\xi} \mapsto p^{\boldsymbol{\xi}}_{sas'}$ that is differentiable for any (s, a, s') . Then, the partial gradient of $J_\rho(\boldsymbol{\pi}, \boldsymbol{\xi})$ on $\boldsymbol{\xi}$ is*

$$\frac{\partial J_\rho(\boldsymbol{\pi}, \boldsymbol{\xi})}{\partial \boldsymbol{\xi}} = \frac{1}{1-\gamma} \mathbb{E}_{\substack{s \sim \boldsymbol{d}^{\boldsymbol{\pi}, \boldsymbol{\xi}}_\rho \\ a \sim \boldsymbol{\pi}_{s\cdot} \\ s' \sim \boldsymbol{p}_{sa\cdot}}} \left[\frac{\partial \log p^{\boldsymbol{\xi}}_{sas'}}{\partial \boldsymbol{\xi}} \left(c_{sas'} + \gamma v^{\boldsymbol{\pi}, \boldsymbol{\xi}}_{s'} \right) \right]. \tag{15}$$

Moreover, when parameterization (14) *is applied, the score function $\frac{\partial \log p^{\boldsymbol{\xi}}_{sas'}}{\partial \boldsymbol{\xi}}$ has the analytical form:*

$$\frac{\partial \log p^{\boldsymbol{\xi}}_{sas'}}{\partial \theta_i} = \frac{\phi_i(s')}{\lambda_{sa}} - \sum_j p^{\boldsymbol{\xi}}_{saj} \cdot \frac{\phi_i(j)}{\lambda_{sa}}, \tag{16}$$

$$\frac{\partial \log p^{\boldsymbol{\xi}}_{sas'}}{\partial \lambda_{sa}} = \sum_j p^{\boldsymbol{\xi}}_{saj} \cdot \frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(j)}{\lambda_{sa}^2} - \frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(s')}{\lambda_{sa}^2}. \tag{17}$$

5. Experiments

In this section, we demonstrate the global convergence of DRPG and verify the robustness of the policies computed

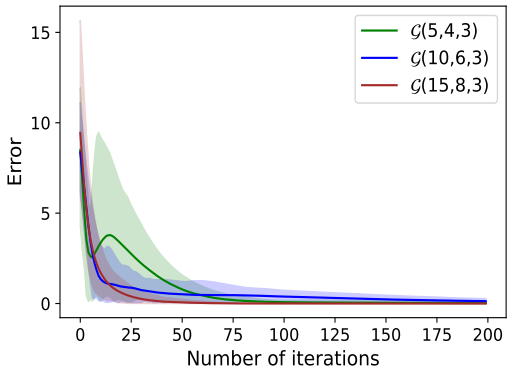

Figure 1. The error of value functions computed by DRPG for three Garnet problems with different sizes.

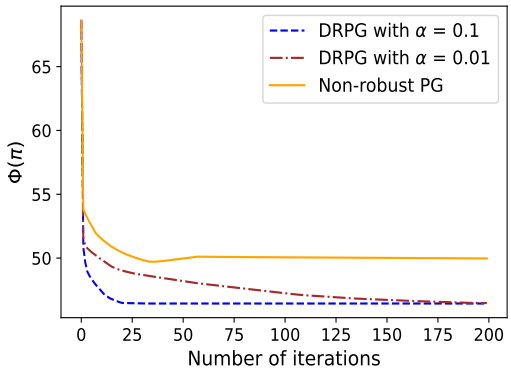

Figure 2. DRPG with parameterization v.s. Non-robust Policy Gradient on the Inventory Management Problem

by DRPG. All algorithms are implemented in Python 3.8.8, and performed on a computer with an i7-11700 CPU with 16GB RAM. We use Gurobi 9.5.2 to solve any linear or quadratic optimization problems involved. To facilitate the reproducibility of the domains, the full source code, which was used to generate them, is available at `https://github.com/JerrisonWang/ICML-DRPG`. The repository also contains CSV files with the precise specification of the RMDPs being solved.

5.1. Experimental Setup

To demonstrate the convergence behavior, we test our algorithm on random GARNET MDPs, one of the widely-used benchmarks for RL algorithms, with three different problem sizes and two settings on the ambiguity sets: (s, a)- and s-rectangular ambiguity sets. We then apply DRPG with inner parameterization on the practical inventory management problem to demonstrate its convergence and robustness.

Garnet MDPs are a class of abstract, but representative, finite MDPs that can be generated randomly (Archibald et al., 1995). A general GARNET $\mathcal{G}(|\mathcal{S}|, |\mathcal{A}|, b)$ is characterized by three parameters, where $|\mathcal{S}|$ is the number of states, $|\mathcal{A}|$ is the number of actions, and b is a branching factor which determines the number of possible next states for each state-action pair and controls the level of connectivity of underlying Markov chains.

In our inventory management problem (Porteus, 2002; Ho et al., 2018), a retailer orders, stores, and sells a single product over an infinite time horizon. The states and actions of the MDP represent the inventory levels and the order quantities in any given time period, respectively. The stochastic demands drive the stochastic state transitions. Any items held in inventory incur deterministic per-period holding costs. The retailer's goal is to find a policy that minimizes the total cost without knowing the exact transition kernel.

More details on the problem settings, parameter choice, and feature selection are available in the Appendix H.

5.2. Results and Discussion

In each of our GARNET problems, we compare the objective values of DRPG at different iterations with the optimal objective value $J^\star$, which is computed by robust value iteration. Robust value iteration solves the robust Bellman equation by iteratively applying robust Bellman updates. For each setup of our GARNET problems, we solve 50 sample instances using both DRPG and robust value iteration. Figure 1 shows how the error (i.e., $|J(\boldsymbol{\pi}_t, \boldsymbol{p}_t) - J^\star|$) decreases when DRPG is performed. The upper and lower envelopes of the curves correspond to the 95 and 5 percentiles of the 50 samples, respectively. As expected, the error decreases to zero as the iteration step increases, which confirms the convergence behavior of DRPG. Similar results are observed for the s-rectangular case.

The results of our numerical study on the inventory management problem are provided in Figure 2. We run DRPG with inner parametrization and compare the performance with the non-robust policy gradient. At each iteration t, we consider the policy $\boldsymbol{\pi}_t$ obtained by DRPG, and then we compute its worst-case expected return $\Phi(\boldsymbol{\pi}_t) := \max_{\boldsymbol{p}\in\mathcal{P}} J(\boldsymbol{\pi}_t, \boldsymbol{p})$. We do the same for the non-robust policy gradient method. As we can see, DRPG obtains a policy that performs much better than the non-robust policy gradient, which demonstrates the robustness of our method. Different step sizes are chosen for DRPG, and they lead to different convergence behaviors; yet, in both cases, DRPGs outperform the non-robust policy gradient method.

6. Conclusion

We proposed a new policy optimization algorithm DRPG to solve RMDPs over general compact ambiguity sets. By

selecting a suitable step size and an adaptive decreasing tolerance sequence, our algorithm converges to the global optimal policy under mild conditions. Moreover, we provide the first gradient-based solution method with a novel parameterization for solving the inner maximization. In our experiments, our results demonstrate the global convergence of DRPG and its reliable performance against the non-robust approach. Future work should address extensions to related models (*e.g.*, distributionally RMDP) and scalable model-free algorithms.

Acknowledgements

We thank the anonymous reviewers for their comments. This work was supported, in part, by NSF grants 2144601 and 1815275, the CityU Start-Up Grant (Project No. 9610481), the National Natural Science Foundation of China (Project No. 72032005), and Chow Sang Sang Group Research Fund sponsored by Chow Sang Sang Holdings International Limited (Project No. 9229076).

References

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

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

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

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

Behzadian, B., Petrik, M., and Ho, C. P. Fast algorithms for l_∞-constrained s-rectangular robust MDPs. *Advances in Neural Information Processing Systems*, 34, 2021a.

Behzadian, B., Russel, R., Ho, C. P., and Petrik, M. Optimizing percentile criterion using robust MDPs. In *International Conference on Artificial Intelligence and Statistics (AIStats)*, 2021b.

Bertsekas, D. P. *Nonlinear Programming*. Athena scientific, 3rd edition, 2016.

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

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., and Lee, M. Natural actor–critic algorithms. *Automatica*, 45(11): 2471–2482, 2009.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019.

Chen, Z., Yu, P., and Haskell, W. B. Distributionally robust optimization for sequential decision-making. *Optimization*, 68(12):2397–2426, 2019.

Condon, A. On algorithms for simple stochastic games. *Advances in computational complexity theory*, 13:51–72, 1990.

Daskalakis, C., Foster, D. J., and Golowich, N. Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems*, 33:5527–5540, 2020.

Davis, D. and Drusvyatskiy, D. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Delgado, K. V., De Barros, L. N., Dias, D. B., and Sanner, S. Real-time dynamic programming for Markov decision processes with imprecise probabilities. *Artificial Intelligence*, 230:192–223, 2016.

Derman, E., Geist, M., and Mannor, S. Twice regularized MDPs and the equivalence between robustness and regularization. *Advances in Neural Information Processing Systems*, 34, 2021.

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

Goyal, V. and Grand-Clément, J. Robust Markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 2022.

Grand-Clément, J. and Kroer, C. First-order methods for Wasserstein distributionally robust MDPs. In *International Conference on Machine Learning*, pp. 2010–2019. PMLR, 2021a.

Grand-Clément, J. and Kroer, C. Scalable first-order methods for robust MDPs. In *AAAI Conference on Artificial Intelligence*, volume 35, pp. 12086–12094, 2021b.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

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

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

Jin, C., Netrapalli, P., and Jordan, M. What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning*, pp. 4880–4889. PMLR, 2020.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*. Citeseer, 2002.

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

Konda, V. and Tsitsiklis, J. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

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

Le Tallec, Y. *Robust, risk-sensitive, and data-driven control of Markov decision processes*. PhD thesis, Massachusetts Institute of Technology, 2007.

Leonardos, S., Overman, W., Panageas, I., and Piliouras, G. Global convergence of multi-agent policy gradient in Markov potential games. *arXiv preprint arXiv:2106.01969*, 2021.

Li, Y., Zhao, T., and Lan, G. First-order policy optimization for robust Markov decision process. *arXiv preprint arXiv:2209.10579*, 2022.

Lim, S. H., Xu, H., and Mannor, S. Reinforcement learning in robust markov decision processes. *Advances in Neural Information Processing Systems*, 26, 2013.

Lin, T., Jin, C., and Jordan, M. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pp. 6083–6093. PMLR, 2020.

Liu, Z., Bai, Q., Blanchet, J., Dong, P., Xu, W., Zhou, Z., and Zhou, Z. Distributionally robust q-learning. In *International Conference on Machine Learning*, pp. 13623–13643. PMLR, 2022.

Luo, L., Ye, H., Huang, Z., and Zhang, T. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Advances in Neural Information Processing Systems*, 33:20566–20577, 2020.

Mai, V. and Johansson, M. Convergence of a stochastic gradient method with momentum for non-smooth nonconvex optimization. In *International conference on machine learning*, pp. 6630–6639. PMLR, 2020.

Mannor, S., Mebel, O., and Xu, H. Robust MDPs with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509, 2016.

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

Mokhtari, A., Ozdaglar, A., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pp. 1497–1507. PMLR, 2020.

Nilim, A. and El Ghaoui, L. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Nouiehed, M., Sanjabi, M., Huang, T., Lee, J. D., and Razaviyayn, M. Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32, 2019.

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

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

Panaganti, K., Xu, Z., Kalathil, D., and Ghavamzadeh, M. Robust reinforcement learning using offline data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Petrik, M. Approximate dynamic programming by minimizing distributionally robust bounds. In *International Conference on Machine Learning*, pp. 497–504, 2012.

Petrik, M. and Subramanian, D. Raam: The benefits of robustness in approximating aggregated MDPs in reinforcement learning. *Advances in Neural Information Processing Systems*, 27, 2014.

Petrik, M., Ghavamzadeh, M., and Chow, Y. Safe policy improvement by minimizing robust baseline regret. *Advances in Neural Information Processing Systems*, 29, 2016.

Pirotta, M., Restelli, M., and Bascetta, L. Policy gradient in Lipschitz Markov decision processes. *Machine Learning*, 100(2):255–283, 2015.

Porteus, E. L. *Foundations of stochastic inventory theory*. Stanford University Press, 2002.

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

Razaviyayn, M., Huang, T., Lu, S., Nouiehed, M., Sanjabi, M., and Hong, M. Nonconvex min-max optimization: Applications, challenges, and recent theoretical advances. *IEEE Signal Processing Magazine*, 37(5):55–66, 2020.

Rockafellar, R. T. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

Rockafellar, R. T. and Wets, R. J.-B. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Roy, A., Xu, H., and Pokutta, S. Reinforcement learning under model mismatch. *Advances in neural information processing systems*, 30, 2017.

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

Russell, R. H. and Petrik, M. Beyond confidence regions: Tight Bayesian ambiguity sets for robust MDPs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Ruszczyński, A. Risk-averse dynamic programming for Markov decision processes. *Mathematical programming*, 125(2):235–261, 2010.

Scherrer, B. Approximate policy iteration schemes: a comparison. In *International Conference on Machine Learning*, pp. 1314–1322. PMLR, 2014.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shapiro, A. Rectangular sets of probability measures. *Operations Research*, 64(2):528–541, 2016.

Shapiro, A. Distributionally robust optimal control and MDP modeling. *Operations Research Letters*, 49(5):809–814, 2021.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *International conference on machine learning*, pp. 387–395. PMLR, 2014.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Thekumparampil, K. K., Jain, P., Netrapalli, P., and Oh, S. Efficient algorithms for smooth minimax optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Vial, J.-P. Strong and weak convexity of sets and functions. *Mathematics of Operations Research*, 8(2):231–259, 1983.

Wang, K., Kumar, N., Zhou, K., Hooi, B., Feng, J., and Mannor, S. The geometry of robust value functions. *arXiv preprint arXiv:2201.12929*, 2022.

Wang, Y. and Zou, S. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34:7193–7206, 2021.

Wang, Y. and Zou, S. Policy gradient method for robust reinforcement learning. In *International Conference on Machine Learning*, pp. 23484–23526, 17–23 Jul 2022.

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

Xu, H. and Mannor, S. Parametric regret in uncertain Markov decision processes. In *IEEE Conference on Decision and Control (CDC)*, pp. 3606–3613. IEEE, 2009.

Xu, H. and Mannor, S. Distributionally robust Markov decision processes. *Advances in Neural Information Processing Systems*, 23, 2010.

Xu, X., Zuo, L., and Huang, Z. Reinforcement learning algorithms with function approximation: Recent advances and applications. *Information Sciences*, 261:1–31, 2014.

Zhang, J., Xiao, P., Sun, R., and Luo, Z. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *Advances in Neural Information Processing Systems*, 33:7377–7389, 2020.

A. Examples of Common Ambiguity Sets

We discuss a particularly popular class of rectangular ambiguity sets which are defined by norm constraints bounding the distance of any feasible transition probabilities from a nominal (average) state distribution. It is usually referred to as an L_1-constrained ambiguity set (Petrik & Subramanian, 2014; Petrik et al., 2016; Ho et al., 2021) or L_∞-constrained ambiguity set (Delgado et al., 2016; Behzadian et al., 2021a). For such rectangular ambiguity sets, problem (10) can be solved efficiently by updating the value function with the robust Bellman operator $\mathcal{T}_{\boldsymbol{\pi}} : \mathbb{R}^S \to \mathbb{R}^S$. Below, we show forms of Bellman operator within different rectangular conditions.

Example B. L_1-constrained (s, a)-rectangular ambiguity sets generally assume the uncertain in transition probabilities is independent for each state-action pair and are defined as

$$\mathcal{P} = \underset{s\in\mathcal{S}, a\in\mathcal{A}}{\times} \mathcal{P}_{s,a} \quad \text{where} \quad \mathcal{P}_{s,a} := \left\{ \boldsymbol{p} \in \Delta^S \mid \|\boldsymbol{p} - \bar{\boldsymbol{p}}_{sa}\|_1 \le \kappa_{sa} \right\}.$$

For (s, a)-rectangular RMDPs constrained by the L_1-norm, $\mathcal{T}_{\boldsymbol{\pi}}$ is defined for each $s \in \mathcal{S}$ as

$$(\mathcal{T}_{\boldsymbol{\pi}} \boldsymbol{v}^{\boldsymbol{\pi},\boldsymbol{p}})_s := \sum_{a\in\mathcal{A}} \left(\pi_{sa} \cdot \max_{\boldsymbol{p}_{sa}\in\mathcal{P}_{sa}} \left\{ \boldsymbol{p}_{sa}^\top (\boldsymbol{c}_{sa} + \gamma \boldsymbol{v}^{\boldsymbol{\pi},\boldsymbol{p}}) \mid \|\boldsymbol{p}_{sa} - \bar{\boldsymbol{p}}_{sa}\|_1 \le \kappa_{sa} \right\} \right).$$

Example C. L_∞-constrained s-rectangular ambiguity sets generally assume the uncertain in transition probabilities is independent for each state-action pair and are defined as

$$\mathcal{P} = \underset{s\in\mathcal{S}}{\times} \mathcal{P}_s \quad \text{where} \quad \mathcal{P}_s := \left\{ (\boldsymbol{p}_{s1}, \ldots, \boldsymbol{p}_{sA}) \in (\Delta^S)^A \mid \sum_{a\in\mathcal{A}} \|\boldsymbol{p}_{sa} - \bar{\boldsymbol{p}}_{sa}\|_\infty \le \kappa_s \right\}.$$

For s-rectangular RMDPs constrained by the L_∞-norm, $\mathcal{T}_{\boldsymbol{\pi}}$ is defined for each $s \in \mathcal{S}$ as

$$(\mathcal{T}_{\boldsymbol{\pi}} \boldsymbol{v}^{\boldsymbol{\pi},\boldsymbol{p}})_s := \max_{\boldsymbol{p}_s\in\mathcal{P}_s} \left\{ \sum_{a\in\mathcal{A}} \pi_{sa} \cdot \boldsymbol{p}_{sa}^\top (\boldsymbol{c}_{sa} + \gamma \boldsymbol{v}^{\boldsymbol{\pi},\boldsymbol{p}}) \mid \sum_{a\in\mathcal{A}} \|\boldsymbol{p}_{sa} - \bar{\boldsymbol{p}}_{sa}\|_\infty \le \kappa_s \right\}.$$

There exists an unique solution to the Bellman equation $\boldsymbol{v}^{\boldsymbol{\pi},\boldsymbol{p}} = \mathcal{T}_{\boldsymbol{\pi}} \boldsymbol{v}^{\boldsymbol{\pi},\boldsymbol{p}}$, which is called the robust value function (Iyengar, 2005; Wiesemann et al., 2013). Specially, both L_1-constrained ambiguity sets and L_∞-constrained ambiguity sets are in fact polyhedral, which implies the worst-case transition probabilities in bellman updates can be computed as the solution of linear programs (LPs). Instead, RMDPs with other distance-type ambiguity sets, such as L_2-constrained ambiguity sets can compute an Bellman update $\mathcal{T}_{\boldsymbol{\pi}}$ by solving convex optimization problems.

D. Technical Lemmas and Definitions

As promised, we first introduce the definition of the Fréchet sub-differential for general functions.

Definition D.1. The Fréchet sub-differential of a function $h : \mathcal{X} \to \mathbb{R}$ at point $\boldsymbol{x} \in \mathcal{X}$ is defined as the set $\partial h(\boldsymbol{x}) = \{u | \liminf_{\boldsymbol{x}'\to\boldsymbol{x}} h(\boldsymbol{x}') - h(\boldsymbol{x}) - \langle \boldsymbol{u}, \boldsymbol{x}' - \boldsymbol{x}\rangle / \|\boldsymbol{x}' - \boldsymbol{x}\| \ge 0\}$.

Then, a common lemma is provided to illustrate a basic property that a smooth function satisfies.

Lemma D.2. *Let $h\colon \mathcal{X} \to \mathbb{R}$ be ℓ-smooth, then it is a ℓ-weakly convex function.*

Proof of Lemma D.2. Let $r(t) := h(x + t(x' - x))$, for any $x, x' \in \mathcal{X}$. The following holds true

$$h(x) = r(0) \text{ and } h(x') = r(1).$$

Then, we observe that

$$h(x') - h(x) = r(1) - r(0) = \int_0^1 \nabla r(t) dt,$$

where

$$\nabla r(t) = \nabla h(x + t(x' - x))^\top (x' - x).$$

We complete the proof as

$$\begin{aligned}
&\|h(x') - h(x) - \nabla h(x)^\top (x' - x)\| \\
&\quad \le \left\| \int_0^1 \nabla r(t) dt - \nabla h(x)^\top (x' - x) \right\| \\
&\quad \le \int_0^1 \left\| \nabla r(t) - \nabla h(x)^\top (x' - x) \right\| dt \\
&\quad = \int_0^1 \left\| \nabla h(x + t(x' - x))^\top (x' - x) - \nabla h(x)^\top (x' - x) \right\| dt \\
&\quad \le \int_0^1 \|\nabla h(x + t(x' - x)) - \nabla h(x)\| \cdot \|(x' - x)\| dt \\
&\quad \le \int_0^1 t\ell \|x' - x\|^2 dt = \frac{\ell}{2} \|x' - x\|^2.
\end{aligned}$$

□

For smooth function $h(x)$, a point $x \in \mathcal{X}$ is defined as the first-order stationary point (FOSP) when $0 \in \partial h(x)$. However, this notion of stationarity can be very restrictive when optimizing nonsmooth functions (Lin et al., 2020). In respond to this issue, an alternative measure of the first-order stationarity is proposed based on the construction of the Moreau envelope (Thekumparampil et al., 2019).

Definition D.3. For function $h : \mathcal{X} \to \mathbb{R}$ and $\lambda > 0$, the Moreau envelope function of h is given by

$$h_\lambda(x) := \min_{x' \in \mathcal{X}} \left\{ h(x') + \frac{1}{2\lambda} \|x - x'\|^2 \right\}. \tag{18}$$

Definition D.4. Given an ℓ-weakly convex function h, we say that $x^\star$ is an ϵ-first order stationary point (ϵ-FOSP) if, $\|\nabla h_{\frac{1}{2\ell}}(x^\star)\| \le \epsilon$, where $h_{\frac{1}{2\ell}}(x)$ is the Moreau envelope function of h with parameter $\lambda = \frac{1}{2\ell}$.

The following lemma connects ℓ-weakly convex function and its Moreau envelope function and will be useful in our proofs.

Lemma D.5. *(Rockafellar & Wets, 2009, Proposition 13.37) Assume $h \colon \mathcal{X} \to \mathbb{R}$ is a ℓ-weakly convex function. Then, for $\lambda < \ell$, the Moreau envelope function h_λ is C^1-smooth with the gradient given by,*

$$\nabla h_\lambda(x) = \lambda^{-1} \left(x - \arg\min_{x'} \left(h(x') + \frac{1}{2\lambda} \|x - x'\|^2 \right) \right).$$

Lemma D.6. *Assume the function $h : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$ is ℓ-weakly convex and not differentiable at any point. Let $\lambda < \frac{1}{\ell}$ and $\hat{\boldsymbol{x}}_\lambda = \arg\min_{\boldsymbol{x}' \in \mathcal{X}} h(\boldsymbol{x}') + \frac{1}{2\lambda} \|\boldsymbol{x} - \boldsymbol{x}'\|^2$. Then we have*

$$\frac{1}{\lambda} \|\hat{\boldsymbol{x}}_\lambda - \boldsymbol{x}\| = \|\nabla h_\lambda(\boldsymbol{x})\|.$$

As a result, $\|\nabla h_\lambda(\boldsymbol{x})\| \le \epsilon$ implies$\|\hat{x}_\lambda - \boldsymbol{x}\| \le \lambda\epsilon$ and $\exists \boldsymbol{\xi} \in \partial h(\hat{\boldsymbol{x}}_\lambda)$ such that

$$-\boldsymbol{\xi} \in \mathcal{N}_{\mathcal{X}}(\hat{\boldsymbol{x}}_\lambda) + \frac{1}{\lambda} (\hat{\boldsymbol{x}}_\lambda - \boldsymbol{x}) \subseteq \mathcal{N}_{\mathcal{X}}(\hat{\boldsymbol{x}}_\lambda) + \frac{1}{\lambda} \|\hat{\boldsymbol{x}}_\lambda - \boldsymbol{x}\| \mathcal{B}(1),$$

where $\mathcal{N}_{\mathcal{X}}(\hat{\boldsymbol{x}}_\lambda)$ denotes the normal cone of $\mathcal{X}$ at $\hat{\boldsymbol{x}}_\lambda$ and $\mathcal{B}(r) := \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| \le r\}$.

Proof. Here, we consider the function $f(\boldsymbol{x}) = h(\boldsymbol{x}) + \mathbb{I}_{\mathcal{X}}(\boldsymbol{x})$ where $\mathbb{I}$ is the indicate function and here $f(\boldsymbol{x}) := \mathbb{R}^n \to \mathbb{R}$. The Moreau envelope function of $f(\boldsymbol{x})$ is defined as

$$\begin{aligned}
f_\lambda(\boldsymbol{x}) &= \min_{\boldsymbol{x}' \in \mathbb{R}^n} \left\{ h(\boldsymbol{x}') + \mathbb{I}_{\mathcal{X}}(\boldsymbol{x}') + \frac{1}{2\lambda} \|\boldsymbol{x} - \boldsymbol{x}'\|^2 \right\}, \quad \forall \boldsymbol{x} \in \mathbb{R}^n \\
&= \min_{\boldsymbol{x}' \in \mathcal{X}} \left\{ h(\boldsymbol{x}') + \frac{1}{2\lambda} \|\boldsymbol{x} - \boldsymbol{x}'\|^2 \right\}, \quad \forall \boldsymbol{x} \in \mathbb{R}^n.
\end{aligned}$$

The gradient of the moreau envelope $f_\lambda(\boldsymbol{x})$ is well defined (Lemma D.5) as

$$\nabla f_\lambda(\boldsymbol{x}) = \lambda^{-1}\left(\boldsymbol{x} - \hat{\boldsymbol{x}}\right),$$

where

$$\begin{aligned}\hat{\boldsymbol{x}} :&= \operatorname*{arg\,min}_{\bar{\boldsymbol{x}}\in\mathbb{R}^n}\left(\underbrace{h(\bar{\boldsymbol{x}}) + \mathbb{I}_{\mathcal{X}}(\bar{\boldsymbol{x}}) + \frac{1}{2\lambda}\left\|\boldsymbol{x}-\bar{\boldsymbol{x}}\right\|^2}_{:=\phi_{\boldsymbol{x}}(\bar{\boldsymbol{x}})}\right)\\ &= \operatorname*{arg\,min}_{\bar{\boldsymbol{x}}\in\mathcal{X}}\left(h(\bar{\boldsymbol{x}}) + \frac{1}{2\lambda}\left\|\boldsymbol{x}-\bar{\boldsymbol{x}}\right\|^2\right)\end{aligned}$$

Then, we consider the optimality of the function $\phi_{\boldsymbol{x}}(\boldsymbol{y}) = h(\boldsymbol{y}) + \mathbb{I}_{\mathcal{X}}(\boldsymbol{y}) + \frac{1}{2\lambda}\left\|\boldsymbol{x}-\boldsymbol{y}\right\|^2$. Notice that, for any $\boldsymbol{x}\in\mathbb{R}^n$, $\hat{\boldsymbol{x}}$ is the optimal solution of $\phi_{\boldsymbol{x}}(\boldsymbol{y})$, then for some $\boldsymbol{\xi}\in\partial h(\hat{\boldsymbol{x}})$, we have

$$\begin{aligned}\phi_{\boldsymbol{x}}(\hat{\boldsymbol{x}}(\boldsymbol{x})) = \min_{\boldsymbol{y}\in\mathbb{R}^n}\phi_{\boldsymbol{x}}(\boldsymbol{y}) &\iff \phi_{\boldsymbol{x}}(\hat{\boldsymbol{x}}(\boldsymbol{x})) = \min_{\boldsymbol{y}\in\mathbb{R}^n} h(\boldsymbol{y}) + \mathbb{I}_{\mathcal{X}}(\boldsymbol{y}) + \frac{1}{2\lambda}\left\|\boldsymbol{x}-\boldsymbol{y}\right\|^2\\ &\iff 0\in\partial\left(h(\boldsymbol{y}) + \mathbb{I}_{\mathcal{X}}(\boldsymbol{y}) + \frac{1}{2\lambda}\left\|\boldsymbol{x}-\boldsymbol{y}\right\|^2\right)\Big|_{\boldsymbol{y}=\hat{\boldsymbol{x}}},\\ &\iff 0\in\boldsymbol{\xi} + \mathcal{N}_{\mathcal{X}}(\hat{\boldsymbol{x}}) + \frac{1}{\lambda}\left(\hat{\boldsymbol{x}}-\boldsymbol{x}\right)\\ &\iff -\boldsymbol{\xi}\in\mathcal{N}_{\mathcal{X}}(\hat{\boldsymbol{x}}) + \frac{1}{\lambda}\left(\hat{\boldsymbol{x}}-\boldsymbol{x}\right).\end{aligned} \tag{19}$$

The above equation (19) implies that, for any $\boldsymbol{z}\in\mathbb{R}^n$,

$$\begin{aligned}\langle\boldsymbol{\xi} + \frac{1}{\lambda}\left(\hat{\boldsymbol{x}}-\boldsymbol{x}\right), \boldsymbol{z}-\hat{\boldsymbol{x}}\rangle \geq 0 &\iff \langle -\boldsymbol{\xi}, \boldsymbol{z}-\hat{\boldsymbol{x}}\rangle \leq \langle\frac{1}{\lambda}\left(\hat{\boldsymbol{x}}-\boldsymbol{x}\right), \boldsymbol{z}-\hat{\boldsymbol{x}}\rangle,\ \forall\boldsymbol{z}\in\mathbb{R}^n\\ &\iff \langle -\boldsymbol{\xi}, \boldsymbol{z}-\hat{\boldsymbol{x}}\rangle \leq \frac{1}{\lambda}\|\hat{\boldsymbol{x}}-\boldsymbol{x}\|\cdot\|\boldsymbol{z}-\hat{\boldsymbol{x}}\|,\ \forall\boldsymbol{z}\in\mathbb{R}^n\\ &\iff \langle -\boldsymbol{\xi}, \boldsymbol{z}-\hat{\boldsymbol{x}}\rangle \leq \frac{1}{\lambda}\|\hat{\boldsymbol{x}}-\boldsymbol{x}\|,\ \forall\boldsymbol{z}\in\mathbb{R}^n,\ \|\boldsymbol{z}-\hat{\boldsymbol{x}}\| = 1.\end{aligned} \tag{20}$$

□

The above Lemma D.6 implies that if $\|\nabla h_\lambda(\boldsymbol{x})\|$ is small enough, then $\boldsymbol{x}$ is an approximate stationary point of the original constrained optimization $\min_{\mathcal{X}} h(x)$, by the definition of ϵ-FOSP. This motivates us to consider the optimality of the Moreau envelope function of $\Phi(\boldsymbol{\pi})$ instead of the optimality of $\Phi(\boldsymbol{\pi})$ directly.

E. Proofs of Section 3

Proof of Lemma 3.1. First, we first derive the form of partial derivative for π_{sa} to obtain (6). While this form was known (Agarwal et al., 2019), we included a proof for the sake of completeness. Notice that,

$$\frac{\partial J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p})}{\partial\pi_{sa}} = \sum_{\hat{s}\in\mathcal{S}}\frac{\partial v_{\hat{s}}^{\boldsymbol{\pi},\boldsymbol{p}}}{\partial\pi_{sa}}\rho_{\hat{s}}.$$

Then, we discuss $\frac{\partial v_{\hat{s}}^{\boldsymbol{\pi},\boldsymbol{p}}}{\partial\pi_{sa}}$ over two cases: $\hat{s}\neq s$ and $\hat{s} = s$

$$\begin{aligned}\frac{\partial v_{\hat{s}}^{\boldsymbol{\pi},\boldsymbol{p}}}{\partial\pi_{sa}}\Big|_{\hat{s}\neq s} &= \frac{\partial}{\partial\pi_{sa}}\left[\sum_{\hat{a}}\pi_{\hat{s}\hat{a}}\sum_{s'\in\mathcal{S}}p_{\hat{s}\hat{a}s'}\left(c_{\hat{s}\hat{a}s'} + \gamma v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}}\right)\right] = \gamma\sum_{\hat{a}}\pi_{\hat{s}\hat{a}}\sum_{s'\in\mathcal{S}}p_{\hat{s}\hat{a}s'}\frac{\partial v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}}}{\partial\pi_{sa}};\\ \frac{\partial v_{\hat{s}}^{\boldsymbol{\pi},\boldsymbol{p}}}{\partial\pi_{sa}}\Big|_{\hat{s}=s} &= \frac{\partial}{\partial\pi_{sa}}\left[\sum_{\hat{a}}\pi_{s\hat{a}}\underbrace{\sum_{s'\in\mathcal{S}}p_{s\hat{a}s'}\left(c_{s\hat{a}s'} + \gamma v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}}\right)}_{q_{s\hat{a}}^{\boldsymbol{\pi},\boldsymbol{p}}}\right] = q_{sa}^{\boldsymbol{\pi},\boldsymbol{p}} + \gamma\sum_{\hat{a}}\pi_{s\hat{a}}\sum_{s'\in\mathcal{S}}p_{s\hat{a}s'}\frac{\partial v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}}}{\partial\pi_{sa}};\end{aligned}$$

Condense the notation

$$\sum_{\hat{a}} \pi_{s\hat{a}} p_{s\hat{a}s'} = p^{\boldsymbol{\pi}}_{ss'}(1) \tag{21}$$

$$p^{\boldsymbol{\pi}}_{ss'}(t-1) \cdot \sum_{a} \pi_{s'a} p_{s'as''} = p^{\boldsymbol{\pi}}_{ss''}(t) \tag{22}$$

Then, combining these two equations, we can obtain,

$$\begin{aligned}
\frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_{\hat{s}}}{\partial \pi_{sa}}\Big|_{\hat{s}\neq s} &= \gamma \sum_{s'\neq s} p^{\boldsymbol{\pi}}_{\hat{s}s'}(1) \frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_{s'}}{\partial \pi_{sa}} + \gamma \sum_{s'=s} p^{\boldsymbol{\pi}}_{\hat{s}s'}(1) \frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_{s'}}{\partial \pi_{sa}} \\
&= \gamma^2 \sum_{s'\neq s} p^{\boldsymbol{\pi}}_{\hat{s}s'}(1) \sum_{\hat{a}} \pi_{s'\hat{a}} \sum_{s''\in\mathcal{S}} p_{s'\hat{a}s''} \frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_{s''}}{\partial \pi_{sa}} \\
&+ \gamma p^{\boldsymbol{\pi}}_{\hat{s}s}(1) \left(q^{\boldsymbol{\pi},\boldsymbol{p}}_{sa} + \gamma \sum_{\hat{a}} \pi_{s\hat{a}} \sum_{s'\in\mathcal{S}} p_{s\hat{a}s'} \frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_{s'}}{\partial \pi_{sa}} \right) \\
&= \gamma p^{\boldsymbol{\pi}}_{\hat{s}s}(1) q^{\boldsymbol{\pi},\boldsymbol{p}}_{sa} + \gamma^2 \sum_{s'} p^{\boldsymbol{\pi}}_{\hat{s}s'}(2) \frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_{s'}}{\partial \pi_{sa}} \\
&= \gamma p^{\boldsymbol{\pi}}_{\hat{s}s}(1) q^{\boldsymbol{\pi},\boldsymbol{p}}_{sa} + \gamma^2 p^{\boldsymbol{\pi}}_{\hat{s}s}(2) q^{\boldsymbol{\pi},\boldsymbol{p}}_{sa} + \gamma^3 \sum_{s'} p^{\boldsymbol{\pi}}_{\hat{s}s'}(3) \frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_{s'}}{\partial \pi_{sa}} \\
&= \cdots \\
&= \sum_{t=1}^{\infty} \gamma^t p^{\boldsymbol{\pi}}_{\hat{s}s}(t) q^{\boldsymbol{\pi},\boldsymbol{p}}_{sa} = \sum_{t=0}^{\infty} \gamma^t p^{\boldsymbol{\pi}}_{\hat{s}s}(t) q^{\boldsymbol{\pi},\boldsymbol{p}}_{sa}.
\end{aligned}$$

The last equality is from the initial assumption $\hat{s} \neq s$, *i.e.*, $p^{\boldsymbol{\pi}}_{\hat{s}s}(0) = 0$, and similarly for the case $\hat{s} = s$ we have,

$$\frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_{\hat{s}}}{\partial \pi_{sa}}\Big|_{\hat{s}=s} = \sum_{t=0}^{\infty} \gamma^t p^{\boldsymbol{\pi}}_{ss}(t) q^{\boldsymbol{\pi},\boldsymbol{p}}_{sa}.$$

Hence, the partial derivative is obtained

$$\frac{\partial J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p})}{\partial \pi_{sa}} = \left(\frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_{s}}{\partial \pi_{sa}} \rho_s + \sum_{\hat{s}\neq s} \frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_{\hat{s}}}{\partial \pi_{sa}} \rho_{\hat{s}} \right) = \frac{1}{1-\gamma} \left(\underbrace{(1-\gamma) \sum_{\hat{s}\in\mathcal{S}} \sum_{t=0}^{\infty} \gamma^t \rho_{\hat{s}} p^{\boldsymbol{\pi}}_{\hat{s}s}(t)}_{d^{\boldsymbol{\pi},\boldsymbol{p}}_{\boldsymbol{\rho}}(s)} \right) q^{\boldsymbol{\pi},\boldsymbol{p}}_{sa}.$$

After deriving the form of partial derivative, we next prove that $J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p})$ is $L_{\boldsymbol{\pi}}$-Lipschitz in $\boldsymbol{\pi}$ by showing the boundedness of $\nabla_{\boldsymbol{\pi}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p})$. The uniformly bounded cost $c_{sas'}$ implies that, the absolute value of the action value function is bounded for any policy $\boldsymbol{\pi}$ and transition kernel $\boldsymbol{p}$,

$$|q^{\boldsymbol{\pi},\boldsymbol{p}}_{sa}| = \left| \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{p}} \left[\sum_{t=0}^{\infty} \gamma^t c_{s_t a_t s_{t+1}} \mid s_0 = s, a_0 = a \right] \right| \leq \sum_{t=0}^{\infty} \gamma^t = \frac{1}{1-\gamma}.$$

Then, by vectorizing the $\boldsymbol{\pi}$ as a SA-dimensional vector, we have

$$\begin{aligned}
\|\nabla_{\boldsymbol{\pi}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p})\| &= \sqrt{\sum_{s,a} \left(\frac{\partial J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p})}{\partial \pi_{sa}} \right)^2} \\
&= \frac{1}{1-\gamma} \sqrt{\sum_{a} \sum_{s} \left(d^{\boldsymbol{\pi},\boldsymbol{p}}_{\boldsymbol{\rho}}(s) q^{\boldsymbol{\pi},\boldsymbol{p}}_{sa} \right)^2} \\
&\leq \frac{1}{(1-\gamma)^2} \sqrt{\sum_{a} \sum_{s} (d^{\boldsymbol{\pi},\boldsymbol{\xi}}_{\boldsymbol{\rho}}(s))^2} \leq \frac{\sqrt{A}}{(1-\gamma)^2},
\end{aligned}$$

where the last inequality holds since the discounted state occupancy measure satisfies

$$\sum_s (d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{\xi}}(s))^2 \leq \left(\sum_s (d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{\xi}}(s))\right)^2 = 1.$$

About the smoothness of $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$, it can be immediately proved by (Agarwal et al., 2021, Lemma 54). Finally, we turn to derive the continuity of $\Phi(\boldsymbol{\pi})$. 1. We first show $\Phi(\boldsymbol{\pi})$ is $L_{\boldsymbol{\pi}}$-Lipschitz if $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$ is $L_{\boldsymbol{\pi}}$-Lipschitz in $\boldsymbol{\pi}$. For any $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2 \in \Pi$, we let $\boldsymbol{p}_1 := \arg\max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_1, \boldsymbol{p})$ and $\boldsymbol{p}_2 := \arg\max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_2, \boldsymbol{p})$, then

$$\begin{aligned}
\Phi(\boldsymbol{\pi}_1) - \Phi(\boldsymbol{\pi}_2) &= \max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_1, \boldsymbol{p}) - \max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_2, \boldsymbol{p}) \\
&= J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_1, \boldsymbol{p}_1) - J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_2, \boldsymbol{p}_2) \\
&\leq J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_1, \boldsymbol{p}_1) - J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_2, \boldsymbol{p}_1) \\
&\leq L_{\boldsymbol{\pi}} \|\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2\|.
\end{aligned}$$

2. Then, (Thekumparampil et al., 2019, Lemma 3) shows that, $\Phi(\boldsymbol{\pi}) = \max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$ is $\ell_{\boldsymbol{\pi}}$-weakly convex if $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$ is $\ell_{\boldsymbol{\pi}}$-smooth. Combining the results of these two parts, this lemma is proved. □

The following lemma is helpful throughout in the convergence analysis of policy optimization.

Lemma E.1. *(The performance difference lemma) For any $\boldsymbol{\pi}, \boldsymbol{\pi}' \in \Pi$, $\boldsymbol{p} \in \mathcal{P}$ and $\boldsymbol{\rho} \in \Delta^S$, we have*

$$J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}) - J_{\boldsymbol{\rho}}(\boldsymbol{\pi}', \boldsymbol{p}) = \frac{1}{1-\gamma} \sum_{s,a} d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}}(s) \pi_{sa} \left(q_{sa}^{\boldsymbol{\pi}',\boldsymbol{p}} - v_s^{\boldsymbol{\pi}',\boldsymbol{p}} \right). \tag{23}$$

Generally, the term $q_{sa}^{\boldsymbol{\pi},\boldsymbol{p}} - v_s^{\boldsymbol{\pi},\boldsymbol{p}}$ is defined as the *advantage function*.

Proof of Lemma E.1. By the definition of $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$ in (2), we have

$$J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}) - J_{\boldsymbol{\rho}}(\boldsymbol{\pi}', \boldsymbol{p}) = \sum_s \rho_s \left(v_s^{\boldsymbol{\pi},\boldsymbol{p}} - v_s^{\boldsymbol{\pi}',\boldsymbol{p}} \right).$$

We introduce the *advantage function* $A_{sa}^{\boldsymbol{\pi},\boldsymbol{p}} := q_{sa}^{\boldsymbol{\pi},\boldsymbol{p}} - v_s^{\boldsymbol{\pi},\boldsymbol{p}}$ for convenience, and observe that, for any $s \in \mathcal{S}$,

$$
\begin{aligned}
& v_s^{\boldsymbol{\pi},\boldsymbol{p}} - v_s^{\boldsymbol{\pi}',\boldsymbol{p}} \\
&= v_s^{\boldsymbol{\pi},\boldsymbol{p}} - \sum_a \pi_{sa} \sum_{s'} p_{sas'} \left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi}',\boldsymbol{p}} \right) + \sum_a \pi_{sa} \sum_{s'} p_{sas'} \left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi}',\boldsymbol{p}} \right) - v_s^{\boldsymbol{\pi}',\boldsymbol{p}} \\
&= \sum_a \pi_{sa} \sum_{s'} p_{sas'} \left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}} \right) - \sum_a \pi_{sa} \sum_{s'} p_{sas'} \left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi}',\boldsymbol{p}} \right) \\
&\quad + \sum_a \pi_{sa} \underbrace{\sum_{s'} p_{sas'} \left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi}',\boldsymbol{p}} \right)}_{q_{sa}^{\boldsymbol{\pi}',\boldsymbol{p}}} - v_s^{\boldsymbol{\pi}',\boldsymbol{p}} \\
&= \gamma \sum_a \pi_{sa} \sum_{s'} p_{sas'} \left(v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}} - v_{s'}^{\boldsymbol{\pi}',\boldsymbol{p}} \right) + \sum_a \pi_{sa} \left(q_{sa}^{\boldsymbol{\pi}',\boldsymbol{p}} - v_s^{\boldsymbol{\pi}',\boldsymbol{p}} \right) \\
&= \gamma \sum_a \pi_{sa} \sum_{s'} p_{sas'} \left(v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}} - v_{s'}^{\boldsymbol{\pi}',\boldsymbol{p}} \right) + \sum_a \pi_{sa} A_{sa}^{\boldsymbol{\pi}',\boldsymbol{p}} \\
&\overset{(a)}{=} \gamma \sum_{s'} p_{ss'}^{\boldsymbol{\pi}}(1) \left(\gamma \sum_{s''} p_{s's''}^{\boldsymbol{\pi}}(1) \left(v_{s''}^{\boldsymbol{\pi},\boldsymbol{p}} - v_{s''}^{\boldsymbol{\pi}',\boldsymbol{p}} \right) + \sum_{a'} \pi_{s'a'} A_{s'a'}^{\boldsymbol{\pi}',\boldsymbol{p}} \right) + \sum_a \pi_{sa} A_{sa}^{\boldsymbol{\pi}',\boldsymbol{p}} \\
&= \sum_a \pi_{sa} A_{sa}^{\boldsymbol{\pi}',\boldsymbol{p}} + \gamma \sum_{s'} p_{ss'}^{\boldsymbol{\pi}}(1) \sum_{a'} \pi_{s'a'} A_{s'a'}^{\boldsymbol{\pi}',\boldsymbol{p}} + \gamma^2 \sum_{s'} p_{ss'}^{\boldsymbol{\pi}}(2) \left(v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}} - v_{s'}^{\boldsymbol{\pi}',\boldsymbol{p}} \right) \\
&= \cdots \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{s'} p_{ss'}^{\boldsymbol{\pi}}(t) \left(\sum_{a'} \pi_{s'a'} A_{s'a'}^{\boldsymbol{\pi}',\boldsymbol{p}} \right),
\end{aligned}
$$

where $p_{ss'}^{\boldsymbol{\pi}}(t)$ is defined in (21), and (a) uses the recursion. We then obtain

$$
\begin{aligned}
J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p}) - J_{\boldsymbol{\rho}}(\boldsymbol{\pi}',\boldsymbol{p}) &= \sum_s \rho_s \left(v_s^{\boldsymbol{\pi},\boldsymbol{p}} - v_s^{\boldsymbol{\pi}',\boldsymbol{p}} \right) \\
&= \sum_s \rho_s \sum_{t=0}^{\infty} \gamma^t \sum_{s'} p_{ss'}^{\boldsymbol{\pi}}(t) \left(\sum_{a'} \pi_{s'a'} A_{s'a'}^{\boldsymbol{\pi}',\boldsymbol{p}} \right) \\
&= \sum_{s'} \left(\sum_s \sum_{t=0}^{\infty} \gamma^t \rho_s p_{ss'}^{\boldsymbol{\pi}}(t) \right) \left(\sum_{a'} \pi_{s'a'} A_{s'a'}^{\boldsymbol{\pi}',\boldsymbol{p}} \right) \\
&= \frac{1}{1-\gamma} \sum_{s,a} d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}}(s) \pi_{sa} A_{sa}^{\boldsymbol{\pi}',\boldsymbol{p}}.
\end{aligned}
$$

The last equality is obtained by the definition of state occupancy measure (See Definition 2.2). □

Then, we introduce the gradient dominance condition for non-RMDPs proposed in (Agarwal et al., 2021), which will be used in the proof of Theorem 3.2.

Lemma E.2 (Gradient dominance)**.** *For any $\boldsymbol{p} \in \mathcal{P}$ and $\boldsymbol{\rho} \in \Delta^S$, we have*

$$
J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p}) - J_{\boldsymbol{\rho}}(\boldsymbol{\pi}^\star,\boldsymbol{p}) \leq \frac{D}{1-\gamma} \max_{\bar{\boldsymbol{\pi}} \in \Pi} (\boldsymbol{\pi} - \bar{\boldsymbol{\pi}})^\top \nabla_{\boldsymbol{\pi}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p}), \tag{24}
$$

where $\boldsymbol{\pi}^\star$ is one of optimal policies over $\boldsymbol{p}$, i.e., $\boldsymbol{\pi}^\star \in \arg\min_{\boldsymbol{\pi} \in \Pi} J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p})$.

Proof of Lemma E.2. From the Lemma E.1, we have

$$\begin{aligned}
J_{\boldsymbol{\rho}}(\boldsymbol{\pi}^\star, \boldsymbol{p}) - J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}) &= \frac{1}{1-\gamma} \sum_{s,a} d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}^\star, \boldsymbol{p}}(s) \pi^\star_{sa} \left(q_{sa}^{\boldsymbol{\pi}, \boldsymbol{p}} - v_s^{\boldsymbol{\pi}, \boldsymbol{p}} \right) \\
&= \frac{1}{1-\gamma} \sum_{s,a} d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}^\star, \boldsymbol{p}}(s) \pi^\star_{sa} A_{sa}^{\boldsymbol{\pi}, \boldsymbol{p}} \\
&\geq \frac{1}{1-\gamma} \sum_{s,a} d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}^\star, \boldsymbol{p}}(s) \pi^\star_{sa} \min_{\bar{a}} A_{s\bar{a}}^{\boldsymbol{\pi}, \boldsymbol{p}} \\
&= \frac{1}{1-\gamma} \sum_{s} d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}^\star, \boldsymbol{p}}(s) \min_{\bar{a}} A_{s\bar{a}}^{\boldsymbol{\pi}, \boldsymbol{p}}.
\end{aligned}$$

Then, we multiply -1 on both sides

$$\begin{aligned}
0 \leq J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}) - J_{\boldsymbol{\rho}}(\boldsymbol{\pi}^\star, \boldsymbol{p}) &\leq \frac{1}{1-\gamma} \sum_{s} d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}^\star, \boldsymbol{p}}(s) - (\min_{\bar{a}} A_{s\bar{a}}^{\boldsymbol{\pi}, \boldsymbol{p}}) \\
&= \frac{1}{1-\gamma} \sum_{s} d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}^\star, \boldsymbol{p}}(s) \max_{\bar{a}} (-A_{s\bar{a}}^{\boldsymbol{\pi}, \boldsymbol{p}}) \\
&= \frac{1}{1-\gamma} \sum_{s} \frac{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}^\star, \boldsymbol{p}}(s)}{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}, \boldsymbol{p}}(s)} d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}, \boldsymbol{p}}(s) \max_{\bar{a}} (-A_{s\bar{a}}^{\boldsymbol{\pi}, \boldsymbol{p}}) \\
&\leq \frac{1}{1-\gamma} \left(\max_{s} \frac{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}^\star, \boldsymbol{p}}(s)}{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}, \boldsymbol{p}}(s)} \right) \sum_{s} d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}, \boldsymbol{p}}(s) \max_{\bar{a}} (-A_{s\bar{a}}^{\boldsymbol{\pi}, \boldsymbol{p}}) \\
&= \left\| \frac{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}^\star, \boldsymbol{p}}}{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}, \boldsymbol{p}}} \right\|_\infty \frac{1}{1-\gamma} \sum_{s} d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}, \boldsymbol{p}}(s) \max_{\bar{a}} (-A_{s\bar{a}}^{\boldsymbol{\pi}, \boldsymbol{p}}) \\
&\leq \frac{D}{1-\gamma} \left(\frac{1}{1-\gamma} \sum_{s} d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}, \boldsymbol{p}}(s) \max_{\bar{a}} (-A_{s\bar{a}}^{\boldsymbol{\pi}, \boldsymbol{p}}) \right) \qquad (25)
\end{aligned}$$

The last inequality (25) is due to the fact (Kakade & Langford, 2002)

$$\left\| \frac{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}^\star, \boldsymbol{p}}}{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}, \boldsymbol{p}}} \right\|_\infty \leq \frac{1}{1-\gamma} \left\| \frac{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}^\star, \boldsymbol{p}}}{\boldsymbol{\rho}} \right\|_\infty \leq \frac{D}{1-\gamma}.$$

Notice that, the term in (25) is equivalent to

$$\begin{aligned}
\frac{1}{1-\gamma} \sum_{s} d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}, \boldsymbol{p}}(s) \max_{\bar{a}} (-A_{s\bar{a}}^{\boldsymbol{\pi}, \boldsymbol{p}}) &= \max_{\bar{\boldsymbol{\pi}} \in \Pi} \frac{1}{1-\gamma} \sum_{s,a} d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}, \boldsymbol{p}}(s) \bar{\pi}_{sa} (-A_{sa}^{\boldsymbol{\pi}, \boldsymbol{p}}) \\
&= \max_{\bar{\boldsymbol{\pi}} \in \Pi} \frac{1}{1-\gamma} \sum_{s,a} d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}, \boldsymbol{p}}(s) (\bar{\pi}_{sa} - \pi_{sa}) (-A_{sa}^{\boldsymbol{\pi}, \boldsymbol{p}}) \\
&= \max_{\bar{\boldsymbol{\pi}} \in \Pi} \frac{1}{1-\gamma} \sum_{s,a} d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}, \boldsymbol{p}}(s) (\pi_{sa} - \bar{\pi}_{sa}) q_{sa}^{\boldsymbol{\pi}, \boldsymbol{p}} \\
&= \max_{\bar{\boldsymbol{\pi}} \in \Pi} (\boldsymbol{\pi} - \bar{\boldsymbol{\pi}})^\top \nabla_{\boldsymbol{\pi}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}).
\end{aligned}$$

The first equality holds since the optimal $\bar{\boldsymbol{\pi}}$ is a deterministic policy, *i.e.*, for some $\bar{a} \in \mathcal{A}$, $\bar{\pi}_{s\bar{a}} = 1$. The second step is supported by the property $\sum_a \pi_{sa} A_{sa}^{\boldsymbol{\pi}, \boldsymbol{p}} = 0$. The third step follows as $\sum_a (\pi_{sa} - \bar{\pi}_{sa}) v_s^{\boldsymbol{\pi}, \boldsymbol{p}} = 0$ and the last equation is obtained from Lemma 3.1. Thus, we obtain that

$$J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}) - J_{\boldsymbol{\rho}}(\boldsymbol{\pi}^\star, \boldsymbol{p}) \leq \frac{D}{1-\gamma} \max_{\bar{\boldsymbol{\pi}} \in \Pi} (\boldsymbol{\pi} - \bar{\boldsymbol{\pi}})^\top \nabla_{\boldsymbol{\pi}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}).$$

□

Before providing the proof of Theorem 3.3, we introduce the below intermediate results which are helpful to our proof. We first introduce a common property for strongly convex functions.

Lemma E.3. *Let $h : \mathcal{X} \to \mathbb{R}$ be a ℓ-strongly convex function. Then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$, $h(\boldsymbol{x})$, we have*

$$h(\boldsymbol{y}) - h(\boldsymbol{x}) \leq \nabla h(\boldsymbol{y})^\top (\boldsymbol{y} - \boldsymbol{x}) - \frac{\ell}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2. \tag{26}$$

Moreover, by taking $\boldsymbol{y} = \boldsymbol{x}^\star := \arg\min_{\boldsymbol{x}\in\mathcal{X}} h(\boldsymbol{x})$ as the minimum point of $h(\boldsymbol{x})$, we get

$$h(\boldsymbol{x}) - \min_{\boldsymbol{x}\in\mathcal{X}} h(\boldsymbol{x}) \geq \frac{\ell}{2}\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2. \tag{27}$$

Proof of Lemma E.3. The inequality (26) is a basic property that the strongly convex function hold, whereas the second inequality is obtained by the first-order optimality condition for the convex optimization problem, *i.e.*, $\nabla h(\boldsymbol{x}^\star)^\top (\boldsymbol{x} - \boldsymbol{x}^\star) \geq 0$. □

We also need to introduce the following Danskin's Theorem, which helps prove our global convergence theorem.

Proposition E.4. *(Bertsekas, 2016, Proposition B.25) Let $\mathcal{Z} \subseteq \mathbb{R}^m$ be a compact set, and let $\phi : \mathbb{R}^n \times \mathcal{Z} \to \mathbb{R}$ be continuous function and such that $\phi(\cdot, z) : \mathbb{R}^n \to \mathbb{R}$ is convex for each $z \in \mathcal{Z}$. If $\phi(\cdot, z)$ is differentiable for all $z \in \mathcal{Z}$ and $\nabla\phi(\boldsymbol{x}, \cdot)$ is continuous on $\mathcal{Z}$ for each $\boldsymbol{x}$, then for $f(\boldsymbol{x}) := \max_{z\in\mathcal{Z}} \phi(\boldsymbol{x}, z)$ and any $\boldsymbol{x} \in \mathbb{R}^n$,*

$$\partial f(\boldsymbol{x}) = \operatorname{conv}\left\{\nabla_x \phi(\boldsymbol{x}, z) \mid z \in \arg\max_{z\in\mathcal{Z}} \phi(\boldsymbol{x}, z)\right\}.$$

Notice that, Lemma 3.1 successfully proves that $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$ is $\ell_{\boldsymbol{\pi}}$-smooth and $L_{\boldsymbol{\pi}}$-Lipschitz in $\boldsymbol{\pi}$. We want to emphasize that, these results also leads to the fact that $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$ is $\ell_{\boldsymbol{\pi}}$-weakly convex in $\boldsymbol{\pi}$ by applying the Lemma D.2.

Now, we are ready to prove Theorem 3.2 and Theorem 3.3.

Proof of Theorem 3.2. Since $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$ is non-concave in $\boldsymbol{p}$ and the ambiguity set $\mathcal{P}$ is only assumed as a compact set, there may exists multiple inner maxima. In particular, we denote $\boldsymbol{p}^{(k)}$ as the k-th element of the set $\arg\max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$ for fixed policy $\boldsymbol{\pi} \in \Pi$. Then, we apply Lemma E.2 to obtain

$$\begin{aligned}
\Phi(\boldsymbol{\pi}) - \Phi(\boldsymbol{\pi}^\star) &= J(\boldsymbol{\pi}, \boldsymbol{p}^{(k)}) - J(\boldsymbol{\pi}^\star, \boldsymbol{p}^\star) \\
&= J(\boldsymbol{\pi}, \boldsymbol{p}^{(k)}) - \min_{\boldsymbol{\pi}\in\Pi} \max_{\boldsymbol{p}\in\mathcal{P}} J(\boldsymbol{\pi}, \boldsymbol{p}) \\
&\leq J(\boldsymbol{\pi}, \boldsymbol{p}^{(k)}) - \min_{\boldsymbol{\pi}\in\Pi} J(\boldsymbol{\pi}, \boldsymbol{p}^{(k)}) \\
&\leq \frac{D}{1-\gamma} \max_{\bar{\boldsymbol{\pi}}} (\boldsymbol{\pi} - \bar{\boldsymbol{\pi}})^\top \nabla_{\boldsymbol{\pi}} J(\boldsymbol{\pi}, \boldsymbol{p}^{(k)}).
\end{aligned} \tag{28}$$

As we mentioned before this proof that, $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$ is $\ell_{\boldsymbol{\pi}}$-weakly convex in $\boldsymbol{\pi}$, it implies that $\tilde{J}_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}) := J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}) + \frac{\ell_{\boldsymbol{\pi}}}{2}\|\boldsymbol{\pi}\|^2$ is convex in $\boldsymbol{\pi}$ and $\nabla_{\boldsymbol{\pi}} \tilde{J}_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}) = \nabla_{\boldsymbol{\pi}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}) + \ell_{\boldsymbol{\pi}}\boldsymbol{\pi}$, referring to (Kruger, 2003, Corollary 1.12.2). Let $\tilde{\Phi}(\boldsymbol{\pi}) := \max_{\boldsymbol{p}\in\mathcal{P}} \tilde{J}_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$. Due to the convexity of $\tilde{J}_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$ and the compactness of $\mathcal{P}$, we can apply Proposition E.4 to attain

$$\begin{aligned}
\partial\tilde{\Phi}(\boldsymbol{\pi}) &= \operatorname{conv}\left\{\nabla_{\boldsymbol{\pi}} \tilde{J}_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}) \mid \boldsymbol{p} \in \arg\max_{\boldsymbol{p}\in\mathcal{P}} \tilde{J}_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})\right\} \\
\implies \partial\Phi(\boldsymbol{\pi}) + \ell_{\boldsymbol{\pi}}\boldsymbol{\pi} &= \operatorname{conv}\left\{\nabla_{\boldsymbol{\pi}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}) + \ell_{\boldsymbol{\pi}}\boldsymbol{\pi} \mid \boldsymbol{p} \in \arg\max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})\right\} \\
\implies \partial\Phi(\boldsymbol{\pi}) &= \operatorname{conv}\left\{\nabla_{\boldsymbol{\pi}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}) \mid \boldsymbol{p} \in \arg\max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})\right\}.
\end{aligned} \tag{29}$$

Assume the set $\arg\max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$ contains N finite components, then, Proposition E.4 implies that, for any $\boldsymbol{\pi} \in \Pi$, there exists a sequence $\{\beta_k\}_{k=1}^N$ with $\sum_k \beta_k = 1$ such that for any sub-gradient $\boldsymbol{\xi} \in \partial\Phi(\boldsymbol{\pi})$, it can be represented by a

convex combination, *i.e.*,

$$\boldsymbol{\xi} = \sum_{k=1}^{N} \beta_k \nabla_{\boldsymbol{\pi}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}^{(k)}) \quad \boldsymbol{p}^{(k)} \in \arg\max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}),\ k = 1, 2, \cdots, N$$

Let us define $\tilde{\boldsymbol{\pi}} = \arg\min_{\hat{\boldsymbol{\pi}}\in\Pi} \Phi(\hat{\boldsymbol{\pi}}) + \ell_{\boldsymbol{\pi}} \|\boldsymbol{\pi} - \hat{\boldsymbol{\pi}}\|^2$ and Lemma D.6 implies that there exists $\bar{\boldsymbol{\xi}} \in \partial\Phi(\tilde{\boldsymbol{\pi}})$ such that it satisfies $-\bar{\boldsymbol{\xi}} \subseteq \mathcal{N}_{\mathcal{X}}(\tilde{\boldsymbol{\pi}}) + 2\ell_{\boldsymbol{\pi}} \|\tilde{\boldsymbol{\pi}} - \boldsymbol{\pi}\| \cdot \mathcal{B}(1)$. Then by assuming $\arg\max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\tilde{\boldsymbol{\pi}}, \boldsymbol{p})$ contains $\bar{N}$ finite components, there exists a specific sequence $\{\bar{\beta}_k\}_{k=1}^{\bar{N}}$ with $\sum_k \bar{\beta}_k = 1$ such that

$$\bar{\boldsymbol{\xi}} = \sum_{k}^{\bar{N}} \bar{\beta}_k \nabla_{\boldsymbol{\pi}} J_{\boldsymbol{\rho}}(\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{p}}^{(k)}), \quad \tilde{\boldsymbol{p}}^{(k)} \in \arg\max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\tilde{\boldsymbol{\pi}}, \boldsymbol{p}),\ k = 1, 2, \cdots, \bar{N} \tag{30}$$

Then, we have

$$\begin{aligned}
\Phi(\tilde{\boldsymbol{\pi}}) - \Phi(\boldsymbol{\pi}^\star) &= \sum_{k=1}^{\bar{N}} \bar{\beta}_k \left(\Phi(\tilde{\boldsymbol{\pi}}) - \Phi(\boldsymbol{\pi}^\star)\right) \\
&\leq \frac{D}{1-\gamma} \sum_{k=1}^{\bar{N}} \bar{\beta}_k \left(\max_{\bar{\boldsymbol{\pi}}\in\Pi} (\tilde{\boldsymbol{\pi}} - \bar{\boldsymbol{\pi}})^\top \nabla_{\boldsymbol{\pi}} J(\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{p}}^{(k)}) \right) \\
&\leq \frac{D}{1-\gamma} \sum_{k=1}^{\bar{N}} \bar{\beta}_k \langle \max_{\bar{\boldsymbol{\pi}}\in\Pi} (\bar{\boldsymbol{\pi}} - \tilde{\boldsymbol{\pi}}), -\nabla_{\boldsymbol{\pi}} J(\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{p}}^{(k)}) \rangle \\
&\leq \frac{D}{1-\gamma} \sum_{k=1}^{\bar{N}} \bar{\beta}_k \langle (\bar{\boldsymbol{\pi}}_k - \tilde{\boldsymbol{\pi}}), -\nabla_{\boldsymbol{\pi}} J(\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{p}}^{(k)}) \rangle,
\end{aligned} \tag{31}$$

where $\bar{\boldsymbol{\pi}}_k := \arg\max_{\bar{\boldsymbol{\pi}}\in\Pi} \langle (\bar{\boldsymbol{\pi}} - \tilde{\boldsymbol{\pi}}), -\nabla_{\boldsymbol{\pi}} J(\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{p}}^{(k)}) \rangle$, and the second step is obtained by using (28). Since the cost function is bounded, *i.e.*, $0 \leq c_{sas'} \leq 1$ for any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, it implies the action value function $q_{s,a}^{\boldsymbol{\pi},\boldsymbol{p}}$ and the partial gradient $\nabla_{\boldsymbol{\pi}} J(\boldsymbol{\pi}, \boldsymbol{p})$ are non-negative. Since $\Phi(\tilde{\boldsymbol{\pi}}) - \Phi(\boldsymbol{\pi}^\star)$ and the partial gradient $\nabla_{\boldsymbol{\pi}} J(\boldsymbol{\pi}, \boldsymbol{p})$ are both non-negative, we can denote the maximum element of the vector sequence $\{\bar{\boldsymbol{\pi}}_k - \tilde{\boldsymbol{\pi}}\}_{k=1}^{\bar{N}}$ as $\bar{\pi}_{sa}$ which satisfies $0 < \bar{\pi_{sa}} \leq 1$. Then we get

$$(31) \leq \frac{D}{1-\gamma} \sum_{k=1}^{\bar{N}} \bar{\beta}_k \langle \bar{\pi}_{sa} \boldsymbol{e}, -\nabla_{\boldsymbol{\pi}} J(\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{p}}^{(k)}) \rangle \tag{32}$$

$$\begin{aligned}
&= \frac{D}{1-\gamma} \langle \bar{\pi}_{sa} \boldsymbol{e}, \sum_{k=1}^{\bar{N}} \bar{\beta}_k \left(-\nabla_{\boldsymbol{\pi}} J(\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{p}}^{(k)}) \right) \rangle \\
&\leq \frac{D}{1-\gamma} \langle \boldsymbol{e}, -\hat{\boldsymbol{\xi}} \rangle \leq \frac{D\sqrt{SA}}{1-\gamma} \|\nabla \Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi})\|.
\end{aligned} \tag{33}$$

Here, the last inequality follows from the definition of $\bar{\boldsymbol{d}}(\tilde{\boldsymbol{\pi}}_t)$ which is mentioned in (30) and $\boldsymbol{e}$ is all-one vector defined in Section 1. Remind that, Lemma 3.1 implies $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$ is $L_{\boldsymbol{\pi}}$-Lipschitz in $\boldsymbol{\pi}$, and Lemma 3.1 also shows that $\Phi(\boldsymbol{\pi})$ is $L_{\boldsymbol{\pi}}$-Lipschitz. Thus, combine this Lipschitz property and the above equation (32), we get

$$\begin{aligned}
\Phi(\boldsymbol{\pi}) - \Phi(\boldsymbol{\pi}^\star) &= \Phi(\boldsymbol{\pi}) - \Phi(\tilde{\boldsymbol{\pi}}) + \Phi(\tilde{\boldsymbol{\pi}}) - \Phi(\boldsymbol{\pi}^\star) \\
&\leq \frac{D\sqrt{SA}}{1-\gamma} \|\nabla \Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi})\| + \Phi(\boldsymbol{\pi}) - \Phi(\tilde{\boldsymbol{\pi}}) \\
&\leq \frac{D\sqrt{SA}}{1-\gamma} \|\nabla \Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi})\| + L_{\boldsymbol{\pi}} \|\boldsymbol{\pi} - \tilde{\boldsymbol{\pi}}\| \\
&= \frac{D\sqrt{SA}}{1-\gamma} \|\nabla \Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi})\| + L_{\boldsymbol{\pi}} \cdot \frac{\|\nabla \Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi})\|}{2\ell_{\boldsymbol{\pi}}},
\end{aligned} \tag{34}$$

where (34) holds by using arguments of Lemma D.5 and Lemma D.6.

□

Proof of Theorem 3.3. The proof is split into two parts. We first show our algorithm can reach a ϵ-first stationary point of $\Phi(\boldsymbol{\pi}) := \max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$. Then, we next prove that this ϵ-first stationary point is close enough to the global minimum of $\Phi(\boldsymbol{\pi})$.

We begin by defining a policy $\tilde{\boldsymbol{\pi}}_t = \arg\min_{\tilde{\boldsymbol{\pi}}\in\Pi} \Phi(\tilde{\boldsymbol{\pi}}) + \ell_{\boldsymbol{\pi}}\|\boldsymbol{\pi}_t - \tilde{\boldsymbol{\pi}}\|^2$ where $\Phi(\boldsymbol{\pi})$ has been well defined as the objective function $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$ taking the worst-case transition probability, then, we have

$$\begin{aligned}
\Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi}_{t+1}) &= \min_{\boldsymbol{\pi}} \Phi(\boldsymbol{\pi}) + \ell_{\boldsymbol{\pi}}\|\boldsymbol{\pi}_{t+1} - \boldsymbol{\pi}\|^2 \\
&\le \Phi(\tilde{\boldsymbol{\pi}}_t) + \ell_{\boldsymbol{\pi}}\|\boldsymbol{\pi}_{t+1} - \tilde{\boldsymbol{\pi}}_t\|^2 \\
&= \Phi(\tilde{\boldsymbol{\pi}}_t) + \ell_{\boldsymbol{\pi}}\|\mathcal{P}_{\Pi}(\boldsymbol{\pi}_t - \alpha\nabla_{\boldsymbol{\pi}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_t, \boldsymbol{p}_t)) - \mathcal{P}_{\Pi}(\tilde{\boldsymbol{\pi}}_t)\|^2 \\
&\overset{(a)}{\le} \Phi(\tilde{\boldsymbol{\pi}}_t) + \ell_{\boldsymbol{\pi}}\|\boldsymbol{\pi}_t - \alpha\nabla_{\boldsymbol{\pi}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_t, \boldsymbol{p}_t) - \tilde{\boldsymbol{\pi}}_t\|^2 \\
&= \Phi(\tilde{\boldsymbol{\pi}}_t) + \ell_{\boldsymbol{\pi}}\|\boldsymbol{\pi}_t - \tilde{\boldsymbol{\pi}}_t\|^2 - 2\ell_{\boldsymbol{\pi}}\alpha\langle\nabla_{\boldsymbol{\pi}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_t, \boldsymbol{p}_t), \boldsymbol{\pi}_t - \tilde{\boldsymbol{\pi}}_t\rangle + \alpha^2\ell_{\boldsymbol{\pi}}\|\nabla_{\boldsymbol{\pi}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_t, \boldsymbol{p}_t)\|^2 \\
&\le \Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi}_t) + 2\ell_{\boldsymbol{\pi}}\alpha\left(\Phi(\tilde{\boldsymbol{\pi}}_t) - \Phi(\boldsymbol{\pi}_t) + \epsilon_t + \frac{\ell_{\boldsymbol{\pi}}}{2}\|\boldsymbol{\pi}_t - \tilde{\boldsymbol{\pi}}_t\|^2\right) + \alpha^2\ell_{\boldsymbol{\pi}}L_{\boldsymbol{\pi}}^2,
\end{aligned} \tag{35}$$

where $(\boldsymbol{\pi}_t, \boldsymbol{p}_t)$ is produced from the DRPG scheme at iteration step t and $\Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}$ is the Moreau envelope fucntion of Φ with parameter $\lambda = \frac{1}{2\ell_{\boldsymbol{\pi}}}$. The inequality (a) follows the basic projection property (Rockafellar, 1976), *i.e.*, for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$,

$$\|\mathcal{P}_{\mathcal{X}}(\boldsymbol{x}_1) - \mathcal{P}_{\mathcal{X}}(\boldsymbol{x}_2)\| \le \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|,$$

and the last inequality holds due to the fact that $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$ is $\ell_{\boldsymbol{\pi}}$-weakly convex in $\boldsymbol{\pi}$, in the sense that, for the $\tilde{\boldsymbol{\pi}}_t$,

$$\begin{aligned}
\Phi(\tilde{\boldsymbol{\pi}}_t) \ge J_{\boldsymbol{\rho}}(\tilde{\boldsymbol{\pi}}_t, \boldsymbol{p}_t) &\ge J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_t, \boldsymbol{p}_t) + \langle\nabla_{\boldsymbol{\pi}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_t, \boldsymbol{p}_t), \tilde{\boldsymbol{\pi}}_t - \boldsymbol{\pi}_t\rangle - \frac{\ell_{\boldsymbol{\pi}}}{2}\|\boldsymbol{\pi}_t - \tilde{\boldsymbol{\pi}}_t\|^2 \\
&\ge \underbrace{\max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_t, \boldsymbol{p})}_{\Phi(\boldsymbol{\pi}_t)} - \epsilon_t + \langle\nabla_{\boldsymbol{\pi}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_t, \boldsymbol{p}_t), \tilde{\boldsymbol{\pi}}_t - \boldsymbol{\pi}_t\rangle - \frac{\ell_{\boldsymbol{\pi}}}{2}\|\boldsymbol{\pi}_t - \tilde{\boldsymbol{\pi}}_t\|^2.
\end{aligned}$$

Next, by summing (35) up over t, we obtain,

$$\Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi}_{T-1}) \le \Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi}_0) + 2\ell_{\boldsymbol{\pi}}\alpha\sum_{t=0}^{T-1}\left(\Phi(\tilde{\boldsymbol{\pi}}_t) - \Phi(\boldsymbol{\pi}_t) + \epsilon_t + \frac{\ell_{\boldsymbol{\pi}}}{2}\|\boldsymbol{\pi}_t - \tilde{\boldsymbol{\pi}}_t\|^2\right) + T\alpha^2\ell_{\boldsymbol{\pi}}L_{\boldsymbol{\pi}}^2.$$

Rearranging this inequality yields

$$\sum_{t=0}^{T-1}\left(\Phi(\boldsymbol{\pi}_t) - \Phi(\tilde{\boldsymbol{\pi}}_t) - \frac{\ell_{\boldsymbol{\pi}}}{2}\|\boldsymbol{\pi}_t - \tilde{\boldsymbol{\pi}}_t\|^2\right) \le \frac{\Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi}_0) - \Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi}_{T-1})}{2\ell_{\boldsymbol{\pi}}\alpha} + \frac{T\alpha L_{\boldsymbol{\pi}}^2}{2} + \sum_{t=0}^{T-1}\epsilon_t. \tag{36}$$

Then, we have

$$\begin{aligned}
&\Phi(\boldsymbol{\pi}_t) - \Phi(\tilde{\boldsymbol{\pi}}_t) - \frac{\ell_{\boldsymbol{\pi}}}{2}\|\boldsymbol{\pi}_t - \tilde{\boldsymbol{\pi}}_t\|^2 \\
&= \Phi(\boldsymbol{\pi}_t) + \ell_{\boldsymbol{\pi}}\|\boldsymbol{\pi}_t - \boldsymbol{\pi}_t\|^2 - \Phi(\tilde{\boldsymbol{\pi}}_t) - \ell_{\boldsymbol{\pi}}\|\boldsymbol{\pi}_t - \tilde{\boldsymbol{\pi}}_t\|^2 + \frac{\ell_{\boldsymbol{\pi}}}{2}\|\boldsymbol{\pi}_t - \tilde{\boldsymbol{\pi}}_t\|^2 \\
&= \Phi(\boldsymbol{\pi}_t) + \ell_{\boldsymbol{\pi}}\|\boldsymbol{\pi}_t - \boldsymbol{\pi}_t\|^2 - \min_{\boldsymbol{\pi}\in\Pi}\left\{\Phi(\boldsymbol{\pi}) + \ell_{\boldsymbol{\pi}}\|\boldsymbol{\pi}_t - \boldsymbol{\pi}\|^2\right\} + \frac{\ell_{\boldsymbol{\pi}}}{2}\|\boldsymbol{\pi}_t - \tilde{\boldsymbol{\pi}}_t\|^2 \\
&\overset{(a)}{\ge} \ell_{\boldsymbol{\pi}}\|\boldsymbol{\pi}_t - \tilde{\boldsymbol{\pi}}_t\|^2 = \frac{1}{4\ell_{\boldsymbol{\pi}}}\|\nabla\Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi}_t)\|^2.
\end{aligned} \tag{37}$$

The inequality (a) is obtained by the Lemma E.3 and the last equality in (37) is obtained by using the gradient of Moreau envelope function proposed in Lemma D.5, *i.e.*,

$$\nabla\Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi}_t) = 2\ell_{\boldsymbol{\pi}}\left(\boldsymbol{\pi}_t - \arg\max_{\boldsymbol{\pi}\in\Pi}\left(\Phi(\boldsymbol{\pi}) + \ell_{\boldsymbol{\pi}}\|\boldsymbol{\pi}_t - \boldsymbol{\pi}\|^2\right)\right) = 2\ell_{\boldsymbol{\pi}}\left(\boldsymbol{\pi}_t - \tilde{\boldsymbol{\pi}}_t\right).$$

Let $\bar{\boldsymbol{\pi}}_1 := \arg\min_{\bar{\boldsymbol{\pi}}\in\Pi} \Phi(\bar{\boldsymbol{\pi}}) + \ell_{\boldsymbol{\pi}}\|\boldsymbol{\pi}_1 - \bar{\boldsymbol{\pi}}\|^2$ and $\bar{\boldsymbol{\pi}}_2 := \arg\min_{\bar{\boldsymbol{\pi}}\in\Pi} \Phi(\bar{\boldsymbol{\pi}}) + \ell_{\boldsymbol{\pi}}\|\boldsymbol{\pi}_2 - \bar{\boldsymbol{\pi}}\|^2$ for any $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2 \in \Pi$, and then we have

$$\begin{aligned}
\Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi}_1) - \Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi}_2) &= \min_{\bar{\boldsymbol{\pi}}\in\Pi} \left(\Phi(\bar{\boldsymbol{\pi}}) + \ell_{\boldsymbol{\pi}}\|\boldsymbol{\pi}_1 - \bar{\boldsymbol{\pi}}\|^2\right) - \min_{\bar{\boldsymbol{\pi}}\in\Pi} \left(\Phi(\bar{\boldsymbol{\pi}}) + \ell_{\boldsymbol{\pi}}\|\boldsymbol{\pi}_2 - \bar{\boldsymbol{\pi}}\|^2\right) && (38)\\
&= \Phi(\bar{\boldsymbol{\pi}}_1) + \ell_{\boldsymbol{\pi}}\|\boldsymbol{\pi}_1 - \bar{\boldsymbol{\pi}}_1\|^2 - \Phi(\bar{\boldsymbol{\pi}}_2) - \ell_{\boldsymbol{\pi}}\|\boldsymbol{\pi}_2 - \bar{\boldsymbol{\pi}}_2\|^2 && (39)\\
&\le \Phi(\bar{\boldsymbol{\pi}}_2) + \ell_{\boldsymbol{\pi}}\|\boldsymbol{\pi}_1 - \bar{\boldsymbol{\pi}}_2\|^2 - \Phi(\bar{\boldsymbol{\pi}}_2) - \ell_{\boldsymbol{\pi}}\|\boldsymbol{\pi}_2 - \bar{\boldsymbol{\pi}}_2\|^2 && (40)\\
&= \ell_{\boldsymbol{\pi}} \left(\|\boldsymbol{\pi}_1 - \bar{\boldsymbol{\pi}}_2\|^2 - \|\boldsymbol{\pi}_2 - \bar{\boldsymbol{\pi}}_2\|^2\right) && (41)\\
&\le 2\ell_{\boldsymbol{\pi}} S. && (42)
\end{aligned}$$

Plug (38) and (37) into (36) and reach the first result that

$$\sum_{t=0}^{T-1} \|\nabla\Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi}_t)\|^2 \le \frac{4\ell_{\boldsymbol{\pi}} S}{\alpha} + 2T\alpha\ell_{\boldsymbol{\pi}} L_{\boldsymbol{\pi}}^2 + 4\ell_{\boldsymbol{\pi}} \sum_{t=0}^{T-1} \epsilon_t. \tag{43}$$

Notice that, when the LHS is smaller than $T\epsilon^2$, *i.e.*,

$$T \cdot \min_t \|\nabla\Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi}_t)\|^2 \le \sum_{t=0}^{T-1} \|\nabla\Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi}_t)\|^2 \le T\epsilon^2,$$

there exists one $\hat{t}$ such that $\|\nabla\Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi}_{\hat{t}})\| \le \epsilon$ and $\boldsymbol{\pi}_{\hat{t}}$ is a ϵ-first order stationary point for $\Phi(\boldsymbol{\pi})$.

We finished the first part of the proof, and the next step is to show this approximate stationary point is close to the global minimum of $\Phi(\boldsymbol{\pi})$. Formally, we next to show there exists some t such that

$$J_{\boldsymbol{\rho}}(\boldsymbol{\pi}^\star, \boldsymbol{p}^\star) - \max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}_t, \boldsymbol{p}) = \Phi(\boldsymbol{\pi}^\star) - \Phi(\boldsymbol{\pi}_t) \le \epsilon. \tag{44}$$

Applying the result in Theorem 3.2 for the iterative policy $\boldsymbol{\pi}_t$, we have

$$J(\boldsymbol{\pi}_t, \boldsymbol{p}_t) - \min_{\boldsymbol{\pi}\in\Pi}\max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}) \le \Phi(\boldsymbol{\pi}_t) - \Phi(\boldsymbol{\pi}^\star) \le \frac{D\sqrt{SA}}{1-\gamma}\|\nabla\Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi}_t)\| + L_{\boldsymbol{\pi}} \cdot \frac{\|\nabla\Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi}_t)\|}{2\ell_{\boldsymbol{\pi}}}. \tag{45}$$

Combined this two parts, we finally state the global convergence guarantee. Equation (45) implies that

$$\begin{aligned}
\min_{t\in\{0,\cdots,T-1\}} \left\{J(\boldsymbol{\pi}_t, \boldsymbol{p}_t) - \min_{\boldsymbol{\pi}\in\Pi}\max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})\right\} &\le \frac{1}{T}\sum_{t=0}^{T-1}\left(J(\boldsymbol{\pi}_t, \boldsymbol{p}_t) - \min_{\boldsymbol{\pi}\in\Pi}\max_{\boldsymbol{p}\in\mathcal{P}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})\right)\\
&\le \frac{1}{T}\sum_{t=0}^{T-1}\left(\Phi(\boldsymbol{\pi}_t) - \Phi(\boldsymbol{\pi}^\star)\right)\\
&\le \frac{1}{T}\left(\frac{D\sqrt{SA}}{1-\gamma} + \frac{L_{\boldsymbol{\pi}}}{2\ell_{\boldsymbol{\pi}}}\right)\sum_{t=0}^{T-1}\left\|\nabla\Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi}_t)\right\| && (46)
\end{aligned}$$

By Cauchy–Schwarz inequality, we can obtain

$$\frac{1}{\sqrt{T}}\sum_{t=0}^{T-1}\left\|\nabla\Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi}_t)\right\| \le \sqrt{\sum_{t=0}^{T-1}\|\nabla\Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi}_t)\|^2}.$$

We then multiply the constant $\frac{D\sqrt{SA}}{1-\gamma} + \frac{L_{\boldsymbol{\pi}}}{2\ell_{\boldsymbol{\pi}}}$ on both sides and combine the inequality (43) to obtain the result that, if the

iteration time T satisfies

$$
\begin{aligned}
(46) &\leq \frac{1}{\sqrt{T}}\left(\frac{D\sqrt{SA}}{1-\gamma}+\frac{L_{\boldsymbol{\pi}}}{2\ell_{\boldsymbol{\pi}}}\right)\sqrt{\sum_{t=0}^{T-1}\|\nabla\Phi_{\frac{1}{2\ell_{\boldsymbol{\pi}}}}(\boldsymbol{\pi}_t)\|^2} \\
&= \frac{1}{\sqrt{T}}\left(\frac{D\sqrt{SA}}{1-\gamma}+\frac{L_{\boldsymbol{\pi}}}{2\ell_{\boldsymbol{\pi}}}\right)\sqrt{\left(\frac{4\ell_{\boldsymbol{\pi}}S}{\alpha}+2T\alpha\ell_{\boldsymbol{\pi}}L_{\boldsymbol{\pi}}^2+4\ell_{\boldsymbol{\pi}}\sum_{t=0}^{T-1}\epsilon_t\right)} \\
&\overset{(a)}{\leq} \frac{1}{\sqrt{T}}\left(\frac{D\sqrt{SA}}{1-\gamma}+\frac{L_{\boldsymbol{\pi}}}{2\ell_{\boldsymbol{\pi}}}\right)\sqrt{\left(\frac{4\ell_{\boldsymbol{\pi}}S\sqrt{T}}{\delta}+2\sqrt{T}\delta\ell_{\boldsymbol{\pi}}L_{\boldsymbol{\pi}}^2+\frac{4\ell_{\boldsymbol{\pi}}\epsilon_0}{1-\gamma}\right)} \\
&\leq \frac{1}{\sqrt{T}}\left(\frac{D\sqrt{SA}}{1-\gamma}+\frac{L_{\boldsymbol{\pi}}}{2\ell_{\boldsymbol{\pi}}}\right)\sqrt{\left(\frac{4\ell_{\boldsymbol{\pi}}S\sqrt{T}}{\delta}+2\sqrt{T}\delta\ell_{\boldsymbol{\pi}}L_{\boldsymbol{\pi}}^2+\frac{4\ell_{\boldsymbol{\pi}}\sqrt{T}}{1-\gamma}\right)} \\
&= \epsilon
\end{aligned}
$$

where the inequality (a) holds due to the adaptive tolerance sequence, in the sense that,

$$
\sum_{t=0}^{T-1}\epsilon_t \leq \sum_{t=0}^{\infty}\epsilon_t \leq \epsilon_0\cdot\left(1+\gamma+\gamma^2+\cdots\right) \leq \frac{\epsilon_0}{1-\gamma},
$$

which implies that

$$
T \geq \frac{\left(\frac{D\sqrt{SA}}{1-\gamma}+\frac{L_{\boldsymbol{\pi}}}{2\ell_{\boldsymbol{\pi}}}\right)^4\left(\frac{4\ell_{\boldsymbol{\pi}}S}{\delta}+2\delta\ell_{\boldsymbol{\pi}}L_{\boldsymbol{\pi}}^2+\frac{4\ell_{\boldsymbol{\pi}}}{1-\gamma}\right)^2}{\epsilon^4} = \mathcal{O}(\epsilon^{-4}),
$$

then, we have

$$
\min_{t\in\{0,\cdots,T-1\}}\left\{J(\boldsymbol{\pi}_t,\boldsymbol{p}_t)-\min_{\boldsymbol{\pi}\in\Pi}\max_{\boldsymbol{p}\in\mathcal{P}}J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p})\right\} \leq \epsilon.
$$

Intuitively, we have

$$
\min_{t\in\{0,\cdots,T-1\}}\left\{\Phi(\boldsymbol{\pi}_t)-\min_{\boldsymbol{\pi}\in\Pi}\Phi(\boldsymbol{\pi})\right\} \leq \epsilon.
$$

□

F. Discussion on R-contamination ambiguity set

Recall that the R-contamination ambiguity set is a kind of (s,a)-rectangular set $\mathcal{P} = \underset{s\in\mathcal{S},a\in\mathcal{A}}{\times}\mathcal{P}_{s,a}$ where $\mathcal{P}_{s,a}$ is defined as

$$
\mathcal{P}_{s,a} := \{(1-R)\bar{\boldsymbol{p}}_{sa}+R\boldsymbol{q} \mid \boldsymbol{q}\in\Delta(S)\}, s\in\mathcal{S}, a\in\mathcal{A}. \tag{47}
$$

We have the following property of the R-contamination sets which illustrates their limited applicability.

Proposition F.1. *Any RMDP with an R-contamination ambiguity set has the same optimal robust policy as a corresponding ordinary MDP with a reduced discount factor.*

Proof of Proposition F.1. The robust optimal bellman operator of a RMDP with R-contamination ambiguity can be written as

$$
\begin{aligned}
(\mathcal{T}^r\boldsymbol{v})_s :&= \min_{a\in\mathcal{A}}\max_{\boldsymbol{p}_{sa}\in\mathcal{P}_{s,a}}(c_{sa}+\gamma\boldsymbol{p}_{sa}^\top\boldsymbol{v}) \\
&= \min_{a\in\mathcal{A}}c_{sa}+\gamma\left[(1-R)\bar{\boldsymbol{p}}_{sa}^\top\boldsymbol{v}+R\max_{s'}v_{s'}\right] \\
&= \left[\min_{a\in\mathcal{A}}c_{sa}+\gamma(1-R)\bar{\boldsymbol{p}}_{sa}^\top\boldsymbol{v}\right]+R\gamma\max_{s'}v_{s'}
\end{aligned}
$$

Consider an ordinary MDP with the same reward function, transition kernel $\boldsymbol{p} := (\bar{\boldsymbol{p}}_{sa})_{s\in\mathcal{S},a\in\mathcal{A}} \in (\Delta^S)^{S\times A}$ and discount factor $\gamma(1-R)$. The optimal bellman operator is defined as

$$(\mathcal{T}\boldsymbol{v})_s := \min_{a\in\mathcal{A}} c_{sa} + \gamma(1-R)\bar{\boldsymbol{p}}_{sa}^\top \boldsymbol{v}.$$

Then, we have that

$$(\mathcal{T}^r\boldsymbol{v})_s = (\mathcal{T}\boldsymbol{v})_s + R\gamma\|\boldsymbol{v}\|_\infty \tag{48}$$

We define optimal value functions for $\mathcal{T}^r$ and $\mathcal{T}$ as follow

$$\mathcal{T}^r\boldsymbol{v}^r = \boldsymbol{v}^r, \quad \mathcal{T}\boldsymbol{v}^{nr} = \boldsymbol{v}^{nr},$$

and consider the value iteration with the given initial value functions $\boldsymbol{v}^r$ first. Then we have that

$$\begin{aligned}
& \mathcal{T}^r\boldsymbol{v}^r = \mathcal{T}\boldsymbol{v}^r + R\gamma\|\boldsymbol{v}^r\|_\infty \boldsymbol{e} \\
\iff \quad & (\mathcal{T}^r)^2\boldsymbol{v}^r = (\mathcal{T})^2\boldsymbol{v}^r + R\gamma\|\mathcal{T}^r\boldsymbol{v}^r\|_\infty \boldsymbol{e} + R\gamma^2(1-R)\|\boldsymbol{v}^r\|_\infty \boldsymbol{e} \\
& \qquad = (\mathcal{T})^2\boldsymbol{v}^r + \left[R\gamma + R\gamma^2(1-R)\right]\cdot\|\boldsymbol{v}^r\|_\infty\cdot\boldsymbol{e} \\
\iff \quad & (\mathcal{T}^r)^k\boldsymbol{v}^r = (\mathcal{T})^k\boldsymbol{v}^r + \left[R\gamma + R\gamma^2(1-R) + R\gamma^3(1-R)^2 + \cdots\right]\cdot\|\boldsymbol{v}^r\|_\infty\cdot\boldsymbol{e} \\
& \qquad = (\mathcal{T})^k\boldsymbol{v}^r + \sum_{n=1}^{k} R\gamma^k(1-R)^{k-1}\cdot\|\boldsymbol{v}^r\|_\infty\cdot\boldsymbol{e}.
\end{aligned}$$

By taking the limitation for both side, we obtain

$$\begin{aligned}
\lim_{k\to\infty}(\mathcal{T}^r)^k\boldsymbol{v}^r &= \boldsymbol{v}^r \\
&= \lim_{k\to\infty}\left[(\mathcal{T})^k\boldsymbol{v}^r + \sum_{n=1}^{k} R\gamma^n(1-R)^{n-1}\cdot\|\boldsymbol{v}^r\|_\infty\cdot\boldsymbol{e}\right] \\
&= \boldsymbol{v}^{nr} + \lim_{k\to\infty}\left[\sum_{n=1}^{k} R\gamma^n(1-R)^{n-1}\cdot\|\boldsymbol{v}^r\|_\infty\cdot\boldsymbol{e}\right] \\
&= \boldsymbol{v}^{nr} + \lim_{k\to\infty}\left[\frac{1-(\gamma(1-R))^k}{1-\gamma(1-R)}\right]\cdot\|\boldsymbol{v}^r\|_\infty\cdot\boldsymbol{e} \\
&= \boldsymbol{v}^{nr} + \frac{1}{1-\gamma(1-R)}\cdot\|\boldsymbol{v}^r\|_\infty\cdot\boldsymbol{e}.
\end{aligned}$$

Each operation $\mathcal{T}^r$ on $\boldsymbol{v}^r$ will take the same optimal action due to the definition of $\boldsymbol{v}^r$, which implies operation $\mathcal{T}^r$ on $\boldsymbol{v}^r$ works with the same action is taken. This intuitive result shows that the RMDP with R-contamination ambiguity and its corresponding ordinary MDP with discount factor $\gamma(1-R)$ has the same optimal policy. □

G. Proofs of Section 4

Proof of Lemma 4.1. Notice that

$$\frac{\partial J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p})}{\partial p_{sas'}} = \sum_{\hat{s}\in\mathcal{S}} \frac{\partial v_{\hat{s}}^{\boldsymbol{\pi},\boldsymbol{p}}}{\partial p_{sas'}}\rho_{\hat{s}}.$$

Then, we discuss $\frac{\partial v_{\hat{s}}^{\boldsymbol{\pi},\boldsymbol{p}}}{\partial p_{sas'}}$ over two cases: $\hat{s}\neq s$ and $\hat{s}=s$

$$\begin{aligned}
\frac{\partial v_{\hat{s}}^{\boldsymbol{\pi},\boldsymbol{p}}}{\partial p_{sas'}}\Big|_{\hat{s}\neq s} &= \frac{\partial}{\partial p_{sas'}}\left[\sum_{\hat{a}}\pi_{\hat{s}\hat{a}}\sum_{\hat{s}'\in\mathcal{S}} p_{\hat{s}\hat{a}\hat{s}'}\left(c_{\hat{s}\hat{a}\hat{s}'} + \gamma v_{\hat{s}'}^{\boldsymbol{\pi},\boldsymbol{p}}\right)\right] = \gamma\sum_{\hat{a}}\pi_{\hat{s}\hat{a}}\sum_{\hat{s}'\in\mathcal{S}} p_{\hat{s}\hat{a}\hat{s}'}\frac{\partial v_{\hat{s}'}^{\boldsymbol{\pi},\boldsymbol{p}}}{\partial p_{sas'}}; \\
\frac{\partial v_{\hat{s}}^{\boldsymbol{\pi},\boldsymbol{p}}}{\partial p_{sas'}}\Big|_{\hat{s}=s} &= \gamma\sum_{\hat{a}}\pi_{s\hat{a}}\sum_{\hat{s}'\in\mathcal{S}} p_{s\hat{a}\hat{s}'}\frac{\partial v_{\hat{s}'}^{\boldsymbol{\pi},\boldsymbol{p}}}{\partial p_{sas'}} + \pi_{sa}\left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}}\right);
\end{aligned}$$

By condensing $\sum_{\hat{a}} \pi_{\hat{s}\hat{a}} p_{\hat{s}\hat{a}\hat{s}'} = p^{\boldsymbol{\pi}}_{\hat{s}\hat{s}'}(1)$, we can obtain,

$$\begin{aligned}
\frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_{\hat{s}}}{\partial p_{sas'}}\Big|_{\hat{s}\neq s} &= \gamma \sum_{\hat{s}'\neq s} p^{\boldsymbol{\pi}}_{\hat{s}\hat{s}'}(1) \frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_{\hat{s}'}}{\partial p_{sas'}} + \gamma \sum_{\hat{s}'= s} p^{\boldsymbol{\pi}}_{\hat{s}\hat{s}'}(1) \frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_{\hat{s}'}}{\partial p_{sas'}} \\
&= \gamma \sum_{\hat{s}'\neq s} p^{\boldsymbol{\pi}}_{\hat{s}\hat{s}'}(1) \cdot \gamma \sum_{\hat{a}} \pi_{\hat{s}'\hat{a}} \sum_{\hat{s}''\in\mathcal{S}} p_{\hat{s}'\hat{a}\hat{s}''} \frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_{\hat{s}''}}{\partial p_{sas'}} \\
&\quad + \gamma p^{\boldsymbol{\pi}}_{\hat{s}s}(1) \cdot \left(\gamma \sum_{\hat{a}} \pi_{s\hat{a}} \sum_{\hat{s}'\in\mathcal{S}} p_{s\hat{a}\hat{s}'} \frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_{\hat{s}'}}{\partial p_{sas'}} + \pi_{sa}\left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{p}}_{s'}\right) \right) \\
&= \gamma p^{\boldsymbol{\pi}}_{\hat{s}s}(1) \pi_{sa}\left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{p}}_{s'}\right) + \gamma^2 \sum_{\hat{s}'} p^{\boldsymbol{\pi}}_{\hat{s}\hat{s}'}(2) \frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_{\hat{s}'}}{\partial p_{sas'}} \\
&= \gamma p^{\boldsymbol{\pi}}_{\hat{s}s}(1) \pi_{sa}\left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{p}}_{s'}\right) + \gamma^2 p^{\boldsymbol{\pi}}_{\hat{s}s}(2) \pi_{sa}\left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{p}}_{s'}\right) + \gamma^3 \sum_{\hat{s}'} p^{\boldsymbol{\pi}}_{\hat{s}\hat{s}'}(3) \frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_{\hat{s}'}}{\partial p_{sas'}} \\
&= \cdots \\
&= \sum_{t=1}^{\infty} \gamma^t p^{\boldsymbol{\pi}}_{\hat{s}s}(t) \pi_{sa}\left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{p}}_{s'}\right) = \sum_{t=0}^{\infty} \gamma^t p^{\boldsymbol{\pi}}_{\hat{s}s}(t) \pi_{sa}\left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{p}}_{s'}\right).
\end{aligned}$$

The last equality is from the initial assumption $\hat{s} \neq s$, *i.e.*, $p^{\boldsymbol{\pi}}_{\hat{s}s}(0) = 0$, and similarly for the case $\hat{s} = s$ we have,

$$\frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_{\hat{s}}}{\partial p_{sas'}}\Big|_{\hat{s}= s} = \sum_{t=0}^{\infty} \gamma^t p^{\boldsymbol{\pi}}_{ss}(t) \pi_{sa}\left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{p}}_{s'}\right).$$

Hence, the partial derivative for transition probability is obtained

$$\begin{aligned}
\frac{\partial J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p})}{\partial p_{sas'}} &= \frac{1}{1-\gamma} \left(\underbrace{(1-\gamma) \sum_{\hat{s}\in\mathcal{S}} \sum_{t=0}^{\infty} \gamma^t \rho_{\hat{s}} p^{\boldsymbol{\pi}}_{\hat{s}s}(t)}_{d^{\boldsymbol{\pi},\boldsymbol{p}}_{\boldsymbol{\rho}}(s)} \right) \pi_{sa}\left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{p}}_{s'}\right) \\
&= \frac{1}{1-\gamma} d^{\boldsymbol{\pi},\boldsymbol{p}}_{\boldsymbol{\rho}}(s) \pi_{sa}\left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{p}}_{s'}\right).
\end{aligned}$$

The uniformly bounded cost $c_{sas'}$ implies that, the absolute value of the value function is bounded for any policy $\boldsymbol{\pi}$ and transition kernel $\boldsymbol{p}$,

$$|v^{\boldsymbol{\pi},\boldsymbol{p}}_{s}| = \left| \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{p}} \left[\sum_{t=0}^{\infty} \gamma^t c_{s_t a_t s_{t+1}} \mid s_0 = s \right] \right| \leq \sum_{t=0}^{\infty} \gamma^t = \frac{1}{1-\gamma},$$

then we obtain that

$$|\pi_{sa}\left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{p}}_{s'}\right)| \leq |\pi_{sa}| \cdot |c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{p}}_{s'}| \leq 1 + \frac{\gamma 1}{1-\gamma} \leq \frac{1}{1-\gamma}.$$

Therefore, by vectorizing the $\boldsymbol{p}$ as a S^2A-dimensional vector, we have

$$\begin{aligned}
\|\nabla_{\boldsymbol{p}} J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p})\| &= \sqrt{\sum_{s,a,s'} \left(\frac{\partial J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p})}{\partial p_{sas'}} \right)^2} \\
&= \frac{1}{1-\gamma} \sqrt{\sum_{s,a,s'} \left[d^{\boldsymbol{\pi},\boldsymbol{p}}_{\boldsymbol{\rho}}(s) \pi_{sa}\left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{p}}_{s'}\right) \right]^2} \\
&\leq \frac{1}{(1-\gamma)^2} \sqrt{\sum_{a,s'} \sum_{s} (d^{\boldsymbol{\pi},\boldsymbol{\xi}}_{\boldsymbol{\rho}}(s))^2} \leq \frac{\sqrt{SA}}{(1-\gamma)^2},
\end{aligned}$$

where the last inequality holds since the discounted state occupancy measure satisfies

$$\sum_s (d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{\xi}}(s))^2 \leq \left(\sum_s (d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{\xi}}(s))\right)^2 = 1.$$

□

Notice that, the objective function $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})$ is twice differentiable on $\boldsymbol{p}$. Hence, to prove the smoothness condition in Lemma 4.2 is equal to show that there exists a constant $L \leq \infty$ such that

$$\nabla_{\boldsymbol{p}}^2 J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}) \preceq L\boldsymbol{I} \iff \forall \boldsymbol{x} \in \mathbb{R}^{AS^2},\ \boldsymbol{x}^\top \nabla_{\boldsymbol{p}}^2 J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p})\boldsymbol{x} \leq L\boldsymbol{x}^\top \boldsymbol{x}.$$

Proof of Lemma 4.2. Denote $\boldsymbol{p}(\alpha) := \boldsymbol{p} + \alpha \boldsymbol{z} \in \mathcal{P}$ where $\alpha \in \mathbb{R}$ is a small scalar, whereas $\boldsymbol{z} \in (\mathbb{R}^{\mathcal{S}})^{\mathcal{S}\times\mathcal{A}}$. Since $J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{p}) = \sum_s \rho_s v_s^{\boldsymbol{\pi},\boldsymbol{p}(\alpha)}$ with a known initial distribution $\boldsymbol{\rho}$, we turn to consider the derivative of value function $v_s^{\boldsymbol{\pi},\boldsymbol{p}(\alpha)}$ of the transition kernel $\boldsymbol{p}(\alpha)$ over α,

$$v_s^{\boldsymbol{\pi},\boldsymbol{p}(\alpha)} = \sum_a \pi_{sa} \sum_{s'} [\boldsymbol{p}(\alpha)]_{sas'} c_{sas'} + \gamma \cdot \sum_a \pi_{sa} \sum_{s'} [\boldsymbol{p}(\alpha)]_{sas'} v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}(\alpha)}, \tag{49}$$

First, let us simplify the form of $v_s^{\boldsymbol{\pi},\boldsymbol{p}(\alpha)}$. We define $\boldsymbol{P}(\alpha) \in (\Delta^S)^S$ as the state transition kernel and for any $s, s' \in \mathcal{S}$,

$$[\boldsymbol{P}(\alpha)]_{ss'} = \sum_a \pi_{sa} [\boldsymbol{p}(\alpha)]_{sas'}, \tag{50}$$

and $\boldsymbol{c}(\alpha) \in \mathbb{R}^S$ where for any $s \in \mathcal{S}$,

$$|[\boldsymbol{c}(\alpha)]_s| = \left| \sum_a \pi_{sa} \sum_{s'} [\boldsymbol{p}(\alpha)]_{sas'} c_{sas'} \right| \leq 1. \tag{51}$$

Then, the value function (49) can be written as,

$$v_s^{\boldsymbol{\pi},\boldsymbol{p}(\alpha)} = \boldsymbol{e}_s^\top \underbrace{(\boldsymbol{I} - \gamma \boldsymbol{P}(\alpha))^{-1}}_{\boldsymbol{M}(\alpha)} \boldsymbol{c}(\alpha), \tag{52}$$

where $\boldsymbol{e}_s := [0, \cdots, 1, \cdots, 0]^\top \in \mathbb{R}^S$ is a vector whose s-th element is 1 and others are 0. By using power series expansion technique (Agarwal et al., 2021; Mei et al., 2020), we can obtain that,

$$\boldsymbol{M}(\alpha) = (\boldsymbol{I} - \gamma \boldsymbol{P}(\alpha))^{-1} = \sum_t^\infty \gamma^t \boldsymbol{P}(\alpha)^t, \tag{53}$$

which implies that, for any $s, s' \in \mathcal{S}$, $[\boldsymbol{M}(\alpha)]_{ss'} \geq 0$, and we have

$$\boldsymbol{e} = \frac{1}{1-\gamma} \cdot (\boldsymbol{I} - \gamma \boldsymbol{P}(\alpha))\, \boldsymbol{e} \iff \boldsymbol{M}(\alpha)\boldsymbol{e} = \frac{1}{1-\gamma} \cdot \boldsymbol{e}, \tag{54}$$

which implies each row of $\boldsymbol{M}(\alpha)$ sums to $1/(1-\gamma)$. Therefore, for any vector $x \in \mathbb{R}^S$, we have

$$\|\boldsymbol{M}(\alpha)\boldsymbol{x}\|_\infty = \max_i |[\boldsymbol{M}(\alpha)\boldsymbol{x}]_i| \leq \frac{1}{1-\gamma} \cdot \|\boldsymbol{x}\|_\infty. \tag{55}$$

Taking derivative with respect to α on $v_s^{\boldsymbol{\pi},\boldsymbol{p}(\alpha)}$ defined in (52),

$$\frac{\partial v_s^{\boldsymbol{\pi},\boldsymbol{p}(\alpha)}}{\partial \alpha} = \boldsymbol{e}_s^\top \boldsymbol{M}(\alpha) \frac{\partial \boldsymbol{c}(\alpha)}{\partial \alpha} + \gamma \boldsymbol{e}_s^\top \boldsymbol{M}(\alpha) \frac{\partial \boldsymbol{P}(\alpha)}{\partial \alpha} \boldsymbol{M}(\alpha) \boldsymbol{c}(\alpha). \tag{56}$$

Then taking the twice derivative with respect to α,

$$\frac{\partial^2 v_s^{\boldsymbol{\pi},\boldsymbol{p}(\alpha)}}{(\partial\alpha)^2} = \boldsymbol{e}_s^\top \boldsymbol{M}(\alpha)\frac{\partial^2 \boldsymbol{c}(\alpha)}{(\partial\alpha)^2} + 2\gamma \boldsymbol{e}_s^\top \boldsymbol{M}(\alpha)\frac{\partial \boldsymbol{P}(\alpha)}{\partial\alpha}\boldsymbol{M}(\alpha)\frac{\partial \boldsymbol{c}(\alpha)}{\partial\alpha} + 2\gamma^2 \boldsymbol{e}_s^\top \boldsymbol{M}(\alpha)\frac{\partial \boldsymbol{P}(\alpha)}{\partial\alpha}\boldsymbol{M}\frac{\partial \boldsymbol{P}(\alpha)}{\partial\alpha}\boldsymbol{M}(\alpha)\boldsymbol{c}(\alpha) + \gamma \boldsymbol{e}_s^\top \boldsymbol{M}(\alpha)\frac{\partial^2 \boldsymbol{P}(\alpha)}{(\partial\alpha)^2}\boldsymbol{M}(\alpha)\boldsymbol{c}(\alpha). \tag{57}$$

Notice that, above two form of derivatives are obtained by using matrix calculus techniques, *i.e.*, for any matrix $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{U}(x)$ and scalar x,

$$\frac{\partial \boldsymbol{A}\boldsymbol{U}(x)\boldsymbol{B}}{\partial x} = \boldsymbol{A}\frac{\partial \boldsymbol{U}(x)}{\partial x}\boldsymbol{B} \quad \text{and} \quad \frac{\partial \boldsymbol{U}(x)^{-1}}{\partial x} = -\boldsymbol{U}(x)^{-1}\frac{\partial \boldsymbol{U}(x)}{\partial x}\boldsymbol{U}(x)^{-1}.$$

So far, we get the derivative form of the value function. Then we'd like to bound $\left|\frac{\partial^2 v_s^{\boldsymbol{\pi},\boldsymbol{p}(\alpha)}}{(\partial\alpha)^2}\Big|_{\alpha=0}\right|$.

For the first term in (57), we have,

$$\begin{aligned}\left|\boldsymbol{e}_s^\top \boldsymbol{M}(\alpha)\frac{\partial^2 \boldsymbol{c}(\alpha)}{(\partial\alpha)^2}\Big|_{\alpha=0}\right| &\le \left\|\boldsymbol{e}_s^\top\right\|_1 \cdot \left\|\boldsymbol{M}(\alpha)\frac{\partial^2 \boldsymbol{c}(\alpha)}{(\partial\alpha)^2}\Big|_{\alpha=0}\right\|_\infty \\ &\le \frac{1}{1-\gamma}\cdot\left\|\frac{\partial^2 \boldsymbol{c}(\alpha)}{(\partial\alpha)^2}\Big|_{\alpha=0}\right\|_\infty \\ &= 0,\end{aligned} \tag{58}$$

where the last but one inequality is obtained from (55) and the last equality holds since for any $\alpha \in \mathbb{R}$,

$$\begin{aligned}\left\|\frac{\partial^2 \boldsymbol{c}(\alpha)}{(\partial\alpha)^2}\right\|_\infty &= \max_s \left|\frac{\partial}{\partial\alpha}\left(\frac{\partial[\boldsymbol{c}(\alpha)]_s}{\partial\alpha}\right)\right| \\ &= \max_s \left|\frac{\partial}{\partial\alpha}\left(\frac{\partial\left(\sum_a \pi_{sa}\sum_{s'}[\boldsymbol{p}(\alpha)]_{sas'}c_{sas'}\right)}{\partial\alpha}\right)\right| \\ &= \max_s \left|\frac{\partial}{\partial\alpha}\left(\sum_a \pi_{sa}\sum_{s'} z_{sas'}c_{sas'}\right)\right| \\ &= 0.\end{aligned} \tag{59}$$

For the second term in (57), we have

$$\begin{aligned}\left|\boldsymbol{e}_s^\top \boldsymbol{M}(\alpha)\frac{\partial \boldsymbol{P}(\alpha)}{\partial\alpha}\boldsymbol{M}(\alpha)\frac{\partial \boldsymbol{c}(\alpha)}{\partial\alpha}\Big|_{\alpha=0}\right| &\le \left\|\boldsymbol{e}_s^\top\right\|_1 \cdot \left\|\boldsymbol{M}(\alpha)\frac{\partial \boldsymbol{P}(\alpha)}{\partial\alpha}\boldsymbol{M}(\alpha)\frac{\partial \boldsymbol{c}(\alpha)}{\partial\alpha}\Big|_{\alpha=0}\right\|_\infty \\ &\le \frac{1}{1-\gamma}\cdot\left\|\frac{\partial \boldsymbol{P}(\alpha)}{\partial\alpha}\boldsymbol{M}(\alpha)\frac{\partial \boldsymbol{c}(\alpha)}{\partial\alpha}\Big|_{\alpha=0}\right\|_\infty.\end{aligned} \tag{60}$$

According to (50), for any $\boldsymbol{x} \in \mathbb{R}^S$ and $s \in \mathcal{S}$, we have,

$$\left[\frac{\partial \boldsymbol{P}(\alpha)}{\partial\alpha}\boldsymbol{x}\right]_s = \sum_{s'}\sum_a \pi_{sa}\frac{\partial[\boldsymbol{p}(\alpha)]_{sas'}}{\partial\alpha}x_{s'},$$

and its ℓ_∞ norm can be upper bounded as

$$\begin{aligned}
\left\| \frac{\partial \boldsymbol{P}(\alpha)}{\partial \alpha}\Big|_{\alpha=0} \boldsymbol{x} \right\|_\infty &= \max_s \left| \sum_{s'} \sum_a \pi_{sa} \frac{\partial [\boldsymbol{p}(\alpha)]_{sas'}}{\partial \alpha}\Big|_{\alpha=0} x_{s'} \right| \\
&\le \max_s \sum_{s'} \sum_a \pi_{sa} \left| z_{sas'} \right| \left| x_{s'} \right| \\
&\le \max_s \sum_{s'} \sum_a \pi_{sa} \left| z_{sas'} \right| \cdot \|x\|_\infty \\
&= \sum_{s'} \sum_a \pi_{\bar{s}a} \left| z_{\bar{s}as'} \right| \cdot \|x\|_\infty \\
&\le \sum_{s'} \sum_a \pi_{\bar{s}a} \max_{s,a,s'} \left| z_{sas'} \right| \cdot \|x\|_\infty \\
&= \max_{s,a,s'} \left| z_{sas'} \right| \cdot \sum_{s'} \|x\|_\infty \\
&\le S \cdot \|\boldsymbol{z}\|_\infty \cdot \|x\|_\infty \\
&\le S \cdot \|\boldsymbol{z}\|_2 \cdot \|x\|_\infty
\end{aligned} \tag{61}$$

Similarly, for any $\alpha \in \mathbb{R}$, we have

$$\begin{aligned}
\left\| \frac{\partial \boldsymbol{c}(\alpha)}{\partial \alpha} \right\|_\infty &= \max_s \left| \frac{\partial \left(\sum_a \pi_{sa} \sum_{s'} [\boldsymbol{p}(\alpha)]_{sas'} c_{sas'} \right)}{\partial \alpha} \right| \\
&= \max_s \left| \sum_a \pi_{sa} \sum_{s'} z_{sas'} c_{sas'} \right| \\
&\le S \cdot \|\boldsymbol{z}\|_2.
\end{aligned} \tag{62}$$

Then, we obtain an upper bound of the second term,

$$\begin{aligned}
\left| \boldsymbol{e}_s^\top \boldsymbol{M}(\alpha) \frac{\partial \boldsymbol{P}(\alpha)}{\partial \alpha} \boldsymbol{M}(\alpha) \frac{\partial \boldsymbol{c}(\alpha)}{\partial \alpha}\Big|_{\alpha=0} \right| &\le \frac{S}{1-\gamma} \cdot \left\| \boldsymbol{M}(\alpha) \frac{\partial \boldsymbol{c}(\alpha)}{\partial \alpha}\Big|_{\alpha=0} \right\|_\infty \cdot \|\boldsymbol{z}\|_2 \\
&\le \frac{S}{(1-\gamma)^2} \cdot \left\| \frac{\partial \boldsymbol{c}(\alpha)}{\partial \alpha}\Big|_{\alpha=0} \right\|_\infty \cdot \|\boldsymbol{z}\|_2 \\
&\le \frac{S^2}{(1-\gamma)^2} \cdot \|\boldsymbol{z}\|_2^2.
\end{aligned} \tag{63}$$

For the third term of in (57), we can similarly bound it as

$$\begin{aligned}
\left| \boldsymbol{e}_s^\top \boldsymbol{M}(\alpha) \frac{\partial \boldsymbol{P}(\alpha)}{\partial \alpha} \boldsymbol{M}(\alpha) \frac{\partial \boldsymbol{P}(\alpha)}{\partial \alpha} \boldsymbol{M}(\alpha) \boldsymbol{c}(\alpha)\Big|_{\alpha=0} \right| &\le \left\| \boldsymbol{M}(\alpha) \frac{\partial \boldsymbol{P}(\alpha)}{\partial \alpha} \boldsymbol{M}(\alpha) \frac{\partial \boldsymbol{P}(\alpha)}{\partial \alpha} \boldsymbol{M}(\alpha) \boldsymbol{c}(\alpha)\Big|_{\alpha=0} \right\|_\infty \\
&\le \frac{1}{1-\gamma} \cdot S \cdot \|\boldsymbol{z}\|_2 \cdot \frac{1}{1-\gamma} \cdot S \cdot \|\boldsymbol{z}\|_2 \cdot \frac{1}{1-\gamma} \\
&= \frac{S^2}{(1-\gamma)^3} \cdot \|\boldsymbol{z}\|_2^2.
\end{aligned} \tag{64}$$

Denote that, for any $\boldsymbol{x} \in \mathbb{R}^S$,

$$\left\| \frac{\partial^2 \boldsymbol{P}(\alpha)}{(\partial \alpha)^2}\Big|_{\alpha=0} \boldsymbol{x} \right\|_\infty = \max_s \left| \sum_{s'} \sum_a \pi_{sa} \frac{\partial^2 [\boldsymbol{p}(\alpha)]_{sas'}}{\partial (\alpha)^2}\Big|_{\alpha=0} x_{s'} \right| = 0. \tag{65}$$

Therefore, we combine (58), (63), (64) and (65),

$$\begin{aligned}
\left|\frac{\partial^2 v_s^{\boldsymbol{\pi},\boldsymbol{p}(\alpha)}}{(\partial\alpha)^2}\Big|_{\alpha=0}\right| &= \left|\boldsymbol{e}_s^\top \boldsymbol{M}(\alpha)\frac{\partial^2 \boldsymbol{c}(\alpha)}{(\partial\alpha)^2}\Big|_{\alpha=0}\right| + 2\gamma^2\cdot\left|\boldsymbol{e}_s^\top \boldsymbol{M}(\alpha)\frac{\partial \boldsymbol{P}(\alpha)}{\partial\alpha}\boldsymbol{M}(\alpha)\frac{\partial \boldsymbol{P}(\alpha)}{\partial\alpha}\boldsymbol{M}(\alpha)\boldsymbol{c}(\alpha)\Big|_{\alpha=0}\right| \\
&\quad + 2\gamma\cdot\left|\boldsymbol{e}_s^\top \boldsymbol{M}(\alpha)\frac{\partial \boldsymbol{P}(\alpha)}{\partial\alpha}\boldsymbol{M}(\alpha)\frac{\partial \boldsymbol{c}(\alpha)}{\partial\alpha}\Big|_{\alpha=0}\right| + \gamma\cdot\left|\boldsymbol{e}_s^\top \boldsymbol{M}(\alpha)\frac{\partial^2 \boldsymbol{P}(\alpha)}{(\partial\alpha)^2}\boldsymbol{M}(\alpha)\boldsymbol{c}(\alpha)\Big|_{\alpha=0}\right| \\
&\le 2\gamma\cdot\frac{S^2}{(1-\gamma)^2}\cdot\|\boldsymbol{z}\|_2^2 + 2\gamma^2\cdot\frac{S^2}{(1-\gamma)^3}\|\boldsymbol{z}\|_2^2 \\
&= \frac{2\gamma S^2}{(1-\gamma)^3}\cdot\|\boldsymbol{z}\|_2^2. \qquad (66)
\end{aligned}$$

Then, for any $\boldsymbol{y}\in\mathbb{R}^{AS^2}$, we have

$$\begin{aligned}
\left|\boldsymbol{y}^\top\nabla_{\boldsymbol{p}}^2 J_\rho(\boldsymbol{\pi},\boldsymbol{p})\boldsymbol{y}\right| &\le \sum_s \rho_s\cdot\left|\boldsymbol{y}^\top\frac{\partial^2 v_s^{\boldsymbol{\pi},\boldsymbol{p}}}{(\partial\boldsymbol{p})^2}\boldsymbol{y}\right| \\
&= \sum_s \rho_s\cdot\left|(\frac{\boldsymbol{y}}{\|\boldsymbol{y}\|_2})^\top\frac{\partial^2 v_s^{\boldsymbol{\pi},\boldsymbol{p}}}{(\partial\boldsymbol{p})^2}(\frac{\boldsymbol{y}}{\|\boldsymbol{y}\|_2})\right|\cdot\|\boldsymbol{y}\|_2^2 \\
&\le \sum_s \rho_s\cdot\max_{\|\boldsymbol{z}\|_2=1}\left|\left\langle\frac{\partial^2 v_s^{\boldsymbol{\pi},\boldsymbol{p}}}{(\partial\boldsymbol{p})^2}\boldsymbol{z},\boldsymbol{z}\right\rangle\right|\cdot\|\boldsymbol{y}\|_2^2 \\
&= \sum_s \rho_s\cdot\max_{\|\boldsymbol{z}\|_2=1}\left|\left\langle\frac{\partial^2 v_s^{\boldsymbol{\pi},\boldsymbol{p}(\alpha)}}{(\partial\boldsymbol{p}(\alpha))^2}\Big|_{\alpha=0}\frac{\partial\boldsymbol{p}(\alpha)}{\partial\alpha},\frac{\partial\boldsymbol{p}(\alpha)}{\partial\alpha}\right\rangle\right|\cdot\|\boldsymbol{y}\|_2^2 \\
&= \sum_s \rho_s\cdot\max_{\|\boldsymbol{z}\|_2=1}\left|\frac{\partial^2 v_s^{\boldsymbol{\pi},\boldsymbol{p}(\alpha)}}{(\partial\alpha)^2}\Big|_{\alpha=0}\right|\cdot\|\boldsymbol{y}\|_2^2 \\
&\le \frac{2\gamma S^2}{(1-\gamma)^3}\cdot\|\boldsymbol{y}\|_2^2. \qquad (67)
\end{aligned}$$

□

Proof of Lemma 4.3. By the definition of $J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p})$, we have

$$J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p}) - J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p}') = \sum_s \rho_s\left(v_s^{\boldsymbol{\pi},\boldsymbol{p}} - v_s^{\boldsymbol{\pi},\boldsymbol{p}'}\right).$$

For any $s\in\mathcal{S}$ and $\boldsymbol{p},\boldsymbol{p}'\in\mathcal{P}$, we have

$$\begin{aligned}
&v_s^{\boldsymbol{\pi},\boldsymbol{p}} - v_s^{\boldsymbol{\pi},\boldsymbol{p}'} \\
&= v_s^{\boldsymbol{\pi},\boldsymbol{p}} - \sum_a \pi_{sa}\sum_{s'} p'_{sas'}\left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}}\right) + \sum_a \pi_{sa}\sum_{s'} p'_{sas'}\left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}}\right) - v_s^{\boldsymbol{\pi},\boldsymbol{p}'} \\
&= \sum_a \pi_{sa}\sum_{s'} p_{sas'}\left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}}\right) - \sum_a \pi_{sa}\sum_{s'} p'_{sas'}\left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}}\right) \\
&\quad + \sum_a \pi_{sa}\sum_{s'} p'_{sas'}\left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}}\right) - \sum_a \pi_{sa}\sum_{s'} p'_{sas'}\left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}'}\right) \\
&= \sum_a \pi_{sa}\sum_{s'}\left(p_{sas'} - p'_{sas'}\right)\left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}}\right) + \gamma\sum_a \pi_{sa}\sum_{s'} p'_{sas'}\left(v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}} - v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}'}\right) \\
&= \cdots \\
&= \sum_{t=0}^{\infty}\gamma^t\sum_{s'} p'^{\boldsymbol{\pi}}_{ss'}(t)\left(\sum_{a'}\pi_{s'a'}\sum_{s''}\left(p_{s'a's''} - p'_{s'a's''}\right)\left(c_{s'a's''} + \gamma v_{s''}^{\boldsymbol{\pi},\boldsymbol{p}}\right)\right).
\end{aligned}$$

Here, the last equation is obtained by the recursion and we then obtain

$$
\begin{aligned}
J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p}) - J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p}') &= \sum_s \rho_s \left(v_s^{\boldsymbol{\pi},\boldsymbol{p}} - v_s^{\boldsymbol{\pi},\boldsymbol{p}'} \right) \\
&= \sum_s \rho_s \sum_{t=0}^{\infty} \gamma^t \sum_{s'} p'^{\boldsymbol{\pi}}_{ss'}(t) \left(\sum_{a'} \pi_{s'a'} \sum_{s''} \left(p_{s'a's''} - p'_{s'a's''}\right)\left(c_{s'a's''} + \gamma v_{s''}^{\boldsymbol{\pi},\boldsymbol{p}}\right) \right) \\
&= \sum_{s'} \left(\sum_s \sum_{t=0}^{\infty} \gamma^t \rho_s p'^{\boldsymbol{\pi}}_{ss'}(t) \right) \left(\sum_{a'} \pi_{s'a'} \sum_{s''} \left(p_{s'a's''} - p'_{s'a's''}\right)\left(c_{s'a's''} + \gamma v_{s''}^{\boldsymbol{\pi},\boldsymbol{p}}\right) \right) \\
&= \frac{1}{1-\gamma} \sum_s d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}'}(s) \left(\sum_a \pi_{sa} \sum_{s'} \left(p_{sas'} - p'_{sas'}\right)\left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}}\right) \right).
\end{aligned}
$$

Let $\boldsymbol{p}' = \boldsymbol{p}^\star$ and then, we have

$$
\begin{aligned}
0 \le J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p}^\star) - J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p}) &= \frac{1}{1-\gamma} \sum_s d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}^\star}(s) \left(\sum_a \pi_{sa} \sum_{s'} \left(p^\star_{sas'} - p_{sas'}\right)\left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}}\right) \right) \\
&= \frac{1}{1-\gamma} \sum_s \frac{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}^\star}(s)}{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}}(s)} \cdot d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}}(s) \left(\sum_a \pi_{sa} \sum_{s'} \left(p^\star_{sas'} - p_{sas'}\right)\left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}}\right) \right) \\
&\overset{(a)}{\le} \frac{1}{1-\gamma} \cdot \left\| \frac{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}^\star}}{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}}} \right\|_\infty \cdot \sum_s d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}}(s) \left(\sum_a \pi_{sa} \sum_{s'} \left(p^\star_{sas'} - p_{sas'}\right)\left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}}\right) \right) \\
&= \left\| \frac{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}^\star}}{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}}} \right\|_\infty \cdot \sum_{s,a,s'} \left(\frac{1}{1-\gamma} d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}}(s) \pi_{sa} \left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}}\right) \right) \cdot \left(p^\star_{sas'} - p_{sas'}\right) \\
&\le \left\| \frac{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}^\star}}{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}}} \right\|_\infty \cdot \max_{\bar{\boldsymbol{p}} \in \mathcal{P}} \left[\sum_{s,a,s'} \left(\frac{1}{1-\gamma} d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}}(s) \pi_{sa} \left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}}\right) \right) \cdot \left(\bar{p}_{sas'} - p_{sas'}\right) \right] \\
&= \left\| \frac{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}^\star}}{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi},\boldsymbol{p}}} \right\|_\infty \cdot \max_{\bar{\boldsymbol{p}} \in \mathcal{P}} \left\langle \bar{\boldsymbol{p}} - \boldsymbol{p}, \frac{\partial J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p})}{\partial \boldsymbol{p}} \right\rangle \qquad \text{(by Lemma 4.1)} \\
&\le \frac{D}{1-\gamma} \max_{\bar{\boldsymbol{p}} \in \mathcal{P}} \left\langle \bar{\boldsymbol{p}} - \boldsymbol{p}, \frac{\partial J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p})}{\partial \boldsymbol{p}} \right\rangle.
\end{aligned}
$$

which completes the proof. The first inequality (a) is obtained due to the fact that for any $s \in \mathcal{S}$,

$$
\sum_a \pi_{sa} \sum_{s'} \left(p^\star_{sas'} - p_{sas'}\right)\left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi},\boldsymbol{p}}\right) \ge 0
$$

holds under the s-rectangularity assumption.

□

Now, we proceed to prove main theorem in section 4. Here we can define $f_{\boldsymbol{\pi}}(\boldsymbol{p}) := J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{p})$ for a fixed policy $\boldsymbol{\pi} \in \Pi$ and define the gradient mapping

$$
G^\beta(\boldsymbol{p}) := \frac{1}{\beta} \left(\mathrm{Proj}_{\mathcal{P}}(\boldsymbol{p} + \beta \nabla f_{\boldsymbol{\pi}}(\boldsymbol{p})) - \boldsymbol{p} \right). \tag{68}
$$

Notice that $\mathcal{P}$ is convex and $f_{\boldsymbol{\pi}}(\boldsymbol{p})$ is $\ell_{\boldsymbol{p}}$-smooth, then the following lemma can be derived directly using existing classic results:

Lemma G.1. *(Beck, 2017, Theorem 10.15) Let $\{\boldsymbol{p}_t\}_{t\ge 0}$ be the sequence generated by Algorithm 2 for solving the inner problem with the constant step size $\beta := \frac{1}{\ell_{\boldsymbol{p}}}$, then*

$$
\min_{t \in \{0,\cdots,T-1\}} \|G^\beta(\boldsymbol{p}_t)\| \le \sqrt{\frac{2\ell_{\boldsymbol{p}} \left(f^\star_{\boldsymbol{\pi}} - f_{\boldsymbol{\pi}}(\boldsymbol{p}_0)\right)}{T}} \tag{69}
$$

Proof of Theorem 4.4. It has been shown in Lemma 3 in (Ghadimi & Lan, 2016) that if $\|G^\beta(\boldsymbol{p})\| \le \epsilon$, then

$$\nabla f_{\boldsymbol{\pi}}(\boldsymbol{p}^+) \in \mathcal{N}_{\mathcal{P}}(\boldsymbol{p}^+) + 2\epsilon\mathcal{B}(1), \tag{70}$$

where $\boldsymbol{p}^+ := \boldsymbol{p} + \beta G^\beta(\boldsymbol{p})$, $\mathcal{N}_{\mathcal{P}}$ is the norm cone of the set $\mathcal{P}$ and $\mathcal{B}(r) := \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| \le r\}$. By the gradient dominance condition established in Lemma 4.3,

$$\begin{aligned}\min_{t\in\{0,\cdots,T-1\}} \{f_{\boldsymbol{\pi}}(\boldsymbol{p}^\star) - f_{\boldsymbol{\pi}}(\boldsymbol{p}_t)\} &\le \frac{D}{1-\gamma} \min_{t\in\{0,\cdots,T-1\}} \max_{\bar{\boldsymbol{p}}\in\mathcal{P}} \langle \bar{\boldsymbol{p}} - \boldsymbol{p}_t, \nabla f_{\boldsymbol{\pi}}(\boldsymbol{p}_t)\rangle \\ &\le \frac{D}{1-\gamma} \max_{\bar{\boldsymbol{p}}\in\mathcal{P}} \langle \bar{\boldsymbol{p}} - \boldsymbol{p}_{\hat{t}}, \nabla f_{\boldsymbol{\pi}}(\boldsymbol{p}_{\hat{t}})\rangle\,,\end{aligned} \tag{71}$$

where $\hat{t} := 1 + \arg\min_{t\le T-1} \|G^\beta(\boldsymbol{p}_t)\|$. Recall Lemma G.1, we showed that

$$\|G^\beta(\boldsymbol{p}_{\hat{t}-1})\| \le \sqrt{\frac{2\ell_{\boldsymbol{p}}\left(f^\star_{\boldsymbol{\pi}} - f_{\boldsymbol{\pi}}(\boldsymbol{p}_0)\right)}{T}} \le \sqrt{\frac{2\ell_{\boldsymbol{p}}}{(1-\gamma)T}},$$

where the last inequality holds due to

$$v^{\boldsymbol{\pi}}_s = \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{p}}\left[\sum_{t=0}^\infty \gamma^t c_{s_t a_t s_{t+1}} \mid s_0 = s\right] \le \sum_{t=0}^\infty \gamma^t = \frac{1}{1-\gamma}. \tag{72}$$

If we set that

$$\sqrt{\frac{2\ell_{\boldsymbol{p}}}{(1-\gamma)T}} \le \frac{(1-\gamma)\epsilon}{4D\sqrt{SA}} \Longleftrightarrow T \ge \frac{32\ell_{\boldsymbol{p}} D^2 SA}{(1-\gamma)^3\epsilon^2} = \mathcal{O}(\epsilon^{-2}),$$

then

$$\|G^\beta(\boldsymbol{p}_{\hat{t}-1})\| \le \frac{(1-\gamma)\epsilon}{4D\sqrt{SA}}.$$

Hence, by applying the equation (70), we have

$$(71) \le \frac{D}{1-\gamma} \max_{\bar{\boldsymbol{p}}\in\mathcal{P}} \|\bar{\boldsymbol{p}} - \boldsymbol{p}_{\hat{t}}\| \cdot 2 \cdot \frac{(1-\gamma)\epsilon}{4D\sqrt{SA}} = \epsilon,$$

where for any $\boldsymbol{p}_1, \boldsymbol{p}_2 \in \mathcal{P}$,

$$\|\boldsymbol{p}_1 - \boldsymbol{p}_2\| \le \|\boldsymbol{p}_1\| + \|\boldsymbol{p}_2\| \le 2\sqrt{SA}. \tag{73}$$

□

Then, we provide the standard proof of Lemma 4.5.

Proof of Lemma 4.5. We first show that the inner problem gradient form. Notice that,

$$\frac{\partial J_{\boldsymbol{\rho}}(\boldsymbol{\pi}, \boldsymbol{\xi})}{\partial \boldsymbol{\xi}} = \sum_{s\in\mathcal{S}} \frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_s}{\partial \boldsymbol{\xi}} \rho_s.$$

Then we consider the $\frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_s}{\partial \boldsymbol{\xi}}$ directly.

$$\begin{aligned}\frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_s}{\partial \boldsymbol{\xi}} &= \frac{\partial}{\partial \boldsymbol{\xi}}\left[\sum_a \pi_{sa} Q^{\boldsymbol{\pi}_{sa},\boldsymbol{\xi}}\right] \\ &= \sum_a \pi_{sa} \frac{\partial}{\partial \boldsymbol{\xi}}\left[\sum_{s'} p^{\boldsymbol{\xi}}_{sas'}\left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s'}\right)\right] \\ &= \sum_a \pi_{sa} \sum_{s'}\left[\frac{\partial p^{\boldsymbol{\xi}}_{sas'}}{\partial \boldsymbol{\xi}}\left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s'}\right) + \gamma p^{\boldsymbol{\xi}}_{sas'} \frac{\partial v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s'}}{\partial \boldsymbol{\xi}}\right] \\ &= \sum_a \pi_{sa} \sum_{s'} \frac{\partial p^{\boldsymbol{\xi}}_{sas'}}{\partial \boldsymbol{\xi}}\left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s'}\right) + \gamma \sum_a \pi_{sa} \sum_{s'} p^{\boldsymbol{\xi}}_{sas'} \frac{\partial v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s'}}{\partial \boldsymbol{\xi}}.\end{aligned}$$

By condensing $\sum_a \pi_{sa} p^{\boldsymbol{\xi}}_{sas'} = p^{\boldsymbol{\pi},\boldsymbol{\xi}}_{ss'}(1)$, we can obtain,

$$
\begin{aligned}
\frac{\partial v^{\boldsymbol{\pi},\boldsymbol{p}}_s}{\partial \boldsymbol{\xi}} &= \sum_a \pi_{sa} \sum_{s'} \frac{\partial p^{\boldsymbol{\xi}}_{sas'}}{\partial \boldsymbol{\xi}} \left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s'} \right) + \gamma \sum_{s'} p^{\boldsymbol{\pi},\boldsymbol{\xi}}_{ss'}(1) \frac{\partial v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s'}}{\partial \boldsymbol{\xi}} \\
&= \sum_a \pi_{sa} \sum_{s'} \frac{\partial p^{\boldsymbol{\xi}}_{sas'}}{\partial \boldsymbol{\xi}} \left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s'} \right) \\
&\quad + \gamma \sum_{s'} p^{\boldsymbol{\pi},\boldsymbol{\xi}}_{ss'}(1) \left[\sum_{a'} \pi_{s'a'} \sum_{s''} \frac{\partial p^{\boldsymbol{\xi}}_{s'a's''}}{\partial \boldsymbol{\xi}} \left(c_{s'a's''} + \gamma v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s''} \right) + \gamma \sum_{s''} p^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s's''}(1) \frac{\partial v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s''}}{\partial \boldsymbol{\xi}} \right] \\
&= \sum_{k=0}^{1} \gamma^k \sum_{s'} p^{\boldsymbol{\pi},\boldsymbol{\xi}}_{ss'}(k) \sum_{a'} \pi_{s'a'} \left[\sum_{s''} \frac{\partial p^{\boldsymbol{\xi}}_{s'a's''}}{\partial \boldsymbol{\xi}} \left(c_{s'a's''} + \gamma v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s''} \right) \right] + \gamma^2 \sum_{s'} p^{\boldsymbol{\pi},\boldsymbol{\xi}}_{ss'}(2) \frac{\partial v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s'}}{\partial \boldsymbol{\xi}} \\
&= \sum_{k=0}^{2} \gamma^k \sum_{s'} p^{\boldsymbol{\pi},\boldsymbol{\xi}}_{ss'}(k) \sum_{a'} \pi_{s'a'} \left[\sum_{s''} \frac{\partial p^{\boldsymbol{\xi}}_{s'a's''}}{\partial \boldsymbol{\xi}} \left(c_{s'a's''} + \gamma v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s''} \right) \right] + \gamma^3 \sum_{s'} p^{\boldsymbol{\pi},\boldsymbol{\xi}}_{ss'}(3) \frac{\partial v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s'}}{\partial \boldsymbol{\xi}} \\
&= \cdots \\
&= \sum_{k=0}^{\infty} \gamma^k \sum_{s'} p^{\boldsymbol{\pi},\boldsymbol{\xi}}_{ss'}(k) \sum_{a'} \pi_{s'a'} \left[\sum_{s''} \frac{\partial p^{\boldsymbol{\xi}}_{s'a's''}}{\partial \boldsymbol{\xi}} \left(c_{s'a's''} + \gamma v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s''} \right) \right].
\end{aligned}
$$

So we have

$$
\begin{aligned}
\frac{\partial J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} &= \sum_{s \in \mathcal{S}} \frac{\partial v^{\boldsymbol{\pi},\boldsymbol{\xi}}_s}{\partial \boldsymbol{\xi}} \rho_s \\
&= \sum_s \rho_s \sum_{k=0}^{\infty} \gamma^k \sum_{s'} p^{\boldsymbol{\pi},\boldsymbol{\xi}}_{ss'}(k) \sum_{a'} \pi_{s'a'} \left[\sum_{s''} \frac{\partial p^{\boldsymbol{\xi}}_{s'a's''}}{\partial \boldsymbol{\xi}} \left(c_{s'a's''} + \gamma v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s''} \right) \right] \\
&= \frac{1}{1-\gamma} \sum_{s'} \underbrace{(1-\gamma) \sum_s \rho_s \sum_{k=0}^{\infty} \gamma^k p^{\boldsymbol{\pi},\boldsymbol{\xi}}_{ss'}(k)}_{d^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s'}} \sum_{a'} \pi_{s'a'} \left[\sum_{s''} \frac{\partial p^{\boldsymbol{\xi}}_{s'a's''}}{\partial \boldsymbol{\xi}} \left(c_{s'a's''} + \gamma v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s''} \right) \right] \\
&= \frac{1}{1-\gamma} \sum_s d^{\boldsymbol{\pi},\boldsymbol{\xi}}_s \sum_a \pi_{sa} \left[\sum_{s'} \frac{\partial p^{\boldsymbol{\xi}}_{sas'}}{\partial \boldsymbol{\xi}} \left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s'} \right) \right] \\
&= \frac{1}{1-\gamma} \sum_s d^{\boldsymbol{\pi},\boldsymbol{\xi}}_s \sum_a \pi_{sa} \sum_{s'} p^{\boldsymbol{\xi}}_{sas'} \left[\frac{\partial p^{\boldsymbol{\xi}}_{sas'}}{\partial \boldsymbol{\xi}} \cdot \frac{1}{p^{\boldsymbol{\xi}}_{sas'}} \cdot \left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s'} \right) \right] \\
&= \frac{1}{1-\gamma} \sum_s d^{\boldsymbol{\pi},\boldsymbol{\xi}}_s \sum_a \pi_{sa} \sum_{s'} p^{\boldsymbol{\xi}}_{sas'} \left[\frac{\partial \log p^{\boldsymbol{\xi}}_{sas'}}{\partial \boldsymbol{\xi}} \left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s'} \right) \right] \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s \sim \boldsymbol{d}^{\boldsymbol{\pi},\boldsymbol{\xi}}_{\boldsymbol{\rho}}} \mathbb{E}_{a \sim \boldsymbol{\pi}_{s\cdot}} \mathbb{E}_{s' \sim \boldsymbol{p}_{sa\cdot}} \left[\frac{\partial \log p^{\boldsymbol{\xi}}_{sas'}}{\partial \boldsymbol{\xi}} \left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s'} \right) \right].
\end{aligned}
$$

Then, we consider the partial derivative on $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ separately. Notice that

$$
\begin{cases}
\frac{\partial J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{\xi})}{\partial \theta_i} = \frac{1}{1-\gamma} \mathbb{E}_{s \sim \boldsymbol{d}^{\boldsymbol{\pi},\boldsymbol{\xi}}_{\boldsymbol{\rho}}} \mathbb{E}_{a \sim \boldsymbol{\pi}_{s\cdot}} \mathbb{E}_{s' \sim \boldsymbol{p}_{sa\cdot}} \left[\frac{\partial \log p^{\boldsymbol{\xi}}_{sas'}}{\partial \theta_i} \left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s'} \right) \right] \\
\frac{\partial J_{\boldsymbol{\rho}}(\boldsymbol{\pi},\boldsymbol{\xi})}{\partial \lambda_{sa}} = \frac{1}{1-\gamma} \mathbb{E}_{s \sim \boldsymbol{d}^{\boldsymbol{\pi},\boldsymbol{\xi}}_{\boldsymbol{\rho}}} \mathbb{E}_{a \sim \boldsymbol{\pi}_{s\cdot}} \mathbb{E}_{s' \sim \boldsymbol{p}_{sa\cdot}} \left[\frac{\partial \log p^{\boldsymbol{\xi}}_{sas'}}{\partial \lambda_{sa}} \left(c_{sas'} + \gamma v^{\boldsymbol{\pi},\boldsymbol{\xi}}_{s'} \right) \right]
\end{cases}
$$

We found that for all $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, θ_i will appear in the parametrization form of $p^{\boldsymbol{\xi}}_{sas'}$. Hence we consider partial

derivative of $\log p^{\boldsymbol{\xi}}_{sas'}$ then.

$$\begin{aligned}
\frac{\partial \log p^{\boldsymbol{\xi}}_{sas'}}{\partial \theta_i} &= \frac{\partial}{\partial \theta_i}\left[\log \bar{\boldsymbol{p}}_{sas'} + \frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{s'})}{\lambda_{sa}}\right] - \frac{\partial}{\partial \theta_i}\left[\log\left(\sum_k \bar{p}_{sak}\cdot \exp(\frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(k)}{\lambda_{sa}})\right)\right] \\
&= \frac{\phi_i(s')}{\lambda_{sa}} - \frac{\sum_k \bar{p}_{sak}\cdot \exp(\frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(k)}{\lambda_{sa}})\cdot \frac{\phi_i(k)}{\lambda_{sa}}}{\sum_k \bar{p}_{sak}\cdot \exp(\frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(k)}{\lambda_{sa}})} \\
&= \frac{\phi_i(s')}{\lambda_{sa}} - \sum_j \frac{\bar{p}_{saj}\cdot \exp(\frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(j)}{\lambda_{sa}})}{\sum_k \bar{p}_{sak}\cdot \exp(\frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(k)}{\lambda_{sa}})}\cdot \frac{\phi_i(j)}{\lambda_{sa}} \\
&= \frac{\phi_i(s')}{\lambda_{sa}} - \sum_j p^{\boldsymbol{\xi}}_{saj}\cdot \frac{\phi_i(j)}{\lambda_{sa}}.
\end{aligned}$$

Now we can obtain that

$$\frac{\partial J_\rho(\boldsymbol{\pi},\boldsymbol{\xi})}{\partial \theta_i} = \frac{1}{1-\gamma}\sum_s d_s^{\boldsymbol{\pi},\boldsymbol{\xi}} \sum_a \pi_{sa} \sum_{s'} p^{\boldsymbol{\xi}}_{sas'}\left[\left(\frac{\phi_i(s')}{\lambda_{sa}} - \sum_j p^{\boldsymbol{\xi}}_{saj}\cdot \frac{\phi_i(j)}{\lambda_{sa}}\right)\cdot\left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi},\boldsymbol{\xi}}\right)\right].$$

Similarly we can derive the partial derivative on λ_{sa} for any state-action pair (s,a). Interestingly, we notice that for $(\bar{s},\bar{a}) \neq (s,a)$, $\frac{\partial \log(p^{\boldsymbol{\xi}}_{\bar{s}\bar{a}s'})}{\partial \lambda_{sa}} = 0$. Therefore, we can consider the case $(\bar{s},\bar{a}) = (s,a)$.

$$\begin{aligned}
\frac{\partial \log p^{\boldsymbol{\xi}}_{sas'}}{\partial \lambda_{sa}} &= \frac{\partial}{\partial \lambda_{sa}}\left[\log \bar{p}_{sas'} + \frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{s'})}{\lambda_{sa}}\right] - \frac{\partial}{\partial \lambda_{sa}}\left[\log\left(\sum_k \bar{p}_{sak}\cdot \exp(\frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(k)}{\lambda_{sa}})\right)\right] \\
&= \frac{\sum_k \bar{p}_{sak}\cdot \exp\left(\frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(k)}{\lambda_{sa}}\right)\cdot \frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(k)}{\lambda_{sa}^2}}{\sum_k \bar{p}_{sak}\cdot \exp\left(\frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(k)}{\lambda_{sa}}\right)} - \frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(s')}{\lambda_{sa}^2} \\
&= \sum_j \frac{\bar{p}_{saj}\cdot \exp\left(\frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(j)}{\lambda_{sa}}\right)}{\sum_k \bar{p}_{sak}\cdot \exp\left(\frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(k)}{\lambda_{sa}}\right)}\cdot \frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(j)}{\lambda_{sa}^2} - \frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(s')}{\lambda_{sa}^2} \\
&= \sum_j p^{\boldsymbol{\xi}}_{saj}\cdot \frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(j)}{\lambda_{sa}^2} - \frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(s')}{\lambda_{sa}^2}.
\end{aligned}$$

Then we can obtain that

$$\frac{\partial J_\rho(\boldsymbol{\pi},\boldsymbol{\xi})}{\partial \lambda_{sa}} = \frac{1}{1-\gamma} d_s^{\boldsymbol{\pi},\boldsymbol{\xi}}\cdot \pi_{sa}\cdot \sum_{s'} p^{\boldsymbol{\xi}}_{sas'}\left[\left(\sum_j p^{\boldsymbol{\xi}}_{saj}\cdot \frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(j)}{\lambda_{sa}^2} - \frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(s')}{\lambda_{sa}^2}\right)\cdot\left(c_{sas'} + \gamma v_{s'}^{\boldsymbol{\pi},\boldsymbol{\xi}}\right)\right].$$

□

H. Experiment Details

H.1. Details on the Garnet problem example

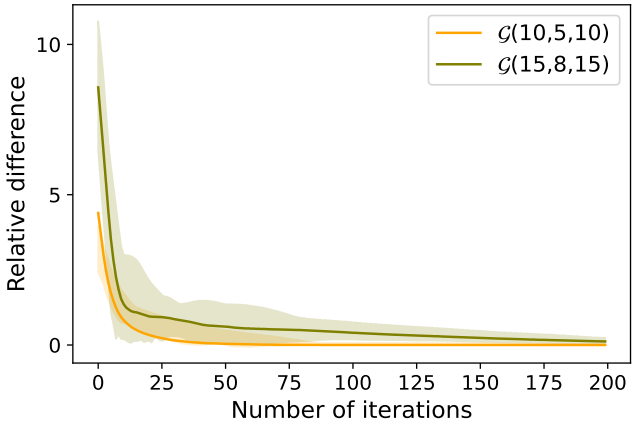

Figure 3. The error of value function computed by non-parametric DRPG for two Garnet problems with s-rectangular ambiguity.

Note that, in our simulations, we test our algorithm for both high connectivity (*i.e.*, b = $|\mathcal{S}|$) in s-rectangular case, and low connectivity (*i.e.*, b = $|\mathcal{S}|/5$) in (s, a)-rectangular case. We also apply DRPG on random RMDPs with L_1-constrained s-rectangular ambiguity, which generally assumes the uncertain in transition probabilities is independent for each state-action pair and are defined as

$$\mathcal{P} = \underset{s\in\mathcal{S}}{\times} \mathcal{P}_s \quad \text{where} \quad \mathcal{P}_s := \left\{ (\boldsymbol{p}_{s1}, \ldots, \boldsymbol{p}_{sA}) \in (\Delta^S)^A \mid \sum_{a\in\mathcal{A}} \|\boldsymbol{p}_{sa} - \bar{\boldsymbol{p}}_{sa}\|_1 \leq \kappa_s \right\}.$$

We run DRPG with a sample size of 50 for 200 iteration times on Garnet problems with three sizes for the (s, a)-rectangular case and two medium sizes for the s-rectangular case. We record the absolute value of gaps between objective values of DRPG and robust value iteration at each iteration time step, and then plot the relative difference under the s-rectangular assumption in Figure 3. The upper and lower envelopes of the curves correspond to the 95 and 5 percentiles of the 50 samples, respectively. From Figure 3, we can obtain similar results with the (s, a)-rectangular case that DRPG converges to a nearly identical value computed by the value iteration computed by the robust value iteration.

H.2. Details on the inventory management example

In our inventory management example, we present a specific example of this problem with eight states and three actions.

We draw the cost for each $(s, a, s) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ at random uniformly in $[0, 1]$, and we fix a discount factor $\gamma = 0.95$. Below we give details about the nominal transitions and the parameter $\boldsymbol{\kappa}$.

The feature function we use is the radial-type features which is introduced in (Sutton & Barto, 2018), *i.e.*, $\boldsymbol{\phi}_i(s) = \exp\left(-\frac{\|s-c_i\|^2}{2\sigma_i^2}\right)$. We define a two-dimension state feature with deterministic c_i and σ_i. Our parameters also share the same dimensions as these two features from our parameterization form.

The ambiguity set Ξ in our problem is simply chosen as a L_1-norm constrained set, that is,

$$\Xi := \{(\boldsymbol{\theta}, \boldsymbol{\lambda}) | \|\boldsymbol{\theta} - \boldsymbol{\theta}_c\|_1 \leq \kappa_\theta, \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_c\|_1 \leq \kappa_\lambda\}. \tag{74}$$

The updating step size for $\boldsymbol{\xi} = (\boldsymbol{\theta}, \boldsymbol{\lambda})$ on the inner problem are taken 0.01. For simplicity, we choose all elements of $\boldsymbol{\lambda}_c$ as one and $\boldsymbol{\theta}_c := [0.4, 0.9]^\top$, and set $\kappa_\theta = 1, \kappa_\lambda = 1$ in this problem. Other parameters are included in the published codes. Note that the instances for a larger number of states are constructed in the same fashion by adding some condition states.