A Single-Loop Robust Policy Gradient Method for Robust Markov Decision Processes

Zhenwei Lin [*1] **Chenyu Xue** [*1] **Qi Deng** [2] **Yinyu Ye** [3]

Abstract

Robust Markov Decision Processes (RMDPs) have recently been recognized as a valuable and promising approach to discovering a policy with creditable performance, particularly in the presence of a dynamic environment and estimation errors in the transition matrix due to limited data. Despite extensive exploration of dynamic programming algorithms for solving RMDPs, there has been a notable upswing in interest in developing efficient algorithms using the policy gradient method. In this paper, we propose the first single-loop robust policy gradient (SRPG) method with the global optimality guarantee for solving RMDPs through its minimax formulation. Moreover, we complement the convergence analysis of the nonconvex-nonconcave min-max optimization problem with the objective function's gradient dominance property, which is not explored in the prior literature. Numerical experiments validate the efficacy of SRPG, demonstrating its faster and more robust convergence behavior compared to its nested-loop counterpart.

1. Introduction

Markov decision processes (MDPs) serve as an important model for sequential decision-making under uncertainty and enjoys wide applications in finance (Deng et al., 2016; Jiang et al., 2017), autonomous driving (Kiran et al., 2021; Sallab et al., 2017), and revenue management (den Boer & Zwart, 2015), etc. However, in most applications, the decision-maker can only estimate model parameters, especially the transition kernel, from noisy and scarce observation data. Consequently, a policy that exhibits poor performance with respect to the true parameters may be employed since the optimal policy obtained from the estimated parameters could be highly sensitive to the small changes in the problem parameters, resulting in suboptimal outcomes (Goyal & Grand-Clément, 2023).

Motivated by recent developments in robust optimization and its practical performance (Ben-Tal & Nemirovski, 2002; Bertsimas & Sim, 2004), robust Markov Decision Processes (RMDPs) have emerged as a valuable and promising approach to overcome this obstacle. It assumes that the transition kernel lies in a pre-determined ambiguity set and then seeks a policy with the best performance under the worst-case transition kernel. Hence, it involves solving a min-max optimization problem. Moreover, it has been demonstrated that the optimal policies of RMDPs display an advantageous performance in out-of-sample scenarios when the transition kernel needs to be estimated from limited data or undergoes changes over time (Xu & Mannor, 2009; Ghavamzadeh et al., 2016; Mannor et al., 2016).

With general uncertainty sets, Wiesemann et al. (2013) prove that it is NP-hard to find the optimal policy for RMDPs. However, under certain rectangular assumptions of the ambiguity set, the problem becomes tractable. For example, when the ambiguity set is (s, a)-rectangular, the dynamic programming techniques apply, and the value iteration is known to achieve linear convergence to optimal robust value (Iyengar, 2005; Nilim & El Ghaoui, 2005). Here, the (s, a)-rectangular ambiguity set allows the adversarial nature to choose the worst-case transition probability vector of each state and action pair independently. Since the (s, a)-rectangular assumption is too restrictive and leads to conservative policies, we consider the more general s-rectangular ambiguity set, which allows the nature to choose the transition kernel for each state without observing the action and also preserves tractability (Wiesemann et al., 2013).

Nowadays, the policy gradient (PG) method has become the workhorse for solving a special case of RMDPs, where there is no ambiguity in the transition kernel. It is scalable, easy to implement, and versatile across various settings, including model-free and continuous state-action spaces (Kakade & Langford, 2002; Schulman et al., 2017). However, the

[*]Equal contribution [1]Shanghai University of Finance and Economics [2]Antai College of Economics and Management, Shanghai Jiao Tong University [3]Stanford University. Correspondence to: Qi Deng <qdeng24@sjtu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

policy gradient approach to solving general RMDPs has been much less investigated. Recently, Wang et al. (2023) propose a double-loop robust policy gradient (DRPG) for solving s-rectangular RMDPs. The outer loop of DRPG is designed for updating policies, which resembles policy gradient updates in non-robust MDPs, while the inner loop is to solve the maximization problem with a given policy over the ambiguity set and update the worst-case transition matrices. They prove that DRPG is guaranteed to converge to a globally optimal policy within $\mathcal{O}(1/\varepsilon^4)$ outer iterations. However, the critical drawback of DRPG is that the inner loop needs to solve a series of nontrivial nonconcave subproblems with increasing accuracy, which brings an unacceptable computation burden and a worse convergence complexity than $\mathcal{O}(1/\varepsilon^4)$.

In this paper, our goal is to develop a more efficient algorithm for RMDPs that significantly improves the convergence complexity of the existing work. Our main contributions are summarized as follows.

First, we present a single-loop algorithm, Single-loop Robust Policy Gradient (SRPG), for RMDPs with an s-rectangular ambiguity set. Our algorithm is adapted from the primal-dual algorithm for min-max optimization (Zhang et al., 2020; Zheng et al., 2023), which employs the Moreau-Yosida smoothing technique to enhance the balance between the primal and dual updates. To our knowledge, SRPG is the first policy gradient method for RMDPs without involving a more complicated nested loop. Consequently, our algorithm not only obtains the best iteration complexity of policy gradient methods for RMDPs with s-rectangular ambiguity sets, but also has a much faster update per iteration than the double-loop algorithm proposed in Wang et al. (2023). Our numerical experiments further show the superior performance of SRPG.

Second, we prove the SRPG converges at rate $\mathcal{O}(1/\varepsilon^4)$ under gradient dominance property. This result appears to be new even for general nonconvex-nonconcave min-max optimization. To our knowledge, such convergence rate for min-max optimization with gradient dominance property was only obtained when the max component is a concave function. While nonconvex-concave min-max optimization has a wide range of applications, unfortunately, it is not satisfied in the RMDP setting.

Finally, we provide a detailed analysis of the varying convergence rates of SRPG when applied to RMDPs. Leveraging the gradient-dominance-like property for nonsmooth weakly convex functions, we demonstrate that the objective, as a function of π, exhibits exponential growth within a specific region. This observation implies an accelerated convergence rate when the iterative sequence is distant from the optimal solution set. While this does not improve the convergence rate in terms of ε, it helps explain the initially faster convergence observed in our experiments.

The paper is organized in the following manner. In the remainder of this section, we review the literature related to our work. In Section 2, we introduce some preliminaries and notations and make some assumptions for the later analysis. Section 3 provides some important properties of RMDPs under s-rectangular ambiguity set. Section 4 presents the specific details of SRPG for RMDPs and the comprehensive convergence analysis for SRPG. In Section 5, we conduct two numerical experiments to examine the convergence performance of SRPG. Finally, we draw the conclusion in Section 5. All omitted proofs and details of numerical experiments can be found in the Appendix.

1.1. Literature Review

Our paper is related to two streams of the literature, including the policy gradient method applied to RMDPs and min-max optimization.

Policy gradient for RMDPs. Recently, there has been a surge of interest in the policy gradient method for RMDPs with different ambiguity sets. Wang & Zou (2022) propose a policy gradient method for solving RMDPs under a r-contamination ambiguity set, which is restrictive compared with s-rectangular ambiguity set, and the problem can be reduced to an ordinary MDP (Wang et al., 2023). For (s,a)-rectangular ambiguity set, Li et al. (2022) develop an algorithm with a linear convergence rate. However, their analysis depends on the (s,a)-rectangular assumption and can not be applied to our s-rectangular set. Wang et al. (2024) focus on the KL ambiguity set and approximate the worst transition kernel at each iteration instead of fully solving the inner problem. There are also some papers considering different structured s-rectangular ambiguity sets (Grand-Clément & Kroer, 2021; Kumar et al., 2023a; Li & Lan, 2023). Grand-Clément & Kroer (2021) propose an algorithm interleaving primal-dual first-order updates with approximate value iteration updates and prove its ergodic convergence for ellipsoidal and Kullback-Leibler s-rectangular ambiguity sets. Kumar et al. (2023a) consider the L_p-ball s-rectangular ambiguity set, give a closed-form expression of the worst-case transition kernel, and propose the robust policy gradient method. However, they mainly discuss the time complexity involved in computing the robust policy gradient without providing the convergence analysis of the overall algorithm. Li & Lan (2023) introduce a new type of s-rectangular ambiguity set: a convex combination of a fixed (possibly unknown) transition kernel and another transition kernel belonging to a pre-specified convex set. They focus on the policy evaluation problem and formulate it as an MDP from the view of nature. For the general s-rectangular ambiguity set, Kumar et al. (2023b) propose an algorithm achieving the $\mathcal{O}(1/\varepsilon)$ convergence rate. However, it relies on a strong

assumption that the objective function of the minimization problem is smooth, which does not necessarily hold for many ambiguity sets. Wang et al. (2023) and Li et al. (2024) propose a nested-loop algorithm for solving s-rectangular and non-rectangular RMDPs, respectively, and they both obtain the $\mathcal{O}(1/\varepsilon^4)$ convergence rate. However, the inner loop needs to optimize over the ambiguity set of transition matrices, which requires high computational cost. To avoid such computation burden, we propose a single-loop robust policy gradient method for the general s-rectangular ambiguity set.

Min-max optimization. Minimax optimization has garnered significant attention across various fields, including robust optimization (Duchi & Namkoong, 2019), game theory (Bailey et al., 2020), and adversarial machine learning (Goodfellow et al., 2014). Although there is a comprehensive range of literature on minimax optimization, most prior research (Hamedani & Aybat, 2021; Ouyang & Xu, 2021; Zhang & Xiao, 2017) predominantly concentrates on the convex-convex setting. Recently, there has been a shift in focus towards nonconvex or nonconcave minimax problems (Zheng et al., 2023; Xu et al., 2023; Lin et al., 2020; Zhang et al., 2020). This emerging interest is primarily due to the prevalent occurrence of such problems in practical applications, as discussed in Razaviyayn et al. (2020). The convergence guarantee of most current algorithms in this field depends on additional information such as one-side convexity, PL condition, weak Minty Variational Inequality (Diakonikolas et al., 2021), positive interaction dominance (Hajizadeh et al., 2023) and KŁ condition (Zheng et al., 2023). However, these properties do not apply in the specific context of RMDPs. The gradient dominance property associated with RMDPs bears similarities to the KŁ property yet extends beyond its scope.

2. Preliminaries and Notations

Throughout the paper, we use Δ^d to denote the d-dimension probability simplex. With slight abuse of notation, we use $\|\cdot\|$ to represent the Frobenius norm for matrix and L_2 norm for vector. We also use $[n]$ to denote the set $\{1, 2, \ldots, n\}$.

We first introduce the notations of MDPs. An ordinary MDP is specified by a tuple $(\mathcal{S}, \mathcal{A}, p, c, \gamma, \rho)$, where $\mathcal{S} := \{1, 2, \cdots, S\}$ is the finite state set of size S, $\mathcal{A} := \{1, 2, \cdots, A\}$ is the action set of size A, $p = (p_{sa})_{s\in\mathcal{S}, a\in\mathcal{A}} \in (\Delta^S)^{S\times A}$ is the transition kernel with $p_{sa} \in \Delta^S$ being the transition probability vector from a current state s to a subsequent state s' after taking an action a, $c = (c_{cas'})_{s\in\mathcal{S}, a\in\mathcal{A}, s'\in\mathcal{S}}$ is the cost of the aforementioned transition, $\gamma \in (0, 1)$ is the discount factor, and $\rho \in \Delta^S$ is the distribution of the initial state. Moreover, a policy maps from state $s \in \mathcal{S}$ to a distribution over action $a \in \mathcal{A}$ is denoted by $\pi := (\pi_s)_{s\in\mathcal{S}}$.

We further define the policy space $\Pi := (\Delta^A)^S$ and transition kernel space $\mathcal{P} := (\Delta^S)^{S\times A}$. We use $\mathcal{N}_\Pi(\pi) := \{\tilde{\pi} \mid \langle\tilde{\pi}, \bar{\pi} - \pi\rangle \le 0, \forall\bar{\pi} \in \Pi\}$ and $\mathcal{N}_\mathcal{P}(p) := \{\tilde{p} \mid \langle\tilde{p}, \bar{p} - p\rangle \le 0, \forall\bar{p} \in \mathcal{P}\}$ to denote the normal cones of Π at policy π and $\mathcal{P}$ at transition kernel p, respectively. We also denote proximal operator of $\phi(\cdot)$ as $\mathrm{prox}_{\phi,r_1}(\pi) := \mathrm{argmin}_{\tilde{\pi}\in\Pi}\, \phi(\tilde{\pi}) + r_1\|\tilde{\pi} - \pi\|^2/2$. We use $\mathrm{dist}(v, \mathcal{V}) := \min_{\bar{v}\in\mathcal{V}} \|\bar{v} - v\|$ to denote distance between vector or matrix v and set $\mathcal{V}$. We also define some notations in Table 1, which will be used frequently in Section 4.

Notation	Meaning
$\phi(\pi)$	$\max_{p\in\mathcal{P}} J_\rho(\pi, p)$
$\chi(\pi, p, \bar{\pi}, \bar{p})$	$J_\rho(\pi, p) + \frac{r_1}{2}\|\pi - \bar{\pi}\|^2 - \frac{r_2}{2}\|p - \bar{p}\|^2$
$\varphi_\pi(p, \bar{\pi}, \bar{p})$	$\min_{\pi\in\Pi} \chi(\pi, p, \bar{\pi}, \bar{p})$
$\varphi_{\pi,p}(\bar{\pi}, \bar{p})$	$\max_{p\in\mathcal{P}} \varphi_\pi(p, \bar{\pi}, \bar{p})$
$\varphi_{\pi,p,\bar{\pi}}(\bar{p})$	$\min_{\bar{\pi}} \varphi_{\pi,p}(\bar{\pi}, \bar{p})$
$\varphi_{\pi,p,\bar{p}}(\bar{\pi})$	$\max_{\bar{p}} \varphi_{\pi,p}(\bar{\pi}, \bar{p})$
$\underline{\chi}$	$\min_{\bar{\pi}} \max_{\bar{p}} \varphi_{\pi,p}(\bar{\pi}, \bar{p})$
$p^+(\bar{\pi}, \bar{p})$	$\mathrm{Proj}_\mathcal{P}(p + \sigma\nabla_p\chi(\pi(p, \bar{\pi}, \bar{p}), p, \bar{\pi}, \bar{p}))$
$\bar{p}^+(\bar{\pi})$	$\bar{p} + \mu\big(p(\pi(\bar{\pi}, \bar{p}), \bar{\pi}, \bar{p}) - \bar{p}\big)$
$p^+(\bar{p})$	$\mathrm{Proj}_\mathcal{P}(p + \sigma\nabla_p J_\rho(\pi(p, \bar{\pi}; r_1), p))$
$\bar{p}(\bar{\pi})$	$\mathrm{argmax}_{\bar{p}} \min_{\pi\in\Pi} \max_{p\in\mathcal{P}} \chi(\pi, p, \bar{\pi}, \bar{p})$
$\pi(\bar{\pi}, \bar{p})$	$\mathrm{argmin}_{\pi\in\Pi} \max_{p\in\mathcal{P}} \chi(\pi, p, \bar{\pi}, \bar{p})$

Table 1. Useful Notations.

We now give some useful definitions in MDP. The discounted state occupancy measure $d_\rho^{\pi,p} : \mathcal{S} \to [0, 1]$ for an initial distribution ρ, a policy π, and a transition kernel p is defined as $d_\rho^{\pi,p}(s') = (1-\gamma)\sum_{s\in\mathcal{S}}\sum_{t=0}^\infty \gamma^t\rho(s)p_{ss'}^\pi(t)$, where $p_{ss'}^\pi(t)$ is the probability of arriving in a state s' after transiting t time steps from state s according to policy π and transition kernel p. The performance of a policy π is measured by the value function $v^{\pi,p} := [v_1^{\pi,p}, \ldots, v_S^{\pi,p}]^\top \in \mathbb{R}^S$, where $v_s^{\pi,p}$ is defined by $v_s^{\pi,p} = \mathbb{E}_{\pi,p}\big[\sum_{t=0}^\infty \gamma^t c_{s_t a_t s_{t+1}} \mid s_0 = s\big]$. The value function of taking action a at state s, namely the action value function, is defined by $q_{sa}^{\pi,p} = \mathbb{E}_{\pi,p}\big[\sum_{t=0}^\infty \gamma^t c_{s_t a_t s_{t+1}} \mid s_0 = s, a_0 = a\big]$. The objective of an ordinary MDP is to find an optimal policy $\pi^* \in \mathbb{R}^{S\times A}$ to minimize the total expected cost: $\pi^* = \mathrm{argmin}_{\pi\in\Pi}\, \mathbb{E}_{\pi,p,s_0\sim\rho}\big[\sum_{t=0}^\infty \gamma^t c_{s_t a_t s_{t+1}}\big]$.

In real applications, the transition kernel is typically unknown, but the decision-maker may have some partial knowledge of it and can build a corresponding ambiguity set. This can be modeled by an RMDP, characterized by $(\mathcal{S}, \mathcal{A}, \mathcal{P}, c, \gamma, \rho)$, where the transition kernel p is replaced by the ambiguity set $\mathcal{P}$. The goal of an RMDP is to find an

optimal policy under the worst-case transition kernel:

$$\min_{\pi\in\Pi}\max_{p\in\mathcal{P}} J_\rho(\pi,p) := \sum_{s\in\mathcal{S}} \rho_s v_s^{\pi,p}. \tag{1}$$

In the rest of this section, we introduce some assumptions used throughout the paper.

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

Assumption 2.2 (s-Rectangular Ambiguity Set)**.** *The ambiguity set $\mathcal{P}$ is convex and s-rectangular, namely the transition probabilities are independent at different states, i.e., $\mathcal{P}$ can be decomposed as $\mathcal{P} = \mathcal{P}_1\times\cdots\times\mathcal{P}_S$, where $\mathcal{P}_s\subseteq\mathbb{R}_+^{S\times A}$ for all $s=1,2,\ldots,S$.*

Assumption 2.3 (Bounded Distribution Mismatch Coefficient)**.** *Define the distribution mismatch coefficient by $D=\|d_\rho^{\pi,p}/\rho\|_\infty$. It is finite for the worst-case transition kernel $p^*(\pi):=\arg\max_{p\in\mathcal{P}} J_\rho(\pi,p)$ for all $\pi\in\Pi$, and for the policy $\pi^*(p):=\arg\min_{\pi\in\Pi} J_\rho(\pi,p)$ for all $p\in\mathcal{P}$.*

The first assumption is standard in the existing literature (Wang et al., 2023), and does not change the optimal policy set. The second assumption suggests that the ambiguity set is representable as a Cartesian product of separate ambiguity sets for the transition probability matrix associated with distinct states, and is widely used. The last assumption implies the bounded distribution mismatch coefficient, which is crucial for Theorem 3.1 in Section 3.

3. Properties of RMDPs

In this section, we discuss several key properties of RMDPs for developing our algorithm design and convergence analysis.

Lemma 3.1. *For general RMDPs, we have:*

i) $J_\rho(\pi,p)$ is L_π-Lipschitz and ℓ_π-smooth in π with $L_\pi:=\sqrt{A}/(1-\gamma)^2$ and $\ell_\pi:=2\gamma A/(1-\gamma)^3$.

ii) $\phi(\pi):=\max_{p\in\mathcal{P}} J_\rho(\pi,p)$ is L_π-Lipschitz and ℓ_π-weakly convex, i.e., $\phi(\pi)+\ell_\pi\|\pi\|^2/2$ is convex.

Furthermore, when the ambiguity set satisfies Assumption 2.2, the following statements also hold:

iii) $J_\rho(\pi,p)$ is L_p-Lipschitz and ℓ_p-smooth in p with $L_p:=\sqrt{SA}/(1-\gamma)^2$ and $\ell_p:=2\gamma S/(1-\gamma)^3$.

iv) There exist two positive constants $\ell_{\pi p},\ell_{p\pi}>0$ such that for $\forall\pi,\pi'\in\Pi$ and $\forall p,p'\in\mathcal{P}$, we have

$$\|\nabla_\pi J_\rho(\pi,p)-\nabla_\pi J_\rho(\pi',p')\|\le\ell_{\pi p}(\|\pi-\pi'\|+\|p-p'\|),$$
$$\|\nabla_p J_\rho(\pi,p)-\nabla_p J_\rho(\pi',p')\|\le\ell_{p\pi}(\|\pi-\pi'\|+\|p-p'\|).$$

The above lemma establishes that the objective function $J_\rho(\pi,p)$ is two-sided Lipschitz and smooth, and the induced function $\phi(\pi)$ is weakly-convex. Moreover, we discover a stronger Lipschitz condition satisfied by the gradient of $J_\rho(\pi,p)$ than prior results.

Remark 3.1. *While statements i)-iii) have been established in existing literature (Agarwal et al., 2021; Wang et al., 2023), the Lipschitz gradient condition iv) is stronger than the results in Wang et al. (2023), which only proves $J_\rho(\pi,\rho)$ to be partially smooth for π and p when fixing the other variable. To interpret statement iv), it indicates the bounded matrix norm of the Hessian of $J_\rho(\pi,p)$, which is critical for the convergence analysis of our single-loop algorithm.*

3.1. Gradient dominance property

In this subsection, we focus on the gradient dominance property satisfied by RMDPs.

For general RMDPs, the objective function exhibits a gradient dominance property with respect to π. Specifically, there exists a constant $\bar{D}_\pi>0$ such that for all $p\in\mathcal{P},\pi\in\Pi$:

$$J_\rho(\pi,p)-\Psi(p)\le\bar{D}_\pi\,\mathrm{dist}\big(0,\nabla_\pi J_\rho(\pi,p)+\mathcal{N}_\Pi(\pi)\big),$$

where $\Psi(p)=\min_{\pi\in\Pi} J_\rho(\pi,p)$. This is the key property to ensure global convergence of policy gradient method for non-robust MDP (Agarwal et al., 2021), and it is also important for our further analysis. Meanwhile, when the ambiguity set in RMDPs satisfies Assumption 2.2, the gradient dominance property also holds for $J_\rho(\pi,p)$ with respect to p, i.e., for all $p\in\mathcal{P},\pi\in\Pi$ with $\bar{D}_p>0$:

$$\phi(\pi)-J_\rho(\pi,p)\le\bar{D}_p\,\mathrm{dist}\big(0,\nabla_p J_\rho(\pi,p)-\mathcal{N}_\mathcal{P}(p)\big). \tag{2}$$

Although the function $\phi(\pi)$ in RMDPs may not have the desired gradient dominance property, its weak convexity, combined with the property of $J_\rho(\pi,p)$, make the Moreau envelope function $\phi_{2\ell_\pi}(\pi)$ satisfy a gradient-dominance-like property (Wang et al., 2023).

Theorem 3.1 (Informal)**.** *Let π^* be a global optimal policy for RMDPs. Then there exists a constant $C_{\ell_\pi}>0$ such that for any policy $\pi\in\Pi$:*

$$\phi_{2\ell_\pi}(\pi)\le\phi(\pi)\le C_{\ell_\pi}\|\nabla\phi_{2\ell_\pi}(\pi)\|+\phi(\pi^*), \tag{3}$$

where $\phi_{2\ell_\pi}(\pi):=\inf_{\pi'\in\Pi}\{\phi(\pi')+\ell_\pi\|\pi-\pi'\|^2\}$.

Remark 3.2. *For weakly convex optimization, it has been shown by Davis & Drusvyatskiy (2019) that iterative algorithms, such as the proximal subgradient method, implicitly optimize the smooth approximation of the weakly convex function given by the Moreau envelope. The gradient-dominance-like property implies that any first-order stationary point of the Moreau envelope function is globally optimal. Hence, it is the key prerequisite for the global optimality established in Theorem 4.2 later.*

Remark 3.3. *We remark that $\phi(\pi) \geq \phi_{2\ell_\pi}(\pi)$ holds for any $\pi \in \Pi$ and $\phi(\pi^*) = \phi_{2\ell_\pi}(\pi^*)$ (Davis & Drusvyatskiy, 2019), which means that $\phi_{2\ell_\pi}(\pi)$ is also gradient-dominated. This inspires the fine-grained characterization of region-dependent convergence rate discussed in Section 4.3.*

Additionally, a crucial and general conclusion arising from the gradient dominance property and the Moreau envelope is presented as follows.

Theorem 3.2. *Suppose that a L_f-Lipschitz continuous, ℓ_f-weakly convex function $f(\cdot)$ also satisfies the gradient dominance property with a constant $C > 0$ on set $\mathcal{X} \subseteq \operatorname{dom} f$, i.e., $f(x) - f(x^*) \leq C \operatorname{dist}\big(\partial f(x) + \mathcal{N}_{\mathcal{X}}(x)\big)$. Then we have the Moreau envelope function of f with parameter $2r$, i.e., $f_{2r}(x) = \inf_{z\in\mathcal{X}} f(z) + r\|z - x\|^2$, also satisfies the gradient dominance property, where $r > \ell_f$.*

The above theorem indicates that the Moreau envelope of a weakly convex and gradient-dominated function is also gradient-dominated. This extends the results in Yu et al. (2022), which shows that the Bregman envelope of a KŁ function is also a KŁ function. We leave more discussion in Appendix A. Theorem 3.2 plays an important role in the following convergence analysis and may be of independent interest for weakly convex optimization.

4. Single-loop Robust Policy Gradient Method

In this section, we introduce our single-loop algorithm SRPG designed for RMDPs and present its convergence analysis. The motivation and details of SRPG are provided in Section 4.1. Section 4.2 gives a standard convergence analysis of SRPG, showing a convergence rate of $\mathcal{O}(1/\varepsilon^4)$ with the gradient dominance condition. In Section 4.3, we reveal the interesting regional exponential growth condition induced by the gradient dominance property and provide a more precise analysis of the convergence rate.

4.1. The detailed algorithm

A simple single-loop algorithm designed for min-max optimization is the well-known gradient descent ascent (GDA), which alternatively performs gradient descent on the minimization problem and gradient ascent on the maximization problem. It has been proved that GDA can converge to an ε-stationary point for nonconvex-strongly-concave problem with an iteration complexity of $O(1/\varepsilon^2)$ (Lin et al., 2020). However, since the objective function of RMDPs is generally nonconvex-nonconcave, GDA may suffer from oscillation and even diverge (Zhang et al., 2020).

To address this problem, some recent works employ the Moreau-Yosida smoothing technique to make the iteration sequence stable and converge (Zhang et al., 2020; Yang et al., 2022; Zheng et al., 2023). Inspired by these works, we consider the following regularized function,

$$\chi(\pi, p, \bar{\pi}, \bar{p}) := J_\rho(\pi, p) + \frac{r_1}{2}\|\pi - \bar{\pi}\|^2 - \frac{r_2}{2}\|p - \bar{p}\|^2.$$

This function introduces two ancillary variables $\bar{\pi}$ and $\bar{p}$ for the primal variable π and dual variable p, respectively, and smooths the primal update and dual update simultaneously by incorporating two quadratic terms. It is crucial to choose appropriate r_1 and r_2 to ensure that this regularized function is convex in π and concave in p. To balance the primal and dual updates, r_1 and r_2 are not necessarily equal.

Algorithm 1 Single-loop Robust Policy Gradient Method

Input: $\pi_0 = \bar{\pi}_0 \in \Pi, p_0 = \bar{p}_0 \in \mathcal{P}$, stepsize $\tau, \sigma > 0$ and $0 < \beta, \mu < 1$
for $k = 0$ **to** $K - 1$ **do**
$\pi_{k+1} = \operatorname{Proj}_\Pi\Big(\pi_k - \tau\nabla_\pi\chi(\pi_k, p_k, \bar{\pi}_k, \bar{p}_k)\Big)$
$p_{k+1} = \operatorname{Proj}_\mathcal{P}\Big(p_k + \sigma\nabla_p\chi(\pi_{k+1}, p_k, \bar{\pi}_k, \bar{p}_k)\Big)$
$\bar{\pi}_{k+1} = \bar{\pi}_k + \beta(\pi_{k+1} - \bar{\pi}_k)$
$\bar{p}_{k+1} = \bar{p}_k + \mu(p_{k+1} - \bar{p}_k)$
end for
Output: return a policy from $\{\bar{\pi}_k\}_{k=1}^K$ uniformly

We provide the details of our SRPG in Algorithm 1. At each iteration, it performs gradient descent on π and gradient ascent on p based on the gradient of χ function. Then, the two auxiliary variables are exponentially averaged to make sure that they do not deviate far from π and p, which contributes to the sequence stability.

4.2. Convergence analysis: an overview

We provide a convergence analysis of SRPG, and a more precise analysis will be presented in Section 4.3. For clarity, we begin with a proof sketch. Firstly, we inherit the Lyapunov function introduced in Zheng et al. (2023) and derive the sufficient descent property between two consecutive iteration of SRPG in Proposition 4.1. Then, we utilize the gradient dominance property of the dual function in Proposition 4.2 to give an upper bound of the negative term presented in the descent property. Finally, we present the convergence rate of SRPG in Theorem 4.1, which further implies the global convergence result of SRPG due to the gradient dominance property with respect to π.

Before delving into the analysis, we define the stationary measure discussed in this paper as follows.

Definition 4.1. *We say point $(\pi, p) \in \Pi \times \mathcal{P}$ is a ε-game stationary point if $\operatorname{dist}(\mathbf{0}, \nabla_\pi J_\rho(\pi, p) + \mathcal{N}_\Pi(\pi)) \leq \varepsilon$ and $\operatorname{dist}(\mathbf{0}, -\nabla_p J_\rho(\pi, p) + \mathcal{N}_\mathcal{P}(p)) \leq \varepsilon$. Moreover, we say point π is a ε-optimization stationary point if $\|\operatorname{prox}_{\phi(\cdot), r_1}(\pi) - \pi\| \leq \varepsilon$ for $r_1 > \ell_\pi$ is a constant.*

The definition above is adapted from the one in Zheng et al. (2023). The ε-game stationary point is extended from the first-order stationary point in the minimization-only problem, and widely used in nonconvex-nonconcave optimization (Diakonikolas et al., 2021; Lee & Kim, 2021). The ε-optimization stationary point is well-studied for weakly convex minimization problem (Davis & Drusvyatskiy, 2019). Since we can regard our minimax problems as weakly convex minimization problem over π, we also consider this stationary measure in the following convergence analysis.

Similar to Zheng et al. (2023), we consider the Lyapunov function as follows:

$$\begin{aligned}&\Phi(\pi,p,\bar{\pi},\bar{p})\\&:=\chi(\pi,p,\bar{\pi},\bar{p})-\varphi_\pi(p,\bar{\pi},\bar{p})+\varphi_{\pi,p}(\bar{\pi},\bar{p})-\varphi_\pi(p,\bar{\pi},\bar{p})\\&\quad+\varphi_{\pi,p,\bar{p}}(\bar{\pi})-\varphi_{\pi,p}(\bar{\pi},\bar{p})+\varphi_{\pi,p,\bar{p}}(\bar{\pi})-\underline{\chi}+\underline{\chi}.\end{aligned}$$

To interpret this Lyapunov function, the primal update and dual update correspond to the primal descent term $\chi(\pi,p,\bar{\pi},\bar{p})-\varphi_\pi(p,\bar{\pi},\bar{p})$ and dual ascent term $\varphi_{\pi,p}(\bar{\pi},\bar{p})-\varphi_\pi(p,\bar{\pi},\bar{p})$. The averaging updates of auxiliary variables $\bar{\pi}$ and $\bar{p}$ can be considered as an approximate gradient descent on $\varphi_{\pi,p,\bar{p}}(\bar{\pi})$ and an approximate gradient ascent on $\varphi_{\pi,p}(\bar{\pi},\bar{p})$. With this function, we establish the following descent property of SRPG.

Proposition 4.1 (Informal)**.** *Assume that parameters τ, σ, β and μ are chosen appropriately, and let $\spadesuit := \|\pi(\bar{\pi}_{k+1},\bar{p}(\bar{\pi}_{k+1}))-\pi(\bar{\pi}_{k+1},\bar{p}_k^+(\bar{\pi}_{k+1}))\|^2$. Then, for any $k>0$, we have*

$$\begin{aligned}\Phi_k-\Phi_{k+1}&\geq\frac{r_1}{32}\|\pi_{k+1}-\pi_k\|^2+\frac{r_2}{15}\|p_k-p_k^+(\bar{\pi}_k,\bar{p}_k)\|^2\\&\quad+\frac{r_1}{5\beta}\|\bar{\pi}_k-\bar{\pi}_{k+1}\|^2+\frac{r_2}{4\mu}\|\bar{p}_k^+(\bar{\pi}_{k+1})-\bar{p}_k\|^2-4r_1\beta\cdot\spadesuit.\end{aligned}$$

Although Proposition 4.1 explicitly characterizes the difference of the Lyapunov function between two consecutive iterates, the presence of the negative term of $\spadesuit$ still prevents the subsequent analysis of the decreasing nature of Φ. In Zheng et al. (2023), they further bound the negative term of $\spadesuit$ by the assumption of the global KŁ property or the concavity of the dual function, however, which is absent in our RMDPs setting. To overcome this challenge, we provide an alternative analysis starting from the gradient dominance property of the dual function, which complements the results in Zheng et al. (2023), and may be of independent interest. We give the following lemma for preparation.

Lemma 4.1. *The Moreau envelope function $\varphi_p(\pi,\bar{\pi},\bar{p}) := \max_{p\in\mathcal{P}}\chi(\pi,\bar{\pi},p,\bar{p})$ of $J_\rho(\pi,p)$ satisfies the gradient dominance property with respect to $\bar{p}$, i.e., there exists a constant $C_{\varphi p}>0$ such that $\max_{\bar{p}}\varphi_p(\pi,\bar{\pi},\bar{p})-\varphi_p(\pi,\bar{\pi},\bar{p})\leq C_{\varphi p}\|\nabla_{\bar{p}}\varphi_p(\pi,\bar{\pi},\bar{p})\|$ for $\forall\pi\in\Pi,\forall p\in\mathcal{P},\forall\bar{\pi}$.*

Lemma 4.1 follows from Theorem 3.2 and $J_\rho(\pi,p)$ is gradient-dominated in p (see (2)). Armed with Lemma 4.1, we give the upper bound of term $\spadesuit$ in the next proposition.

Proposition 4.2 (Informal)**.** *With the above definitions, there exists a positive constant ω such that $\spadesuit \leq \omega\|\bar{p}_+^k(\bar{\pi}_{k+1})-\bar{p}_k\|$.*

Equipped with Proposition 4.1 and Proposition 4.2, we further establish the convergence result of SRPG.

Theorem 4.1. *Suppose that $\beta=\mathcal{O}(K^{-1/2})$. Then for any $K>0$, we have: 1. There exists a positive integer $k_1<K$ such that (π_{k_1+1},p_{k_1+1}) is an $\mathcal{O}\big((D_\Pi^{1/2}+D_\mathcal{P}^{1/2})/K^{1/4}\big)$-game stationary point for problem (1); 2. There exists another positive integer $k_2<K$ such that $\bar{\pi}_{k_2+1}$ is an $\mathcal{O}\big((D_\Pi^{1/2}+D_\mathcal{P}^{1/2})/K^{1/4}\big)$-optimization stationary point for problem (1). Here $D_\Pi := \max_{\pi_1\in\Pi,\pi_2\in\Pi}\|\pi_1-\pi_2\|$ and $D_\mathcal{P} := \max_{p_1\in\mathcal{P},p_2\in\mathcal{P}}\|p_1-p_2\|$ are the diameter of set Π and $\mathcal{P}$, respectively.*

Remark 4.1. *While we mainly focus on the RMDP setting, Theorem 4.1 holds for the general nonconvex-nonconcave min-max optimization problem $\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}f(x,y)$ as long as the dual function $\max_{y\in\mathcal{Y}}f(x,y)$ satisfies the gradient dominance property, that is, there exists a positive constant C such that $\max_{y'\in\mathcal{Y}}f(x,y')-f(x,y)\leq C\,\mathrm{dist}(0,-\nabla_y f(x,y)+\mathcal{N}_\mathcal{Y}(y))$ for all $y\in\mathcal{Y}$. This marks a novel finding for min-max optimization.*

Remark 4.2. *In general, the ε-game stationary point is not the same as the ε-optimization stationary point. If (π,p) is a ε-game stationary point, then it is an $\mathcal{O}(\varepsilon^{1/2})$-optimal stationary point (Zheng et al., 2023). However, Theorem 4.1 suggests that, when the last iterate π_{k+1} is only an $\mathcal{O}(\varepsilon^{1/2})$-optimization stationary point, the exponential average point $\bar{\pi}_{k+1}$ is an ε-optimization stationary point.*

We have established so far the convergence result of SRPG. In the rest of the subsection, we strengthen the result and further derive the global convergence result of SRPG. Before proceeding, we provide the definitions of ε-global optimal solution and (ε,δ)-global optimal solution as follows.

Definition 4.2. *A solution π is an ε-global optimal solution for problem (1) if $\phi(\pi)-\phi(\pi^*)\leq\varepsilon$. Moreover, a solution π is an (ε,δ)-global optimal solution if $\mathrm{dist}(\pi,\Pi_{\ell_\pi}^*(\varepsilon))\leq\delta$, where $\Pi_{\ell_\pi}^*(\varepsilon) := \{\pi\mid\phi_{2\ell_\pi}(\pi)-\phi(\pi^*)<\varepsilon\}$.*

Remark 4.3. *The (ε,δ)-global optimal solution is crucial for the analysis in Section 4.3, and we define it here for compactness. We emphasize that an $(\varepsilon,0)$-global optimal solution is not an ε-global optimal solution, since there exists a gap between $\phi_{2\ell_\pi}(\cdot)$ and $\phi(\cdot)$.*

With this preparation, we state our main global convergence result of SRPG in the next theorem.

Theorem 4.2. *Under the same setting of Theorem 4.1, for the sequence $\{\bar{\pi}_k\}_{k=1}^K$ generated by SRPG, there exists a*

positive integer $k < K$ such that $\bar{\pi}_{k+1}$ is a $\mathcal{O}\big((D_\Pi^{1/2} + D_\mathcal{P}^{1/2})/K^{1/4}\big)$-global optimal solution.

4.3. Convergence analysis: a deeper dive

We provide a more precise analysis of the convergence rate of SRPG. More specifically, we show that, when the algorithm proceeds and the sequence $\{\bar{\pi}_k\}$ approaches to the optimal solution, the convergence rate of SRPG gradually slows down and degenerates to $\mathcal{O}(1/\varepsilon^4)$ stated in Theorem 4.2. Although this observation does not enhance the iteration complexity with respect to ε, it offers a noteworthy explanation for the phenomena observed later in our numerical experiments in Section 5.

To begin the analysis, we introduce the regional exponential growth property as follows.

Lemma 4.2 (Regional Exponential Growth)**.** *Under the same setting of Theorem 3.1, for ℓ_π-weakly convex function $\phi(\pi)$, we have*

$$\phi(\pi) - \phi^* \geq \zeta \exp\big(\operatorname{dist}(\pi, \Pi^*_{\ell_\pi}(\zeta))/C_{\ell_\pi}\big), \forall \pi \in \Pi \backslash \Pi^*_{\ell_\pi}(\zeta),$$

where $\zeta > 0$ is any given constant.

Lemma 4.2 suggests that within region $\Pi \setminus \Pi^*_{\ell_\pi}(\zeta)$, the function value of $\phi(\cdot)$ exhibits the exponential growth property, in contrast to the general quadratic growth property guaranteed by the PŁ condition (Necoara et al., 2019). We highlight that this region is not the optimal solution set to $\min_{\pi\in\Pi} \phi(\pi)$, but is instead associated with the Moreau envelope function $\phi_{2\ell_\pi}(\cdot)$ and a pre-determined constant ζ. In fact, we can prove that $\Pi^*(\zeta) \subseteq \Pi^*_{\ell_\pi}(\zeta)$, where $\Pi^*(\zeta) := \{\pi \mid \phi(\pi) - \phi^* \leq \zeta\}$. After taking $\zeta = \varepsilon$, we obtain the following proposition:

Proposition 4.3. *Under the same setting of Theorem 3.1, for the sequence $\{\bar{\pi}_k\}$ generated by SRPG and $\bar{\pi}_k \in \Pi \setminus \Pi^*_{\ell_\pi}(\varepsilon)$, we have $\operatorname{dist}(\bar{\pi}_k, \Pi^*_{\ell_\pi}(\varepsilon)) = \mathcal{O}\big(\log\big((D_\Pi^{1/2} + D_\mathcal{P}^{1/2})k^{-0.25}\varepsilon^{-1}\big)\big)$.*

Remark 4.4. *Proposition 4.3 implies the following interesting observation that the iteration complexity to obtain any (ε, δ)-optimal solution can be bounded by $\mathcal{O}\big((D_\Pi^2 + D_\mathcal{P}^2)\varepsilon^{-4}\exp(-4\delta)\big)$. When δ approaches 0, the rate degenerates to $\mathcal{O}(1/\varepsilon^4)$, which is the same as we proved in Theorem 4.2. For a comprehensive illustration, one may refer to Figure 1. The iteration sequence within the white area indicates that $\delta > 0$. A larger δ corresponds to a faster convergence performance. Upon entering the small blue region, Lemma 4.2 and Proposition 4.3 do not hold anymore, leading to a worse convergence rate $\mathcal{O}(1/\varepsilon^4)$. Ultimately, the sequence reaches the orange dashed region, signifying the attainment of an ε-optimal solution.*

Remark 4.5. *An average convergence rate for $(\varepsilon, 0)$ global optimal solution can be estimated by*

$$\frac{(D_\Pi^2 + D_\mathcal{P}^2)}{\tilde{D}_\Pi \cdot \varepsilon^4} \int_0^{\tilde{D}_\Pi} \exp(-4\delta) d\delta \leq \frac{D_\Pi^2 + D_\mathcal{P}^2}{4\tilde{D}_\Pi \cdot \varepsilon^4},$$

where $\tilde{D}_\Pi$ is the distance between initial point π_0 and set $\Pi^*_{\ell_\pi}(\varepsilon)$. This finding indicates that an unfavorable initial point (where $\tilde{D}_\Pi$ is large) may not significantly impact the algorithm's performance.

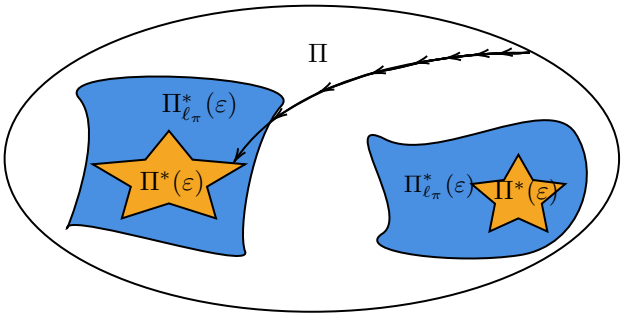

Figure 1. Illustration of the changing convergence rate: a denser arrow implies a faster convergence rate.

5. Numerical Experiments

We conduct several experiments to investigate the performance of SRPG compared with DRPG (Wang et al., 2023). In particular, we consider two different problems, including GARNET MDPs and an inventory management problem. To measure the performance of SRPG and DRPG, we use the worst-case expected return for a given π_k, i.e., $\phi(\pi_k)$.

We apply the projected gradient method (PGM) to solve the inner maximization problem in DRPG. Following the setup of Wang et al. (2023), we terminate PGM when the relative error between two iterations is smaller than 10^{-4} or the iteration number reaches 200. Note that PGM is also used to evaluate $\phi(\pi_k)$ for SRPG and each projection operator needed in the two algorithms. To solve the quadratic subproblems in PGM, we resort to the state-of-the-art commercial solver GUROBI (Gurobi Optimization, LLC, 2023). We provide the code in this link.

5.1. GARNET MDPs

In the first experiment, we consider the GARNET MDPs, which are introduced by Archibald et al. (1995) and widely recognized as a key benchmark in RMDPs (Wang et al., 2023; Li et al., 2024). Specifically, a standard GARNET MDP, denoted by GARNET(S, A, b), consists of three different parameters. Here, S and A indicate the finite numbers of states and actions, respectively. The branching factor b determines the number of states that are reachable from any given state-action pair in one transition.

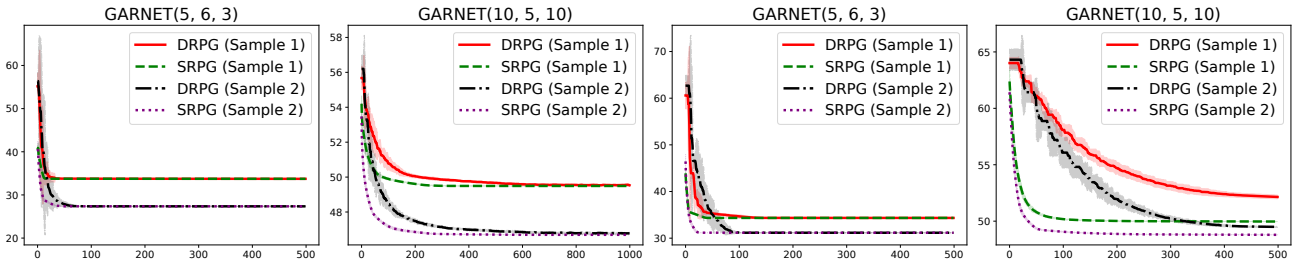

Figure 2. The global convergence behavior of SRPG and DRPG on the GARNET MDPs. The solid curves depict the average value of $\phi(\pi_k)$, represented by y-axis, while the shaded areas correspond to its 95% confidence interval. The x-axis represents the total number of updates on π and p. The left two figures are the results for s-rectangular ambiguity set, and the right two figures are the results for (s,a)-rectangular ambiguity set. We remark that each sample in each figure corresponds to a different task.

Problem setup. We randomly generate the nominal transition kernel $\bar{p}$ according to two different GARNET MDPs: GARNET$(5,6,3)$ and GARNET$(10,5,10)$. We let the discount factor $\gamma = 0.95$, and sample the cost $c_{sas'}$ i.i.d. from the uniform distribution supported on $[0,5]$. Each element in the initial state distribution $\rho \in \Delta^S$ is sampled from $[0,5]$ uniformly, then we project ρ into the probability space. For each $\bar{p}$, we construct both L_1-normed s- and (s,a)-rectangular ambiguity sets. The s-rectangular ambiguity set is defined by $\mathcal{P} = \times_{s\in[S]}\mathcal{P}_s$, where $\mathcal{P}_s$ is

$$\mathcal{P}_s := \Big\{(p_{s1},\ldots,p_{sA}) \in (\Delta^S)^A \mid \sum_{a\in\mathcal{A}} \|p_{sa}-\bar{p}_{sa}\|_1 \le \kappa_s\Big\}.$$

Similarly, the (s,a)-rectangular ambiguity set is formulated as $\mathcal{P} = \times_{s\in[S],a\in[A]}\mathcal{P}_{sa}$, with each $\mathcal{P}_{sa}$ being

$$\mathcal{P}_{sa} := \big\{p_{sa} \in \Delta^A \mid \|p_{sa}-\bar{p}_{sa}\|_1 \le \kappa_{sa}\big\}.$$

Here, κ_s and κ_{sa} control the uncertainty included in the ambiguity sets, and are sampled uniformly from $[0.1, 0.5]$. For each generated $\bar{p}$ and each ambiguity set, we conduct two independent experiments, which brings 8 different tasks.

Experiment configuration. Throughout all the tasks, we tune SRPG as follows. We choose the primal stepsize τ and dual stepsize σ from $\{0.01, 0.05, 0.1\}$. We also choose the extrapolation parameters β and μ from $\{0.01, 0.05, 0.1, 0.2, 0.4\}$ for SRPG. For DRPG, we also tune its primal and dual stepsize from $\{0.01, 0.05, 0.1\}$. To facilitate a fair comparison between the two algorithms, we run each algorithm with an identical total number of updates on π and p, and choose the number from $\{500, 1000\}$. For each task, we conduct 10 random experiments with different initial points, and record the average values of $\phi(\pi_k)$.

Comparison. Figure 2 shows the global convergence behavior of two algorithms by plotting the average value of the sequence $\{\phi(\pi_k)\}$ against the total number of updates on π and p. It demonstrates that our SRPG converges consistently and significantly faster than DRPG over all 8 tasks, which indicates the superior performance and robustness of SRPG. Generally speaking, both algorithms exhibit a significant decrease in the value of $\phi(\pi_k)$ within the initial several iterations. However, as the algorithms proceed, the convergence rate slows down notably. This validates the regional exponential growth property discussed in Section 4.3, which states that when the iterate is near the optimal solution, the convergence rate will gradually become $\mathcal{O}(1/\varepsilon^4)$. Furthermore, DRPG displays an oscillation pattern, especially in the last figure. The underlying reason of this phenomenon could be attributed to the policy π not being updated within the inner loop of DRPG, resulting in significant changes to the dual variable during the update of the primal variable π. This also highlights the advantage of our single-loop algorithm. The convergence behavior of SRPG is smoother, implying a steady and consistent improvement without significant fluctuations.

5.2. Inventory management problem

Our second experiment considers the inventory management problem, in which a retailer engages in the ordering, storage, and sale of a single product over an infinite time horizon (Porteus, 2002; Ho et al., 2018). It can be formulated as an RMDP as follows. The inventory levels and order quantities correspond to the states and actions of the RMDP in any given time step, respectively. The distribution of demand, which is unknown to the retailer, gives the transition kernel. Moreover, an item held in inventory will incur a deterministic per time step holding cost. We aim to find a policy that minimizes the worst-case total cost.

Because the numerical results in the last subsection have already demonstrated the superior performance of SRPG over DRPG in the tabular setting, we now go beyond the assumption of s-rectangular ambiguity set and test the performance of two algorithms in the parameterization setting. This is more suitable and practical for the large-scale RMDP. Specifically, we first test with the parameterization of transition kernel, which is also employed by Wang et al. (2023),

and then test with parameterization of both policy and transition kernel. The detailed parameterization methods are provided in Appendix D.

Experiment configuration. Similar to the experiments on GARNET MDPs, we tune SRPG by choosing the primal stepsize τ and dual stepsize σ from $\{0.01, 0.05, 0.1\}$ and selecting the extrapolation parameters β and μ from $\{0.1, 0.2, 0.3\}$. For DRPG, we also tune its primal and dual stepsize from $\{0.01, 0.05, 0.1\}$. When testing with transition kernel parameterization, we run the algorithms with an identical total number of updates on both primal and dual variables and conduct ten random experiments with different initial points. We use the same performance measure as in the last subsection. When considering parameterization on both policy and transition kernel, we run DRPG on each instance with 15,000 total updates, recording the time and the obtained objective value. Then, we run SRPG and record the time required to obtain the same objective value as DRPG.

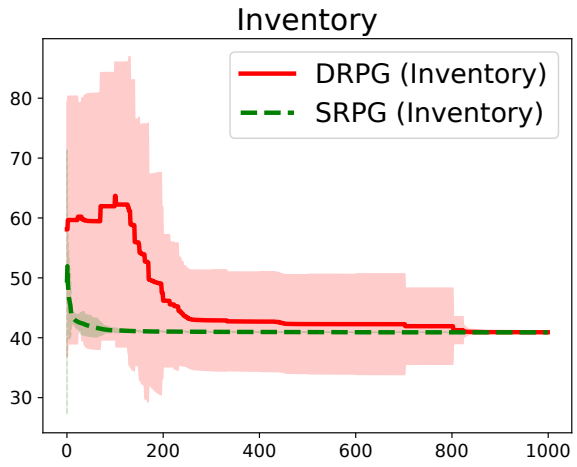

Figure 3. The global convergence behavior SRPG and DRPG on the inventory management problem. The figure is plotted in the same way as Figure 2.

Discussion on computational cost for projection operator. Let the computational cost for computing the projection operator of primal variable π be M_π, and the one of dual variable p be M_p. Then, for SRPG, the total computational cost for the projection operator is $\mathcal{O}(\epsilon^{-4}(M_\pi + M_p))$. For DRPG, it is a double-loop algorithm and needs to compute the projection operator $\mathcal{O}(\epsilon_k^{-2})$ for the k-th outer iteration. It at least needs the computational cost of $\mathcal{O}(\epsilon^{-4}(M_\pi + M_p\epsilon^{-2}))$ when we take $\epsilon_k = \epsilon$ for all k. Therefore, DRPG is much more computationally expensive than ours in terms of projection operator, as well as gradient computation.

Comparison. When only parameterizing the transition kernel, Figure 3 shows the average values of sequence $\{\phi(\pi_k)\}$ for the two algorithms. It can be observed that SRPG demonstrates superior robustness compared to DRPG, and converges faster. Despite minor fluctuations in the initial phase, SRPG quickly stabilizes and converges. In contrast, DRPG oscillates and faces formidable challenges in converging to the optimal policy. When parameterizing both policy and transition kernel, Table 2 demonstrates that SRPG significantly decreases the time needed to find a solution when compared to DRPG, and thus more efficient. This is also consistent with the theoretical analysis.

S	A	DRPG	SRPG
200	20	9,377.126	**1,517.999**
300	20	12,758.126	**4,719.497**
500	20	35,108.616	**3,737.307**
200	30	7,605.747	**1,764.330**
300	30	16,725.184	**5,852.938**
200	50	12,910.183	**5,351.032**
300	50	32,342.525	**2,114.342**

Table 2. Comparison of DRPG and SRPG on both parameterization on policy and transition kernel. The smaller figure indicates less runtime and is bolded.

6. Conclusion

This paper introduces the first single-loop robust policy gradient method (SRPG) for RMDPs. We provide its convergence analysis, demonstrating its ability to converge to the globally optimal policy with complexity $\mathcal{O}(1/\varepsilon^4)$. Additionally, we unveil the regional exponential growth property and leverage it for a more precise convergence analysis of SRPG. Our numerical experiments showcase that SRPG exhibits faster and more stable convergence behavior when compared to its double-loop counterpart. For future work, a promising direction is to incorporate the mirror descent method into our sing-loop framework and attempt to obtain a better convergence rate for solving RMDPs. It is also interesting to capture the noise gradient and function approximation in the analysis, which is more suitable for practical applications.

Acknowledgement

The authors are grateful to the Area Chairs and the anonymous reviewers for their constructive comments. This research is partially supported by the Major Program of National Natural Science Foundation of China (Grant 72394360, 72394364).

Impact Statement

This paper presents work whose goal is to advance the field of Machine learning. There are many potential societal

consequences of our work, but none of which we feel must be specifically highlighted here.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 22(1):4431–4506, 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.

Bailey, J. P., Gidel, G., and Piliouras, G. Finite regret and cycles with fixed step-size via alternating gradient descent-ascent. In *Conference on Learning Theory*, pp. 391–407. PMLR, 2020.

Ben-Tal, A. and Nemirovski, A. Robust optimization–methodology and applications. *Mathematical programming*, 92:453–480, 2002.

Bertsimas, D. and Sim, M. The price of robustness. *Operations research*, 52(1):35–53, 2004.

Bolte, J., Nguyen, T. P., Peypouquet, J., and Suter, B. W. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165:471–507, 2017.

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

den Boer, A. V. and Zwart, B. Dynamic pricing and learning with finite inventories. *Operations research*, 63(4):965–978, 2015.

Deng, Y., Bao, F., Kong, Y., Ren, Z., and Dai, Q. Deep direct reinforcement learning for financial signal representation and trading. *IEEE transactions on neural networks and learning systems*, 28(3):653–664, 2016.

Diakonikolas, J., Daskalakis, C., and Jordan, M. I. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2746–2754. PMLR, 2021.

Duchi, J. and Namkoong, H. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019.

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

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv :1412.6572*, 2014.

Goyal, V. and Grand-Clément, J. Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 48(1):203–226, 2023.

Grand-Clément, J. and Kroer, C. Scalable first-order methods for robust mdps. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):12086–12094, 2021.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2023. URL https://www.gurobi.com.

Hajizadeh, S., Lu, H., and Grimmer, B. On the linear convergence of extragradient methods for nonconvex–nonconcave minimax problems. *INFORMS Journal on Optimization*, 2023.

Hamedani, E. Y. and Aybat, N. S. A primal-dual algorithm with line search for general convex-concave saddle point problems. *SIAM Journal on Optimization*, 31(2):1299–1329, 2021.

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.

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

Jiang, Z., Xu, D., and Liang, J. A deep reinforcement learning framework for the financial portfolio management problem. *arXiv :1706.10059*, 2017.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274, 2002.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pp. 795–811. Springer, 2016.

Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., and Pérez, P. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23 (6):4909–4926, 2021.

Kumar, N., Derman, E., Geist, M., Levy, K., and Mannor, S. Policy gradient for s-rectangular robust markov decision processes. *arXiv :2301.13589*, 2023a.

Kumar, N., Usmanova, I., Levy, K. Y., and Mannor, S. Towards faster global convergence of robust policy gradient methods. In *Sixteenth European Workshop on Reinforcement Learning*, 2023b.

Lee, S. and Kim, D. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems*, 34: 22588–22600, 2021.

Li, M., Kuhn, D., and Sutter, T. Policy gradient algorithms for robust mdps with non-rectangular uncertainty sets, 2024.

Li, Y. and Lan, G. First-order policy optimization for robust policy evaluation. *arXiv :2307.15890*, 2023.

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

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.

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

Necoara, I., Nesterov, Y., and Glineur, F. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175:69–107, 2019.

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

Ouyang, Y. and Xu, Y. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1-2):1–35, 2021.

Pang, J.-S. A posteriori error bounds for the linearly-constrained variational inequality problem. *Mathematics of Operations Research*, 12(3):474–484, 1987.

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

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.

Sallab, A. E., Abdou, M., Perot, E., and Yogamani, S. Deep reinforcement learning framework for autonomous driving. *arXiv :1704.02532*, 2017.

Schulman, J., Chen, X., and Abbeel, P. Equivalence between policy gradients and soft q-learning. *arXiv :1704.06440*, 2017.

Sion, M. On general minimax theorems. *Pacific Journal of Mathematics*, 1958.

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

Wang, K., Gadot, U., Kumar, N., Levy, K., and Mannor, S. Bring your own (non-robust) algorithm to solve robust mdps by estimating the worst kernel, 2024.

Wang, Q., Ho, C. P., and Petrik, M. Policy gradient in robust mdps with global convergence guarantee. In *International Conference on Machine Learning*, pp. 35763–35797. PMLR, 2023.

Wang, Y. and Zou, S. Policy gradient method for robust reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 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 *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, pp. 3606–3613. IEEE, 2009.

Xu, Z., Zhang, H., Xu, Y., and Lan, G. A unified single-loop alternating gradient projection algorithm for nonconvex–concave and convex–nonconcave minimax problems. *Mathematical Programming*, pp. 1–72, 2023.

Yang, J., Orvieto, A., Lucchi, A., and He, N. Faster single-loop algorithms for minimax optimization without strong concavity. In *International Conference on Artificial Intelligence and Statistics*, pp. 5485–5517. PMLR, 2022.

Yu, P., Li, G., and Pong, T. K. Kurdyka–łojasiewicz exponent via inf-projection. *Foundations of Computational Mathematics*, 22(4):1171–1217, 2022.

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.

Zhang, Y. and Xiao, L. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *Journal of Machine Learning Research*, 18(84):1–42, 2017.

Zheng, T., Zhu, L., So, A. M.-C., Blanchet, J., and Li, J. Universal gradient descent ascent method for nonconvex-nonconcave minimax optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Appendix

Table of Contents

Structure of the Appendix

The appendix is organized as follows. Appendix A introduces some definitions and notations essential for understanding the proof process. It also includes more discussion about the difference between gradient dominance and KŁ property. Appendix B presents some crucial results that underpin our convergence analysis. Detailed proofs for convergence results are thoroughly outlined in Appendix C, and we provide an illustration in Figure 4. Furthermore, Appendix D offers more extensive details on our experiments.

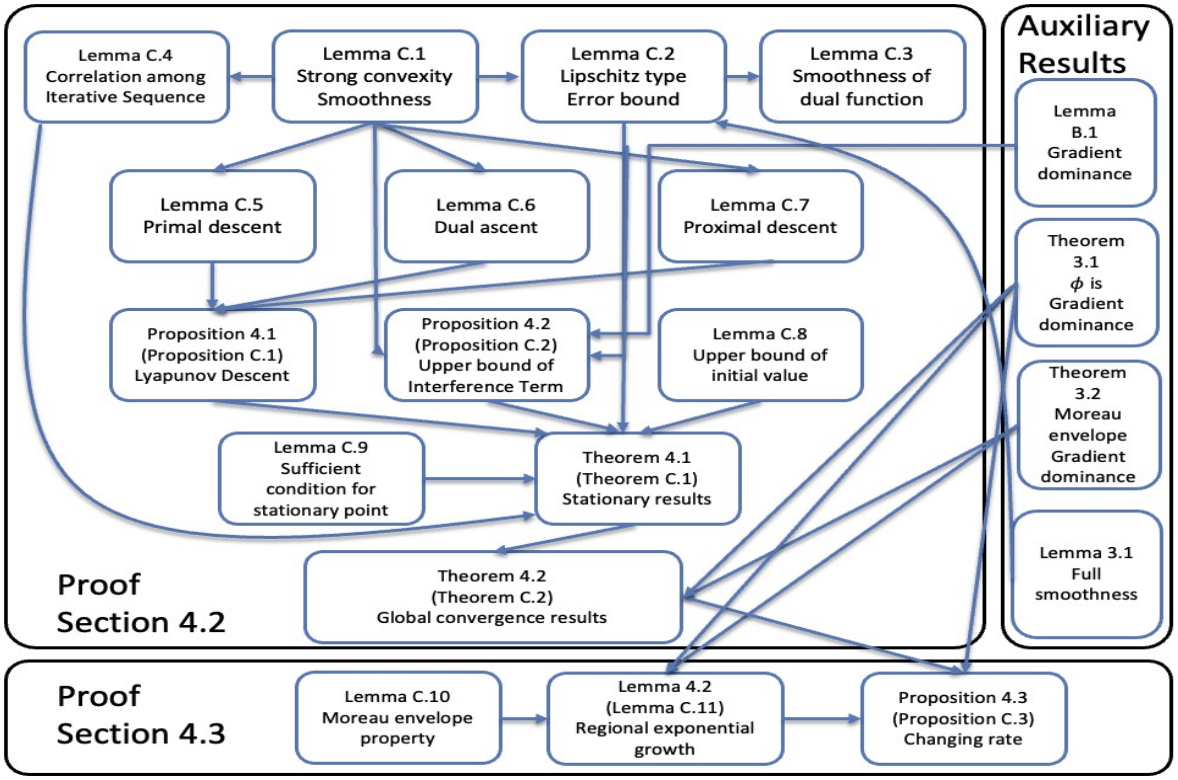

Figure 4. Flowchart of proof process.

A. Preliminaries and Notations

We first give some definitions used in this paper, and list some useful notations in Table 3.

Definition A.1 (L-Lipschitz Function)**.** *A function f is called L-Lipschitz with respect to norm $\|\cdot\|$ if we have*

$$|f(x) - f(y)| \leq L\|x - y\| \quad \forall x, y \in \mathbb{R}^n.$$

Definition A.2 (L-smooth Function)**.** *A function f is called L-smooth if it is continuously differentiable and its gradient is Lipschitz continuous with Lipschitz constant L,*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y \in \mathbb{R}^n.$$

Definition A.3 (Weakly Convex Function)**.** *A function f is called weakly convex if there exists $\rho \geq 0$ such that $f(x) + \frac{\rho}{2}\|x\|^2$ is a convex function.*

Definition A.4 (Normal Cone)**.** *Let $C \subset \mathbb{R}^n$ be a convex set with $\bar{x} \in C$. The normal cone to C at $\bar{x}$ is*

$$\mathcal{N}_C(\bar{x}) := \{v \in \mathbb{R}^n \mid \langle v, x - \bar{x}\rangle \leq 0, \forall x \in C\}.$$

Definition A.5 (Proximal Operator)**.** *Let $r > 0$, then the proximal operator of a function f is given by*

$$prox_{f,r}(x) = \arg\min_y \left\{ f(y) + \frac{r\|x - y\|^2}{2} \right\}.$$

Definition A.6 (Moreau envelope)**.** *The Moreau envelope of a function f is given by*

$$f_\mu(x) = \inf_y \left\{ f(y) + \frac{\mu}{2}\|x - y\|_2^2 \right\}.$$

Definition A.7 (Bergman Divergence)**.** *Let $\psi : \Omega \to \mathbb{R}$ be a strictly convex and continuously differentiable function defined on a closed convex set Ω. Then the Bregman divergence is defined as*

$$\mathcal{D}(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y\rangle, \quad \forall x, y \in \Omega.$$

Definition A.8 (Bregman Envelope)**.** *The Bregman envelope of a proper lower semi-continuous convex function f is defined as*

$$f_{\mathcal{D},\lambda}(\bar{x}) := \inf_{x \in X}\{f(x) + \lambda\mathcal{D}(x, \bar{x})\}.$$

where $\mathcal{D}(x, \bar{x})$ is the Bergman divergence.

Difference between gradient dominance property and KŁ property

We now discuss the difference between gradient dominance property and KŁ property. We recall the definition of KŁ property.

Definition A.9 (KŁ property (Yu et al., 2022))**.** *We say that a proper closed function $h : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ satisfies the KŁ property at $\hat{x} \in \operatorname{dom}\partial h$ if there are $c \in (0, \infty]$, a neighborhood V of $\hat{x}$ and a continuous concave function $\Omega : [0, c) \to [0, \infty)$ with $\Omega(0) = 0$ such that*

- *Ω is continuously differentiable on $(0, c)$ with $\Omega' > 0$ on $(0, c)$;*
- *For any $x \in V$ with $h(\hat{x}) < h(x) < h(\hat{x}) + c$, it holds that*

$$\Omega'(h(x) - h(\hat{x}))\operatorname{dist}(0, \partial h(x)) \geq 1. \tag{4}$$

When we take $\Omega(z) = \bar{c}z^{1-\theta}$ for some $\bar{c} > 0$ and $\theta \in [0, 1)$. Then (4) can be written as

$$\bar{c}(1 - \theta)(h(x) - h(\hat{x}))^{-\theta}\operatorname{dist}(0, \partial h(x)) \geq 1.$$

One may view gradient dominance property as a special case of KŁ property with $\theta = 1$. However, this will make the first condition and (4) invalid. Hence, it is challenging to make KŁ property reduce to the gradient dominance property. This also implies that some interesting properties induced by KŁ property fail in the context of gradient dominance property. Whereas in this paper, we combine the gradient dominance property with the weakly convex function, and derive that the Moreau envelope of a weakly convex and gradient-dominated function is also gradient-dominated.

Notation	Meaning	Remark
$\chi(\pi,p,\bar{\pi},\bar{p})$	$J_\rho(\pi,p)+\frac{r_1}{2}\|\pi-\bar{\pi}\|^2-\frac{r_2}{2}\|p-\bar{p}\|^2$	-
$\varphi_\pi(p,\bar{\pi},\bar{p})$	$\min_{\pi\in\Pi}\chi(\pi,p,\bar{\pi},\bar{p})$	$\pi(p,\bar{\pi},\bar{p})\in\operatorname{argmin}_{\pi\in\Pi}\chi(\pi,p,\bar{\pi},\bar{p})$
$\varphi_p(\pi,\bar{\pi},\bar{p})$	$\max_{p\in\mathcal{P}}\chi(\pi,p,\bar{\pi},\bar{p})$	$p(\pi,\bar{\pi},\bar{p})\in\operatorname{argmax}_{p\in\mathcal{P}}\chi(\pi,p,\bar{\pi},\bar{p})$
$\varphi_{\pi,p}(\bar{\pi},\bar{p})$	$\min_{\pi\in\Pi}\max_{p\in\mathcal{P}}\chi(\pi,p,\bar{\pi},\bar{p})$	-
$\varphi_{\pi,p}(\bar{\pi},\bar{p})$	$\min_{\pi\in\Pi}\varphi_p(\pi,\bar{\pi},\bar{p})$	$\pi(\bar{\pi},\bar{p})=\pi\big(p(\bar{\pi},\bar{p}),\bar{\pi},\bar{p}\big)$
$\varphi_{\pi,p}(\bar{\pi},\bar{p})$	$\max_{p\in\mathcal{P}}\varphi_\pi(p,\bar{\pi},\bar{p})$	$p(\bar{\pi},\bar{p})=p\big(\pi(\bar{\pi},\bar{p}),\bar{\pi},\bar{p}\big)$
$\varphi_{\pi,p,\bar{\pi}}(\bar{p})$	$\min_{\bar{\pi}}\varphi_{\pi,p}(\bar{\pi},\bar{p})$	$\bar{\pi}(\bar{p})\in\operatorname{argmin}_{\bar{\pi}\in\mathbb{R}^{\|\mathcal{S}\|\times\|\mathcal{A}\|}}\varphi_{\pi,p}(\bar{\pi},\bar{p})$
$\varphi_{\pi,p,\bar{p}}(\bar{\pi})$	$\max_{\bar{p}}\varphi_{\pi,p}(\bar{\pi},\bar{p})$	$\bar{p}(\bar{\pi})\in\operatorname{argmax}_{\bar{p}}\varphi_{\pi,p}(\bar{\pi},\bar{p})$
$p^+(\bar{\pi},\bar{p})$	$\operatorname{Proj}_{\mathcal{P}}(p+\sigma\nabla_p\chi(\pi(p,\bar{\pi},\bar{p}),p,\bar{\pi},\bar{p}))$	-
$\bar{p}^+(\bar{\pi})$	$\bar{p}+\mu\big(p(\pi(\bar{\pi},\bar{p}),\bar{\pi},\bar{p})-\bar{p}\big)$	-
$\psi_p(\pi;\bar{\pi})$	$\max_{p\in\mathcal{P}}J_\rho(\pi,p)+\frac{r_1}{2}\|\pi-\bar{\pi}\|^2$	-
$\psi(\pi,p,\bar{\pi})$	$J_\rho(\pi,p)+\frac{r_1}{2}\|\pi-\bar{\pi}\|^2$	$\pi(p,\bar{\pi};r_1)\in\operatorname{argmin}_{\pi\in\Pi}\psi(\pi,p,\bar{\pi})$
		$\pi^*(\bar{\pi})\in\operatorname{argmin}_{\pi\in\Pi}\psi_p(\pi;\bar{\pi})$
$p^+(\bar{p})$	$\operatorname{Proj}_{\mathcal{P}}(p+\sigma\nabla_pJ_\rho(\pi(p,\bar{\pi};r_1),p))$	-

Table 3. Notations.

B. Auxillary Results and Proof of Results in Section 3

Lemma B.1 (Lemma 4.3 and E.2 in Wang et al. (2023))**.** *We have two excellent properties:*

1. For any fixed $p\in\mathcal{P}$, $J_\rho(\pi,p)$ satisfies the following condition:

$$J_\rho(\pi,p)-J_\rho(\pi^*(p),p)\le\bar{D}_\gamma\max_{\bar{\pi}\in\Pi}\langle\pi-\bar{\pi},\nabla_\pi J_\rho(\pi,p)\rangle\le\bar{D}_\gamma\max_{\bar{\pi}\in\Pi}\langle\pi-\bar{\pi},\nabla_\pi J_\rho(\pi,p)+\varkappa\rangle,\forall\varkappa\in\mathcal{N}_\Pi(\pi),\quad(5)$$

   where $J_\rho(\pi^*(p),p):=\min_{\pi\in\Pi}J_\rho(\pi,p)$ and $\mathcal{N}_\Pi(\pi):=\{\varkappa\mid\langle\varkappa,\bar{\pi}-\pi\rangle\le0,\forall\bar{\pi}\in\Pi\}$.

2. Assume the ambiguity set is s rectangular. Then for any $p\in\mathcal{P}$, we have

$$J_\rho(\pi,p^*(\pi))-J_\rho(\pi,p)\le\bar{D}_\gamma\max_{\bar{p}\in\mathcal{P}}\langle\bar{p}-p,\nabla_pJ_\rho(\pi,p)\rangle\le\bar{D}_\gamma\max_{\bar{p}\in\mathcal{P}}\langle\bar{p}-p,\nabla_pJ_\rho(\pi,p)-\varkappa\rangle,\forall\varkappa\in\mathcal{N}_\mathcal{P}(p),\quad(6)$$

   where $J_\rho(\pi,p^*(\pi)):=\max_{p\in\mathcal{P}}J_\rho(\pi,p)$, $\bar{D}_\gamma=D(1-\gamma)^{-1}$ and $\mathcal{N}_\mathcal{P}(p):=\{\varkappa\mid\langle\varkappa,\bar{p}-p\rangle\le0,\forall\bar{p}\in\mathcal{P}\}$.

Remark B.1. *The above Lemma B.1 implies the following two inequalities*

$$J_\rho(\pi,p^*(\pi))-J_\rho(\pi,p)\le\bar{D}_p\operatorname{dist}\big(\nabla_pJ_\rho(\pi,p)-\mathcal{N}_\mathcal{P}(p)\big)\quad(7)$$
$$J_\rho(\pi,p)-J_\rho(\pi^*(p),p)\le\bar{D}_\pi\operatorname{dist}\big(\nabla_\pi J_\rho(\pi,p)+\mathcal{N}_\Pi(\pi)\big)\quad(8)$$

Proof of Lemma 3.1

Now, we prepare to give the proof of Lemma 3.1.

Proof. Item i), Item iii) and Item ii) can be found in Lemmas 3.1, 4.1, 4.2 in Wang et al. (2023). Now, we give a proof of Item iv). Define

$$p^\pi_{ss'}(1)=\sum_{a\in\mathcal{A}}\pi_{sa}p_{sas'},\ \ p^\pi_{ss'}(t-1)p^\pi_{s's''}(1)=p^\pi_{ss''}(t),\ \ q^{\pi,p}_{sa}=\sum_{s'\in\mathcal{S}}p_{sas'}c_{sas'}+\gamma p_{sas'}(v^{\pi,p}_{s'})$$

Then we first give some middle results

$$\frac{\partial p^{\pi}_{ss'}(1)}{\partial p_{sas'}} = \pi_{sa},\ \frac{\partial p^{\pi}_{ss'}(1)}{\partial p_{s''as'}} \mid_{s\neq s''} = 0,\ \ \frac{\partial q^{\pi,p}_{sa}}{\partial p_{\hat{s}\hat{a}\hat{s}'}} \mid_{(s,a)\neq(\hat{s},\hat{a})} = \gamma \sum_{s'\in\mathcal{S}} p_{sas'} \frac{\partial v^{\pi,p}_{s'}}{\partial p_{\hat{s}\hat{a}\hat{s}'}},$$
$$\frac{\partial q^{\pi,p}_{sa}}{\partial p_{\hat{s}\hat{a}\hat{s}'}} \mid_{(s,a)=(\hat{s},\hat{a})} = c_{sa\hat{s}'} + \gamma \sum_{s'\in\mathcal{S}} \frac{\partial p_{sas'}}{\partial p_{\hat{s}\hat{a}\hat{s}'}} v^{\pi,p}_{s'} + \gamma \sum_{s'\in\mathcal{S}} p_{sas'} \frac{\partial v^{\pi,p}_{s'}}{\partial p_{\hat{s}\hat{a}\hat{s}'}} = c_{sa\hat{s}'} + \gamma v^{\pi,p}_{\hat{s}'} + \gamma \sum_{s'\in\mathcal{S}} p_{sas'} \frac{\partial v^{\pi,p}_{s'}}{\partial p_{\hat{s}\hat{a}\hat{s}'}},$$

which will be used in the following process, repeatly. Since $J_\rho(\pi,p)$ is ℓ_π,ℓ_p-smooth w.r.t. π and p, respectively. Hence, if all there exist constant C and U such that $|\partial v^{\pi,p}_{s_1}/(\partial \pi_{s_2a_1}\partial p_{s_3a_2s})| \le C < \infty$ and $\left\| \frac{\partial J_\rho(\pi,p)}{\partial \pi \partial p} \right\| \le U < \infty$ hold for any $s_1, s_2, s_3, a_2, a_3, s$ then we complete our proof of Item iv). Now, we give the details of how to prove the existence of U.

The first order derivatives are shown as follows:

$$\begin{aligned}
\Big[\frac{\partial v^{\pi,p}_{\hat{s}}}{\partial \pi_{sa}}\Big]_{\hat{s}\neq s} &= \gamma \sum_{s'\in\mathcal{S}} p^{\pi}_{\hat{s}s'}(1) \frac{\partial v^{\pi,p}_{s'}}{\partial \pi_{sa}} \\
\Big[\frac{\partial v^{\pi,p}_{\hat{s}}}{\partial \pi_{sa}}\Big]_{\hat{s}=s} &= q^{\pi,p}_{sa} + \gamma \sum_{s'\in\mathcal{S}} p^{\pi,p}_{ss'}(1) (\frac{\partial v^{\pi,p}_{s'}}{\partial \pi_{sa}}) \\
\Big[\frac{\partial v^{\pi,p}_{\hat{s}}}{\partial p_{sas'}}\Big]_{\hat{s}\neq s} &= \gamma \sum_{\hat{s}'\in\mathcal{S}} p^{\pi}_{\hat{s}\hat{s}'}(1) \frac{\partial v^{\pi,p}_{\hat{s}'}}{\partial p_{sas'}} \\
\Big[\frac{\partial v^{\pi,p}_{\hat{s}}}{\partial p_{sas'}}\Big]_{\hat{s}=s} &= \gamma \sum_{\hat{s}'} p^{\pi}_{s\hat{s}'}(1) \frac{\partial v^{\pi,p}_{\hat{s}'}}{\partial p_{sas'}} + \pi_{sa}\big(c_{sas'} + \gamma v^{\pi,p}_{s'}\big).
\end{aligned} \tag{9}$$

Consider $\partial v^{\pi,p}_{s_1}/(\partial \pi_{s_2a_1}\partial p_{s_3a_2s})$, for simplicity, we abbreviate it as $\partial(s_1,s_2,a_2,s_3,a_3)$:

$$\begin{aligned}
\partial(s_1,s_1,a_2,s_1,a_2) = c_{s_1a_2s} + \gamma v^{\pi,p}_{s} &+ \gamma \sum_{s'\in\mathcal{S}} p_{s_1a_2s'} \frac{\partial v^{\pi,p}_{s'}}{\partial p_{s_1a_2s}} \\
&+ \gamma \sum_{s'\in\mathcal{S}} \Big(\pi_{s_1a_2} \cdot \frac{\partial v^{\pi,p}_{s'}}{\partial \pi_{s_1a_2}} + p^{\pi,p}_{s_1s'}(1) \frac{\partial v^{\pi,p}_{s'}}{\partial \pi_{s_1a_2} \partial p_{s_1a_2s}} \Big)
\end{aligned} \tag{10}$$

$$\partial(s_1,s_1,a_2,s_1,a_3) = \gamma \sum_{s'\in\mathcal{S}} p_{s_1a_2s'} \frac{\partial v^{\pi,p}_{s'}}{\partial p_{s_1a_3s}} + \gamma \sum_{s'\in\mathcal{S}} \Big(\pi_{s_1a_3} \cdot \frac{\partial v^{\pi,p}_{s'}}{\partial \pi_{s_1a_2}} + p^{\pi,p}_{s_1s'}(1) \partial(s',s_1,a_2,s_1,a_3) \Big) \tag{11}$$

$$\partial(s_1,s_2,a_2,s_2,a_2) = \gamma \sum_{s'\in\mathcal{S}} \Big(p^{\pi}_{s_1s'}(1) \partial(s',s_2,a_2,s_2,a_2) \Big) \tag{12}$$

$$\partial(s_1,s_2,a_2,s_2,a_3) = \gamma \sum_{s'\in\mathcal{S}} \Big(p^{\pi}_{s_1s'}(1) \partial(s',s_2,a_2,s_2,a_3) \Big) \tag{13}$$

$$\partial(s_1,s_2,a_2,s_1,a_2) = \gamma \sum_{s'\in\mathcal{S}} \Big(\pi_{s_1a_2} \cdot \frac{\partial v^{\pi,p}_{s'}}{\partial \pi_{s_2a_2}} + p^{\pi}_{s_1s'}(1) \partial(s',s_2,a_2,s_1,a_2) \Big) \tag{14}$$

$$\partial(s_1,s_2,a_2,s_1,a_3) = \gamma \sum_{s'\in\mathcal{S}} \Big(\pi_{s_1a_2} \cdot \frac{\partial v^{\pi,p}_{s'}}{\partial \pi_{s_2a_2}} + p^{\pi}_{s_1s'}(1) \partial(s',s_2,a_2,s_1,a_3) \Big) \tag{15}$$

$$\partial(s_1,s_1,a_2,s_3,a_2) = \gamma \sum_{s'\in\mathcal{S}} p_{s_1a_2s'} \frac{\partial v^{\pi,p}_{s'}}{\partial p_{s_3a_2s}} + \gamma \sum_{s'\in\mathcal{S}} \Big(p^{\pi}_{s_1s'}(1) \partial(s',s_1,a_2,s_3,a_2) \Big) \tag{16}$$

$$\partial(s_1,s_1,a_2,s_3,a_3) = \gamma \sum_{s'\in\mathcal{S}} p_{s_1a_2s'} \frac{\partial v^{\pi,p}_{s'}}{\partial p_{s_3a_3s}} + \gamma \sum_{s'\in\mathcal{S}} \Big(p^{\pi}_{s_1s'}(1) \partial(s',s_1,a_2,s_3,a_3) \Big) \tag{17}$$

$$\partial(s_1,s_2,a_2,s_3,a_2) = \gamma \sum_{s'\in\mathcal{S}} \Big(p^{\pi}_{s_1s'}(1) \partial(s',s_2,a_2,s_3,a_2) \Big) \tag{18}$$

$$\partial(s_1,s_2,a_2,s_3,a_3) = \gamma \sum_{s'\in\mathcal{S}} \Big(p^{\pi}_{s_1s'}(1) \partial(s',s_2,a_2,s_3,a_3) \Big) \tag{19}$$

Now, based on the above results, we proceed as follows:

$$
\begin{aligned}
&\partial(s_1,s_1,a_2,s_1,a_2)-(c_{s_1a_2s}+\gamma v_s^{\pi,p})\\
&=\gamma\sum_{s'\in\mathcal{S}}p_{s_1a_2s'}\frac{\partial v_{s'}^{\pi,p}}{\partial p_{s_1a_2s}}\\
&\quad+\gamma\sum_{s'\in\mathcal{S}}\Big(\pi_{s_1a_2}\cdot\frac{\partial v_{s'}^{\pi,p}}{\partial\pi_{s_1a_2}}+p_{s_1s'}^{\pi,p}(1)\frac{\partial v_{s'}^{\pi,p}}{\partial\pi_{s_1a_2}\partial p_{s_1a_2s}}\Big)\\
&=\gamma\sum_{s'\neq s_1}p_{s_1a_2s'}\cdot\gamma\sum_{s''\in\mathcal{S}}p_{s's''}^{\pi}(1)\frac{\partial v_{s''}^{\pi,p}}{\partial p_{s_1a_2s}}+\gamma p_{s_1a_2s_1}\Big(\gamma\sum_{s''}p_{s's''}^{\pi}(1)\frac{\partial v_{s''}^{\pi,p}}{\partial p_{s_1a_2s}}+\pi_{s_1a_2}\big(c_{s_1a_2s}+\gamma v_s^{\pi,p}\big)\Big)\\
&\quad+\gamma\Big(\pi_{s_1a_2}\cdot\frac{\partial v_{s_1}^{\pi,p}}{\partial\pi_{s_1a_2}}+p_{s_1s_1}^{\pi,p}(1)\frac{\partial v_{s_1}^{\pi,p}}{\partial\pi_{s_1a_2}\partial p_{s_1a_2s}}\Big)+\gamma\sum_{s'\neq s_1}\Big(\pi_{s_1a_2}\cdot\frac{\partial v_{s'}^{\pi,p}}{\partial\pi_{s_1a_2}}+p_{s_1s'}^{\pi,p}(1)\frac{\partial v_{s'}^{\pi,p}}{\partial\pi_{s_1a_2}\partial p_{s_1a_2s}}\Big)\\
&\overset{(a)}{=}\gamma\pi_{s_1a_2}\cdot q_{s_1a_2}^{\pi,p}+\gamma p_{s_1a_2s_1}\pi_{s_1a_2}\big(c_{s_1a_2s}+\gamma v_s^{\pi,p}\big)+\gamma p_{s_1s_1}^{\pi,p}(1)\big(c_{s_1a_2s}+\gamma v_s^{\pi,p}\big)\\
&\quad+\gamma^2\sum_{s',s''}\Big(\pi_{s_1a_2}p_{s's''}^{\pi,p}(1)\frac{\partial v_{s''}^{\pi,p}}{\partial\pi_{s_1a_2}}+p_{s_1a_2s'}p_{s's''}^{\pi}(1)\frac{\partial v_{s''}^{\pi,p}}{\partial p_{s_1a_2s}}\Big)\\
&\quad+\gamma^2p_{s_1s_1}^{\pi,p}(1)\sum_{s'}\Big(p_{s_1a_2s'}\frac{\partial v_{s'}^{\pi,p}}{\partial p_{s_1a_2s}}+\big(\pi_{s_1a_2}\cdot\frac{\partial v_{s'}^{\pi,p}}{\partial\pi_{s_1a_2}}\big)\Big)\\
&\quad+\gamma^2\sum_{s'}p_{s_1s'}^{\pi,p}(2)\Big(\frac{\partial v_{s''}^{\pi,p}}{\partial\pi_{s_1a_2}\partial p_{s_1a_2s}}\Big),
\end{aligned}
$$

where (a) holds by (9) and (19).

Since $J_\rho(\pi,p)$ is L_π, L_p continuous, then there exist two constants c_p, c_π such that

$$
|\pi_{s_1a_2}p_{s's''}^{\pi,p}(1)\frac{\partial v_{s''}^{\pi,p}}{\partial\pi_{s_1a_2}}|\le c_p,\ \ |p_{s_1a_2s'}p_{s's''}^{\pi}(1)\frac{\partial v_{s''}^{\pi,p}}{\partial p_{s_1a_2s}}|\le c_\pi,\forall s'',s_1,a_2.
$$

Since $c_{s_ta_ts_{t+1}}\in[0,1]$, then we have

$$
|v_s^{\pi,p}|=\mathbb{E}_{\pi,p}\Big[\sum_{t=0}^{\infty}\gamma^t c_{s_ta_ts_{t+1}}\mid s_0=s\Big]\le\frac{1}{1-\gamma},\tag{20}
$$

$$
|q_{sa}^{\pi,p}|=\Big|\mathbb{E}_{\pi,p}\big[\sum_{t=0}^{\infty}\gamma^t c_{s_ta_ts_{t+1}}\mid s_0=s,a_0=a\big]\Big|\le\frac{1}{1-\gamma}.\tag{21}
$$

Now we have

$$
\begin{aligned}
&\Big|\partial(s_1,s_1,a_2,s_1,a_2)\Big|=c_{s_1a_2s}+\gamma v_s^{\pi,p}+\gamma|h(1)|\\
&\le c_{s_1a_2s}+\gamma v_s^{\pi,p}+\gamma\big(\frac{3}{1-\gamma}\big)+\gamma^2|S|^2\big(c_p+c_\pi\big)+\gamma^2|h(2)|\\
&=c_{s_1a_2s}+\gamma v_s^{\pi,p}+\gamma C_1+\gamma^2C_2+\gamma^2|h(2)|\le c_{s_1a_2s}+\gamma v_s^{\pi,p}+\gamma^2(C_2+C_1)+\gamma^2|h(2)|,
\end{aligned}
$$

where $h(t)=p_{s_1s_1}^{\pi,p}(t-1)\sum_{s'}\Big(p_{s_1a_2s'}\frac{\partial v_{s'}^{\pi,p}}{\partial p_{s_1a_2s}}+\big(\pi_{s_1a_2}\cdot\frac{\partial v_{s'}^{\pi,p}}{\partial\pi_{s_1a_2}}\big)\Big)+\sum_{s'}p_{s_1s'}^{\pi,p}(t)\Big(\frac{\partial v_{s''}^{\pi,p}}{\partial\pi_{s_1a_2}\partial p_{s_1a_2s}}\Big)$, $C_1=\frac{3}{1-\gamma}$ and $C_2=|\mathcal{S}|^2\big(c_p+c_\pi\big)$, and (a) holds by (21) and (20), $c_{s_ta_ts_{t+1}}\le1$, $p_{s_1s_1}^{\pi,p}(1)\le1$ and $\pi_{s_1a_2}\le1$. Hence, we conclude that

$$
\Big|\partial(s_1,s_1,a_2,s_1,a_2)\Big|\le c_{s_1a_2s}+\gamma v_s^{\pi,p}+(C_2+C_1)\sum_{t=0}^{\infty}\gamma^t\le\frac{1+C_1+C_2}{1-\gamma}.
$$

Similar results hold for the remaining 9 cases, then we have there exists a constant $C<\infty$ such that

$$
\Big|\partial(s_1,s_1,a_2,s_1,a_2)\Big|\le C,\forall s_1,s_2,s_3\in\mathcal{S},\forall a_2,a_3\in\mathcal{A}.
$$

By the equivalence of the matrix norm, we have there exists a constant U such that

$$\left\| \frac{\partial J_\rho(\pi, p)}{\partial \pi \partial p} \right\| \leq U < \infty.$$

Here the $\|\cdot\|$ is the spectral norm. □

Proof of Theorem 3.1

Theorem B.1. *Let π^* be a global optimal policy for RMDPs. Denote $C_{\ell_\pi} = \frac{D\sqrt{SA}}{1-\gamma} + \frac{L_\pi}{2\ell_\pi}$, where $D := \sup_{\pi\in\Pi, p\in\mathcal{P}} \|d_\rho^{\pi,p}/\rho\| < \infty$. Then for any policy $\pi \in \Pi$, we have*

$$\phi_{2\ell_\pi}(\pi) \leq \phi(\pi) \leq C_{\ell_\pi} \|\nabla \phi_{2\ell_\pi}(\pi)\| + \phi(\pi^*), \tag{22}$$

where $\phi_{2\ell_\pi}(\pi) := \inf_{\pi'\in\Pi}\{\phi(\pi') + \ell_\pi \|\pi - \pi'\|^2\}$.

Proof. Proof can be found in Wang et al. (2023) or Li et al. (2024). □

Remark B.2. *Mismatch coefficient $D < \infty$ can be guaranteed by Assumption 2.3.*

Proof of Theorem 3.2

Proof. Let $\tilde{x}(x) \in \operatorname{argmin}_{z\in\mathcal{X}} f(z) + r\|z - x\|^2$. It follows from the definition of $f_{2r}(x)$ and $\min_{x\in\mathcal{X}} f(x) = \min_{x\in\mathcal{X}} f_{2r}(x) = f^*$ that

$$\begin{aligned} & f_{2r}(x) - f^* \\ & = \inf_{z\in\mathcal{X}} f(z) + r\|z - x\|^2 - f^* \\ & \leq f(x) - f^* = f(x) - f(\tilde{x}(x)) + f(\tilde{x}(x)) - f^* \\ & \overset{(a)}{\leq} L_f \|\tilde{x}(x) - x\| + C \operatorname{dist}\big(\partial f(\tilde{x}(x)) + \mathcal{N}_{\mathcal{X}}(\tilde{x}(x))\big), \end{aligned} \tag{23}$$

where (a) holds by $f(x)$ is L_f continuity and the gradient dominance property of f. By the definition of $f_{2r}(x)$, we have

$$0 \in \partial f(\tilde{x}(x)) + 2r(\tilde{x}(x) - x) + \mathcal{N}_{\mathcal{X}}(\tilde{x}(x)). \tag{24}$$

Furthermore, by the Moreau envelope property, we have

$$\nabla f_{2r}(x) \in \partial f(\tilde{x}(x)) + \mathcal{N}_{\mathcal{X}}(\tilde{x}(x)), \quad \nabla f_{2r}(x) = 2r(\tilde{x}(x) - x). \tag{25}$$

Combining (23), (24) and (25) yields

$$f_{2r}(x) - f^* \leq \Big(\frac{L_f}{2r} + C\Big)\|\nabla f_{2r}(x)\|.$$

□

C. Convergence Analysis

C.1. Proof of Results in Section 4.2

Lemma C.1. *For any $\pi, \pi' \in \Pi$, $p, p' \in \mathcal{P}$, $\bar{\pi}$ and $\bar{p}$, we have*

$$\begin{aligned} & \frac{r_1 - \ell_{\pi p}}{2}\|\pi - \pi'\|^2 \leq \chi(\pi', p, \bar{\pi}, \bar{p}) - \chi(\pi, p, \bar{\pi}, \bar{p}) - \langle \nabla_\pi \chi(\pi, p, \bar{\pi}, \bar{p}), \pi' - \pi\rangle \leq \frac{\ell_{\pi p} + r_1}{2}\|\pi - \pi'\|^2, \\ & -\frac{\ell_{p\pi} + r_2}{2}\|p - p'\|^2 \leq \chi(\pi, p', \bar{\pi}, \bar{p}) - \chi(\pi, p, \bar{\pi}, \bar{p}) - \langle \nabla_p \chi(\pi, p, \bar{\pi}, \bar{p}), p' - p\rangle \leq \frac{\ell_{p\pi} - r_2}{2}\|p - p'\|^2. \end{aligned}$$

Proof. Since $J_\rho(\pi, p)$ is $\ell_{\pi p}$, $\ell_{p\pi}$ smooth, respectively, we have

$$\begin{aligned} -\frac{\ell_{\pi p}}{2}\|\pi-\pi'\|^2 \le J_\rho(\pi',p) - J_\rho(\pi,p) - \langle \nabla_\pi J_\rho(\pi,p), \pi'-\pi\rangle \le \frac{\ell_{\pi p}}{2}\|\pi-\pi'\|^2, \\ -\frac{\ell_{p\pi}}{2}\|p-p'\|^2 \le J_\rho(\pi,p') - J_\rho(\pi,p) - \langle \nabla_p J_\rho(\pi,p), p'-p\rangle \le \frac{\ell_{p\pi}}{2}\|p-p'\|^2. \end{aligned} \tag{26}$$

Hence, consider function $\chi(\pi, p, \bar\pi, \bar p)$, we know that

$$\begin{aligned} &\chi(\pi', p, \bar\pi, \bar p) - \chi(\pi, p, \bar\pi, \bar p) - \langle \nabla_\pi \chi(\pi, p, \bar\pi, \bar p), \pi'-\pi\rangle \\ &= J_\rho(\pi', p) - J_\rho(\pi, p) - \langle \nabla_\pi J_\rho(\pi,p) + r_1(\pi-\bar\pi), \pi'-\pi\rangle + \frac{r_1}{2}\|\pi'-\bar\pi\|^2 - \frac{r_1}{2}\|\pi-\bar\pi\|^2 \\ &= J_\rho(\pi', p) - J_\rho(\pi, p) - \langle \nabla_\pi J_\rho(\pi,p), \pi'-\pi\rangle + \frac{r_1}{2}\|\pi'-\pi\|^2 \end{aligned} \tag{27}$$

and similarly

$$\begin{aligned} &\chi(\pi, p', \bar\pi, \bar p) - \chi(\pi, p, \bar\pi, \bar p) - \langle \nabla_p \chi(\pi, p, \bar\pi, \bar p), p'-p\rangle \\ &= J_\rho(\pi, p') - J_\rho(\pi, p) - \langle \nabla_p J_\rho(\pi,p) - r_2(p-\bar p), p'-p\rangle - \frac{r_2}{2}\|p'-\bar p\|^2 + \frac{r_2}{2}\|p-\bar p\|^2 \\ &= J_\rho(\pi, p') - J_\rho(\pi, p) - \langle \nabla_p J_\rho(\pi,p), p'-p\rangle - \frac{r_2}{2}\|p'-p\|^2. \end{aligned} \tag{28}$$

Combing (26), (27) and (28), we directly obtain the desired results. □

Lemma C.2 (Lemma 2 in Zheng et al. (2023))**.** *Suppose that* $r_2 > (\frac{\ell_{p\pi}}{r_1-\ell_{\pi p}} + 2)\ell_{p\pi}$*, then for any* $\pi, \pi' \in \Pi$*,* $p, p' \in \mathcal{P}$*,* $\bar\pi, \bar\pi' \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ *and* $p, p' \in \mathbb{R}^{|\mathcal{S}|}$*. Then the following inequalities hold:*

1. $\|\pi(p', \bar\pi, \bar p) - \pi(p, \bar\pi, \bar p)\| \le \kappa_1\|p'-p\|$,
2. $\|\pi(p, \bar\pi', \bar p) - \pi(p, \bar\pi, \bar p)\| \le \kappa_2\|\bar\pi-\bar\pi'\|$,
3. $\|\pi(\bar\pi', \bar p) - \pi(\bar\pi, \bar p)\| \le \kappa_2\|\bar\pi-\bar\pi'\|$,
4. $\|p(\bar\pi, \bar p) - p(\bar\pi', \bar p)\| \le \kappa_3\|\bar\pi-\bar\pi'\|$,
5. $\|p(\pi, \bar\pi, \bar p) - p(\pi', \bar\pi, \bar p)\| \le \kappa_4\|\pi-\pi'\|$
6. $\|p(\pi, \bar\pi, \bar p) - p(\pi, \bar\pi, \bar p')\| \le \kappa_5\|\bar p-\bar p'\|$
7. $\|p(\bar\pi, \bar p) - p(\bar\pi, \bar p')\| \le \kappa_5\|\bar p-\bar p'\|$,

where $\kappa_1 = \frac{\ell_{p\pi}+r_1-\ell_{\pi p}}{r_1-\ell_{\pi p}}$, $\kappa_2 = \frac{r_1}{r_1-\ell_{\pi p}}$, $\kappa_3 = \frac{r_1\kappa_1}{r_2-\ell_{p\pi}} + \frac{\kappa_2}{\kappa_1}$, $\kappa_4 = \frac{\ell_{\pi p}+r_2-\ell_{p\pi}}{r_2-\ell_{p\pi}}$, and $\kappa_5 = \frac{r_2}{r_2-\ell_{p\pi}}$.

Proof. Item 1: It follows from Lemma C.1 that

$$\chi(\pi(p,\bar\pi,\bar p), p', \bar\pi, \bar p) - \chi(\pi(p',\bar\pi,\bar p), p', \bar\pi, \bar p) \ge \frac{r_1-\ell_{\pi p}}{2}\|\pi(p,\bar\pi,\bar p) - \pi(p',\bar\pi,\bar p)\|^2, \tag{29}$$

$$\chi(\pi(p,\bar\pi,\bar p), p', \bar\pi, \bar p) - \chi(\pi(p,\bar\pi,\bar p), p, \bar\pi, \bar p) \le \langle \nabla_p \chi(\pi(p,\bar\pi,\bar p), p, \bar\pi, \bar p), p'-p\rangle + \frac{\ell_{p\pi}-r_2}{2}\|p-p'\|^2, \tag{30}$$

$$\chi(\pi(p',\bar\pi,\bar p), p, \bar\pi, \bar p) - \chi(\pi(p',\bar\pi,\bar p), p', \bar\pi, \bar p) \le \langle \nabla_p \chi(\pi(p',\bar\pi,\bar p), p, \bar\pi, \bar p), p-p'\rangle + \frac{\ell_{p\pi}+r_2}{2}\|p-p'\|^2, \tag{31}$$

$$\chi(\pi(p,\bar\pi,\bar p), p, \bar\pi, \bar p) - \chi(\pi(p',\bar\pi,\bar p), p, \bar\pi, \bar p) \le \frac{\ell_{\pi p}-r_1}{2}\|\pi(p,\bar\pi,\bar p) - \pi(p',\bar\pi,\bar p)\|^2. \tag{32}$$

Summing up (30) and (31) and combining it with (29) yields

$$\begin{aligned}
&(r_1-\ell_{\pi p})\|\pi(p,\bar\pi,\bar p)-\pi(p',\bar\pi,\bar p)\|^2\\
&\le\langle\nabla_p\chi(\pi(p,\bar\pi,\bar p),p,\bar\pi,\bar p)-\nabla_p\chi(\pi(p',\bar\pi,\bar p),p,\bar\pi,\bar p),p'-p\rangle+\ell_{p\pi}\|p-p'\|^2\\
&=\langle\big(\nabla_pJ(\pi(p,\bar\pi,\bar p),p)-r_2\big(p-\bar p\big)\big)-\big(\nabla_pJ(\pi(p',\bar\pi,\bar p),p)-r_2\big(p-\bar p\big)\big),p'-p\rangle+\ell_{p\pi}\|p-p'\|^2\\
&\overset{(a)}{\le}\ell_{p\pi}\|\pi(p',\bar\pi,\bar p)-\pi(p,\bar\pi,\bar p)\|\|p'-p\|+\ell_{p\pi}\|p-p'\|^2,
\end{aligned}\tag{33}$$

where (a) holds by Cauchy Schwarz inequality and Item iv) in Lemma 3.1. Let $\Lambda:=\|\pi(p',\bar\pi,\bar p)-\pi(p,\bar\pi,\bar p)\|/\|p'-p\|$. Then (33) can be rewritten as

$$\begin{aligned}
&\Lambda^2\le\frac{\ell_{p\pi}}{r_1-\ell_{\pi p}}+\frac{\ell_{p\pi}}{r_1-\ell_{\pi p}}\Lambda\overset{(a)}{\le}\frac{\ell_{p\pi}}{r_1-\ell_{\pi p}}+\frac12\Big(\frac{\ell_{p\pi}}{r_1-\ell_{\pi p}}\Big)^2+\frac12\Lambda^2,\\
&\Rightarrow\Lambda\le\sqrt{\Big(\frac{\ell_{p\pi}}{r_1-\ell_{\pi p}}\Big)^2+2\Big(\frac{\ell_{p\pi}}{r_1-\ell_{\pi p}}\Big)}\le\frac{\ell_{p\pi}}{r_1-\ell_{\pi p}}+1:=\kappa_1.
\end{aligned}\tag{34}$$

where (a) holds by $\langle\mathbf{a},\mathbf{b}\rangle\le(\|\mathbf{a}\|^2+\|\mathbf{b}\|^2)/2$.

Item 2 and Item 3: It follows from from Lemma C.1 that

$$\begin{aligned}
\chi(\pi(p,\bar\pi,\bar p),p,\bar\pi',\bar p)-\chi(\pi(p,\bar\pi',\bar p),p,\bar\pi',\bar p)&\ge\frac{r_1-\ell_{\pi p}}{2}\|\pi(p,\bar\pi,\bar p)-\pi(p,\bar\pi',\bar p)\|^2,\\
\chi(\pi(p,\bar\pi,\bar p),p,\bar\pi,\bar p)-\chi(\pi(p,\bar\pi',\bar p),p,\bar\pi,\bar p)&\le\frac{\ell_{\pi p}-r_1}{2}\|\pi(p,\bar\pi,\bar p)-\pi(p,\bar\pi',\bar p)\|^2.
\end{aligned}\tag{35}$$

By the definition of χ, we have

$$\begin{aligned}
\chi(\pi(p,\bar\pi,\bar p),p,\bar\pi',\bar p)-\chi(\pi(p,\bar\pi,\bar p),p,\bar\pi,\bar p)&=\frac{r_1}{2}\langle\bar\pi'+\bar\pi-2\pi(p,\bar\pi,\bar p),\bar\pi'-\bar\pi\rangle,\\
\chi(\pi(p,\bar\pi',\bar p),p,\bar\pi,\bar p)-\chi(\pi(p,\bar\pi',\bar p),p,\bar\pi',\bar p)&=\frac{r_1}{2}\langle\bar\pi+\bar\pi'-2\pi(p,\bar\pi',\bar p),\bar\pi-\bar\pi'\rangle.
\end{aligned}\tag{36}$$

Putting (35) and (36) together and using the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
&(r_1-\ell_{\pi p})\|\pi(p,\bar\pi,\bar p)-\pi(p,\bar\pi',\bar p)\|^2\\
&\le r_1\langle\pi(p,\bar\pi',\bar p)-\pi(p,\bar\pi,\bar p),\bar\pi'-\bar\pi\rangle\\
&\le r_1\|\pi(p,\bar\pi',\bar p)-\pi(p,\bar\pi,\bar p)\|\|\bar\pi'-\bar\pi\|,
\end{aligned}$$

which completes the proof of Item 2. Since $\varphi_p(\pi,\bar\pi,\bar p)$ is $\big(r_1-\ell_{\pi p}\big)-$strongly convex in π, then we have

$$\begin{aligned}
\varphi_p(\pi(\bar\pi,\bar p),\bar\pi',\bar p)-\varphi_p(\pi(\bar\pi',\bar p),\bar\pi',\bar p)&\ge\frac{r_1-\ell_{\pi p}}{2}\|\pi(\bar\pi,\bar p)-\pi(\bar\pi',\bar p)\|^2\\
\varphi_p(\pi(\bar\pi,\bar p),\bar\pi,\bar p)-\varphi_p(\pi(\bar\pi',\bar p),\bar\pi,\bar p)&\le\frac{\ell_{\pi p}-r_1}{2}\|\pi(\bar\pi,\bar p)-\pi(\bar\pi',\bar p)\|^2
\end{aligned}\tag{37}$$

By the definition of φ_p, we have

$$\begin{aligned}
\varphi_p(\pi(\bar\pi,\bar p),\bar\pi',\bar p)-\varphi_p(\pi(\bar\pi,\bar p),\bar\pi,\bar p)&=\frac{r_1}{2}\langle\bar\pi+\bar\pi'-2\pi(\bar\pi,\bar p),\bar\pi'-\bar\pi\rangle\\
\varphi_p(\pi(\bar\pi',\bar p),\bar\pi,\bar p)-\varphi_p(\pi(\bar\pi',\bar p),\bar\pi',\bar p)&=\frac{r_1}{2}\langle\bar\pi+\bar\pi'-2\pi(\bar\pi',\bar p),\bar\pi-\bar\pi'\rangle.
\end{aligned}\tag{38}$$

Putting (37) and (38) together and using the Cauchy-Schwarz inequality, we obtain the result Item 3.

Item 4: Since $-\varphi_\pi(p,\bar\pi,\bar p)$ is $(r_2-\ell_{p\pi})$ strongly convex w.r.t. p, then we have

$$\begin{aligned}
\varphi_\pi(p(\bar\pi,\bar p),\bar\pi,\bar p)-\varphi_\pi(p(\bar\pi',\bar p),\bar\pi,\bar p)&\ge\frac{r_2-\ell_{p\pi}}{2}\|p(\bar\pi,\bar p)-p(\bar\pi',\bar p)\|^2,\\
\varphi_\pi(p(\bar\pi',\bar p),\bar\pi',\bar p)-\varphi_\pi(p(\bar\pi,\bar p),\bar\pi',\bar p)&\ge\frac{r_2-\ell_{p\pi}}{2}\|p(\bar\pi',\bar p)-p(\bar\pi,\bar p)\|^2,
\end{aligned}\tag{39}$$

By the definition of φ_π, we have

$$\begin{aligned}
&\varphi_\pi(p(\bar{\pi},\bar{p}),\bar{\pi},\bar{p}) - \varphi_\pi(p(\bar{\pi},\bar{p}),\bar{\pi}',\bar{p}) \\
&= \chi(\pi(p(\bar{\pi},\bar{p}),\bar{\pi},\bar{p}),p(\bar{\pi},\bar{p}),\bar{\pi},\bar{p}) - \chi(\pi(p(\bar{\pi},\bar{p}),\bar{\pi}',\bar{p}),p(\bar{\pi},\bar{p}),\bar{\pi}',\bar{p}) \\
&\le \chi(\pi(p(\bar{\pi},\bar{p}),\bar{\pi}',\bar{p}),p(\bar{\pi},\bar{p}),\bar{\pi},\bar{p}) - \chi(\pi(p(\bar{\pi},\bar{p}),\bar{\pi}',\bar{p}),p(\bar{\pi},\bar{p}),\bar{\pi}',\bar{p}) \\
&= \frac{r_1}{2}\langle \bar{\pi}+\bar{\pi}' - 2\pi(p(\bar{\pi},\bar{p}),\bar{\pi}',\bar{p}),\bar{\pi}-\bar{\pi}'\rangle
\end{aligned} \tag{40}$$

and

$$\begin{aligned}
&\varphi_\pi(p(\bar{\pi}',\bar{p}),\bar{\pi}',\bar{p}) - \varphi_\pi(p(\bar{\pi}',\bar{p}),\bar{\pi},\bar{p}) \\
&= \chi(\pi(p(\bar{\pi}',\bar{p}),\bar{\pi}',\bar{p}),p(\bar{\pi}',\bar{p}),\bar{\pi}',\bar{p}) - \chi(\pi(p(\bar{\pi}',\bar{p}),\bar{\pi},\bar{p}),p(\bar{\pi}',\bar{p}),\bar{\pi},\bar{p}) \\
&\le \chi(\pi(p(\bar{\pi}',\bar{p}),\bar{\pi},\bar{p}),p(\bar{\pi}',\bar{p}),\bar{\pi}',\bar{p}) - \chi(\pi(p(\bar{\pi}',\bar{p}),\bar{\pi},\bar{p}),p(\bar{\pi}',\bar{p}),\bar{\pi},\bar{p}) \\
&= \frac{r_1}{2}\langle \bar{\pi}'+\bar{\pi} - 2x(p(\bar{\pi}',\bar{p}),\bar{\pi},\bar{p}),\bar{\pi}'-\bar{\pi}\rangle.
\end{aligned} \tag{41}$$

Putting (39), (40) and (41) together yields

$$\begin{aligned}
&(r_2-\ell_{p\pi})\|p(\bar{\pi},\bar{p}) - p(\bar{\pi}',\bar{p})\|^2 \\
&\le r_1\langle \pi(p(\bar{\pi}',\bar{p}),\bar{\pi},\bar{p}) - \pi(p(\bar{\pi},\bar{p}),\bar{\pi}',\bar{p}),\bar{\pi}-\bar{\pi}'\rangle \\
&\overset{(a)}{\le} r_1\|\bar{\pi}-\bar{\pi}'\|\big(\kappa_1\|\pi(p(\bar{\pi}',\bar{p}) - \pi(p(\bar{\pi},\bar{p})\| + \kappa_2\|\bar{\pi}-\bar{\pi}'\|\big) \\
&= r_1\kappa_1\|\bar{\pi}-\bar{\pi}'\|\|\pi(p(\bar{\pi}',\bar{p}) - \pi(p(\bar{\pi},\bar{p})\| + r_1\kappa_2\|\bar{\pi}-\bar{\pi}'\|^2
\end{aligned} \tag{42}$$

where (a) holds by Item 1 and Item 2. Let $\Lambda = \|p(\bar{\pi},\bar{p}) - p(\bar{\pi}',\bar{p})\|/\|\bar{\pi}-\bar{\pi}'\|$ here, then (42) implies

$$\begin{aligned}
&\Lambda^2 \le \frac{r_1\kappa_1}{r_2-\ell_{p\pi}}\Lambda + \frac{r_1\kappa_2}{r_2-\ell_{p\pi}} \le \frac{1}{2}\Big(\frac{r_1\kappa_1}{r_2-\ell_{p\pi}}\Big)^2 + \frac{1}{2}\Lambda^2 + \frac{r_1\kappa_2}{r_2-\ell_{p\pi}} \\
\Rightarrow &\Lambda^2 \le \Big(\frac{r_1\kappa_1}{r_2-\ell_{p\pi}}\Big)^2 + \frac{2r_1\kappa_2}{r_2-\ell_{p\pi}} \le \Big(\frac{r_1\kappa_1}{r_2-\ell_{p\pi}} + \frac{\kappa_2}{\kappa_1}\Big)^2 =: \kappa_3^2,
\end{aligned} \tag{43}$$

which complete our proof of Item 4.

Item 5. It follows from $\chi(\cdot,p,\cdot,\cdot)$ is $(r_2-\ell_{p\pi})-$strongly concave that

$$\begin{aligned}
\chi(\pi,p(\pi,\bar{\pi},\bar{p}),\bar{\pi},\bar{p}) - \chi(\pi,p(\pi',\bar{\pi},\bar{p}),\bar{\pi},\bar{p}) &\ge \frac{r_2-\ell_{p\pi}}{2}\|p(\pi,\bar{\pi},\bar{p}) - p(\pi',\bar{\pi},\bar{p})\|^2, \\
\chi(\pi',p(\pi,\bar{\pi},\bar{p}),\bar{\pi},\bar{p}) - \chi(\pi',p(\pi',\bar{\pi},\bar{p}),\bar{\pi},\bar{p}) &\le \frac{\ell_{p\pi}-r_2}{2}\|p(\pi,\bar{\pi},\bar{p}) - p(\pi',\bar{\pi},\bar{p})\|^2, \\
\chi(\pi,p(\pi,\bar{\pi},\bar{p}),\bar{\pi},\bar{p}) - \chi(\pi',p(\pi,\bar{\pi},\bar{p}),\bar{\pi},\bar{p}) &\le \langle \nabla_\pi\chi(\pi',p(\pi,\bar{\pi},\bar{p}),\bar{\pi},\bar{p}),\pi-\pi'\rangle + \frac{\ell_{\pi p}+r_1}{2}\|\pi-\pi'\|^2, \\
\chi(\pi',p(\pi',\bar{\pi},\bar{p}),\bar{\pi},\bar{p}) - \chi(\pi,p(\pi',\bar{\pi},\bar{p}),\bar{\pi},\bar{p}) &\le \langle \nabla_\pi\chi(\pi',p(\pi',\bar{\pi},\bar{p}),\bar{\pi},\bar{p}),x'-x\rangle + \frac{\ell_{\pi p}-r_1}{2}\|\pi-\pi'\|^2.
\end{aligned}$$

Summing them up, we derive that

$$(r_2-\ell_{p\pi})\|p(\pi,\bar{\pi},\bar{p}) - p(\pi',\bar{\pi},\bar{p})\|^2 \le \ell_{\pi p}\|\pi-\pi'\|^2 + \ell_{\pi p}\|\pi-\pi'\|\|p(\pi,\bar{\pi},\bar{p}) - p(\pi',\bar{\pi},\bar{p})\|.$$

Let $\Lambda = \|p(\pi,\bar{\pi},\bar{p}) - p(\pi',\bar{\pi},\bar{p})\|/\|\pi-\pi'\|$, then we have

$$\Lambda^2 \le \frac{\ell_{\pi p}}{r_2-\ell_{p\pi}} + \frac{\ell_{\pi p}}{r_2-\ell_{p\pi}}\Lambda \overset{(a)}{\Rightarrow} \Lambda \le \frac{\ell_{\pi p}+r_2-\ell_{p\pi}}{r_2-\ell_{p\pi}} =: \kappa_4,$$

where (a) holds by similary argument in (34) and (43).

Item 6 and Item 7: By the definition of χ and the $(r_2 - \ell_{p\pi})$ strongly conevx of $\chi(\cdot, p, \cdot, \cdot)$ w.r.t. p, we have

$$\chi(\pi, p(\pi,\bar{\pi},\bar{p}),\bar{\pi},\bar{p}) - \chi(\pi, p(\pi,\bar{\pi},\bar{p}),\bar{\pi},\bar{p}') = \frac{r_2}{2}\langle \bar{p}' + \bar{p} - 2p(\pi,\bar{\pi},\bar{p}), \bar{p}' - \bar{p}\rangle,$$
$$\chi(\pi, p(\pi,\bar{\pi},\bar{p}'),\bar{\pi},\bar{p}') - \chi(\pi, p(\pi,\bar{\pi},\bar{p}'),\bar{\pi},\bar{p}) = \frac{r_2}{2}\langle v + \bar{p}' - 2p(\pi,\bar{\pi},\bar{p}'), \bar{p} - \bar{p}'\rangle,$$
$$\chi(\pi, p(\pi,\bar{\pi},\bar{p}),\bar{\pi},\bar{p}) - \chi(\pi, p(\pi,\bar{\pi},\bar{p}'),\bar{\pi},\bar{p}) \geq \frac{r_2 - \ell_{p\pi}}{2}\|p(\pi,\bar{\pi},\bar{p}) - p(\pi,\bar{\pi},\bar{p}')\|^2,$$
$$\chi(\pi, p(\pi,\bar{\pi},\bar{p}),\bar{\pi},\bar{p}') - \chi(\pi, p(\pi,\bar{\pi},\bar{p}'),\bar{\pi},\bar{p}') \leq \frac{\ell_{p\pi} - r_2}{2}\|p(\pi,\bar{\pi},\bar{p}) - p(\pi,\bar{\pi},\bar{p}')\|^2.$$

Armed with these inequalities and Cauchy Schwarz inequality, we conclude

$$(r_2 - \ell_{p\pi})\|p(\pi,\bar{\pi},\bar{p}) - p(\pi,\bar{\pi},\bar{p}')\|^2 \leq r_2\|p(\pi,\bar{\pi},\bar{p}') - p(\pi,\bar{\pi},\bar{p})\|\|\bar{p} - \bar{p}'\|,$$

which complete our proof of Item 6. By the definition of φ_π and similar argument, we can obtain result Item 7. □

Lemma C.3 (Lemma 3 in Zheng et al. (2023))**.** *The function $\varphi_\pi(p,\cdot,\cdot)$ is continuously differentiable with the gradient $\nabla_p\varphi_\pi(p,\bar{\pi},\bar{p}) = \nabla_p\chi(\pi(p,\bar{\pi},\bar{p}),p,\bar{\pi},\bar{p})$ and*

$$\|\nabla_p\varphi_\pi(p,\bar{\pi},\bar{p}) - \nabla_p\varphi_\pi(p',\bar{\pi},\bar{p})\| \leq L_{\phi\pi}\|p - p'\|,$$

where $L_{\phi\pi} := \ell_{p\pi}\kappa_1 + \ell_{p\pi} + r_2$.

Proof. Using Danskin's theorem, we know that $\varphi_\pi(\cdot,\bar{\pi},\bar{p})$ is differentiable with $\nabla_p\varphi_\pi(p,\bar{\pi},\bar{p}) = \nabla_p\chi(\pi(p,\bar{\pi},\bar{p}),p,\bar{\pi},\bar{p})$. Also, it follows from Lemma 3.1 that

$$\begin{aligned}
&\|\nabla_p\varphi_\pi(p,\bar{\pi},\bar{p}) - \nabla_p\varphi_\pi(p',\bar{\pi},\bar{p})\| \\
&= \|\nabla_p\chi(\pi(p,\bar{\pi},\bar{p}),\bar{\pi},\bar{p}) - \nabla_p\chi(\pi(p',\bar{\pi},\bar{p}),p',\bar{\pi},\bar{p})\| \\
&\leq \|\nabla_p\chi(\pi(p,\bar{\pi},\bar{p}),\bar{\pi},\bar{p}) - \nabla_p\chi(\pi(p',\bar{\pi},\bar{p}),\bar{\pi},\bar{p})\| \\
&\quad + \|\nabla_p\chi(\pi(p',\bar{\pi},\bar{p}),\bar{\pi},\bar{p}) - \nabla_p\chi(\pi(p',\bar{\pi},\bar{p}),p',\bar{\pi},\bar{p})\| \\
&\leq \ell_{p\pi}\|\pi(p',\bar{\pi},\bar{p}) - \pi(p,\bar{\pi},\bar{p})\| + (\ell_{p\pi} + r_2)\|p - p'\| \\
&\overset{(a)}{\leq} (\ell_{p\pi}\kappa_1 + \ell_{p\pi} + r_2)\|p - p'\| = L_{\phi\pi}\|p - p'\|,
\end{aligned}$$

where (a) holds by Item 1 in Lemma C.2. □

Lemma C.4 (Lemma 4 in Zheng et al. (2023))**.** *For any $k \geq 0$, the following inequlities holds*

a) $\|\pi_{k+1} - \pi(p_k,\bar{\pi}_k,\bar{p}_k)\| \leq \kappa_6\|\pi_{k+1} - \pi_k\|$

b) $\|p_{k+1} - p(\pi_{k+1},\bar{\pi}_k,\bar{p}_k)\| \leq \kappa_7\|p_{k+1} - p_k\|$

c) $\|p(\bar{\pi}_k,\bar{p}_k) - p_k\| \leq \kappa_8\|p_k - p_k^+(\bar{\pi}_k,\bar{p}_k)\|$

d) $\|p_{k+1} - p_k^+(\bar{\pi}_k,\bar{p}_k)\| \leq \ell_{p\pi}\sigma\kappa_6\|\pi_k - \pi_{k+1}\|$,

where $\kappa_6 = (2\tau r_1 + 1)/(\tau r_1 - \tau\ell_{\pi p})$, $\kappa_7 = (2\sigma r_2 + 1)/(\sigma r_2 - \sigma\ell_{p\pi})$, and $\kappa_8 = (1 + \sigma L_{\phi\pi})/(\sigma(r_2 - \ell_{p\pi}))$.

Proof. Item a): It follows from Lemma C.1 that $\chi(\pi,p,\bar{\pi},\bar{p})$ is $(r_1 - \ell_{\pi p})$ strongly convex and $(r_1 + \ell_{\pi p})$ smooth w.r.t. π on Π for any $p \in \mathcal{P}, \bar{\pi} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}, \bar{p} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|\times|\mathcal{S}|}$. Adopting the proof in Theorem 3.1 in Pang (1987), we have

$$\|\pi_k - \pi(p_k,\bar{\pi}_k,\bar{p}_k)\| \leq \frac{\tau\ell_{\pi p} + \tau\cdot r_1 + 1}{\tau\cdot r_1 - \tau\ell_{\pi p}}\|\pi_{k+1} - \pi_k\|,$$

which implies that

$$\|\pi_{k+1} - \pi(p_k,\bar{\pi}_k,\bar{p}_k)\| \leq \|\pi_{k+1} - \pi_k\| + \|\pi_k - \pi(p_k,\bar{\pi}_k,\bar{p}_k)\| \leq \frac{2\tau\cdot r_1 + 1}{\tau\cdot r_1 - \tau\ell_{\pi p}}\|\pi_{k+1} - \pi_k\|.$$

Item b): Similar to Item a), Item b) holds by $-\chi(\pi, p, \bar{\pi}, \bar{p})$ is $(r_2 - \ell_{p\pi})$-strongly convex and $(r_2 + \ell_{p\pi})$ smooth w.r.t. p.

Item c): Similar to Item a), Item c) holds by $-\varphi_\pi(p, \bar{\pi}, \bar{p})$ is $(r_2 - \ell_{p\pi})$-strongly convex and $L_{\phi\pi}$ smooth w.r.t. p.

Item d): It follows from the non-expansivity of projection operator that

$$\begin{aligned}
&\|p_{k+1} - p_k^+(\bar{\pi}_k, \bar{p}_k)\| \\
&= \|\operatorname{Proj}_{\mathcal{P}}(p_k + \sigma\nabla_p F(\pi_{k+1}, p_k, \bar{\pi}_k, \bar{p}_k)) - \operatorname{Proj}_{\mathcal{P}}(p_k + \sigma\nabla_p \chi(\pi(p_k, \bar{\pi}_k, \bar{p}_k), p_k, \bar{\pi}_k, \bar{p}_k))\| \\
&\le \sigma\|\nabla_p F(\pi_{k+1}, p_k, \bar{\pi}_k, \bar{p}_k) - \nabla_p \chi(\pi(p_k, \bar{\pi}_k, \bar{p}_k), p_k, \bar{\pi}_k, \bar{p}_k)\| \\
&\le \sigma\ell_{p\pi}\|\pi_{k+1} - \pi(p_k, \bar{\pi}_k, \bar{p}_k)\| \le \ell_{p\pi}\sigma\kappa_6\|\pi_k - \pi_{k+1}\|.
\end{aligned}$$

□

Lemma C.5 (Lemma 5 in Zheng et al. (2023))**.** *For any $k \ge 0$, the following inequality holds:*

$$\begin{aligned}
&\chi(\pi_k, p_k, \bar{\pi}_k, \bar{p}_k) \\
&\ge \chi(\pi_{k+1}, p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_{k+1}) + \left(\frac{1}{\tau} - \frac{\ell_{\pi p} + r_1}{2}\right)\|\pi_{k+1} - \pi_k\|^2 \\
&\quad + \langle \nabla_p F(\pi_{k+1}, p_k, \bar{\pi}_k, \bar{p}_k), p_k - p_{k+1}\rangle + \frac{r_2 - \ell_{p\pi}}{2}\|p_{k+1} - p_k\|^2 \\
&\quad + \frac{2-\beta}{2\beta} r_1 \|\bar{\pi}_{k+1} - \bar{\pi}_k\|^2 + \frac{\mu - 2}{2\mu} r_2 \|\bar{p}_{k+1} - \bar{p}_k\|^2.
\end{aligned}$$

Proof. We split the target as follows:

$$\begin{aligned}
&\chi(\pi_k, p_k, \bar{\pi}_k, \bar{p}_k) - \chi(\pi_{k+1}, p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_{k+1}) \\
&= \underbrace{\chi(\pi_k, p_k, \bar{\pi}_k, \bar{p}_k) - \chi(\pi_{k+1}, p_k, \bar{\pi}_k, \bar{p}_k)}_{\text{I}} + \underbrace{\chi(\pi_{k+1}, p_k, \bar{\pi}_k, \bar{p}_k) - \chi(\pi_{k+1}, p_{k+1}, \bar{\pi}_k, \bar{p}_k)}_{\text{II}} \\
&\quad + \underbrace{\chi(\pi_{k+1}, p_{k+1}, z^t, \bar{p}_k) - \chi(\pi_{k+1}, p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_k)}_{\text{III}} + \underbrace{\chi(\pi_{k+1}, p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_k) - \chi(\pi_{k+1}, p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_{k+1})}_{\text{IV}}.
\end{aligned}$$

Term I: By the optimality condition $\langle \pi_k - \tau\nabla_\pi \chi(\pi_k, p_k, \bar{\pi}_k, \bar{p}_k) - \pi_{k+1}, \pi_k - \pi_{k+1}\rangle \le 0$ and χ is $\ell_{\pi p} + r_1$ smooth w.r.t. π, we have

$$\begin{aligned}
\chi(\pi_k, p_k, \bar{\pi}_k, \bar{p}_k) - \chi(\pi_{k+1}, p_k, \bar{\pi}_k, \bar{p}_k) &\ge \langle \nabla_\pi \chi(\pi_k, p_k, \bar{\pi}_k, \bar{p}_k), \pi_k - \pi_{k+1}\rangle - \frac{\ell_{\pi p} + r_1}{2}\|\pi_{k+1} - \pi_k\|^2 \\
&\ge \left(\frac{1}{\tau} - \frac{\ell_{\pi p} + r_1}{2}\right)\|\pi_{k+1} - \pi_k\|^2.
\end{aligned}$$

Term II: By χ is $r_2 - \ell_{p\pi}$ strongly convex w.r.t. p, we have

$$\chi(\pi_{k+1}, p_k, \bar{\pi}_k, \bar{p}_k) - \chi(\pi_{k+1}, p_{k+1}, \bar{\pi}_k, \bar{p}_k) \ge \langle \nabla_p \chi(\pi_{k+1}, p_k, \bar{\pi}_k, \bar{p}_k), p_k - p_{k+1}\rangle + \frac{r_2 - \ell_{p\pi}}{2}\|p_{k+1} - p_k\|^2.$$

Term III: It follows from $\bar{\pi}_{k+1} = \bar{\pi}_k + \beta(\pi_{k+1} - \bar{\pi}_k)$ that

$$\chi(\pi_{k+1}, p_{k+1}, \bar{\pi}_k, \bar{p}_k) - \chi(\pi_{k+1}, p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_k) = \frac{2-\beta}{2\beta} r_1 \|\bar{\pi}_{k+1} - \bar{\pi}_k\|^2.$$

Term IV: By $\bar{p}_{k+1} = \bar{p}_k + \mu(p_{k+1} - \bar{p}_k)$, we have

$$\chi(\pi_{k+1}, p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_k) - \chi(\pi_{k+1}, p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_{k+1}) = \frac{\mu - 2}{2\mu} r_2 \|\bar{p}_{k+1} - \bar{p}_k\|^2.$$

Combining all the above bounds leads to the conclusion. □

Lemma C.6 (Lemma 6 in (Zheng et al., 2023)). *For any $k \geq 0$, the following inequality holds:*

$$\begin{aligned}\varphi_\pi(p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_{k+1}) \geq \varphi_\pi(p_k, \bar{\pi}_k, \bar{p}_k) &+ \frac{(2-\mu)r_2}{2\mu}\|\bar{p}_{k+1} - \bar{p}_k\|^2 \\ &+ \frac{r_1}{2}\langle \bar{\pi}_{k+1} + \bar{\pi}_k - 2\pi(p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_k), \bar{\pi}_{k+1} - \bar{\pi}_k\rangle \\ &+ \langle \nabla_p \chi(\pi(p_k, \bar{\pi}_k, \bar{p}_k), p_k, \bar{\pi}_k, \bar{p}_k), p_{k+1} - p_k\rangle - \frac{L_{\phi\pi}}{2}\|p_{k+1} - p_k\|^2.\end{aligned}$$

Proof. We split the target as follows:

$$\begin{aligned}&\varphi_\pi(p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_{k+1}) - \varphi_\pi(p_k, \bar{\pi}_k, \bar{p}_k) \\ &= \underbrace{\varphi_\pi(p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_{k+1}) - \varphi_\pi(p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_k)}_{\text{I}} + \underbrace{\varphi_\pi(p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_k) - \varphi_\pi(p_{k+1}, \bar{\pi}_k, \bar{p}_k)}_{\text{II}} \\ &+ \underbrace{\varphi_\pi(p_{k+1}, \bar{\pi}_k, \bar{p}_k) - d(y^t, \bar{\pi}_k, \bar{p}_k)}_{\text{III}}.\end{aligned}$$

Term I: By the update rule of $\bar{p}_{k+1}$, we have

$$\text{I} = \frac{r_2}{2}\left(\|p_{k+1} - \bar{p}_k\|^2 - \|p_{k+1} - \bar{p}_{k+1}\|^2\right) = \frac{(2-\mu)r_2}{2\mu}\|\bar{p}_{k+1} - \bar{p}_k\|^2.$$

Term II :

$$\begin{aligned}\text{II} &= \chi(\pi(p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_k), p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_k) - \chi(\pi(p_{k+1}, \bar{\pi}_k, \bar{p}_k), p_{k+1}, \bar{\pi}_k, \bar{p}_k) \\ &\geq \chi(\pi(p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_k), p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_k) - \chi(\pi(p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_k), p_{k+1}, \bar{\pi}_k, \bar{p}_k) \\ &= \frac{r_1}{2}\langle \bar{\pi}_{k+1} + \bar{\pi}_k - 2\pi(p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_k), \bar{\pi}_{k+1} - \bar{\pi}_k\rangle.\end{aligned}$$

Term III : It follows from φ_π is $L_{\phi\pi}$-smooth w.r.t. p that

$$\text{III} = \langle \nabla_p \chi(\pi(p_k, \bar{\pi}_k, \bar{p}_k), p_k, \bar{\pi}_k, \bar{p}_k), p_{k+1} - p_k\rangle - \frac{L_{\phi\pi}}{2}\|p_{k+1} - p_k\|^2.$$

Combining the above inequalities finishes the proof. □

Lemma C.7 (Lemma 7 in Zheng et al. (2023)). *For all $k \geq 0$, the following inequality holds:*

$$q(\bar{\pi}_k) \geq q(\bar{\pi}_{k+1}) + \frac{r_1}{2}\langle \bar{\pi}_k + \bar{\pi}_{k+1} - 2\pi(\bar{\pi}_k, v(\bar{\pi}_{k+1})), \bar{\pi}_k - \bar{\pi}_{k+1}\rangle.$$

Proof. From Sion's minimax theorem (Sion, 1958), we have

$$\begin{aligned}\varphi_{\pi,p,\bar{p}}(\bar{\pi}) &= \max_{\bar{p}} \varphi_{\pi,p}(\bar{\pi}, \bar{p}) = \max_{\bar{p}} \min_{\pi\in\Pi} \max_{p\in\mathcal{P}} \chi(\pi, p, \bar{\pi}, \bar{p}) \\ &= \max_{\bar{p}} \varphi_p(\pi(\bar{\pi}, \bar{p}), \bar{\pi}, \bar{p}) = \max_{\bar{p}} \chi(\pi(p(\bar{\pi}, \bar{p}), \bar{\pi}, \bar{p}), p(\bar{\pi}, \bar{p}), \bar{\pi}, \bar{p}).\end{aligned}$$

Thus, we have

$$\begin{aligned}&\varphi_{\pi,p,\bar{p}}(\bar{\pi}_k) - \varphi_{\pi,p,\bar{p}}(\bar{\pi}_{k+1}) \\ &= \varphi_p(\pi(\bar{\pi}_k, \bar{p}(\bar{\pi}_k)), \bar{\pi}_k, \bar{p}(\bar{\pi}_k)) - \varphi_p(\pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})), \bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) \\ &\geq \varphi_p(\pi(\bar{\pi}_k, \bar{p}(\bar{\pi}_{k+1})), \bar{\pi}_k, \bar{p}(\bar{\pi}_{k+1})) - \varphi_p(\pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})), \bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) \\ &\geq \varphi_p(\pi(\bar{\pi}_k, \bar{p}(\bar{\pi}_{k+1})), \bar{\pi}_k, \bar{p}(\bar{\pi}_{k+1})) - \varphi_p(\pi(\bar{\pi}_k, \bar{p}(\bar{\pi}_{k+1})), \bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) \\ &\geq \chi(\pi(\bar{\pi}_k, \bar{p}(\bar{\pi}_{k+1})), p(\pi(\bar{\pi}_k, \bar{p}(\bar{\pi}_{k+1})), \bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})), \bar{\pi}_k, \bar{p}(\bar{\pi}_{k+1})) \\ &\quad - \chi(\pi(\bar{\pi}_k, \bar{p}(\bar{\pi}_{k+1})), p(\pi(\bar{\pi}_k, \bar{p}(\bar{\pi}_{k+1})), \bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})), \bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) \\ &= \frac{r_1}{2}\langle \bar{\pi}_k + \bar{\pi}_{k+1} - 2\pi(\bar{\pi}_k, \bar{p}(\bar{\pi}_{k+1})), \bar{\pi}_k - \bar{\pi}_{k+1}\rangle.\end{aligned}$$

□

Proof of Proposition 4.1

Proposition C.1 (Proposition 1 in Zheng et al. (2023)). *Define a Lyapunov function*

$$\Phi(\pi,p,\bar{\pi},\bar{p}) := \underbrace{\chi(\pi,p,\bar{\pi},\bar{p})-\varphi_{\pi}(p,\bar{\pi},\bar{p})}_{\textit{Primal descent}}+\underbrace{\varphi_{\pi,p}(\bar{\pi},\bar{p})-\varphi_{\pi}(p,\bar{\pi},\bar{p})}_{\textit{Dual ascent}}$$
$$+\underbrace{\varphi_{\pi,p,\bar{p}}(\bar{\pi})-\varphi_{\pi,p}(\bar{\pi},\bar{p})}_{\textit{Proximal ascent}}+\underbrace{\varphi_{\pi,p,\bar{p}}(\bar{\pi})-\underline{\chi}}_{\textit{Proximal descent}}+\underline{\chi},$$

and let parameters satisfy the following properties:

$$\begin{aligned}
&r_1 \geq 2L,\ \ r_2 \geq 2\lambda L,\ \ \ell_{p\pi}=\lambda\ell_{\pi p}=\lambda L\\
&0<\tau\leq\min\left\{\frac{4}{3(L+r_1)},\frac{1}{6\lambda L}\right\}, 0<\sigma\leq\min\left\{\frac{2}{3\lambda L\kappa_6^2},\frac{1}{6L_{\phi\pi}},\frac{1}{5\lambda\sqrt{\lambda+5}L}\right\},\\
&0<\beta\leq\min\left\{\frac{24r_1}{360r_1+5r_1^2\lambda+(2\lambda L+5r_1)^2},\frac{\sigma\lambda^2L^2}{384r_1(\lambda+5)(\lambda+1)^2}\right\},\\
&0<\mu\leq\min\left\{\frac{(\lambda+5)}{2(\lambda+5)+\lambda^2L^2},\frac{(\lambda+5)}{\sigma\lambda^2L^2},\frac{\sigma\lambda^2L^2}{64r_2(\lambda+5)}\right\}.
\end{aligned}$$

Then for any $k>0$,

$$\begin{aligned}
\Phi_k-\Phi_{k+1}\geq\ &\frac{r_1}{32}\|\pi_{k+1}-\pi_k\|^2+\frac{r_2}{15}\|p_k-p_k^+(\bar{\pi}_k,\bar{p}_k)\|^2+\frac{r_1}{5\beta}\|\bar{\pi}_k-\bar{\pi}_{k+1}\|^2\\
&+\frac{r_2}{4\mu}\|\bar{p}_k^+(\bar{\pi}_{k+1})-\bar{p}_k\|^2-4r_1\beta\|\pi(\bar{\pi}_{k+1},\bar{p}(\bar{\pi}_{k+1}))-\pi(\bar{\pi}_{k+1},\bar{p}_k^+(\bar{\pi}_{k+1}))\|^2
\end{aligned}\tag{44}$$

where $\kappa_6 \coloneqq (2\tau\cdot r_1+1)/(c(r_1-L))$.

Proof. It follows from Lemma C.5, Lemma C.6, Lemma C.7 that

$$\begin{aligned}
&\Phi(\pi_k,p_k,\bar{\pi}_k,\bar{p}_k)\\
&\geq\Phi(\pi_{k+1},p_{k+1},\bar{\pi}_{k+1},\bar{p}_{k+1})+\left(\frac{1}{\tau}-\frac{\ell_{\pi p}+r_1}{2}\right)\|\pi_{k+1}-\pi_k\|^2\\
&+\left(\frac{r_2-\ell_{p\pi}}{2}-L_{\phi\pi}\right)\|p_{k+1}-p_k\|^2+\frac{(2-\beta)r_1}{2\beta}\|\bar{\pi}_{k+1}-\bar{\pi}_k\|^2\\
&+\frac{(2-\mu)r_2}{2\mu}\|\bar{p}_{k+1}-\bar{p}_k\|^2+\underbrace{\langle\nabla_p\chi(\pi_{k+1},p_k,\bar{\pi}_k,\bar{p}_k),p_{k+1}-p_k\rangle}_{\text{I}}\\
&+\underbrace{2\langle\nabla_p\chi(\pi(p_k,\bar{\pi}_k,\bar{p}_k),p_k,\bar{\pi}_k,\bar{p}_k)-\nabla_p\chi(\pi_{k+1},p_k,\bar{\pi}_k,\bar{p}_k),p_{k+1}-p_k\rangle}_{\text{II}}\\
&+\underbrace{2r_1\langle\pi(\bar{\pi}_k,\bar{p}(\bar{\pi}_{k+1}))-\pi(p_{k+1},\bar{\pi}_{k+1},\bar{p}_k),\bar{\pi}_{k+1}-\bar{\pi}_k\rangle}_{\text{III}}.
\end{aligned}$$

Term I: using the projection update of p_{k+1}, we have

$$\langle\nabla_p\chi(\pi_{k+1},p_k,\bar{\pi}_k,\bar{p}_k),p_{k+1}-p_k\rangle\geq\frac{1}{\sigma}\|p_k-p_{k+1}\|^2.$$

Term II: Due to the Lipschitz gradient property and error bound Item a) in Lemma C.4:

$$\begin{aligned}
&2\langle\nabla_p\chi(\pi(p_k,\bar{\pi}_k,\bar{p}_k),p_k,\bar{\pi}_k,\bar{p}_k)-\nabla_p\chi(\pi_{k+1},p_k,\bar{\pi}_k,\bar{p}_k),p_{k+1}-p_k\rangle\\
&\geq-2\ell_{p\pi}\|\pi(p_k,\bar{\pi}_k,\bar{p}_k)-\pi_{k+1}\|\|p_{k+1}-p_k\|\\
&\geq-\ell_{p\pi}\kappa_6^2\|p_{k+1}-p_k\|^2-\ell_{p\pi}\kappa_6^{-2}\|\pi(p_k,\bar{\pi}_k,\bar{p}_k)-\pi_{k+1}\|^2\\
&\geq-\ell_{p\pi}\kappa_6^2\|p_{k+1}-p_k\|^2-\ell_{p\pi}\|\pi_{k+1}-\pi_k\|^2.
\end{aligned}$$

Term III: for any $\kappa > 0$, note that

$$\begin{aligned}
&2r_1\langle \pi(\bar{\pi}_k, \bar{p}(\bar{\pi}_{k+1})) - \pi(p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_k), \bar{\pi}_{k+1} - \bar{\pi}_k\rangle \\
&= 2r_1\langle \pi(\bar{\pi}_k, \bar{p}(\bar{\pi}_{k+1})) - \pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})), \bar{\pi}_{k+1} - \bar{\pi}_k\rangle + 2r_1\langle \pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) - \pi(p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_k), \bar{\pi}_{k+1} - \bar{\pi}_k\rangle \\
&\geq -2r_1\kappa_2\|\bar{\pi}_{k+1} - \bar{\pi}_k\|^2 - \frac{r_1}{\kappa}\|\bar{\pi}_{k+1} - \bar{\pi}_k\|^2 - r_1\kappa\|\pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) - \pi(p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_k)\|^2,
\end{aligned}$$

where the inequality holds by Cauchy-Schwarz inequality, Young's inequality and Item 3 in Lemma C.2. Hence,

$$\begin{aligned}
&\Phi(\pi_k, p_k, \bar{\pi}_k, \bar{p}_k) - \Phi(\pi_{k+1}, p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_{k+1}) \\
&\geq \left(\frac{1}{\tau} - \frac{\ell_{\pi p} + r_1}{2} - \ell_{p\pi}\right)\|\pi_{k+1} - \pi_k\|^2 + \left(\frac{1}{\sigma} + \frac{r_2 - \ell_{p\pi}}{2} - L_{\phi\pi} - \ell_{p\pi}\kappa_6^2\right)\|p_{k+1} - p_k\|^2 \\
&\quad + r_1\left(\frac{2-\beta}{2\beta} - 2\kappa_2 - \frac{1}{\kappa}\right)\|\bar{\pi}_{k+1} - \bar{\pi}_k\|^2 + \frac{(2-\mu)r_2}{2\mu}\|\bar{p}_{k+1} - \bar{p}_k\|^2 \\
&\quad - r_1\kappa\underbrace{\|\pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) - \pi(p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_k)\|^2}_{\text{IV}}.
\end{aligned} \tag{45}$$

Now, we give a upper bound of IV as follows. Note that

$$\begin{aligned}
&\|\pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) - \pi(p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_k)\|^2 \\
&\leq 2\|\pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) - \pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1}))\|^2 + 2\|\pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1})) - \pi(p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_k)\|^2 \\
&\overset{(a)}{\leq} 2\|\pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) - \pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1}))\|^2 + 2\kappa_1^2\|p_{k+1} - p(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1}))\|^2,
\end{aligned} \tag{46}$$

where (a) holds by Item 1 in Lemma C.2, and

$$\begin{aligned}
&\|p_{k+1} - p(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1}))\|^2 \\
&\leq 3\|p_{k+1} - p(\bar{\pi}_k, \bar{p}_k)\|^2 + 3\|p(\bar{\pi}_k, \bar{p}_k) - p(\bar{\pi}_{k+1}, \bar{p}_k)\|^2 + 3\|p(\bar{\pi}_{k+1}, \bar{p}_k) - p(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1}))\|^2 \\
&\overset{(b)}{\leq} 6\ell_{p\pi}^2\sigma^2\kappa_6^2\|\pi_k - \pi_{k+1}\|^2 + 6(\kappa_8+1)^2\|p_k - p_k^+(\bar{\pi}_k, \bar{p}_k)\|^2 + 3\kappa_3^2\|\bar{\pi}_k - \bar{\pi}_{k+1}\|^2 + 3\kappa_5^2\|\bar{p}_k - \bar{p}_+^t(\bar{\pi}_{k+1})\|^2,
\end{aligned} \tag{47}$$

where (b) follows from Item 4 and Item 7 in Lemma C.2 and Item c) and Item d) in Lemma C.4. Putting (46) and (47) together, we have

$$\begin{aligned}
&\|\pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) - \pi(p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_k)\|^2 \\
&\leq 2\|\pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) - \pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1}))\|^2 + 12\ell_{p\pi}^2\sigma^2\kappa_1^2\kappa_6^2\|\pi_k - \pi_{k+1}\|^2 \\
&\quad + 12\kappa_1^2(\kappa_8+1)^2\|p_k - p_k^+(\bar{\pi}_k, \bar{p}_k)\|^2 + 6\kappa_1^2\kappa_3^2\|\bar{\pi}_k - \bar{\pi}_{k+1}\|^2 + 6\kappa_1^2\kappa_5^2\|\bar{p}_k - \bar{p}_k^+(\bar{\pi}_{k+1})\|^2.
\end{aligned} \tag{48}$$

Combining (45) and (48) and letting $z_1 := \frac{1}{\tau} - \frac{\ell_{\pi p}+r_1}{2} - \ell_{p\pi}$, $z_2 := \frac{1}{\sigma} + \frac{r_2-\ell_{p\pi}}{2} - L_{\phi\pi} - \ell_{p\pi}\kappa_6^2$, and $z_3 := r_1\left(\frac{2-\beta}{2\beta} - 2\kappa_2 - \frac{1}{\kappa}\right)$, yields

$$\begin{aligned}
&\Phi(\pi_k, p_k, \bar{\pi}_k, \bar{p}_k) - \Phi(\pi_{k+1}, p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_{k+1}) \\
&\geq z_1\|\pi_{k+1} - \pi_k\|^2 + z_2\|p_{k+1} - p_k\|^2 + z_3\|\bar{\pi}_{k+1} - \bar{\pi}_k\|^2 + \frac{(2-\mu)r_2}{2\mu}\|\bar{p}_{k+1} - \bar{p}_k\|^2 \\
&\quad - 12r_1\kappa\ell_{p\pi}^2\sigma^2\kappa_1^2\kappa_6^2\|\pi_k - \pi_{k+1}\|^2 - 12r_1\kappa\kappa_1^2(\kappa_8+1)^2\|p_k - p_k^+(\bar{\pi}_k, \bar{p}_k)\|^2 - 6r_1\kappa\kappa_1^2\kappa_3^2\|\bar{\pi}_k - \bar{\pi}_{k+1}\|^2 \\
&\quad - 6r_1\kappa\kappa_1^2\kappa_5^2\|\bar{p}_k - \bar{p}_k^+(\bar{\pi}_{k+1})\|^2 - 2r_1\kappa\|\pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) - \pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1}))\|^2 \\
&= \left(z_1 - 12r_1\kappa\ell_{p\pi}^2\sigma^2\kappa_1^2\kappa_6^2\right)\|\pi_k - \pi_{k+1}\|^2 + z_2\|p_{k+1} - p_k\|^2 + \left(z_3 - 6r_1\kappa\kappa_1^2\kappa_3^2\right)\|\bar{\pi}_{k+1} - \bar{\pi}_k\|^2 \\
&\quad + \frac{(2-\mu)r_2}{2\mu}\|\bar{p}_{k+1} - \bar{p}_k\|^2 - 6r_1\kappa\kappa_1^2\kappa_5^2\|\bar{p}_k - \bar{p}_k^+(\bar{\pi}_{k+1})\|^2 - 12r_1\kappa\kappa_1^2(\kappa_8+1)^2\|p_k - p_k^+(\bar{\pi}_k, \bar{p}_k)\|^2 \\
&\quad - 2r_1\kappa\|\pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) - \pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1}))\|^2.
\end{aligned} \tag{49}$$

Note that, by the Item d) in Lemma C.4, we have

$$\begin{aligned}\|p_{k+1}-p_k\|^2 &\geq \frac{1}{2}\|p_k - p_k^+(\bar{\pi}_k,\bar{p}_k)\|^2 - \|p_{k+1}-p_k^+(\bar{\pi}_k,\bar{p}_k)\|^2 \\ &\geq \frac{1}{2}\|p_k - p_k^+(\bar{\pi}_k,\bar{p}_k)\|^2 - \ell_{p\pi}^2\sigma^2\kappa_6^2\|\pi_k-\pi_{k+1}\|^2.\end{aligned} \tag{50}$$

Similarly, we provide a lower bound for $\|\bar{p}_{k+1}-\bar{p}_k\|^2$ as follows:

$$\begin{aligned}\|\bar{p}_{k+1}-\bar{p}_k\|^2 &\geq \frac{1}{2}\|\bar{p}_k - \bar{p}_k^+(\bar{\pi}_{k+1})\|^2 - \|\bar{p}_{k+1}-\bar{p}_k^+(\bar{\pi}_{k+1})\|^2 \\ &\geq \frac{1}{2}\|\bar{p}_k - \bar{p}_k^+(\bar{\pi}_{k+1})\|^2 - \mu^2\Big(4\ell_{p\pi}^2\sigma^2\kappa_6^2\|\pi_k-\pi_{k+1}\|^2 + 4(\kappa_8+1)^2\|p_k-p_k^+(\bar{\pi}_k,\bar{p}_k)\|^2 \\ &\quad + 2\kappa_3^2\|\bar{\pi}_k-\bar{\pi}_{k+1}\|^2\Big).\end{aligned} \tag{51}$$

Recalling that $\bar{p}_k^+(\bar{\pi}_{k+1}) = \bar{p}_k + \mu(p(\bar{\pi}_{k+1},\bar{p}_k)-\bar{p}_k)$, the last inequality can be obtained by using Item 4 in Lemma C.2 and Item c) and Item d) in Lemma C.5, i.e.,

$$\begin{aligned}&\|\bar{p}_{k+1}-\bar{p}_k^+(\bar{\pi}_{k+1})\|^2 = \mu^2\|p_{k+1}-p(\bar{\pi}_{k+1},\bar{p}_k)\|^2 \\ &\leq\ \mu^2\left(2\|p_{k+1}-p(\bar{\pi}_k,\bar{p}_k)\|^2 + 2\|p(\bar{\pi}_k,\bar{p}_k)-p(\bar{\pi}_{k+1},\bar{p}_k)\|^2\right) \\ &\leq\ \mu^2\left(4\ell_{p\pi}^2\sigma^2\kappa_6^2\|\pi_k-\pi_{k+1}\|^2 + 4(\kappa_8+1)^2\|p_k-p_k^+(\bar{\pi}_k,\bar{p}_k)\|^2 + 2\kappa_3^2\|\bar{\pi}_k-\bar{\pi}_{k+1}\|^2\right).\end{aligned}$$

Substituting (50) and (51) to (49) yields

$$\begin{aligned}&\Phi(\pi_k,p_k,\bar{\pi}_k,\bar{p}_k) - \Phi(\pi_{k+1},p_{k+1},\bar{\pi}_{k+1},\bar{p}_{k+1}) \\ &\geq \underbrace{\left(z_1 - (12r_1\kappa\kappa_1^2 + s_2 + 2\mu(2-\mu)r_2)\ell_{p\pi}^2\sigma^2\kappa_6^2\right)}_{\text{III}}\|\pi_{k+1}-\pi_k\|^2 \\ &\quad + \underbrace{\left(\frac{z_2}{2} - (12r_1\kappa\kappa_1^2 + 2\mu(2-\mu)r_2)(1+\kappa_8)^2\right)}_{\text{II}}\|p_k-p_k^+(\bar{\pi}_k,\bar{p}_k)\|^2 \\ &\quad + \underbrace{\left(z_3 - (\mu(2-\mu)r_2 + 6r_1\kappa\kappa_1^2)\kappa_3^2\right)}_{\text{I}}\|\bar{\pi}_k-\bar{\pi}_{k+1}\|^2 + \left(\frac{(2-\mu)r_2}{4\mu} - 6r_1\kappa\kappa_1^2\kappa_5^2\right)\|\bar{p}_k-\bar{p}_k^+(\bar{\pi}_{k+1})\|^2 \\ &\quad - 2r_1\kappa\|\pi(\bar{\pi}_{k+1},\bar{p}(\bar{\pi}_{k+1})) - \pi(\bar{\pi}_{k+1},\bar{p}_k^+(\bar{\pi}_{k+1}))\|^2.\end{aligned}$$

Now, we give the upper bound of I, II, III, separately.

Term I: Based on the parameters setting, we have

$$\begin{aligned}r_1 \geq 2\ell_{\pi p} \Rightarrow \kappa_1 = \frac{\ell_{p\pi}}{r_1-\ell_{\pi p}} + 1 \leq \frac{\ell_{p\pi}}{\ell_{\pi p}} + 1 = \lambda + 1, &\text{and } \kappa_2 = \frac{r_1}{r_1-\ell_{\pi p}} \leq 2, \\ r_2 \geq 2\ell_{p\pi} \Rightarrow \kappa_3 = \frac{r_1\kappa_1}{r_2-\ell_{p\pi}} + \frac{\kappa_2}{\kappa_1} \leq \frac{r_1(\lambda+1)}{\ell_{p\pi}} + 2, &\text{and } \kappa_5 = \frac{r_2}{r_2-\ell_{p\pi}} \leq 2.\end{aligned}$$

Then we obtain:

$$\frac{(2-\mu)r_2}{4\mu} - 6r_1\kappa\kappa_1^2\kappa_5^2 \overset{(a)}{\geq} \frac{r_2}{2\mu} - \frac{r_2}{4} - 48(\lambda+1)^2r_1\beta \overset{(b)}{\geq} \frac{r_2}{\mu}\left(\frac{1}{2} - \frac{\mu}{4} - \frac{\ell_{p\pi}^2\sigma\mu}{8(\lambda+5)}\right) \overset{(c)}{\geq} \frac{r_2}{4\mu},$$

where (a) holds by $\kappa_2,\kappa_5 \leq 2, \kappa = 2\beta$ and $\kappa_1 \leq \lambda+1$, (b) holds by the definition of β and $\ell_{p\pi} = \lambda L$ and (c) follows

from the defintion of μ and $\mu \leq \frac{1}{2}$. Similarly, we have

$$\begin{aligned}
&\mu(2-\mu)r_2 + 6r_1\kappa\kappa_1^2 \overset{(a)}{\leq} 2r_2\mu + 12r_1\beta(\lambda+1)^2 \overset{(b)}{\leq} \frac{\sigma\ell_{p\pi}^2}{16(\lambda+5)} \\
\Rightarrow \mathrm{I} &= z_3 - \left(\mu(2-\mu)r_2 + 6r_1\kappa\kappa_1^2\right)\kappa_3^2 = r_1\left(\frac{2-\beta}{2\beta} - 2\kappa_2 - \frac{1}{\kappa}\right) - \left(\mu(2-\mu)r_2 + 6r_1\kappa\kappa_1^2\right)\kappa_3^2 \\
&\overset{(c)}{\geq} r_1\left(\frac{1}{\beta} - \frac{1}{2} - 4 - \frac{1}{2\beta}\right) - \frac{\ell_{p\pi}^2}{16(\lambda+5)}\left(\frac{r_1^2(\lambda+1)^2}{\ell_{p\pi}^2} + 4 + \frac{4r_1(\lambda+1)}{\ell_{p\pi}}\right) \\
&\overset{(d)}{\geq} \frac{r_1}{\beta}\left(\frac{1}{2} - \left(\frac{9}{2} + \frac{r_1(\lambda+1)}{16} + \frac{\ell_{p\pi}^2}{20r_1} + \frac{\ell_{p\pi}}{4}\right)\beta\right) \overset{(f)}{\geq} \frac{r_1}{5\beta},
\end{aligned}$$

where (a) holds by $\kappa = 2\beta$ and $\kappa_1 \leq \lambda + 1$, (b) follows from $\mu \leq (\sigma\lambda^2 L^2)/(64r_2(\lambda+5))$, (c) is by $\kappa_2 \leq 2$, (b) and the definition of κ_3 and $(d), (f)$ follow from the definition of β.

Term II: Based on the parameters setting, we have

$$\begin{aligned}
&\tau^{-1} \geq \max\left\{\frac{3}{4}(\ell_{\pi p} + r_1), 6\ell_{p\pi}\right\} \Rightarrow z_1 \geq \frac{1}{6\tau}, \text{and } \kappa_6 = \frac{2r_1 + \frac{1}{\tau}}{r_1 - \ell_{\pi p}} \geq 2 + \frac{1}{\tau \cdot r_1} \geq 2 + \frac{\frac{3}{4}(\ell_{\pi p} + r_1)}{r_1} \geq \frac{11}{4} \\
&\frac{1}{\sigma} \geq \max\left\{\frac{3}{2}\ell_{p\pi}\kappa_6^2, 6L_{\phi\pi}, 5\sqrt{\lambda+5}\ell_{p\pi}\right\} \Rightarrow z_2 \geq \frac{1}{6\sigma} + \frac{r_2 - \ell_{p\pi}}{2} \text{and } \kappa_8 = \frac{\frac{1}{\sigma} + L_{\phi\pi}}{r_2 - \ell_{p\pi}} \leq \frac{2}{\sigma r_2} + \lambda + 4
\end{aligned}$$

Then we have:

$$\begin{aligned}
\mathrm{II} &= \frac{z_2}{2} - \left(12r_1\kappa\kappa_1^2 + 2\mu(2-\mu)r_2\right)(1+\kappa_8)^2 \\
&\geq \frac{1}{12\sigma} + \frac{r_2 - \ell_{p\pi}}{4} - \frac{\sigma\ell_{p\pi}^2}{8(\lambda+5)}\left(\frac{4}{\sigma^2 r_2^2} + (\lambda+5)^2 + \frac{4(\lambda+5)}{\sigma r_2}\right) \\
&\geq \frac{1}{12\sigma} + \frac{\ell_{p\pi}}{4} - \frac{1}{8(\lambda+5)\sigma} - \frac{(\lambda+5)\sigma\ell_{p\pi}^2}{8} - \frac{\ell_{p\pi}}{4} \\
&\geq \frac{1}{60\sigma} - \frac{(\lambda+5)\sigma\ell_{p\pi}^2}{8} \geq \frac{1}{90\sigma} \geq \frac{r_2}{15}
\end{aligned}$$

Term III:

$$\begin{aligned}
\mathrm{III} &= z_1 - (12r_1\kappa\kappa_1^2 + s_2 + 2\mu(2-\mu)r_2)\ell_{p\pi}^2\sigma^2\kappa_6^2 \geq \frac{1}{6\tau} - \frac{\sigma^3\ell_{p\pi}^4\kappa_6^2}{8(\lambda+5)} - \frac{2\ell_{p\pi}}{3} \\
&\geq \frac{1}{6\tau} - \frac{13\ell_{p\pi}}{2500} - \frac{2\ell_{p\pi}}{3} \geq \frac{1}{6\tau} - \frac{7\ell_{p\pi}}{10} \geq \frac{1}{24c} \geq \frac{r_1}{32},
\end{aligned}$$

where the first inequality is due to $2/(3\sigma) \geq \ell_{p\pi}\kappa_6^2$ and

$$s_2 \leq \frac{1}{\sigma} + \frac{r_2 - \ell_{p\pi}}{2} - \ell_{p\pi}\kappa_6^2 - L_{\phi\pi} = \frac{1}{\sigma} - \frac{r_2 + \ell_{p\pi}}{2} - \ell_{p\pi}\kappa_6^2 - (\kappa_1 + 1)\ell_{p\pi} \leq \frac{1}{\sigma}.$$

Putting together all the pieces, we get

$$\begin{aligned}
&\Phi(\pi_k, p_k, \bar{\pi}_k, \bar{p}_k) - \Phi(\pi_{k+1}, p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_{k+1}) \\
&\geq \frac{r_1}{32}\|\pi_{k+1} - \pi_k\|^2 + \frac{r_2}{15}\|p_k - p_k^+(\bar{\pi}_k, \bar{p}_k)\|^2 + \frac{r_1}{5\beta}\|\bar{\pi}_k - \bar{\pi}_{k+1}\|^2 + \frac{r_2}{4\mu}\|\bar{p}_k^+(\bar{\pi}_{k+1}) - \bar{p}_k\|^2 \\
&\quad - 4r_1\beta\|\pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) - \pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1}))\|^2.
\end{aligned}$$

□

Proof of Proposition 4.2

Proposition C.2. *Under the setting of Proposition 4.1, then we have*

$$\|\pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1})) - \pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1}))\|^2 \leq \omega_0 \|\bar{p}_k^+(\bar{\pi}_{k+1}) - \bar{p}_k\|, \tag{52}$$

where $\omega_0 := \frac{2}{(r_1 - L)\tau}\left(\frac{r_2(1-\mu)}{\mu} + \frac{r_2^2}{r_2 - \lambda L}\right)$. *Moreover, we have*

$$\|\pi^*(\bar{\pi}) - \pi(\bar{\pi}, \bar{p})\|^2 \leq \omega_1 \|\bar{p} - p(\bar{\pi}, \bar{p})\|, \tag{53}$$

where $\omega_1 := \frac{2r_2}{\tau(r_1 - \ell_{\pi p})}$.

Proof. Inequality (52): Since $\varphi_p(\pi, \cdot, \cdot)$ is $(r_1 - \ell_{\pi p})$-strongly convex, we have

$$\begin{aligned}
&\max_{\bar{p}} \varphi_p(\pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1})), \bar{\pi}_{k+1}, \bar{p}) - \varphi_p(\pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})), \bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) \\
&\geq \varphi_p(\pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1})), \bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) - \varphi_p(\pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})), \bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) \\
&\geq \frac{r_1 - \ell_{\pi p}}{2} \|\pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1})) - \pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1}))\|^2.
\end{aligned} \tag{54}$$

Since $J_\rho(\cdot, p)$ is $\ell_{p\pi}$-smooth, namely $-J_\rho(\cdot, p)$ is $\ell_{p\pi}$-weakly convex w.r.t. p and $-J_\rho(\cdot, p)$ is gradient dominace (Lemma B.1). By Lemma 4.1, there exists a constant $C_{\varphi p} > 0$ such that

$$\begin{aligned}
&\max_{\bar{p}} \varphi_p(\pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1})), \bar{\pi}_{k+1}, \bar{p}) - \varphi_p(\pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})), \bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) \\
&\leq \max_{\bar{p}} \varphi_p(\pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1})), \bar{\pi}_{k+1}, \bar{p}) - \varphi_p(\pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1})), \bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1})) \\
&\leq C_{\varphi p} \|\nabla_{\bar{p}} \varphi_p(\pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1})), \bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1}))\|.
\end{aligned} \tag{55}$$

Next, we further bound the right-hand part.

$$\begin{aligned}
&\|\nabla_{\bar{p}} \varphi_p(\pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1})), \bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1}))\| \\
&= r_2 \|\bar{p}_k^+(\bar{\pi}_{k+1}) - p(\pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1})), \bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1}))\| \\
&= r_2 \|(1-\mu)\bar{p}_k + \mu p(\bar{\pi}_{k+1}, \bar{p}_k) - p(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1}))\| \\
&\leq r_2(1-\mu)\|\bar{p}_k - p(\bar{\pi}_{k+1}, \bar{p}_k)\| + r_2 \|p(\bar{\pi}_{k+1}, \bar{p}_k) - p(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1}))\| \\
&\overset{(a)}{\leq} r_2 \left(\frac{1-\mu}{\mu} + \kappa_5\right) \|\bar{p}_k^+(\bar{\pi}_{k+1}) - \bar{p}_k\|,
\end{aligned} \tag{56}$$

where (a) holds by Item 6 in Lemma C.2 and the definition of $\bar{p}_k^+(\bar{\pi}_{k+1})$. Combing (54), (55), and (56), we have

$$\|\pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1})) - \pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1}))\|^2 \leq \frac{2}{(r_1 - \ell_{\pi p})C_{\varphi p}} \left(\frac{r_2(1-\mu)}{\mu} + r_2\kappa_5\right) \|\bar{p}_k^+(\bar{\pi}_{k+1}) - \bar{p}_k\|.$$

Inequality (53): Since $\phi_p(\pi; \bar{\pi})$ is $(r_1 - \ell_{\pi p})$-strongly convex w.r.t. π, we have

$$\psi_p(\pi(\bar{\pi}, \bar{p}); \bar{\pi}) - \psi_p(\pi^*(\bar{\pi}); \bar{\pi}) \geq \frac{r_1 - \ell_{\pi p}}{2} \|\pi^*(\bar{\pi}) - \pi(\bar{\pi}, \bar{p})\|^2.$$

On the other hand, we have

$$\begin{aligned}
&\psi_p(\pi(\bar{\pi}, \bar{p}); \bar{\pi}) - \psi_p(\pi^*(\bar{\pi}); \bar{\pi}) \\
&= \psi_p(\pi(\bar{\pi}, \bar{p}); \bar{\pi}) - \left(\max_{p \in \mathcal{P}} J_\rho(\pi^*(\bar{\pi}), p) + \frac{r_1}{2}\|\pi^*(\bar{\pi}) - \bar{\pi}\|^2\right) \\
&\overset{(a)}{\leq} \psi_p(\pi(\bar{\pi}, \bar{p}); \bar{\pi}) - \left(J_\rho(\pi(\bar{\pi}, \bar{p}), p(\bar{\pi}, p)) + \frac{r_1}{2}\|\pi(\bar{\pi}, \bar{p}) - \bar{\pi}\|^2\right) \\
&= \max_{p \in \mathcal{P}} J_\rho(\pi(\bar{\pi}, \bar{p}), p) - J_\rho(\pi(\bar{\pi}, \bar{p}), p(\bar{\pi}, p)),
\end{aligned}$$

where (a) holds by the definition of $\pi(\bar{\pi},\bar{p})$ and $\pi^*(\bar{\pi})$. It follows from Lemma B.1 that

$$\max_{p\in\mathcal{P}} J_\rho(\pi(\bar{\pi},\bar{p}),p) - J_\rho(\pi(\bar{\pi},\bar{p}),p(\bar{\pi},\bar{p})) \le \bar{D}_p \operatorname{dist}\left(\nabla_p J_\rho(\pi,p(\bar{\pi},\bar{p})) - \mathcal{N}_{\mathcal{P}}(p(\bar{\pi},\bar{p}))\right) \overset{(a)}{\le} \bar{D}_p \|r_2(\bar{p}-p(\bar{\pi},\bar{p}))\|,$$

where (a) holds by $p(\bar{\pi},\bar{p}) \in \operatorname{argmax}_{p\in\mathcal{P}} J_\rho(\pi(\bar{\pi},\bar{p}),p) - \frac{r_1}{2}\|p-\bar{p}\|^2$. □

Lemma C.8. *For Lyapunov function Φ, we have*

$$\Phi(\pi_0,p_0,\bar{\pi}_0,\bar{p}_0) - \underline{\chi} \le L_\pi\|\pi_0-\pi^*\| + L_p\|p_0-p^*\| + L_\pi\|\pi(\bar{\pi}_0)-\pi(p_0,\bar{\pi}_0,\bar{p}_0)\| + L_p\|p(\bar{\pi}_0)-p_0\|. \tag{57}$$

Proof. Since $\Phi(\pi_0,p_0,\bar{\pi}_0,\bar{p}_0) - \underline{\chi} = \chi(\pi_0,p_0,\bar{\pi}_0,\bar{p}_0) - 2\varphi_\pi(p_0,\bar{\pi}_0,\bar{p}_0) + 2\varphi_{\pi,p,\bar{p}}(\bar{\pi}_0) - \chi(\pi^*,p^*,\bar{\pi}^*,\bar{p}^*)$, then we have

$$\begin{aligned}
&\Phi(\pi_0,p_0,\bar{\pi}_0,\bar{p}_0) - \underline{\chi}\\
&= \chi(\pi_0,p_0,\bar{\pi}_0,\bar{p}_0) - \chi(\pi^*,p^*,\bar{\pi}^*,\bar{p}^*)\\
&\quad + 2\big(\varphi_{\pi,p,\bar{p}}(\bar{\pi}_0) - \varphi_\pi(p_0,\bar{\pi}_0,\bar{p}_0)\big)\\
&= J_\rho(\pi_0,p_0) - J_\rho(\pi^*,p^*) + \frac{r_1}{2}\big(\|\pi_0-\bar{\pi}_0\|^2 - \|\pi^*-\bar{\pi}^*\|^2\big)\\
&\quad - \frac{r_2}{2}\big(\|p_0-\bar{p}_0\|^2 - \|p^*-\bar{p}^*\|^2\big)\\
&\quad + 2\Big(\chi\big(\pi(\bar{\pi}_0),p(\bar{\pi}_0),\bar{\pi}_0,\bar{p}(\bar{\pi}_0)\big) - \chi\big(\pi(p_0,\bar{\pi}_0,\bar{p}_0),p_0,\bar{\pi}_0,\bar{p}_0\big)\Big)
\end{aligned}$$

It is easy to know that $\pi^*=\bar{\pi}^*, p^*=\bar{p}^*$ since $\bar{\pi}^* = \operatorname{argmin}_{\bar{\pi}} J_\rho(\pi^*,p^*) + \frac{r_1}{2}\|\pi^*-\bar{\pi}\|^2$ and $\bar{p}^* = \operatorname{argmax}_{\bar{p}} J_\rho(\pi^*,p^*) - \frac{r_2}{2}\|p^*-\bar{p}^*\|^2$. Hence, we have

$$\begin{aligned}
&\Phi(\pi_0,p_0,\bar{\pi}_0,\bar{p}_0) - \underline{\chi}\\
&\le L_\pi\|\pi_0-\pi^*\| + L_p\|p_0-p^*\| + \frac{r_1}{2}\|\pi_0-\bar{\pi}_0\|^2\\
&\quad + 2\Big(\chi\big(\pi(\bar{\pi}_0),p(\bar{\pi}_0),\bar{\pi}_0,\bar{p}(\bar{\pi}_0)\big) - \chi\big(\pi(p_0,\bar{\pi}_0,\bar{p}_0),p_0,\bar{\pi}_0,\bar{p}_0\big)\Big)
\end{aligned} \tag{58}$$

For term $\chi\big(\pi(\bar{\pi}_0),p(\bar{\pi}_0),\bar{\pi}_0,\bar{p}(\bar{\pi}_0)\big) - \chi\big(\pi(p_0,\bar{\pi}_0,\bar{p}_0),p_0,\bar{\pi}_0,\bar{p}_0\big)$, we give upper bound as follows:

$$\begin{aligned}
&\chi\big(\pi(\bar{\pi}_0),p(\bar{\pi}_0),\bar{\pi}_0,\bar{p}(\bar{\pi}_0)\big) - \chi\big(\pi(p_0,\bar{\pi}_0,\bar{p}_0),p_0,\bar{\pi}_0,\bar{p}(\bar{\pi}_0)\big)\\
&= \chi\big(\pi(\bar{\pi}_0),p(\bar{\pi}_0),\bar{\pi}_0,\bar{p}(\bar{\pi}_0)\big) - \chi\big(\pi(p_0,\bar{\pi}_0,\bar{p}_0),p(\bar{\pi}_0),\bar{\pi}_0,\bar{p}(\bar{\pi}_0)\big)\\
&\quad + \chi\big(\pi(p_0,\bar{\pi}_0,\bar{p}_0),p(\bar{\pi}_0),\bar{\pi}_0,\bar{p}(\bar{\pi}_0)\big) - \chi\big(\pi(p_0,\bar{\pi}_0,\bar{p}_0),p_0,\bar{\pi}_0,\bar{p}(\bar{\pi}_0)\big)\\
&\le L_\pi\|\pi(\bar{\pi}_0)-\pi(p_0,\bar{\pi}_0,\bar{p}_0)\| + L_p\|p(\bar{\pi}_0)-p_0\|,
\end{aligned} \tag{59}$$

and

$$\begin{aligned}
&\chi\big(\pi(p_0,\bar{\pi}_0,\bar{p}_0),p_0,\bar{\pi}_0,\bar{p}(\bar{\pi}_0)\big) - \chi\big(\pi(p_0,\bar{\pi}_0,\bar{p}_0),p_0,\bar{\pi}_0,\bar{p}_0\big)\\
&= -\frac{r_2}{2}\|p_0-\bar{p}(\bar{\pi}_0)\|^2 + \frac{r_2}{2}\|p_0-\bar{p}_0\|^2.
\end{aligned} \tag{60}$$

Without loss of generality, taking $\bar{p}_0=p_0, \bar{\pi}_0=\pi_0$ and combining (58), (59) and (60) yields

$$\Phi(\pi_0,p_0,\bar{\pi}_0,\bar{p}_0) - \underline{\chi} \le L_\pi\|\pi_0-\pi^*\| + L_p\|p_0-p^*\| + L_\pi\|\pi(\bar{\pi}_0)-\pi(p_0,\bar{\pi}_0,\bar{p}_0)\| + L_p\|p(\bar{\pi}_0)-p_0\|.$$

□

Lemma C.9 (Lemma 8 in Zheng et al. (2023))**.** *Let $\varepsilon \ge 0$. Suppose that*

$$\max\left\{\frac{\|\pi_k-\pi_{k+1}\|}{\tau}, \frac{\|p_k-p_k^+(\bar{\pi}_k,\bar{p}_k)\|}{\sigma}, \frac{\|\bar{\pi}_k-\bar{\pi}_{k+1}\|}{\beta}, \frac{\|\bar{p}_k-\bar{p}_{k+1}\|}{\mu}\right\} \le \varepsilon.$$

Then, there exists a $\varpi>0$ such that (π_{k+1},p_{k+1}) is a $\varpi\varepsilon-$game stationary point.

Proof of Theorem 4.1

Theorem C.1. *Under the setting of Proposition 4.1 and suppose $\beta = \mathcal{O}(K^{-1/2})$, then for any $K > 0$, there exists a $k \leq K$ such that (π^{k+1}, p^{k+1}) is an $\mathcal{O}\big((D_\Pi^{1/2} + D_\mathcal{P}^{1/2})/K^{1/4}\big)$ game stationary point and $\bar{\pi}^{k+1}$ is an $\mathcal{O}\big((D_\Pi^{1/2} + D_\mathcal{P}^{1/2})/K^{1/4}\big)$ optimal stationary point, where D_Π and $D_\mathcal{P}$ are the diameter of set Π and $\mathcal{P}$, respectively.*

Proof. It is easy to know that $\Phi(\pi, p, \bar{\pi}, \bar{p})$ is lower bounded by $\underline{\chi}$. Let

$$\iota := \max\left\{\frac{r_1}{32}\|\pi_{k+1} - \pi_k\|^2, \frac{r_2}{15}\|p_k - p_k^+(\bar{\pi}_k, \bar{p}_k)\|^2, \frac{r_1}{5\beta}\|\bar{\pi}_k - \bar{\pi}_{k+1}\|^2, \frac{r_2}{4\mu}\|\bar{p}_k^+(\bar{\pi}_{k+1}) - \bar{p}_k\|^2\right\}.$$

Then, we consider the following two cases separately:

Case1: There exists $k \in [K-1]$ such that

$$\frac{1}{2}\iota \leq 4r_1\beta\|\pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) - \pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1}))\|^2.$$

It follows from Proposition C.2 that

$$\|\bar{p}_k^+(\bar{\pi}_{k+1}) - \bar{p}_k\|^2 \leq \frac{32r_1\mu\beta}{r_2}\|\pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) - \pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1}))\|^2 \leq \frac{32r_1\mu\beta\omega_0}{r_2}\|\bar{p}_k^+(\bar{\pi}_{k+1}) - \bar{p}_k\|,$$

which implies that $\|\bar{p}_k^+(\bar{\pi}_{k+1}) - \bar{p}_k\| \leq \varpi_1\beta$, where $\varpi_1 := 32r_1\mu\omega_0/r_2$. Armed with this, we can bound other terms as follows:

$$\begin{aligned}
\frac{\|\pi_{k+1} - \pi_k\|^2}{\tau^2} &\leq \frac{256\beta}{\tau^2}\|\pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) - \pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1}))\|^2 \\
&\leq \frac{256\beta\omega_0}{\tau^2}\|\bar{p}_k^+(\bar{\pi}_{k+1}) - \bar{p}_k\| = \varpi_2\beta^2, \\
\frac{\|p_k - p_k^+(\bar{\pi}_k, \bar{p}_k)\|^2}{\sigma^2} &\leq \frac{120r_1\beta}{r_2\sigma^2}\|\pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) - \pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1}))\|^2 \\
&\leq \frac{120r_1\beta\omega_0}{r_2\sigma^2}\|\bar{p}_k^+(\bar{\pi}_{k+1}) - \bar{p}_k\| = \varpi_3\beta, \\
\frac{\|\bar{\pi}_{k+1} - \bar{\pi}_k\|^2}{\beta^2} &\leq 40\|\pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) - \pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1}))\|^2 \leq 40\omega_0\|z_+^t(\bar{p}_{k+1}) - \bar{\pi}_k\| = \varpi_4\beta, \\
\frac{\|\bar{p}_{k+1} - \bar{p}_k\|^2}{\mu^2} &\leq \frac{2\|\bar{p}_k^+(\bar{\pi}_{k+1}) - \bar{p}_k\|^2}{\mu^2} + 2\Big(4\ell_{p\pi}^2\sigma^2\kappa_6^2\|\pi_k - \pi_{k+1}\|^2 + 4(\kappa_8+1)^2\|p_k - p_k^+(\bar{\pi}_k, \bar{p}_k)\|^2\Big) \\
&\quad + 4\kappa_3^2\|\bar{\pi}_k - \bar{\pi}_{k+1}\|^2 \\
&\leq \frac{2\varpi_1^2}{\mu^2}\beta^2 + 8\ell_{p\pi}^2\sigma^2\kappa_6^2c^2\varpi_2\beta^2 + 8(\kappa_8+1)^2\sigma^2\varpi_3\beta^2 + 4\kappa_3^2\varpi_3\beta^3 \leq \varpi_5\beta^2,
\end{aligned}$$

where $\varpi_2 := \frac{256\omega_0}{\tau^2}\varpi_1$, $\varpi_3 := \frac{120r_1\omega_0}{r_2\sigma^2}\varpi_1$, $\varpi_4 := 40\omega_0\varpi_1$ and $\varpi_5 := \frac{2\varpi_1^2}{\mu^2} + 8\ell_{p\pi}^2\sigma^2\kappa_6^2\tau^2\varpi_2 + 8(\kappa_8+1)^2\sigma^2\varpi_3 + 4\kappa_3^2\varpi_3$. According to Lemma C.9, there exists a $\varpi > 0$ such that (π_{k+1}, p_{k+1}) is a $\varpi\varepsilon$-game stationary point, where $\varepsilon = \max\{\sqrt{\varpi_2}\beta, \sqrt{\varpi_3}\beta, \sqrt{\varpi_4}\beta, \sqrt{\varpi_5}\beta^{\frac{1}{2}}\} = \mathcal{O}(\beta^{1/2})$. Next, we show that $\bar{\pi}_{k+1}$ is an $\mathcal{O}(\beta^{1/2})$-optimization stationary point. Note that

$$\begin{aligned}
&\|\bar{\pi}_{k+1} - \pi^*(\bar{\pi}_{k+1})\| \\
&\leq \|\bar{\pi}_{k+1} - \bar{\pi}_k\| + \|\bar{\pi}_k - \pi_{k+1}\| + \|\pi_{k+1} - \pi(p_k, \bar{\pi}_k, \bar{p}_k)\| + \|\pi(p_k, \bar{\pi}_k, \bar{p}_k) - \pi(p(\bar{\pi}_k, \bar{p}_k), \bar{\pi}_k, \bar{p}_k)\| \\
&\quad + \|\pi(\bar{\pi}_k, \bar{p}_k) - \pi(\bar{\pi}_{k+1}, \bar{p}_k)\| + \|\pi(\bar{\pi}_{k+1}, \bar{p}_k) - \pi^*(\bar{\pi}_{k+1})\| \\
&\overset{(a)}{\leq} (1+\kappa_2)\|\bar{\pi}_{k+1} - \bar{\pi}_k\| + \frac{\|\bar{\pi}_k - \bar{\pi}_{k+1}\|}{\beta} + \kappa_6\|\pi_k - \pi_{k+1}\| + \kappa_1\kappa_8\|p_k - p_k^+(\bar{\pi}_k, \bar{p}_k)\| \\
&\quad + \omega_1\|\bar{p}_k - p(\bar{\pi}_{k+1}, \bar{p}_k)\|^{\frac{1}{2}} \\
&\overset{(b)}{\leq} (1+\kappa_2)\sqrt{\varpi_4}\mathcal{O}(\beta^{3/2}) + \sqrt{\varpi_4}\mathcal{O}(\beta^{1/2}) + \left(\kappa_6\sqrt{\varpi_2} + \kappa_1\kappa_8\sqrt{\varpi_3}\right)\mathcal{O}(\beta) + \omega_1\varpi_1\mathcal{O}(\beta^{1/2}) = \mathcal{O}(\beta^{1/2}),
\end{aligned} \tag{61}$$

where (a) holds by Item 3 Item 1 in Lemma C.2, Item a) and Item c) in Lemma C.4 and Proposition C.2, and (b) holds by (π_{k+1}, p_{k+1}) is a $\varpi\varepsilon$ game stationary point.

Case2: For any $k \in [K-1]$, we have

$$\frac{1}{2}\iota \geq 4r_1\beta\|\pi(\bar{\pi}_{k+1}, \bar{p}(\bar{\pi}_{k+1})) - \pi(\bar{\pi}_{k+1}, \bar{p}_k^+(\bar{\pi}_{k+1}))\|^2.$$

Since

$$\begin{aligned}
&\Phi(\pi_k, p_k, \bar{\pi}_k, \bar{p}_k) - \Phi(\pi_{k+1}, p_{k+1}, \bar{\pi}_{k+1}, \bar{p}_{k+1}) \\
&\geq \frac{r_1}{64}\|\pi_{k+1} - \pi_k\|^2 + \frac{r_2}{30}\|p_k - p_k^+(\bar{\pi}_k, \bar{p}_k)\|^2 + \frac{r_1}{10\beta}\|\bar{\pi}_k - \bar{\pi}_{k+1}\|^2 + \frac{r_2}{8\mu}\|\bar{p}_k^+(\bar{\pi}_{k+1}) - \bar{p}_k\|^2
\end{aligned}$$

holds for $k \in [K-1]$ and $\Phi(\pi, p, \bar{\pi}, \bar{p}) \geq \underline{\chi}$, we know that

$$\begin{aligned}
&\Phi(\pi_0, p_0, \bar{\pi}_0, \bar{p}_0) - \underline{\chi} \\
&\geq \sum_{k=0}^{K-1} \frac{r_1}{64}\|\pi_{k+1} - \pi_k\|^2 + \frac{r_2}{30}\|p_k - p_k^+(\bar{\pi}_k, \bar{p}_k)\|^2 + \frac{r_1}{10\beta}\|\bar{\pi}_k - \bar{\pi}_{k+1}\|^2 + \frac{r_2}{8\mu}\|\bar{p}_k^+(\bar{\pi}_{k+1}) - \bar{p}_k\|^2 \\
&\geq K\min\left\{\frac{r_1\tau^2}{64}, \frac{r_2\sigma^2}{30}, \frac{r_1}{10}, \frac{r_2\mu}{8}\right\}\left(\frac{\|\pi_{k+1} - \pi_k\|^2}{\tau^2} + \frac{\|p_k - p_k^+(\bar{\pi}_k, \bar{p}_k)\|^2}{\sigma^2}\right) \\
&\quad + K\min\left\{\frac{r_1\tau^2}{64}, \frac{r_2\sigma^2}{30}, \frac{r_1}{10}, \frac{r_2\mu}{8}\right\}\left(\frac{\|\bar{\pi}_k - \bar{\pi}_{k+1}\|^2}{\beta} + \frac{\|\bar{p}_k^+(\bar{\pi}_{k+1}) - \bar{p}_k\|^2}{\mu^2}\right).
\end{aligned}$$

Since $\Phi(\pi, p, \bar{\pi}, \bar{p}) \geq \underline{\chi}$, there exists a $k \in [K-1]$ such that

$$\begin{aligned}
&\max\left\{\frac{\|\pi_k - \pi_{k+1}\|^2}{\tau^2}, \frac{\|p_k - p_k^+(\bar{\pi}_k, \bar{p}_k)\|^2}{\sigma^2}, \frac{\|\bar{\pi}_k - \bar{\pi}_{k+1}\|^2}{\beta}, \frac{\|\bar{p}_k - \bar{p}_{k+1}^+(\bar{\pi}_{k+1})\|^2}{\mu^2}\right\} \\
&\leq \frac{\Phi(\pi_0, p_0, \bar{\pi}_0, \bar{p}_0) - \underline{\chi}}{K\min\left\{\frac{r_1\tau^2}{64}, \frac{r_2\sigma^2}{30}, \frac{r_1}{10}, \frac{r_2\mu}{8}\right\}} \overset{(a)}{\leq} \frac{\eta}{K},
\end{aligned}$$

where (a) holds by taking

$$\begin{aligned}
\eta &:= \left(\min\left\{\frac{r_1\tau^2}{64}, \frac{r_2\sigma^2}{30}, \frac{r_1}{10}, \frac{r_2\mu}{8}\right\}\right)^{-1}\left(L_\pi\|\pi_0 - \pi^*\| + L_p\|p_0 - p^*\| + L_\pi\|\pi(\bar{\pi}_0) - \pi(p_0, \bar{\pi}_0, \bar{p}_0)\| + L_p\|p(\bar{\pi}_0) - p_0\|\right) \\
&= \mathcal{O}(D_\Pi + D_\mathcal{P})
\end{aligned}$$

and (57). Note that

$$\begin{aligned}
\frac{\|\bar{p}_{k+1} - \bar{p}_k\|^2}{\mu^2} &\leq \frac{2\|\bar{p}_k^+(\bar{\pi}_{k+1}) - \bar{p}_k\|^2}{\mu^2} + 4\kappa_3^2\|\bar{\pi}_k - \bar{\pi}_{k+1}\|^2 \\
&\quad + 2\Big(4\ell_{p\pi}^2\sigma^2\kappa_6^2\|\pi_k - \pi_{k+1}\|^2 + 4(\kappa_8+1)^2\|p_k - p_k^+(\bar{\pi}_k, \bar{p}_k)\|^2\Big) \\
&\leq \frac{\eta\left(2 + 8\ell_{p\pi}^2\sigma^2\kappa_6^2\tau^2 + 8(\kappa_8+1)^2\sigma^2 + 4\kappa_3^2\beta\right)}{K} \overset{(a)}{\leq} \frac{\varpi_6}{K},
\end{aligned}$$

where $\varpi_6 = \eta\left(2 + 8\ell_{p\pi}^2\sigma^2\kappa_6^2\tau^2 + 8(\kappa_8+1)^2\sigma^2 + 4\kappa_3^2\right)$ and (a) holds by $\beta < 1$. Thus, we know there exists a $\varpi > 0$ such that (π_{k+1}, p_{k+1}) is a $\varpi\varepsilon$ game stationary point, where $\varepsilon = \sqrt{\frac{\varpi_6}{K\beta}}$. Similar argument as (61) shows that $\bar{\pi}^{k+1}$ is an $\mathcal{O}(\omega_6^{1/2}K^{-1/2}\beta^{-1/2})$ optimal stationary point.

Combining the two cases, we have the following conclusion: if we choose $\beta = \mathcal{O}(K^{-1/2})$, then optimal stationary point and game stationary point are coincide as the same rate $\mathcal{O}(\omega_6^{1/2}K^{-1/4})$. □

Proof of Theorem 4.2

Theorem C.2. *Under all settings of Theorem C.1. Then for the sequence $\{\bar{\pi}_k\}_{k=1}^K$ generated by SRPG, there exists a $k < K$ such that $\bar{\pi}_{k+1}$ is $\mathcal{O}\big((D_\Pi^{1/2} + D_\mathcal{P}^{1/2})/K^{1/4}\big)$ global optimal solution.*

Proof. It follows from Theorem C.1 that there exists a $\bar{\pi}_{k+1}$ such that $\|\mathrm{prox}_{\phi,r_1}(\bar{\pi}_{k+1}) - \bar{\pi}_{k+1}\| = \mathcal{O}\big((D_\Pi^{1/2} + D_\mathcal{P}^{1/2})/K^{1/4}\big)$. Recall the definition of $\nabla\phi_{2\ell_\pi}(\pi) = \ell_\pi(\mathrm{prox}_{\phi,2\ell_\pi}(\pi) - \pi)$, then combining the result and Theorem 3.1, we have there exists a $k < K$ such that $\phi(\bar{\pi}_{k+1}) - \phi^* = \mathcal{O}\big((D_\Pi^{1/2} + D_\mathcal{P}^{1/2})/K^{1/4}\big)$. □

C.2. Proof of Results in Section 4.3

Lemma C.10. *For function $\phi(\pi)$ and the corresponding Moreau Envelope function $\phi_{1/\ell_\pi}(\pi)$, we have*

A) $\phi(\pi) - \phi_{2\ell_\pi}(x) \geq \ell_\pi\|\pi - \tilde{\pi}(\pi)\|^2/2$, *where* $\tilde{\pi}(\pi) \in \mathrm{argmin}_{\pi'\in\Pi}\{\phi(\pi') + \ell_\pi\|\pi' - \pi\|^2\}$.

B) $\Pi^* = \Pi^*_{\ell_\pi}$ *where* $\Pi^* := \mathrm{argmin}_{\pi\in\Pi}\,\phi(\pi)$ *and* $\Pi^*_{\ell_\pi} := \mathrm{argmin}_{\pi\in\Pi}\,\phi_{2\ell_\pi}(\pi)$

C) $\phi(\pi^*) = \phi_{2\ell_\pi}(\pi^*)$ *holds for any* $\pi^* \in \Pi^*$.

Proof. Item A): Note that:

$$\begin{aligned}&\phi(\pi) - \phi_{2\ell_\pi}(\pi)\\ &= \phi(\pi) + \ell_\pi\|\pi - \pi\|^2 - (\phi(\tilde{\pi}(\pi)) + \ell_\pi\|\tilde{\pi}(\pi) - \pi\|^2)\\ &\overset{(a)}{\geq} \frac{\ell_\pi}{2}\|\tilde{\pi}(\pi) - x\|^2,\end{aligned} \tag{62}$$

where (a) holds by $\phi(\pi') + \ell_\pi\|\pi' - \pi\|^2$ is ℓ_π-strongly convex with respect to π'.

Item B): we consider $\Pi^*_{\ell_\pi} \subseteq \Pi^*$ and $\Pi^* \subseteq \Pi^*_{\ell_\pi}$, separately. $\Pi^*_{\ell_\pi} \subseteq \Pi^*$: for any $\pi^* \in \mathrm{argmin}_{\pi\in\Pi}\,\phi_{2\ell_\pi}(\pi)$, we have

$$\|\nabla\phi_{2\ell_\pi}(\pi^*)\| = 4\ell_\pi^2\|\tilde{\pi}(\pi^*) - \pi^*\|^2 = 0,$$

i.e., $\tilde{\pi}(\pi^*) = \pi^*$, where $\tilde{\pi}(\pi^*) \in \mathrm{argmin}_{\pi'\in\Pi}\{\phi(\pi') + \ell_\pi\|\pi' - \pi^*\|^2\}$, which implies that

$$\tilde{\pi}(\pi^*) \in \operatorname*{argmin}_{\pi'\in\Pi}\{\phi(\pi')\}.$$

Hence $\mathrm{argmin}_{\pi\in\Pi}\,\phi_{2\ell_\pi}(\pi) \subseteq \mathrm{argmin}_{\pi\in\Pi}\,\phi(\pi)$.

$\Pi^* \subseteq \Pi^*_{\ell_\pi}$: For any $\pi^* \in \mathrm{argmin}_{\pi\in\Pi}\,\phi(\pi)$, we have

$$\phi(\pi) \geq \phi(\pi^*) - \ell_\pi\|\pi - \pi^*\|/2, \tag{63}$$

Taking $\pi = \tilde{\pi}(\pi^*)$, where $\tilde{\pi}(\pi^*) \in \mathrm{argmin}_{\pi'\in\Pi}\{\phi(\pi') + L\|\pi' - \pi^*\|^2\}$, then we have

$$\phi(\tilde{\pi}(\pi^*)) \geq \phi(\pi^*) - \ell_\pi\|\tilde{\pi}(\pi^*) - \pi^*\|/2. \tag{64}$$

It follows from the definition of $\tilde{\pi}(\pi^*)$ that $\phi(\pi^*) \geq \phi(\tilde{\pi}(\pi^*)) + \ell_\pi\|\tilde{\pi}(\pi^*) - \pi^*\|^2$. Putting (63) and (64) together, we have

$$\phi(\pi^*) - \ell_\pi\|\tilde{\pi}(\pi^*) - \pi^*\|/2 \leq \phi(\tilde{\pi}(\pi^*)) \leq \phi(\pi^*) - \ell_\pi\|\tilde{\pi}(\pi^*) - \pi^*\|^2.$$

Hence we have $\tilde{\pi}(\pi^*) = \pi^*$, which implies that $\|\nabla\phi_{2\ell_\pi}(\tilde{\pi}(\pi^*))\| = \|\nabla\phi_{2\ell_\pi}(\pi^*)\| = 4L^2\|\tilde{\pi}(\pi^*) - \pi^*\|^2 = 0$. Note that

$$\phi_{2\ell_\pi}(\pi^*) = \phi(\tilde{\pi}(\pi^*)) + \ell_\pi\|\tilde{\pi}(\pi^*) - \pi^*\|^2 = \phi(\pi^*), \tag{65}$$

and it follows from Lemma 3.1 that

$$\phi(\pi) - \phi(\pi^*) \leq C_{\ell\pi}\|\nabla\phi_{2\ell_\pi}(\pi)\|, \ \ \forall\pi \in \Pi. \tag{66}$$

Combining (65), $\phi(\pi^*) \leq \phi(\tilde{\pi}(\pi)), \forall\pi \in \Pi$ and the definition of $\phi_{2\ell_\pi}(\pi)$ yields

$$0 \leq \phi(\tilde{\pi}(\pi)) + \ell_\pi\|\tilde{\pi}(\pi) - \pi\|^2 - \phi(\pi^*) = \phi_{2\ell_\pi}(\pi) - \phi_{2\ell_\pi}(\pi^*), \ \ \forall\pi \in \Pi,$$

which implies that $\pi^* \in \mathrm{argmin}_{\pi\in\Pi}\,\phi_{2\ell_\pi}(\pi)$.

Item C): It follows from (65) that Item C) holds. □

Proof of Lemma 4.2

Lemma C.11 (Regional Exponential Growth)**.** *Suppose all assumptions of Theorem 3.1 holds. Then for ℓ_π-weakly convex function $\phi(\pi)$, we have*

$$\phi(\pi) - \phi^* \geq \zeta \exp\Big(\operatorname{dist}(\pi, \Pi^*_{\ell_\pi}(\zeta))/C_{\ell\pi}\Big), \;\; \pi \in \Pi \setminus \Pi^*_{\ell_\pi}(\zeta)$$

where $\Pi^*_{\ell_\pi}(\zeta) = \{x \mid \phi_{2\ell_\pi}(\pi) - \phi^* < \zeta\}$ and $\zeta > 0$ is a constant.

Proof. Our proof is similar to Bolte et al. (2017); Karimi et al. (2016). Define the function $g(\pi) = \ln\Big(\phi_{2\ell_\pi}(\pi) - \phi^*\Big), \pi \in \Pi \setminus \Pi^*_{\ell_\pi}(\zeta)$, we have

$$\|\nabla g(\pi)\| = \Big\|\frac{\nabla \phi_{2\ell_\pi}(\pi)}{\phi_{2\ell_\pi}(\pi) - \phi^*}\Big\| \overset{(a)}{\geq} \Big\|\frac{\nabla \phi_{2\ell_\pi}(\pi)}{\phi(\pi) - \phi^*}\Big\| \overset{(b)}{\geq} \frac{1}{C_{\ell\pi}}, \tag{67}$$

where (a) holds by Item A) and (b) follows from Theorem 3.1. For any point $\pi \in \Pi \setminus \Pi^*_{\ell_\pi}(\zeta)$, consider the following differential equation:

$$\frac{d\pi(t)}{dt} = -\nabla g(\pi(t)),$$
$$x(t=0) = \pi_0, \;\; \pi(t) \in \Pi \setminus \Pi^*_{\ell_\pi}(\zeta)$$

By (67), $\|\nabla g\|$ is bounded from below, and as g is also bounded from below $\ln \zeta$, since $\pi \in \Pi \setminus \Pi^*_{\ell_\pi}(\zeta)$. Thus, by moving along the path defined by above, we are sufficiently reducing and will eventually reach the set $\Pi^*_{\ell_\pi}(\zeta)$. Hence, there exists a T such that $\pi(t) \in \Pi^*_{\ell_\pi}(\zeta), \forall t \leq T$. Now, we can show this by using the steps

$$\begin{aligned} g(\pi_0) - g(\pi_T) &= \int_{\pi_T}^{\pi_0} \langle \nabla g(x), dx\rangle = -\int_{\pi_0}^{\pi_T} \langle \nabla g(x), dx\rangle \\ &= -\int_0^T \langle \nabla g(\pi(t)), \frac{d\pi(t)}{dt}\rangle dt = \int_0^T \|\nabla g(\pi(t))\|^2 dt \end{aligned}$$

The length of the orbit $\pi(t)$ starting at π_0, which will be denoted by $\mathcal{L}(\pi_0)$ is given by

$$\mathcal{L}(\pi_0) = \int_0^T \|d\pi(t)/dt\| dt = \int_0^T \|\nabla g(\pi(t))\| dt \geq \operatorname{dist}(\pi_0, \Pi^*_{\ell_\pi}(\zeta)).$$

Hence, we have

$$g(\pi_0) - g(\pi_T) = \int_0^T \|\nabla g(\pi(t))\|^2 dt \geq \frac{1}{C_{\ell\pi}} \int_0^T \|\nabla g(\pi(t))\| dt \geq \operatorname{dist}(\pi_0, \Pi^*_{\ell_\pi}(\zeta))/C_{\ell\pi}.$$

Since $g(\pi_T) = \ln \zeta$, this yields

$$\ln\Big(\phi_{2\ell_\pi}(\pi_0) - \phi^*\Big) = g(\pi_0) \geq \ln \zeta + \operatorname{dist}(\pi_0, \Pi^*_{\ell_\pi}(\zeta))/C_{\ell\pi}.$$

The above result implies that for any $\pi \in \Pi \setminus \Pi^*_{\ell_\pi}(\zeta)$, we have

$$\phi(\pi) - \phi^* \geq \phi_{2\ell_\pi}(\pi) - \phi^* \geq \zeta \exp\Big(\operatorname{dist}(\pi, \Pi^*_{\ell_\pi}(\zeta))/C_{\ell\pi}\Big).$$

□

Proof of Proposition 4.3

Proposition C.3. *Suppose all the assumptions of Lemma 3.1 holds. Then for the sequence $\{\bar{\pi}_k\}$ generated by SRPG, we have*

$$\operatorname{dist}(\bar{\pi}_k, \Pi^*_{\ell_\pi}(\varepsilon)) = \mathcal{O}(\log((D_\pi^{1/2} + D_{\mathcal{P}}^{1/2})k^{-0.25}\varepsilon^{-1})), \forall \bar{\pi}_k \in \Pi \setminus \Pi^*_{\ell_\pi}(\varepsilon).$$

Proof. Take $\zeta = \varepsilon$ in Lemma C.11, and combine it with the result in Theorem C.2, we have there exist a constant $\mathfrak{C}$ such that

$$\varepsilon \cdot \exp\Big(\operatorname{dist}(\bar{\pi}_k, \Pi^*_{\ell_\pi}(\varepsilon))/C_{\ell_\pi}\Big) \leq \frac{\mathfrak{C} \cdot (D_\Pi^{1/2} + D_{\mathcal{P}}^{1/2})}{k^{0.25}}, \tag{68}$$

which implies the finally result $\operatorname{dist}(\bar{\pi}_k, \Pi^*_{\ell_\pi}(\varepsilon)) = \mathcal{O}\Big(\log\big((D_\Pi^{1/2} + D_{\mathcal{P}}^{1/2})k^{-0.25}\varepsilon^{-1}\big)\Big)$ for $\bar{\pi}_k \in \Pi \setminus \Pi^*_{\ell_\pi}(\varepsilon)$. □

D. Experiment Details

D.1. Inventory management problem

We parameterize nominal transition kernel by $\theta \in \mathbb{R}^m$ and $\lambda \in \mathbb{R}^n_{++}$. The L_1-norm ambiguity set is defined by

$$\Xi = \left\{(\theta,\lambda)| \begin{array}{c} \theta \in \mathbb{R}^m, \lambda \in \mathbb{R}^n_{++}, \\ \|\theta-\theta_c\|_1 \le \kappa_\theta, \|\lambda-\lambda_c\|_1 \le \kappa_\lambda \end{array} \right\},$$

where $\theta_c, \lambda_c, \kappa_\theta$ and κ_λ are pre-specified parameters. For any (s,a,s'), we parameterize transition kernel with:

$$p^\xi_{sas'} := \frac{\bar{p}_{sas'} \cdot \exp(\theta^\top \phi_\theta(s')/(\lambda^\top \phi_\lambda(s,a)))}{\sum_{s'\in\mathcal{S}} \bar{p}_{sas'} \cdot \exp(\theta^\top \phi_\theta(s')/(\lambda^\top \phi_\lambda(s,a)))},$$

where $\xi := (\theta,\lambda)$, $\theta \in \mathbb{R}^m, \phi_\theta(s) \in \mathbb{R}^m, \lambda \in \mathbb{R}^n_{++}, \phi_\lambda(s,a) \in \mathbb{R}^n$, m and n are hyperparameters. We term function $\phi_\theta(\cdot)$ as state featuring function and $\phi_\lambda(\cdot)$ as state-action featuring function. We use radial-type features (Sutton & Barto, 2018):

$$[\phi_\theta(s)]_i = \frac{1}{\sqrt{2\pi}\sigma_\theta} \exp\Big\{-\frac{\|s-[c_\theta]_i\|^2}{2\sigma_\theta^2}\Big\}, \forall i \in [m],$$
$$[\phi_\lambda(s,a)]_i = \frac{1}{\sqrt{2\pi}\sigma_\lambda} \exp\Big\{-\frac{\|s-[c_{\lambda,s}]_i\|^2 + \|a-[c_{\lambda,a}]_i\|^2}{2\sigma_\lambda^2}\Big\}, \forall i \in [n].$$

Based on the above model, we give the specific parameter setting in our experiments: $S = 8, A = 3, b = 5, \gamma = 0.95, \theta_c = [0.4, 0.9]^\top \in \mathbb{R}^2, \lambda_c = [0.7, 0.6]^\top \in \mathbb{R}^2$, every state is a two dimension vector (storing and selling), we set $\mathcal{S} := \big\{(0.25,1.3),(0.5,-2.1),(0.75,3.4),(1,-1),(0.25,2.5),(0.5,0.5),(0.75,1.8),(1,-0.8)\big\}$. Every action (ordering) is one dimension, we set $\mathcal{A} := \{-3,-1,5\}$. We set $\kappa_\lambda = \kappa_\theta = 1, [c_\theta]_1 = [-1,2]^\top, [c_\theta]_2 = [0.3,-0.6]^\top, [c_{\lambda,s}]_1 = [1.3,2.1]^\top, [c_{\lambda,s}]_2 = [-0.7,1.5]^\top, [c_{\lambda,a}]_1 = 1, [c_{\lambda,a}]_2 = 0.5, \sigma_\theta = 1, \sigma_\lambda = 2$.

Now, we give more details on the gradient. Let $\lambda_{sa} = (\lambda^\top \phi_\lambda(s,a)), \partial\lambda_{sa}/\partial\lambda_i = [\phi_\lambda(s,a)]_i$, then the partial gradient are

$$\begin{aligned} \frac{\partial \log p^\xi_{sas'}}{\partial \theta_i} &= \Big(\frac{\phi_{\theta i}(s')}{\lambda_{sa}} - \sum_k p^\xi_{sak} \cdot \frac{\phi_{\theta i}(k)}{\lambda_{sa}}\Big) \\ \frac{\partial \log p^\xi_{sas'}}{\partial \lambda_i} &= \Big(\Big(\sum_k p^\xi_{sak} \cdot \frac{\theta^\top \phi_\theta(k)}{\lambda^2_{sa}}\Big) - \frac{\theta^\top \phi(s')}{\lambda^2_{sa}}\Big) \cdot [\phi_\lambda(s,a)]_i \end{aligned} \tag{69}$$

Moreover, we have

$$\frac{J_\rho(\pi,\xi)}{\partial \xi} = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^{\pi,\xi}_\rho, a\sim\pi_s, s'\sim p_{sas'}} \Big[\frac{\partial \log p^\xi_{sas'}}{\partial \xi}\big(c_{sas'} + \gamma v^{\pi,\xi}_{s'}\big)\Big], \tag{70}$$

where $d^{\pi,\xi}_\rho = (1-\gamma)\sum_{s\in\mathcal{S}} \sum_{t=0}^\infty \gamma^t \rho(s) p^\pi_{ss'}(t)$ and $\xi := (\theta,\lambda)$. Combining (69) and (70) yields

$$\begin{aligned} \frac{\partial J_\rho(\pi,\xi)}{\partial \theta_i} &= \frac{1}{1-\gamma} \sum_{s\in\mathcal{S}} d^{\pi,\xi}_s \sum_a \pi_{sa} \sum_{s'\in\mathcal{S}} p^\xi_{sas'} \Big[\Big(\frac{\phi_{\theta i}(s')}{\lambda_{sa}} - \sum_{k\in\mathcal{S}} p^\xi_{sak} \cdot \frac{\phi_{\theta i}(k)}{\lambda_{sa}}\Big) \cdot \big(c_{sas'} + \gamma v^{\pi,\xi}_{s'}\big)\Big] \\ \frac{\partial J_\rho(\pi,\xi)}{\partial \lambda_{sa}} &= \frac{1}{1-\gamma} \sum_{s\in\mathcal{S}} d^{\pi,\xi}_s \sum_a \pi_{sa} \sum_{s'\in\mathcal{S}} p^\xi_{sas'} \Big[\Big(\Big(\sum_{k\in\mathcal{S}} p^\xi_{sak} \cdot \frac{\theta^\top \phi_\theta(k)}{\lambda^2_{sa}}\Big) - \frac{\theta^\top \phi(s')}{\lambda^2_{sa}}\Big) \cdot \phi_{\lambda i}(sa) \cdot \big(c_{sas'} + \gamma v^{\pi,\xi}_{s'}\big)\Big] \end{aligned}$$

On the other hand, we have $\frac{\partial J_\rho(\pi,p)}{\partial \pi_{sa}} = \frac{1}{1-\gamma} \cdot d^{\pi,p}_\rho(s) \cdot q^{\pi,p}_{sa}$, where $d^{\pi,p}_\rho$ and $q^{\pi,p}$ are calculated by the following:

$$\begin{aligned} d^{\pi,p}_\rho &= (1-\gamma) \cdot (I - \gamma P_\pi)^{-1} \rho, \ P_\pi \text{ is the state transition matrix under policy } \pi \\ q^{\pi,p}_{sa} &= \mathbb{E}_{\pi,p}\Big[\frac{1}{1-\gamma} \cdot c_{sas} \mid s_0 = s, a_0 = a\Big]. \end{aligned} \tag{71}$$

Furthermore, for parameterized policy experiments, we consider the following softmax parameterized method:

$$\pi_{sa} = \frac{\exp(w^\top \phi_\lambda(s,a))}{\sum_{a'\in\mathcal{A}} \exp(w^\top \phi_\lambda(s,a'))}. \tag{72}$$