Accelerated Policy Gradient for s-Rectangular Robust MDPs with Large State Spaces

Ziyi Chen [1] **Heng Huang** [1]

Abstract

Robust Markov decision process (robust MDP) is an important machine learning framework to make a reliable policy that is robust to environmental perturbation. Despite empirical success and popularity of policy gradient methods, existing policy gradient methods require at least iteration complexity $\mathcal{O}(\epsilon^{-4})$ to converge to the global optimal solution of s-rectangular robust MDPs with ϵ-accuracy and are limited to deterministic setting with access to exact gradients and small state space that are impractical in many applications. In this work, we propose an accelerated policy gradient algorithm with iteration complexity $\mathcal{O}(\epsilon^{-3}\ln\epsilon^{-1})$ in the deterministic setting using entropy regularization. Furthermore, we extend this algorithm to stochastic setting with access to only stochastic gradients and large state space which achieves the sample complexity $\mathcal{O}(\epsilon^{-7}\ln\epsilon^{-1})$. In the meantime, our algorithms are also the first scalable policy gradient methods to entropy-regularized robust MDPs, which provide an important but underexplored machine learning framework.

1. Introduction

Reinforcement Learning (RL) modeled by Markov decision processes (MDP) is a broadly used machine learning framework where an agent learns and makes decisions by interacting with a dynamic environment. RL has many applications including robotics (Kober et al., 2013; Peng et al., 2018), energy flow control (Perera and Kamalaruban, 2021), production scheduling (Wang and Usher, 2005), flight control (Abbeel et al., 2006), etc. RL system is usually trained in a simulated environment to avoid deployment cost (Zhou et al., 2023). However, the simulated environment usually differs from real-world environment, which may degrade the performance of the trained RL system on real-world environment (Peng et al., 2018; Zhou et al., 2023). To make RL robust to this simulation-to-reality gap, robust Markov decision process (robust MDP) (Iyengar, 2005; Nilim and El Ghaoui, 2005; Wiesemann et al., 2013) has been proposed, which aims to find the optimal robust policy that optimizes the performance under the worst possible environment from a certain ambiguity set.

Robust MDP problem with a general ambiguity set is proved to be NP-hard (Wiesemann et al., 2013). To make it computationally tractable, various structural conditions on ambiguity set have been used including (s,a)-rectangularity (Nilim and El Ghaoui, 2005; Iyengar, 2005; Wiesemann et al., 2013; Wang and Zou, 2022; Li et al., 2023c; Zhou et al., 2023) and s-rectangularity (Wiesemann et al., 2013; Ho et al., 2021; Wang et al., 2023; Kumar et al., 2023a;c). This work focuses on the more general s-rectangularity which allows the nature to select an adversarial environment before observing the learning agent's action (Wang et al., 2023) and yields less conservative policies (Kumar et al., 2023c).

Various methods have been adopted to solve robust MDP, including value-iteration (Iyengar, 2005; Nilim and El Ghaoui, 2005; Wiesemann et al., 2013; Grand-Clément and Kroer, 2021; Kumar et al., 2023b), policy-iteration (Iyengar, 2005; Badrinath and Kalathil, 2021; Ho et al., 2021; Kumar et al., 2022) and policy gradient (Wang and Zou, 2022; Li et al., 2023c; Wang et al., 2023; Zhou et al., 2023; Kumar et al., 2023c; Li et al., 2023b; Guha and Lee, 2023). Among these methods, policy gradient has gained significant attention due to its simple implementation (Silver et al., 2014), excellent real-world performance (Silver et al., 2014; Xu et al., 2014; Wang et al., 2023) and scalability to large state and action spaces (Silver et al., 2014; Wang et al., 2023). Moreover, policy gradient methods also have provable global convergence guarantee on non-robust MDP (Agarwal et al., 2021; Bhandari and Russo, 2021; Xiao, 2022). However, the global convergence of policy gradient methods for robust MDP is much harder to obtain since the robust value function is not differentiable. For (s,a)-rectangular case, Wang

[1]Department of Computer Science, University of Maryland College Park. Correspondence to: Ziyi Chen <zc286@umd.edu>, Heng Huang <heng@umd.edu>.

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

Table 1: Comparison of policy gradient works for s-rectangular robust MDPs. The measures include the number of updates on policy π as well as transition kernel p and the complexity to achieve ϵ-optimal robust policy defined in Definition 1. The complexity denotes iteration complexity (the total number of updates on π and p) in deterministic setting with access to exact gradients, and sample complexity (total number of required samples) in stochastic setting with access to only stochastic gradients. See Appendix A for more explanation of this table.

WORKS	# π UPDATES	# p UPDATES	COMPLEXITY	STOCHASTIC	LARGE SPACE
(WANG ET AL., 2023)	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-4}+\epsilon^{4}\gamma^{-\mathcal{O}(\epsilon^{-4})})$	$\mathcal{O}(\epsilon^{-4}+\epsilon^{4}\gamma^{-\mathcal{O}(\epsilon^{-4})})$	$\times$	$\times$
(LI ET AL., 2023B)	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-6})$	$\mathcal{O}(\epsilon^{-6})$	$\times$	$\times$
(KUMAR ET AL., 2023C)	$\mathcal{O}(\epsilon^{-1})$	-	-	$\times$	$\times$
(GUHA AND LEE, 2023)	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-4})$	$\times$	$\times$
OUR ALGORITHM 1	$\mathcal{O}(\epsilon^{-3}\ln\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3}\ln\epsilon^{-1})$	$\times$	$\times$
OUR ALGORITHM 2	$\mathcal{O}(\epsilon^{-3}\ln\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-7}\ln\epsilon^{-1})$	$\checkmark$	$\times$
OUR ALGORITHM 3	$\mathcal{O}(\epsilon^{-3}\ln\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-7}\ln\epsilon^{-1})$	$\checkmark$	$\checkmark$

and Zou (2022); Li et al. (2023c); Zhou et al. (2023) tackle this challenge by evaluating and using the uniquely-defined robust Q function.

For the more general s-rectangular case, this robust Q function is not well-defined (Li and Lan, 2023), which makes the global convergence even more challenging. Among the existing policy gradient methods (Wang et al., 2023; Li et al., 2023b; Kumar et al., 2023c; Guha and Lee, 2023) on s-rectangular case, the state of the art iteration complexity (defined as the total number of updates on both transition kernel and policy) is $\mathcal{O}(\epsilon^{-4})$ (Guha and Lee, 2023) as shown in Table 1. Moreover, Kumar et al. (2023c) has oracle access to the sub-gradient of the robust optimal return that is assumed to be Lipschitz-smooth, which does not hold in many cases. Hence, we are motivated to ask:

> ***Q1:*** *Can we propose a policy gradient algorithm with lower iteration complexity to achieve global optimal solution of a generic s-rectangular robust MDP?*

Moreover, existing policy gradient algorithms are analyzed in deterministic setting with access to exact gradients and need to obtain policy or transition kernel over all states and actions, which are impractical in many applications where only stochastic samples are available and the state space is very large. Though Wang et al. (2023); Li et al. (2023b); Guha and Lee (2023) mention transition kernel parameterization to mitigate this issue, their global convergence results still involve enumeration over all states and actions and thus do not apply to large state space. As a result, we want to ask:

> ***Q2:*** *Can we extend policy gradient algorithms to stochastic setting and large state space for s-rectangular case?*

1.1. Our Contributions

We answer affirmatively to these questions by proposing an accelerated policy gradient algorithm (Algorithm 1) for s-rectangular robust MDPs in the deterministic setting with access to exact gradients, and extending this algorithm to stochastic setting with access to only stochastic gradients (Algorithm 2) and then to large state space (Algorithm 3). We summarize the advantages of our algorithms and their global convergence results as follows and also in Table 1.

Acceleration: To accelerate existing policy gradient algorithms which directly optimize the non-differentiable objective function (Li et al., 2023b; Guha and Lee, 2023; Kumar et al., 2023c), we apply entropy regularization to the policy which provides a smooth approximation to the non-differentiable objective function. The approximation error can be arbitrarily small by using a sufficiently small regularization coefficient. Moreover, the entropy regularization not only ensures exponential convergence of our inner policy update, but also yields the Lipschitz-smoothness and gradient dominance of the objective function that guarantees efficient global convergence of the outer transition update. Hence, we obtain $\mathcal{O}(\epsilon^{-3}\ln\epsilon^{-1})$ iteration complexity, faster than the state of the art $\mathcal{O}(\epsilon^{-4})$ (Guha and Lee, 2023).

We are the first to obtain the above Lipschitz-smoothness of the entropy regularized objective (Proposition 2), in which we adopt two novel techniques to tackle entropy regularizer. First, as the log-policy $\ln\pi(a|s)$ may approach $-\infty$, we use the Lipschitz property of $\pi(a|s)\ln\pi(a|s)$ with respect to $\ln\pi(a|s)$. Second, the optimal log-policy $\ln\pi_p$ given the transition kernel p involves a certain Q function Q_p as the unique fixed point of a Bellman operator T_p. Hence, we also need to obtain the Lipschitz property of T_p and accordingly obtain a recursive bound for the Lipschitz property of Q_p.

Stochastic Setting: We extend Algorithm 1 to the practical stochastic setting, by applying temporal difference (TD) method and sample-average to approximate the Q function and transition gradient respectively. In this way, we obtain

the first stochastic policy gradient algorithm (Algorithm 2) with provable global convergence and sample complexity $\mathcal{O}(\epsilon^{-7}\ln\epsilon^{-1})$ for s-rectangular robust MDP.

In this sample complexity analysis, the entropy regularized cost involves $\tau\ln\pi(a|s)$ where $\ln\pi(a|s)$ might approach $-\infty$. To bound $|\tau\ln\pi(a|s)|$, we prove that the optimal log-policy $\ln\pi_p(a|s)\geq\mathcal{O}(-1/\tau)$ for any p. As $\ln\pi\to\ln\pi_p$ exponentially fast with policy optimization, we prove that $\ln\pi(a|s)\geq\mathcal{O}(-1/\tau)$ for any π involved in the algorithm and thus $|\tau\ln\pi(a|s)|\leq\mathcal{O}(1)$.

Large State Space: We further extend Algorithm 2 to large state space, by using linear transition kernel parameterization and linear Q function approximation to reduce state enumeration. We prove that linear kernel parameterization preserves Lipschitz property as well as gradient dominance, which avoids parameterization error in the global convergence result. The obtained Algorithm 3 retains the sample complexity $\mathcal{O}(\epsilon^{-7}\ln\epsilon^{-1})$ to converge to the global optimal solution up to a function approximation error term $\zeta>0$.

Entropy Regularized Robust MDP: All our algorithms are also the first policy gradient methods to solve entropy regularized robust MDP (Mankowitz et al., 2019; Mai and Jaillet, 2021; Eysenbach and Levine, 2021), an important but under-explored learning framework. Entropy regularized robust MDP is important because it combines both the advantage of robustness, and the advantage of entropy regularization in encouraging exploration and prohibiting early convergence to sub-optimal policies (Mankowitz et al., 2019; Mai and Jaillet, 2021), which is suitable for application to inverse reinforcement learning (IRL) (Mai and Jaillet, 2021).

1.2. Related Works

Robust Policy Evaluation: While this work focuses on policy optimization problem, i.e., to find the optimal policy, Li and Lan (2023); Kumar et al. (2023a) focused on robust policy evaluation problem of s-rectangular robust MDP, i.e., to evaluate the value of a policy under the worst-case environment. Li and Lan (2023) considered the nature's choice of transition kernel as a policy and proposes policy gradient methods which achieve linear global convergence in the deterministic setting and $\widetilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity in the stochastic setting. Kumar et al. (2023a) studied a robust MDP where both transition kernel and reward are uncertain and range in L_p-ball constrained s-rectangular ambiguity sets, which has closed-form optimal solution and yields linear convergence in the deterministic case.

Policy Gradient for Non-robust MDP: Policy gradient based algorithms including policy gradient (Sutton et al., 1999), natural policy gradient (Kakade, 2001), actor-critic (Konda and Tsitsiklis, 1999) and natural actor-critic (Bhatnagar et al., 2009) are also very popular for policy optimization in non-robust MDP. Agarwal et al. (2021) provided sub-linear global convergence rates and complexity results of policy gradient methods for various policy parameterizations including tabular, softmax, log-linear and neural policies. Bhandari and Russo (2021); Xiao (2022) accelerated the global convergence to linear rate for tabular policy by relating policy gradient methods to policy-iteration.

Policy Gradient for (s,a)-rectangular Robust MDP: Wang and Zou (2022) presented a smoothed robust policy gradient algorithm for robust MDP with a specific R-contamination ambiguity set and achieves $\mathcal{O}(\epsilon^{-3})$ iteration complexity in the deterministic setting and $\mathcal{O}(\epsilon^{-7})$ sample complexity in the stochastic setting. Li et al. (2023c) introduced a robust policy mirror descent algorithm for robust MDP with a more general (s,a)-rectangular ambiguity set and obtains linear global convergence in the deterministic setting and $\widetilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity in the stochastic setting. Zhou et al. (2023) proposed a robust stochastic natural actor-critic algorithm with linear function approximation and two specific (s,a)-rectangular ambiguity sets which applies to robust MDPs with large state spaces and also achieves $\widetilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity.

Entropy Regularized Robust MDP: Mankowitz et al. (2019) extended the Maximum A-Posteriori Policy Optimization algorithm (Abdolmaleki et al., 2018) to entropy regularized robust MDP. Mai and Jaillet (2021) proposed a value-iteration algorithm for entropy regularized robust MDP with provable worst-case complexity. Eysenbach and Levine (2021) proved that entropy regularized MDP provides a lower bound of the robust MDP objective.

2. Problem Settings

2.1. Robust MDP

A vanilla MDP is characterized by a tuple $(\mathcal{S},\mathcal{A},p,c,\gamma,\rho)$. $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces respectively. $\gamma\in(0,1)$ is the discount factor. p is the state transition kernel where $p(\cdot|s,a)\in\Delta^{\mathcal{S}}$ is a distribution on $\mathcal{S}$ for any state $s\in\mathcal{S}$ and action $a\in\mathcal{A}$. $c:\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to[0,1]$ is the cost function. $\rho\in\Delta^{\mathcal{S}}$ is the distribution of the environment's initial state s_0. At time t, an agent observes the environment's current state s_t and takes a random action $a_t\sim\pi(\cdot|s)$ based on the agent's stationary policy $\pi\in\Pi:=(\Delta^{\mathcal{A}})^{\mathcal{S}}$. Then the environment transitions to the next state $s_{t+1}\sim p(\cdot|s_t,a_t)$ and gives a cost $c_t:=c(s_t,a_t,s_{t+1})$ to the agent. We define the following value function, which characterizes the long-term expected cost under the policy π.

$$J_\rho(\pi,p):=\mathbb{E}_{\pi,p}\Big[\sum_{t=0}^{\infty}\gamma^t c_t\Big|s_0\sim\rho\Big]. \tag{1}$$

The aim of vanilla MDP is to find the optimal policy π that minimizes the expected cost $J_\rho(\pi,p)$ for a given transition

kernel p. However, in practice, p is usually unknown and thus has to be estimated from data. The estimation error often degrades the performance after deployment. To make the performance robust to this estimation error, robust MDP (Iyengar, 2005; Nilim and El Ghaoui, 2005; Wiesemann et al., 2013) has been proposed where the transition kernel p ranges in a certain ambiguity set $\mathcal{P}$ which can be selected to contain the true transition kernel. The aim of robust MDP is to find the optimal robust policy that minimizes the robust value function $\Phi_\rho(\pi) := \max_{p\in\mathcal{P}} J_\rho(\pi, p)$ under the worst-case transition kernel $p \in \mathcal{P}$, as formulated by the following minimax optimization problem.

$$\min_{\pi\in\Pi} \max_{p\in\mathcal{P}} J_\rho(\pi, p). \tag{2}$$

Definition 1. *A policy π is called an ϵ-optimal robust policy if $\Phi_\rho(\pi) \le \min_{\pi'\in\Pi} \Phi_\rho(\pi') + \epsilon$ for a certain precision $\epsilon \ge 0$.*

The robust MDP problem (2) is in general NP-hard (Wiesemann et al., 2013). To make it tractable, $\mathcal{P}$ is often assumed to be (s, a)-rectangular (Nilim and El Ghaoui, 2005; Iyengar, 2005; Wiesemann et al., 2013; Wang and Zou, 2022; Li et al., 2023c; Zhou et al., 2023) and s-rectangular (Wiesemann et al., 2013; Ho et al., 2021; Wang et al., 2023; Kumar et al., 2023a;c). An (s, a)-rectangular $\mathcal{P}$ is defined as a Cartesian product of sets $\mathcal{P}_{s,a} \subset \Delta^{\mathcal{S}}$ for all s, a, i.e.,

$$\mathcal{P} = \{p \in (\Delta^{\mathcal{S}})^{\mathcal{S}\times\mathcal{A}} : p(\cdot|s, a) \in \mathcal{P}_{s,a}, \forall s \in \mathcal{S}, a \in \mathcal{A}\}.$$

An s-rectangular $\mathcal{P}$ is defined as a Cartesian product of sets $\mathcal{P}_s \subset (\Delta^{\mathcal{S}})^{\mathcal{A}}$ for all s, i.e.,

$$\mathcal{P} = \{p \in (\Delta^{\mathcal{S}})^{\mathcal{S}\times\mathcal{A}} : p(\cdot|s, \cdot) \in \mathcal{P}_s, \forall s \in \mathcal{S}\}.$$

We adopt the following assumption throughout this work.

Assumption 1. *$\mathcal{P}$ is s-rectangular, compact and convex.*

2.2. Entropy Regularized Robust MDP

Entropy regularized robust MDP (Mankowitz et al., 2019; Mai and Jaillet, 2021; Eysenbach and Levine, 2021) is also an important but underexplored learning framework. Its objective is shown below.

$$\min_{\pi\in\Pi} \max_{p\in\mathcal{P}} J_{\rho,\tau}(\pi, p) := \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t c_{\tau,\pi,t} \Big| s_0 \sim \rho\Big], \tag{3}$$

where $c_{\tau,\pi,t} := c_t + \tau \ln \pi(a_t|s_t)$ is the entropy regularized cost with $\tau \in [0, 1]$. The above objective adds entropy regularizer to the robust MDP objective (1). Hence, entropy regularized robust MDP has both the advantage of policy robustness, and the advantage of entropy regularization in encouraging exploration and prohibiting early convergence to sub-optimal policies (Mankowitz et al., 2019; Mai and Jaillet, 2021), which is suitable for application to inverse reinforcement learning (IRL) (Mai and Jaillet, 2021).

Define the following V and Q functions (Cayci et al., 2022):

$$V_\tau(\pi, p; s) := \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t c_{\tau,\pi,t} \Big| s_0 = s\Big], \tag{4}$$

$$Q_\tau(\pi, p; s, a) := \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t c_{\tau,\pi,t} \Big| s_0 = s, a_0 = a\Big]. \tag{5}$$

A good solution to the entropy regularized problem (3) can be an approximate Nash equilibrium defined as follows.

Definition 2. *A policy-transition pair $(\pi, p) \in \Pi \times \mathcal{P}$ is called an (ϵ, τ)-Nash equilibrium if it satisfies*

$$J_{\rho,\tau}(\pi, p) - \min_{\pi'\in\Pi} J_{\rho,\tau}(\pi', p) \le \epsilon,$$
$$\max_{p'\in\mathcal{P}} J_{\rho,\tau}(\pi, p') - J_{\rho,\tau}(\pi, p) \le \epsilon.$$

Proposition 1. *Under Assumption 1, for any $\epsilon \ge 0$ and $\tau > 0$, (ϵ, τ)-Nash equilibrium exists. If $(\pi, p) \in \Pi \times \mathcal{P}$ is an (ϵ, τ)-Nash equilibrium, then π is a $\big(2\epsilon + \frac{\tau \ln|\mathcal{A}|}{1-\gamma}\big)$-optimal robust policy to the optimization problem (2).*

Proposition 1 indicates that by letting $\tau = \mathcal{O}(\epsilon)$, the (ϵ, τ)-Nash equilibrium also solves the robust MDP problem (2) with precision $\mathcal{O}(\epsilon)$. Hence, we can solve robust MDP by solving entropy-regularized robust MDP. In the next section, we will provide the first policy gradient algorithm for entropy-regularized robust MDP, which also solves s-rectangular robust MDP with lower iteration complexity than the the existing policy gradient algorithms (Wang et al., 2023; Li et al., 2023b; Kumar et al., 2023c; Guha and Lee, 2023) that directly aim at the robust MDP objective (2).

3. Accelerated Robust Policy Gradient Algorithm

In this section, we will provide a robust policy gradient algorithm (Algorithm 1) for both entropy-regularized robust MDP and robust MDP, obtain convergence results under the determinsitic case. Then we will extend Algorithm 1 to stochastic case and obtain sample complexity result.

3.1. Accelerated Robust Policy Gradient Algorithm

A major challenge to solve robust MDP is that its objective $\Phi_\rho(\pi)$ defined by eq. (2) is non-differentiable since the optimal transition kernel $\arg\max_p J_\rho(\pi, p)$ is non-unique. In contrast, the entropy-regularized robust MDP (3) is equivalent to its dual form below (Mai and Jaillet, 2021) (see Lemma 2 for the proof of equivalence):

$$\max_{p\in\mathcal{P}} \min_{\pi\in\Pi} J_{\rho,\tau}(\pi, p), \tag{6}$$

for which the optimal policy $\pi_p := \arg\min_\pi J_{\rho,\tau}(\pi, p)$ is unique given the transition kernel p for any $\tau > 0$ (Cen et al., 2022). Hence, based on the Danskin's Theorem (Bernhard and Rapaport, 1995), $F_{\rho,\tau}(p) := \min_{\pi\in\Pi} J_{\rho,\tau}(\pi, p)$ is differentiable with $\nabla F_{\rho,\tau}(p) = \nabla_2 J_{\rho,\tau}(\pi_p, p)$ (See Lemma 7 and its proof in Appendix C)[1]. Furthermore, we will show that $F_{\rho,\tau}(p)$ is Lipschitz smooth in Section 3.2.

As a result, a natural idea is to apply projected gradient ascent $p_{t+1} = \text{proj}_{\mathcal{P}}\big(p_t + \alpha_t \nabla F_{\rho,\tau}(p_t)\big)$ to the Lipschitz smooth objective (6) where $\nabla F_{\rho,\tau}(p_t) = \nabla_2 J_{\rho,\tau}(\pi_{p_t}, p_t)$. The unique optimal policy $\pi_{p_t} := \arg\min_{\pi\in\Pi} J_{\rho,\tau}(\pi, p_t)$ can be efficiently approximated using the following natural policy gradient (NPG) step (Cen et al., 2022):

$$\pi_{t,k+1}(\cdot|s) \propto \pi_{t,k}(\cdot|s) \exp\Big[-\frac{\eta \widehat{Q}_{t,k}(s,\cdot)}{1-\gamma}\Big], \tag{7}$$

where $\eta > 0$ is the stepsize and $\widehat{Q}_{t,k}$ approximates the Q function $Q_\tau(\pi_{t,k}, p_t) := Q_\tau(\pi_{t,k}, p_t; \cdot, \cdot) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ defined by eq. (5) [2]. After T' NPG steps (7), $\pi_t := \pi_{t,T'}$ converges to π_{p_t} exponentially fast with T' (Cen et al., 2022) (also see Lemma 9). Hence, we can replace π_{p_t} with π_t when computing $\nabla F_{\rho,\tau}(p_t) = \nabla_2 J_{\rho,\tau}(\pi_{p_t}, p_t)$, which yields the following projected gradient ascent rule:

$$p_{t+1} = \text{proj}_{\mathcal{P}}\big(p_t + \beta \widehat{\nabla}_p J_{\rho,\tau}(\pi_t, p_t)\big), \tag{8}$$

where $\beta > 0$ is the stepsize and $\widehat{\nabla}_p J_{\rho,\tau}(\pi_t, p_t) \approx \nabla_p J_{\rho,\tau}(\pi_t, p_t)$ is the estimated gradient. $\nabla_p J_{\rho,\tau}(\pi_t, p_t) \in \mathbb{R}^{|\mathcal{S}|^2|\mathcal{A}|}$ is the exact gradient and its (s, a, s')-th entry is given below [3]:

$$\nabla_p J_{\rho,\tau}(\pi, p)(s, a, s') = \frac{d_\rho^{\pi,p}(s)\pi(a|s)}{1-\gamma} \Big[c(s, a, s') + \tau \ln \pi(a|s) + \gamma V_\tau(\pi, p; s')\Big]. \tag{9}$$

where the occupancy measure $d_\rho^{\pi,p}(s)$ is defined as:

$$d_\rho^{\pi,p}(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{\pi,p}(s_t = s | s_0 \sim \rho). \tag{10}$$

Our accelerated robust policy gradient algorithm is shown in Algorithm 1, which updates the policy π_t in the inner loop and transition kernel p_{t+1} in the outer loop. The lines 6 and 10 assume access to $\widehat{Q}_{t,k} \approx Q_\tau(\pi_{t,k}, p_t)$ and $\widehat{\nabla}_p J_{\rho,\tau}(\pi_t, p_t) \approx \nabla_p J_{\rho,\tau}(\pi_t, p_t)$ with arbitrary predefined precisions $\epsilon_1, \epsilon_2 \geq 0$ respectively. This covers both exact case where exact Q-functions and gradients can be easily computed (i.e., $\epsilon_1 = \epsilon_2 = 0$) and inexact case where exact computation is intractable or requires much more computation than inexact estimation. We will show how to obtain these inexact estimations in Section 3.3.

Algorithm 1 Accelerated Robust Policy Gradient

1: **Inputs:** $\tau, T, T', \eta, \beta, \epsilon_1, \epsilon_2$.
2: **Initialize:** p_0.
3: **for** transition update steps $t = 0, 1, \ldots, T-1$ **do**
4: Let $\pi_{t,0}(a|s) \equiv 1/|\mathcal{A}|$.
5: **for** policy update steps $k = 0, 1, \ldots, T'-1$ **do**
6: Obtain $\widehat{Q}_{t,k} \approx Q_\tau(\pi_{t,k}, p_t)$ such that $\|\widehat{Q}_{t,k} - Q_\tau(\pi_{t,k}, p_t)\|_\infty \leq \epsilon_1$.
7: Obtain $\pi_{t,k+1}$ using the NPG step (7).
8: **end for**
9: Let $\pi_t := \pi_{t,T'}$.
10: Obtain approximate gradient $\widehat{\nabla}_p J_{\rho,\tau}(\pi_t, p_t)$ such that $\|\widehat{\nabla}_p J_{\rho,\tau}(\pi_t, p_t) - \nabla_p J_{\rho,\tau}(\pi_t, p_t)\| \leq \epsilon_2$
11: Obtain p_{t+1} by projected gradient ascent step (8).
12: **end for**
13: **Output:** $\pi_{\widetilde{T}}, p_{\widetilde{T}}$ where $\widetilde{T} \in \underset{0\leq t\leq T-1}{\text{argmin}} \|p_{t+1} - p_t\|$.

3.2. Iteration Complexity of Algorithm 1

We will first show two amenable geometric properties of the regularized problem (6) that yields faster global convergence of Algorithm 1 than existing policy gradient works. Throughout this work, $\|\cdot\|_p$ ($p \in [1, \infty]$) is Euclidean p-norm and $\|\cdot\| = \|\cdot\|_2$ is 2-norm by default.

Proposition 2. *Under Assumption 1, $F_{\rho,\tau}(p)$ is Lipschitz smooth with parameter $\ell_F := \frac{8|\mathcal{S}||\mathcal{A}|(1+\gamma\tau\ln|\mathcal{A}|)^2}{\tau(1-\gamma)^5}$, i.e., for any $p, p' \in \mathcal{P}$,*

$$\|\nabla F_{\rho,\tau}(p') - \nabla F_{\rho,\tau}(p)\| \leq \ell_F \|p' - p\|. \tag{11}$$

Technical Novelty: The Lipschitz property of $\nabla F_{\rho,\tau}(p)$ for entropy regularized robust MDP has not been obtained in existing literature to our knowledge. We use two novel techniques to tackle the entropy regularizer. First, we need to prove the Lipschitz continuity of $J_{\rho,\tau}$ and $\nabla_p J_{\rho,\tau}$ (see Lemma 6). $J_{\rho,\tau}$ contains the regularized cost $c(s, a, s') + \tau \ln \pi(a|s)$ which goes to $-\infty$ as $\pi(a|s) \to +0$. To solve this, we control $\ln \pi(a|s)$ by multiplying it with $\pi(a|s)$ and use $|\pi'(a|s)\ln\pi'(a|s) - \pi(a|s)\ln\pi(a|s)| \leq |\ln \pi'(a|s) - \ln \pi(a|s)|$ (see Lemma 16), so we obtained the Lipschitz properties of $J_{\rho,\tau}$ and $\nabla_p J_{\rho,\tau}$ with respect

[1] $\nabla_2 J_{\rho,\tau}(\pi_p, p)$ denotes the gradient with respect to the second argument p. We do not use $\nabla_p J_{\rho,\tau}(\pi_p, p)$ since π_p depends on p.

[2] The NPG step (7) is equivalent to the update rule $\pi_{t,k+1}(\cdot|s) \propto \pi_{t,k}(\cdot|s)^{1-\frac{\eta\tau}{1-\gamma}} \exp\big[\frac{\eta \widehat{Q}_{t,k}(s,\cdot)}{1-\gamma}\big]$ in (Cen et al., 2022), since they define $Q_\tau(\pi, p; s, a) := \mathbb{E}_{s'\sim p(\cdot|s,a)}[c(s, a, s') + \gamma V_\tau(s')]$ which corresponds to our $Q_\tau(\pi, p; s, a) - \tau \ln(a|s)$, and our objective $\min_{\pi\in\Pi} J_{\rho,\tau}(\pi, p)$ corresponds to their objective $\max_{\pi\in\Pi} J_{\rho,\tau}(\pi, p)$

[3] The gradient (9) is obtained by using Lemma 4.1 of (Wang et al., 2023) with the cost $c(s, a, s')$ replaced by regularized cost $c(s, a, s') + \tau \ln \pi(a|s)$

to $\ln \pi$. Second, as $\nabla F_{\rho,\tau} = \nabla_2 J_\rho(\pi_p, p)$, we also need to obtain the Lipschitz property of $\ln \pi_p$ with respect to p. This is not straightforward, since $\ln \pi_p$ is implicitly defined by $Q_p := \widetilde{Q}_\tau(\pi_p, p)$ ($\widetilde{Q}_\tau$ is defined in eq. (30)), the unique fixed point of a Bellman operator T_p defined in eq. (72). As a result, we also need to obtain Lipschitz property of T_p which yields the recursive bound below for any $p, p' \in \mathcal{P}$:

$$\begin{aligned}
&\|Q_{p'} - Q_p\|_\infty = \|T_{p'} Q_{p'} - T_p Q_p\|_\infty \\
&\le \|T_{p'} Q_{p'} - T_p Q_{p'}\|_\infty + \|T_p Q_{p'} - T_p Q_p\|_\infty \\
&\le c \max_{s,a} \|p'(\cdot|s,a) - p(\cdot|s,a)\|_1 + \gamma \|Q_{p'} - Q_p\|_\infty,
\end{aligned}$$

where $c > 0$ is a constant. This implies that $\|Q_{p'} - Q_p\|_\infty \le \frac{c}{1-\gamma} \max_{s,a} \|p'(\cdot|s,a) - p(\cdot|s,a)\|_1$ and thus the Lipschitz continuity of $\ln \pi_p$.

Proposition 2 guarantees that Algorithm 1, which can be seen as approximate gradient ascent on $F_{\rho,\tau}(p)$, converges to a stationary point of $F_{\rho,\tau}(p)$. Such a stationary point also provides a global optimal solution as shown in the following gradient dominance property. Throughout, we define $D_{\mathcal{P}} := \sup_{p,\widetilde{p} \in \mathcal{P}} \|\widetilde{p} - p\|$ as the diameter of $\mathcal{P}$ and $D := \sup_{\pi \in \Pi, p \in \mathcal{P}} \|d_\rho^{\pi,p}/\rho\|_\infty < \infty$ as the distribution mismatch coefficient which has also been used in (Agarwal et al., 2021; Leonardos et al., 2022; Wang et al., 2023).

Proposition 3 (Gradient dominance)**.** *Under Assumption 1, the function $J_{\rho,\tau}$ satisfies the following gradient dominance property for any $\pi \in \Pi$ and $p \in \mathcal{P}$,*

$$\begin{aligned}
&\max_{p' \in \mathcal{P}} J_{\rho,\tau}(\pi, p') - J_{\rho,\tau}(\pi, p) \\
&\le \frac{D}{1-\gamma} \max_{p' \in \mathcal{P}} (p' - p)^\top \nabla_p J_{\rho,\tau}(\pi, p). \qquad (12)
\end{aligned}$$

Based on the two properties, we obtain the following convergence rates of Algorithm 1.

Theorem 1. *Implement Algorithm 1 with $\beta \le \frac{1}{2\ell_F}$, $\eta = \frac{1-\gamma}{\tau}$. Then the output $(\pi_{\widetilde{T}}, p_{\widetilde{T}})$ satisfies the following rates under Assumption 1.*

$$J_{\rho,\tau}(\pi_{\widetilde{T}}, p_{\widetilde{T}}) - \min_{\pi \in \Pi} J_{\rho,\tau}(\pi, p_{\widetilde{T}}) \le \mathcal{O}\Big(\frac{\gamma^{T'} + \epsilon_1}{\tau}\Big), \qquad (13)$$

$$\begin{aligned}
&\max_{p \in \mathcal{P}} J_{\rho,\tau}(\pi_{\widetilde{T}}, p) - J_{\rho,\tau}(\pi_{\widetilde{T}}, p_{\widetilde{T}}) \\
&\le \mathcal{O}\Big[(1 + \tau\epsilon_2)\Big(\frac{\gamma^{T'} + \epsilon_1}{\tau} + \epsilon_2 + \frac{1}{\sqrt{T\beta}}\Big)\Big]. \qquad (14)
\end{aligned}$$

Proof Sketch of Theorem 1. The rate (13) is straightforward since $\pi_{\widetilde{T}} := \pi_{\widetilde{T},T'} \to \pi_{p_{\widetilde{T}}}$ exponentially fast as $T' \to \infty$. The rate (14) relies on Propositions 2 and 3 which apply to different functions $F_{\rho,\tau}$ and $J_{\rho,\tau}$ respectively. We tackle this challenge by using the connection $\widehat{\nabla}_p J_{\rho,\tau}(\pi_t, p_t) \approx \nabla F_{\rho,\tau}(p_t)$ between the two functions, which implies that

Algorithm 2 Accelerated Stochastic Robust Policy Gradient

Inputs: $\tau, T, T', T_1, \alpha, \eta, \beta, N, H$.
Initialize: p_0.
for transition update steps $t = 0, 1, \ldots, T-1$ **do**
  Initialize $\pi_{t,0}(a|s) \equiv 1/|\mathcal{A}|$.
  for policy update steps $k = 0, 1, \ldots, T'-1$ **do**
    For $\pi = \pi_{t,k}$ and $p = p_t$, perform the TD update rule (15) for T_1 iterations.
    Assign $\widehat{Q}_{t,k} \leftarrow \overline{q}_{T_1} := \frac{1}{T_1} \sum_{n=1}^{T_1} q_n$.
    Obtain $\pi_{t,k+1}$ by the NPG update step (7).
  end for
  Let $\pi_t := \pi_{t,T'}$.
  Obtain $\widehat{\nabla}_p J_{\rho,\tau}(\pi_t, p_t)$ using eq. (16).
  Obtain p_{t+1} by the projected gradient ascent step (8).
end for
Output: $\pi_{\widetilde{T}}, p_{\widetilde{T}}$ where $\widetilde{T} \in \operatorname{argmin}_{0 \le t \le T-1} \|p_{t+1} - p_t\|$.

Algorithm 1 is essentially a projected gradient ascent algorithm on the ℓ_F-smooth objective $\max_p F_{\rho,\tau}(p)$. Hence, we can apply the standard convergence analysis to the projected gradient $G_t = (p_{t+1} - p_t)/\beta$ and obtain the convergence rate of $\|G_{\widetilde{T}}\|$. As G_t is defined by the projection step (8) involving $\widehat{\nabla}_p J_{\rho,\tau}(\pi_t, p_t)$, we can apply properties about projection which implies that $\max_{p' \in \mathcal{P}} (p' - p_{\widetilde{T}})^\top \widehat{\nabla}_p J_{\rho,\tau}(\pi_{\widetilde{T}}, p_{\widetilde{T}}) \le \mathcal{O}(\|G_{\widetilde{T}}\|)$. This bound along with Proposition 3 implies the convergence rate (14). □

Under deterministic setting where we can access exact Q function Q_τ and gradient $\nabla_p J_{\rho,\tau}$, we have $\epsilon_1 = \epsilon_2 = 0$ in Algorithm 1. In this case, we obtain the following iteration complexity result based on Theorem 1.

Corollary 1 (Iteration Complexity of Algorithm 1)**.** *Implement Algorithm 1 under deterministic setting ($\epsilon_1 = \epsilon_2 = 0$). For any $\epsilon > 0$, select hyperparameters $\tau = \min\big(\frac{\epsilon(1-\gamma)}{3 \ln |\mathcal{A}|}, 1\big)$, $T = \mathcal{O}(\epsilon^{-3})$, $T' = \mathcal{O}[\ln(\epsilon^{-1})]$, $\eta = \frac{1-\gamma}{\tau}$, $\beta = \frac{1}{2\ell_F}$. Then the output $(\pi_{\widetilde{T}}, p_{\widetilde{T}})$ is both ϵ-optimal robust policy and (ϵ, τ)-Nash equilibrium under Assumption 1. This requires $T = \mathcal{O}(\epsilon^{-3})$ transition kernel updates, $TT' = \mathcal{O}(\epsilon^{-3} \ln \epsilon^{-1})$ policy updates and iteration complexity $T + TT' = \mathcal{O}(\epsilon^{-3} \ln \epsilon^{-1})$.*

Comparison with Existing Works: With the aforementioned amenable geometric properties given by the entropy regularization, our iteration complexity $\mathcal{O}(\epsilon^{-3} \ln \epsilon^{-1})$ is lower than the state of the art $\mathcal{O}(\epsilon^{-4})$ (Guha and Lee, 2023) among the existing policy gradient methods for s-rectangular robust MDP under deterministic setting. In addition, our Algorithm 1 is also the first policy gradient algorithm with global convergence to both the robust MDP and the entropy regularized robust MDP.

3.3. Stochastic Estimation and Sample Complexity

We will extend Algorithm 1 to the stochastic setting where the Q function $Q_\tau(\pi,p) := Q_\tau(\pi,p;\cdot,\cdot) \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ and $\nabla_p J_{\rho,\tau}(\pi,p)$ for fixed π and p can only be estimated by stochastic samples, and obtain the sample complexity result.

We estimate $Q_\tau(\pi,p)$ via the following temporal difference (TD) update rule (Bhandari et al., 2018; Xu et al., 2021; Li et al., 2023a; Samsonov et al., 2023).

$$q_{n+1}(s_n,a_n) = q_n(s_n,a_n) + \alpha[c(s_n,a_n,s_n') + \tau \ln \pi(a_n|s_n) + \gamma q_n(s_n',a_n') - q_n(s_n,a_n)]; n = 0,1,\dots,T_1-1, \quad (15)$$

where $s_n \sim \mu_{\pi,p}$ (the state stationary distribution under π,p), $a_n \sim \pi(\cdot|s_n)$, $s_n' \sim p(\cdot|s_n,a_n)$, $a_n' \sim \pi(\cdot|s_n')$, and $c(s_n,a_n,s_n')+\tau\ln\pi(a_n|s_n)$ can be seen as regularized cost. We use $\bar{q}_{T_1} := \frac{1}{T_1}\sum_{n=1}^{T_1} q_n$ as the output which provably converges to $Q_\tau(\pi,p)$ (Li et al., 2023a).

A stochastic estimation of $\nabla_p J_{\rho,\tau}(\pi,p)$ defined in eq. (9) can be obtained as follows.

$$\widehat{\nabla}_p J_{\rho,\tau}(\pi,p)(s,a,s') = \frac{1}{N(1-\gamma)}\sum_{i=1}^{N}\pi(a|s)\mathbb{1}\{s_{i,H_i}=s\} \Big[c(s,a,s') + \tau\ln\pi(a|s) + \gamma\sum_{a'}\pi(a'|s')q_{T_1}(s',a')\Big], \quad (16)$$

where $\mathbb{1}(\cdot)$ is an indicator function, H_i is generated by $\mathbb{P}(H_i=h) \propto \gamma^h (h=0,1,\dots,H-1)$, a geometric distribution truncated at level H, and then every i-th trajectory $\{s_{i,h},a_{i,h}\}_{h=0}^{H_i}$ is generated via $s_{i,0}\sim\rho$, $a_{i,h}\sim\pi(\cdot|s_{i,h})$, $s_{i,h+1}\sim p(\cdot|s_{i,h},a_{i,h})$ such that the distribution of s_{i,H_i} is $\mathcal{O}(\gamma^H)$-close to $d_\rho^{\pi,p}$.

By estimating $Q_\tau(\pi,p)$ and $\nabla_p J_{\rho,\tau}(\pi,p)$ using the TD rule (15) and stochastic gradient (16) respectively, we obtain a stochastic implementation of Algorithm 1 in Algorithm 2, the first stochastic policy gradient method for s-rectangular robust MDP to our knowledge. Due to entropy regularization, a major challenge to obtain the sample complexity of Algorithm 2 is to bound the regularized cost $c(s,a,s')+\tau\ln\pi(a|s)$ involved in the estimations (15) and (16), where $\pi(a|s)$ may approach 0. To tackle this, we can prove that the policy π obtained by the NPG step (7) satisfies $|\tau\ln\pi(a|s)| \le \mathcal{O}(1)$ by the following Lemma.

Lemma 1. *The NPG step (7) with $\|\widehat{Q}_{t,k} - Q_\tau(\pi_{t,k},p_t)\|_\infty \le \epsilon_1$, stepsize $\eta = \frac{1-\gamma}{\tau}$ and initial policy $\pi_{t,0}(a|s) \equiv 1/|\mathcal{A}|$ always guarantees*

$$0 \ge \ln\pi_{t,k}(a|s) \ge -\frac{3\ln|\mathcal{A}|+3/\tau}{1-\gamma} - \frac{4\epsilon_1}{\tau(1-\gamma)^2}. \quad (17)$$

Proof Sketch of Lemma 1: The proof is shown in the proof of Lemma 9 in Appendix D. Since $\ln\pi_{t,k} \to \ln\pi_t^* := \ln\pi_{p_t}$ exponentially fast as $k\to\infty$, we only need to lower bound $\ln\pi_t^*$, which can be obtained by using the analytical solution of $\ln\pi_t^*$ (see Lemma 4 in Appendix B).

With this bounded cost, it is well known that TD (15) achieves $\|q_{T_1} - Q_\tau(\pi,p)\|_\infty \le \epsilon_1$ with $T_1 = \mathcal{O}(\epsilon_2^{-2}\ln\epsilon_2^{-1})$ iterations (Li et al., 2023a), and the stochastic gradient (16) achieves $\|\widehat{\nabla}_p J_{\rho,\tau}(\pi_t,p_t) - \nabla_p J_{\rho,\tau}(\pi_t,p_t)\| \le \epsilon_2$ with $\mathcal{O}(\epsilon_2^{-2}\ln\epsilon_2^{-1})$ stochastic samples (see Lemma 14). These results along with the convergence rates in Theorem 1 yield the following sample complexity result of Algorithm 2.

Corollary 2 (Sample Complexity of Algorithm 2)**.** *For any $\epsilon > 0$ and $\delta \in (0,1)$, implement Algorithm 2 with hyperparameters $\tau = \min\left(\frac{\epsilon(1-\gamma)}{3\ln|\mathcal{A}|},1\right)$, $T = \mathcal{O}(\epsilon^{-3})$, $T' = \mathcal{O}[\ln(\epsilon^{-1})]$, $T_1 = \mathcal{O}(\epsilon^{-4})$, $\alpha = \mathcal{O}[\ln^{-1}(\epsilon^{-1})]$, $\eta = \frac{1-\gamma}{\tau}$, $\beta = \frac{1}{2\ell_F}$, $N = \mathcal{O}(\epsilon^{-2})$, $H = \mathcal{O}[\ln(\epsilon^{-1})]$. The output $(\pi_{\widetilde{T}}, p_{\widetilde{T}})$ is both ϵ-optimal robust policy and (ϵ,τ)-Nash equilibrium with probability at least $1-\delta$ under Assumption 1. Furthermore, the sample complexity is $T(T'T_1+NH) = \mathcal{O}(\epsilon^{-7}\ln\epsilon^{-1})$.*

4. Extension to Large State Space

Algorithms 1-2 and the existing policy gradient algorithms for s-rectangular robust MDPs need to compute $\pi(a|s)$, $p(s'|s,a)$ or $Q(s,a)$ for all states $s,s' \in \mathcal{S}$ and actions $a \in \mathcal{A}$. This is intractable in many practical applications where the state space is very large. We will introduce two key techniques to reduce such state enumeration, namely, transition kernel parameterization and Q function approximation. Based on these techniques, we extend Algorithm 1 to large space and obtain sample complexity result.

4.1. Transition Kernel Parameterization

To reduce the dimensionality $|\mathcal{S}|^2|\mathcal{A}|$ of the transition kernel p, we adopt a linear transition kernel parameterization which has also been used in linear mixture MDP (Ayoub et al., 2020; Zhou et al., 2021; Zhang et al., 2023) and robust MDP (Li et al., 2023b). Linear kernel parameterization can be written as $p_\xi(s'|s,a) = \psi(s,a,s')^\top\xi$ with fixed and known features $\psi(s,a,s') \in \mathbb{R}^{d_p}$ and unkonwn parameter $\xi \in \Xi \subset \mathbb{R}^{d_p}$ with $d_p \ll |\mathcal{S}|^2|\mathcal{A}|$. The ambiguity set Ξ can be defined as a neighborhood of a nominal parameter $\bar{\xi}$ which can be estimated from data. For simplicity, the linear kernel parameterization can be rewritten as $p_\xi = \Psi\xi$ where the (s,a,s')-th row of the feature matrix $\Psi \in \mathbb{R}^{|\mathcal{S}|^2|\mathcal{A}|\times d_p}$ is $\psi(s,a,s')^\top$.

The gradient for $p_\xi := \Psi\xi$ has the following expression [4].

$$\nabla_\xi J_{\rho,\tau}(\pi,p_\xi) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi,p_\xi},a\sim\pi(\cdot|s),s'\sim p_\xi(\cdot|s,a)}$$

[4] The gradient (18) can be obtained by applying Lemma 4.5 of (Wang et al., 2023) with the cost $c(s,a,s')$ replaced by $c(s,a,s')+\tau\ln\pi(a|s)$ and using $\nabla\ln p_\xi(s'|s,a) = \frac{\psi(s,a,s')}{p_\xi(s'|s,a)}$

$$\left[\frac{\psi(s,a,s')}{p_\xi(s'|s,a)}[c(s,a,s')+\tau\ln\pi(a|s)+\gamma V_\tau(\pi,p_\xi;s')]\right] \quad (18)$$

Similar to the stochastic gradient (16), a stochastic sample-based estimation of the above gradient is shown below.

$$\widehat{\nabla}_\xi J_{\rho,\tau}(\pi,p_\xi)=\frac{1}{N(1-\gamma)}\sum_{i=1}^{N}\frac{\psi(s_{i,H_i},a_{i,H_i},s_{i,H_i+1})}{p_\xi(s_{i,H_i+1}|s_{i,H_i},a_{i,H_i})}$$
$$\big[c(s_{i,H_i},a_{i,H_i},s_{i,H_i+1})+\tau\ln\pi(a_{i,H_i}|s_{i,H_i})$$
$$+\gamma\phi(s_{i,H_i+1},a_{i,H_i+1})^\top\overline{w}_{T_1}\big], \quad (19)$$

where H_i is generated by $\mathbb{P}(H_i=h)\propto\gamma^h(h=0,1,\ldots,H-1)$ and then the trajectory $\{s_{i,h},a_{i,h}\}_{h=0}^{H_i+1}$ is generated via $s_{i,0}\sim\rho$, $a_{i,h}\sim\pi(\cdot|s_{i,h})$, $s_{i,h+1}\sim p_\xi(\cdot|s_{i,h},a_{i,h})$. In large state space, this computation of $\widehat{\nabla}_\xi J_{\rho,\tau}(\pi,p_\xi)\in\mathbb{R}^{d_p}$ with $d_p\ll|\mathcal{S}|^2|\mathcal{A}|$ is less intensive than to compute $\widehat{\nabla}_p J_{\rho,\tau}(\pi,p)(s,a,s')$ for every (s,a,s').

With the stochastic gradient (16), we can apply projected stochastic gradient ascent to ξ as follows.

$$\xi_{t+1}=\text{proj}_\Xi\big(\xi_t+\beta\widehat{\nabla}_\xi J_{\rho,\tau}(\pi_t,p_{\xi_t})\big). \quad (20)$$

where $\pi_t\approx\pi_{p_{\xi_t}}$ can be obtained by policy optimization in the next subsection.

4.2. Q Function Approximation

To avoid direct evaluation of $Q_\tau(\pi,p;s,a)$ for every s,a, we adopt the popular linear Q function approximation $Q_\tau(\pi,p;s,a)\approx\phi(s,a)^\top w$ with fixed and known features $\phi(s,a)\in\mathbb{R}^d$ and parameter $w\in\mathbb{R}^d$ (Huh and Lee, 2018; Zou et al., 2019; Wang et al., 2021; Zhou et al., 2023).

The parameter w can be estimated by the following TD algorithm (Huh and Lee, 2018; Li et al., 2023a).

$$w_{n+1}=w_n+\alpha\phi(s_n,a_n)[c(s_n,a_n,s_n')+\tau\ln\pi(a_n|s_n)$$
$$+\gamma\phi(s_n',a_n')^\top w_n-\phi(s_n,a_n)^\top w_n];n=0,1,\ldots,T_1-1, \quad (21)$$

where $s_n\sim\mu_{\pi,p}$ (stationary distribution), $a_n\sim\pi(\cdot|s_n)$, $s_n'\sim p(\cdot|s_n,a_n)$, $a_n'\sim\pi(\cdot|s_n')$. We take $\overline{w}_{T_1}:=\frac{1}{T_1}\sum_{n=1}^{T_1}w_n$ as the output which has provable convergence to the optimal parameter (Li et al., 2023a).

Suppose we got $\widehat{Q}_{t,k}=\phi(s,a)^\top w_{t,k}$ via the above TD algorithm. Then the NPG step (7) can be rewritten below.

$$\pi_{t,k}(\cdot|s)\propto\exp\left[-\frac{\phi(s,\cdot)^\top u_{t,k}}{1-\gamma}\right], \quad (22)$$
$$\text{where } u_{t,k+1}=u_{t,k}+\eta w_{t,k}. \quad (23)$$

In this way, the policy $\pi_{t,k}$ is implicitly parameterized by $u_{t,k}$. Instead of computing $\pi_{t,k}(a|s)$ for every s,a, we only need to compute $\pi_{t,k}(\cdot|s)$ above to obtain action samples given the state samples s, which significantly reduces the computation for large state space.

Based on the above discussions, we extend Algorithm 2 to Algorithm 3, the first stochastic policy gradient algorithm for s-rectangular robust MDP with large state space, by changing the outer update of p_{t+1} to eq. (20) under linear kernel parameterization, and changing the inner update rule of π_t to eq. (22) under linear Q function approximation.

4.3. Sample Complexity of Algorithm 3

The global convergence of Algorithm 3 is largely guaranteed by linear kernel parameterization, which preserves the Lipschitz smoothness and gradient dominance. To elaborate, $\nabla_\xi F_{\rho,\tau}(p)=\Psi^\top\nabla F_{\rho,\tau}(p)$ is $\ell_F\|\Psi\|$-smooth based on Proposition 2. Similar to Proposition 3, we have the following gradient dominance property.

Proposition 4. *Under Assumption 1, the function $J_{\rho,\tau}$ satisfies the following gradient dominance property for any $\pi\in\Pi$, $\xi\in\Xi$.*

$$\max_{p'\in\mathcal{P}}J_{\rho,\tau}(\pi,p')-J_{\rho,\tau}(\pi,p_\xi)$$
$$\le\frac{D}{1-\gamma}\max_{\xi'\in\Xi}(\xi'-\xi)^\top\nabla_\xi J_{\rho,\tau}(\pi,p_\xi). \quad (24)$$

Define $\zeta:=\sup_{\pi\in\Pi,p\in\mathcal{P},s\in\mathcal{S},a\in\mathcal{A},\tau\in[0,1]}|\phi(s,a)^\top w^*_{\pi,p}-Q_\tau(\pi,p;s,a)|^2$ as the linear Q function approximation error where $w^*_{\pi,p}$ is the optimal critic parameter [5] (Xu et al., 2020; Chen et al., 2022). Then we obtain the sample complexity of Algorithm 3 as follows.

Theorem 2 (Sample Complexity of Algorithm 3)**.** *For any $\epsilon>0$ and $\delta\in(0,1)$, implement Algorithm 3 with hyperparameters $\tau=\min\big[\mathcal{O}(\sqrt{\zeta}+\epsilon),1\big]$, $T=\mathcal{O}(\epsilon^{-3})$, $T'=\mathcal{O}[\ln(\epsilon^{-1})]$, $T_1:=\mathcal{O}(\epsilon^{-4})$, $\alpha=\mathcal{O}[\ln^{-1}(\zeta+\epsilon^2)^{-1}]$, $\eta=\frac{1-\gamma}{\tau}$, $\beta=\frac{1}{2\ell_F\|\Psi\|}$, $N=\mathcal{O}(\epsilon^{-4})$, $H=\mathcal{O}[\ln(\epsilon^{-1})]$. Then under Assumption 1 and the assumption that $\inf_{s,a,s'}p_\xi(s'|s,a)>p_{\min}$ for a constant $p_{\min}>0$, $(\pi_{\widetilde{T}},p_{\widetilde{T}})$ is both $\big(\mathcal{O}(\sqrt{\zeta}+\zeta+\epsilon),\tau\big)$-Nash equilibrium and $\mathcal{O}(\sqrt{\zeta}+\zeta+\epsilon)$-optimal robust policy with probability at least $1-\delta$. The required sample complexity is $T(T'T_1+NH)=\mathcal{O}(\epsilon^{-7}\ln\epsilon^{-1})$.*

The sample complexity above is the same as that in Corollary 2 for small state space. The major difference is that the convergence error $\epsilon>0$ becomes $\mathcal{O}(\sqrt{\zeta}+\zeta+\epsilon)$ due to the linear Q function approximation error term ζ. The linear kernel parameterization does not cause additional error terms since we have proved that linear kernel parameterization preserves the amenable geometric properties of Lipschitz smoothness and gradient dominance.

[5] See eq. (52) of Appendix E for the expression of the optimal critic parameter $w^*_{\pi,p}$.

Algorithm 3 Accelerated Stochastic Robust Policy Gradient for Large State Space

Inputs: $\tau, T, T', T_1, \alpha, \eta, \beta, N, H$.
Initialize: ξ_0.
for transition update steps $t = 0, 1, \ldots, T-1$ **do**
  Initialize $u_{t,0} = 0$.
  for policy update steps $k = 0, 1, \ldots, T'-1$ **do**
    For $\pi = \pi_{t,k}$ defined by eq. (22) and $p = p_{\xi_t}$, perform the TD update rule (21) for T_1 iterations.
    Assign $w_{t,k} \leftarrow \overline{w}_{T_1} := \frac{1}{T_1}\sum_{n=1}^{T_1} w_n$.
    Obtain $u_{t,k+1}$ by eq. (23).
  end for
  For $\pi_t := \pi_{t,T'}$ defined by eq. (22), obtain stochastic gradient $\widehat{\nabla}_\xi J_{\rho,\tau}(\pi_t, p_{\xi_t})$ using eq. (19).
  Obtain ξ_{t+1} by projected gradient ascent step (20).
end for
Output: $\pi_{\widetilde{T}}, \xi_{\widetilde{T}}$ where $\widetilde{T} \in \arg\min_{0\le t\le T-1} \|\xi_{t+1} - \xi_t\|$.

5. Conclusion

This work proposes a policy gradient algorithm with faster global convergence than existing policy gradient algorithms on s-rectangular robust MDP, by solving an entropy regularized robust MDP. We further extend this algorithm to stochastic setting and large state space, and obtain the first sample complexity results for policy gradient on s-rectangular robust MDP. Moreover, our algorithms are also the first policy gradient methods that can solve entropy regularized robust MDP problem, which is an important but underexplored area. Since $F_{\rho,\tau}(p)$ is Lipschitz smooth with parameter $\ell_F = \mathcal{O}(\tau^{-1})$ (see Proposition 2) while $\tau = \mathcal{O}(\epsilon)$ is required for ϵ-accuracy (see Proposition 1), our algorithm requires small stepsize $\beta = \mathcal{O}(\epsilon)$ and thus the iteration complexity and sample complexities are not minimax optimal. An interesting future direction is to further accelerate our algorithms using techniques such as Nesterov's acceleration and variance reduction, etc.

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, none which we feel must be specifically highlighted here.

Acknowledgements

This work was partially supported by NSF IIS 2347592, 2347604, 2348159, 2348169, DBI 2405416, CCF 2348306, CNS 2347617.

References

Abbeel, P., Coates, A., Quigley, M., and Ng, A. Y. (2006). An application of reinforcement learning to aerobatic helicopter flight. In *Proceedings of the International Conference on Neural Information Processing Systems (Neurips)*, pages 1–8.

Abdolmaleki, A., Springenberg, J. T., Tassa, Y., Munos, R., Heess, N., and Riedmiller, M. (2018). Maximum a posteriori policy optimisation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2021). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 22(1):4431–4506.

Altman, E. (2004). *Constrained Markov decision processes*. CRC press. https://www-sop.inria.fr/members/Eitan.Altman/PAPERS/h.pdf.

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

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. (2020). Model-based reinforcement learning with value-targeted regression. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 463–474.

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

Bernhard, P. and Rapaport, A. (1995). On a theorem of danskin with an application to a theorem of von neumann-sion. *Nonlinear Analysis: Theory, Methods & Applications*, 24(8):1163–1181.

Bhandari, J. and Russo, D. (2021). On the linear convergence of policy gradient methods for finite mdps. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2386–2394.

Bhandari, J., Russo, D., and Singal, R. (2018). A finite time analysis of temporal difference learning with linear function approximation. In *Proceedings of the Conference on learning theory (COLT)*, pages 1691–1692.

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

Cayci, S., He, N., and Srikant, R. (2022). Finite-time analysis of entropy-regularized neural natural actor-critic algorithm. *ArXiv:2206.00833*.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. (2022). Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578.

Chen, Z., Zhou, Y., Chen, R.-R., and Zou, S. (2022). Sample and communication-efficient decentralized actor-critic algorithms with finite-time analysis. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3794–3834.

Eysenbach, B. and Levine, S. (2021). Maximum entropy rl (provably) solves some robust rl problems. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

Guha, E. K. and Lee, J. D. (2023). Solving robust mdps through no-regret dynamics. *ArXiv:2305.19035*.

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

Huh, J. and Lee, D. D. (2018). Efficient sampling with q-learning to guide rapidly exploring random trees. *IEEE Robotics and Automation Letters*, 3(4):3868–3875.

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

Kakade, S. (2001). A natural policy gradient. In *Proceedings of the International Conference on Neural Information Processing Systems (Neurips)*, pages 1531–1538.

Kober, J., Bagnell, J. A., and Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274.

Konda, V. R. and Tsitsiklis, J. N. (1999). Actor-citic agorithms. In *Proceedings of the International Conference on Neural Information Processing Systems (Neurips)*, pages 1008–1014.

Kumar, N., Derman, E., Geist, M., Levy, K., and Mannor, S. (2023a). Policy gradient for rectangular robust markov decision processes. In *Proceedings of the International Conference on Neural Information Processing Systems (Neurips)*.

Kumar, N., Levy, K., Wang, K., and Mannor, S. (2022). Efficient policy iteration for robust markov decision processes via regularization. *ArXiv:2205.14327*.

Kumar, N., Levy, K., Wang, K., and Mannor, S. (2023b). An efficient solution to s-rectangular robust markov decision processes. *ArXiv:2301.13642*.

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

Leonardos, S., Overman, W., Panageas, I., and Piliouras, G. (2022). Global convergence of multi-agent policy gradient in markov potential games. In *ICLR 2022 Workshop on Gamification and Multiagent Solutions*.

Li, G., Wu, W., Chi, Y., Ma, C., Rinaldo, A., and Wei, Y. (2023a). Sharp high-probability sample complexities for policy evaluation with linear function approximation. *ArXiv:2305.19001*.

Li, M., Sutter, T., and Kuhn, D. (2023b). Policy gradient algorithms for robust mdps with non-rectangular uncertainty sets. *ArXiv:2305.19004*.

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

Li, Y., Lan, G., and Zhao, T. (2023c). First-order policy optimization for robust markov decision process. *ArXiv:2209.10579*.

Mai, T. and Jaillet, P. (2021). Robust entropy-regularized markov decision processes. *ArXiv:2112.15364*.

Mankowitz, D. J., Levine, N., Jeong, R., Abdolmaleki, A., Springenberg, J. T., Shi, Y., Kay, J., Hester, T., Mann, T., and Riedmiller, M. (2019). Robust reinforcement learning for continuous control with model misspecification. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

Peng, X. B., Andrychowicz, M., Zaremba, W., and Abbeel, P. (2018). Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3803–3810. IEEE.

Perera, A. and Kamalaruban, P. (2021). Applications of reinforcement learning in energy systems. *Renewable and Sustainable Energy Reviews*, 137:110618.

Samsonov, S., Tiapkin, D., Naumov, A., and Moulines, E. (2023). Finite-sample analysis of the temporal difference learning. *ArXiv:2310.14286*.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. (2014). Deterministic policy gradient algorithms. In *Proceedings of the International conference on machine learning (ICML)*, pages 387–395.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the International Conference on Neural Information Processing Systems (Neurips)*, pages 1057–1063.

Wang, Q., Ho, C. P., and Petrik, M. (2023). Policy gradient in robust mdps with global convergence guarantee. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 202, pages 35763–35797.

Wang, T., Zhou, D., and Gu, Q. (2021). Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. In *Proceedings of the International Conference on Neural Information Processing Systems (Neurips)*.

Wang, Y. and Zou, S. (2022). Policy gradient method for robust reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 23484–23526.

Wang, Y.-C. and Usher, J. M. (2005). Application of reinforcement learning for agent-based production scheduling. *Engineering applications of artificial intelligence*, 18(1):73–82.

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

Xiao, L. (2022). On the convergence rates of policy gradient methods. *The Journal of Machine Learning Research*, 23(1):12887–12922.

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

Xu, T., Wang, Z., and Liang, Y. (2020). Improving sample complexity bounds for (natural) actor-critic algorithms. In *Proceedings of the International Conference on Neural Information Processing Systems (Neurips)*, pages 4358–4369.

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

Zhang, J., Zhang, W., and Gu, Q. (2023). Optimal horizon-free reward-free exploration for linear mixture MDPs. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 202, pages 41902–41930.

Zhou, D., Gu, Q., and Szepesvari, C. (2021). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 4532–4576.

Zhou, R., Liu, T., Cheng, M., Kalathil, D., Kumar, P., and Tian, C. (2023). Natural actor-critic for robust reinforcement learning with function approximation. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems (Neurips)*.

Zou, S., Xu, T., and Liang, Y. (2019). Finite-sample analysis for sarsa with linear function approximation. In *Proceedings of the International Conference on Neural Information Processing Systems (Neurips)*, pages 8668–8678.

Appendix

Table of Contents

A. Existing Complexity Results of Robust Policy Gradient

In this section, we will explain the existing complexity results listed in Table 1. We will also explain about the claim of (Li et al., 2023b; Guha and Lee, 2023) that their complexity results do not require rectangularity assumption.

A.1. Complexity Results of (Wang et al., 2023)

(Wang et al., 2023) proposes a double-loop robust policy gradient (DRPG) algorithm for robust MDP with s-rectangular ambiguity set, which applies projected gradient descent steps to update the policy π in the outer loop and projected gradient ascent steps to update the transition kernel p in the inner loop. Based on Theorem 3.3 of (Wang et al., 2023), to obtain an ϵ-optimal robust policy, DRPG requires $T = \mathcal{O}(\epsilon^{-4})$ outer iterations of the policy updates, and the t-th outer iteration requires a transition kernel p_t such that $J_\rho(\pi_t, p_t) \geq \max_{p\in\mathcal{P}} J_\rho(\pi_t, p) - \epsilon_t$, where the precision $\epsilon_t > 0$ satisfies $\epsilon_{t+1} \leq \gamma\epsilon_t$ and $\epsilon_0 \leq \sqrt{T}$. Such an ϵ_t-accurate p_t further requires $T_t = \mathcal{O}(\epsilon_t^{-2})$ iterations of the transition kernel updates based on Theorem 4.4 of (Wang et al., 2023). T_t is exponentially growing as $T_t \geq \mathcal{O}(\gamma^t\epsilon_0)^{-2} \geq \mathcal{O}(T^{-1}\gamma^{-2t})$. Hence, the required number of policy updates is

$$\begin{aligned}\max\Big(T, \sum_{t=0}^{T} T_t\Big) &\geq \max\Big(T, T^{-1}\sum_{t=0}^{T-1}\mathcal{O}(\gamma^{-2t})\Big)\\ &= \max\big(T, \mathcal{O}(T^{-1}\gamma^{-2T})\big)\\ &= \max\big(\mathcal{O}(\epsilon^{-4}), \mathcal{O}(\epsilon^4\gamma^{-\mathcal{O}(\epsilon^{-4})})\big)\\ &= \mathcal{O}(\epsilon^{-4} + \epsilon^4\gamma^{-\mathcal{O}(\epsilon^{-4})}).\end{aligned}$$

Therefore, the iteration complexity (the total number of updates on both transition kernel and policy) is $T + \mathcal{O}(\epsilon^{-4} + \epsilon^4\gamma^{-\mathcal{O}(\epsilon^{-4})}) = \mathcal{O}(\epsilon^{-4} + \epsilon^4\gamma^{-\mathcal{O}(\epsilon^{-4})})$, which exponentially increases as $\epsilon \to +0$.

A.2. Complexity Results of (Li et al., 2023b)

(Li et al., 2023b) proposes an actor-critic algorithm (see their Algorithm 4.1) which has similar double loop structure as the DRPG algorithm (Wang et al., 2023). To achieve an ϵ-optimal robust policy, this actor-critic algorithm requires $\mathcal{O}(\epsilon^{-4})$ outer policy updates and $\mathcal{O}(\epsilon^{-2})$ inner transition kernel updates per outer iteration, based on Theorems 4.5 and 3.8 of (Li et al., 2023b) respectively. Therefore, the total number of transition kernel updates is $\mathcal{O}(\epsilon^{-4})\mathcal{O}(\epsilon^{-2}) = \mathcal{O}(\epsilon^{-6})$, so the iteration complexity is $\mathcal{O}(\epsilon^{-4}) + \mathcal{O}(\epsilon^{-6}) = \mathcal{O}(\epsilon^{-6})$.

A.3. Extension to Non-rectangularity in (Li et al., 2023b) Also Applies to Our Work

(Li et al., 2023b) extends their complexity results to non-rectangular ambiguity set $\mathcal{P}$ by defining the following degree of non-rectangularity.

$$\delta_{\mathcal{P}} := \max_{p'\in\mathcal{P}}\Big[\max_{p_s\in\mathcal{P}_s}\langle\nabla_p J_\rho(\pi, p'), p_s\rangle - \max_{p\in\mathcal{P}}\langle\nabla_p J_\rho(\pi, p'), p\rangle\Big] \tag{25}$$

where $\mathcal{P}_s$ denotes the smallest s-rectangular ambiguity set containing $\mathcal{P}$.

Then the proof of the convergence for the inner transition kernel updates in (Li et al., 2023b) (see their Theorem 3.8) uses the following gradient dominance property.

$$\max_{p'\in\mathcal{P}} J_\rho(\pi, p') - J_\rho(\pi, p) \leq \Big\|\frac{d_\rho^{\pi,p_s^*}}{d_\rho^{\pi,p}}\Big\|_\infty \max_{p_s\in\mathcal{P}_s}\langle\nabla_p J_\rho(\pi, p), p_s - p\rangle \overset{(i)}{\leq} \frac{D}{1-\gamma}\Big[\delta_{\mathcal{P}} + \max_{p'\in\mathcal{P}}\langle\nabla_p J_\rho(\pi, p), p' - p\rangle\Big], \tag{26}$$

where $p_s^* \in \arg\max_{p'\in\mathcal{P}_s} J_\rho(\pi, p')$ denotes the optimal transition kernel and (i) uses $D := \sup_{\pi\in\Pi, p\in\mathcal{P}} \|d_\rho^{\pi,p}/\rho\|_\infty < \infty$ and $d_\rho^{\pi,p}(s) \geq (1-\gamma)\rho(s)$. Compared with the gradient dominance property (Proposition 3) used in our convergence proof for s-rectangular case, the above gradient dominance property involves the degree of non-rectangularity $\delta_{\mathcal{P}} > 0$ defined in eq. (25). Hence, we can also extend our convergence result to non-rectangular robust MDPs in the same way by replacing Proposition 3 with the above gradient dominance property (26), where the objective function J_ρ should be changed to $J_{\rho,\tau}$ to fit our entropy regularized case.

A.4. Complexity Results of (Kumar et al., 2023c)

(Kumar et al., 2023c) aims to solve the following robust MDP problem,

$$\max_\pi \min_{(P,R)\in\mathcal{U}} \rho_{P,R}^\pi := \mathbb{E}_{\pi,p}\Big[\sum_{t=0}^{\infty}\gamma^t R(s_t, a_t)\Big| s_0 \sim \rho\Big],$$

where $\mathcal{U}$ denotes the ambiguity set and $\rho^{\pi}_{P,R}$ denotes the value function under policy π, transition kernel P and reward function R. This work assumes oracle access to the optimal transition kernel and reward $(P^{\pi}_{\mathcal{U}}, R^{\pi}_{\mathcal{U}}) \in \arg\inf_{(P,R)\in\mathcal{U}} \rho^{\pi}_{(P,R)}$ and assumes that the robust value function $\rho^{\pi}_{\mathcal{U}} := \min_{(P,R)\in\mathcal{U}} \rho^{\pi}_{P,R}$ has Lipschitz-continuous gradient $\nabla_{\pi}\rho^{\pi}_{\mathcal{U}} = \nabla_{\pi}\rho^{\pi}_{P^{\pi}_{\mathcal{U}},R^{\pi}_{\mathcal{U}}}$. Under these assumptions which are not practical in many applications, (Kumar et al., 2023c) proved that it takes $\mathcal{O}(\epsilon^{-1})$ iterations of the following projected gradient ascent steps to obtain an ϵ-robust optimal policy.

$$\pi_{t+1} := \text{proj}_{\Pi}\left(\pi_k + \eta\nabla_{\pi}\rho^{\pi_k}_{\mathcal{U}}\right).$$

However, (Kumar et al., 2023c) has not discussed how to obtain the optimal transition kernel and reward $(P^{\pi}_{\mathcal{U}}, R^{\pi}_{\mathcal{U}})$, so the total iteration complexity defined as the updates of all the variables (π, P and R) is unknown.

A.5. Complexity Results of (Guha and Lee, 2023)

(Guha and Lee, 2023) proposes a gradient-based no-regret RL algorithm, which has T time steps. In each time step, both the policy and transition kernels are updated using $T_{\mathcal{O}}$ projected gradient descent steps. The convergence rate is $\mathcal{O}(T^{-1/2} + T_{\mathcal{O}}^{-1/2})$ based on Theorem 7.2 of (Guha and Lee, 2023). Hence, to obtain ϵ-optimal robust policy, it requires $T, T_{\mathcal{O}} = \mathcal{O}(\epsilon^{-2})$, which means both policy and transition kernels are updated $TT_{\mathcal{O}} = \mathcal{O}(\epsilon^{-4})$ times, so the iteration complexity is also $\mathcal{O}(\epsilon^{-4})$.

A.6. Why Does (Guha and Lee, 2023) Require s-rectangularity

The gradient-based no-regret RL algorithm (Guha and Lee, 2023) claims to globally converge without rectangularity condition. However, their global convergence relies on the following gradient-dominance condition (see their Lemma 6.5), which requires s-rectangularity as will be elaborated soon.

$$V_W(\mu) - V_{W^*}(\mu) \le \frac{-1}{1-\gamma}\left\|\frac{d^W_{\mu}}{\mu}\right\|_{\infty} \min_{\bar{W}\in\mathcal{W}} \left[(\bar{W} - W)^{\top}\nabla_W V_W(\mu)\right], \tag{27}$$

where W, μ, $\mathcal{W}$, $V_W(\mu)$, d^W_{μ}, $\left\|\frac{d^W_{\mu}}{\mu}\right\|_{\infty}$ correspond to our transition kernel p, initial state distribution ρ, ambiguity set $\mathcal{P}$, objective function $J_{\rho}(\pi, p)$ (with fixed policy π), occupancy measure $d^{\pi,p}_{\rho}$ and constant $D := \sup_{\pi\in\Pi, p\in\mathcal{P}} \|d^{\pi,p}_{\rho}/\rho\|_{\infty} < \infty$ respectively.

Their proof of the above gradient dominance property (27) made the following mistake at the beginning of page 16.

$$\sum_{s',a,s}\left[\gamma^t d^W_{\mu}(s)\pi(a \mid s)\min_{s'}\left(A^W(s',a,s)\right)\right] = \min_{\bar{W}\in\mathcal{W}}\sum_{s',a,s}\left[\gamma^t d^W_{\mu}(s)\pi(a \mid s)\mathbb{P}_{\bar{W}}(s',a,s)\left(A^W(s',a,s)\right)\right] \tag{28}$$

where $\mathbb{P}_{\bar{W}} = \bar{W}$, $A^W(s',a,s) := \gamma V_W(s') + r(s,a) - V_W(s)$ ((Guha and Lee, 2023) uses reward function r instead of our cost c). The above equality uses the fact that the right side is minimized when $\mathbb{P}_{\overline{W}}(s',a,s) = 1$ for $s' \in \arg\min_{s'} A^W(s',a,s)$. However, this is not true since $A^W(s',a,s) < 0$ is possible. Furthermore, even if $\inf_{s,a,s'} A^W(s',a,s) \ge 0$, such a deterministic choice of $\mathbb{P}_{\overline{W}}$ does not necessarily satisfy the constraint that $W \in \mathcal{W}$.

The correct proof of the above gradient dominance condition (27) is shown in the proof of Lemma 4.3 in (Wang et al., 2023). At the end of their proof, they use an inequality that requires s-rectangularity condition.

B. Basic Properties of Entropy Regularized Robust MDP

We quote the perfect duality result of entropy regularized robust MDP as follows from Theorem 3.2 of (Mai and Jaillet, 2021).

Lemma 2. *Under Assumption 1, the following perfect duality holds for $J_{\rho,\tau}(\pi, p)$, the objective function of entropy regularized robust MDP defined in eq. (3).*

$$\min_{\pi\in\Pi}\max_{p\in\mathcal{P}} J_{\rho,\tau}(\pi,p) = \max_{p\in\mathcal{P}}\min_{\pi\in\Pi} J_{\rho,\tau}(\pi,p) \tag{29}$$

Proof. Assumption 1 says $\mathcal{P}$ is s-rectangular, compact and convex. Hence, each $\mathcal{P}_s$ is compact and convex, which means all the conditions of Theorem 3.2 of (Mai and Jaillet, 2021) holds, and thus its conclusion of perfect duality follows. □

To facilitate further discussion, we follow (Cen et al., 2022) and define a variant of Q_τ (defined in eq. (5)) as follows.

$$\widetilde{Q}_\tau(\pi, p; s, a) := \mathbb{E}_{s' \sim p(\cdot|s,a)}[c(s, a, s') + \gamma V_\tau(s')] \tag{30}$$

It can be directly seen that $\widetilde{Q}_\tau(\pi, p; s, a) = Q_\tau(\pi, p; s, a) - \tau \ln \pi(a|s)$.

Lemma 3. *If $c(s, a, s') \in [0, 1]$, then the functions $J_{\rho,\tau}$, V_τ, $F_{\rho,\tau}$ and $\widetilde{Q}_\tau$ defined by eqs. (3), (4), (6) and (30) respectively have the following ranges for any $\pi \in \Pi$, $p \in \mathcal{P}$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$.*

$$J_{\rho,\tau}(\pi, p), V_\tau(\pi, p; s), F_{\rho,\tau}(p) \in \Big[-\frac{\tau \ln |\mathcal{A}|}{1-\gamma}, \frac{1}{1-\gamma} \Big] \tag{31}$$

$$\widetilde{Q}_\tau(\pi, p; s, a) \in \Big[-\frac{\gamma\tau \ln |\mathcal{A}|}{1-\gamma}, \frac{1}{1-\gamma} \Big] \tag{32}$$

Proof. We rewrite the function V_τ as follows.

$$\begin{aligned} V_\tau(\pi, p; s, a) &\overset{(i)}{=} \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t [c_t + \tau \ln \pi(a_t|s_t)] \Big| s_0 = s \Big] \\ &\overset{(ii)}{=} \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t \Big(c_t + \tau \sum_a \pi(a|s_t) \ln \pi(a|s_t) \Big) \Big| s_0 = s \Big], \end{aligned}$$

where (i) uses eq. (4) and (ii) uses $a_t \sim \pi(\cdot|s_t)$ conditioned on s_t. Since $c(s, a, s') \in [0, 1]$ and the negative entropy $\sum_a \pi(a|s_t) \ln \pi(a|s_t) \in [-\ln |\mathcal{A}|, 0]$, the range (31) holds for the function V_τ, and thus also holds for the functions $J_{\rho,\tau}(\pi, p) = \mathbb{E}_{s\sim\rho} V_\tau(\pi, p; s)$ and $F_{\rho,\tau}(p) = \min_{\pi\in\Pi} J_{\rho,\tau}(\pi, p)$.

Then the range (32) can be proved as follows.

$$\begin{aligned} \widetilde{Q}_\tau(\pi, p; s, a) &\overset{(i)}{=} \mathbb{E}_{s' \sim p(\cdot|s,a)}[c(s, a, s') + \gamma V_\tau(\pi, p; s')] \\ &\overset{(ii)}{\in} \Big[0 + \gamma \Big(-\frac{\tau \ln |\mathcal{A}|}{1-\gamma} \Big), 1 + \frac{\gamma}{1-\gamma} \Big] = \Big[-\frac{\gamma\tau \ln |\mathcal{A}|}{1-\gamma}, \frac{1}{1-\gamma} \Big], \end{aligned}$$

where (i) uses eq. (30) and (ii) uses $c(s, a, s') \in [0, 1]$ and the range (31). □

Lemma 4. *For any $p \in \mathcal{P}$, the optimal policy $\pi_p := \arg\min_{\pi\in\Pi} J_{\rho,\tau}(\pi, p)$ is unique and has the following lower bound.*

$$\ln \pi_p(a|s) \geq -\frac{\ln |\mathcal{A}| + 1/\tau}{1-\gamma}; \forall s \in \mathcal{S}, a \in \mathcal{A}. \tag{33}$$

Proof. Based on (Cen et al., 2022), $\pi_p := \arg\min_{\pi\in\Pi} J_{\rho,\tau}(\pi, p)$ is unique with the following expression.

$$\pi_p(a|s) = \frac{\exp[-\widetilde{Q}_\tau(\pi_p, p; s, a)/\tau]}{\sum_{a'} \exp[-\widetilde{Q}_\tau(\pi_p, p; s, a')/\tau]}, \tag{34}$$

where $\widetilde{Q}_\tau$ is defined by eq. (30). Therefore, eq. (33) can be proved as follows.

$$\ln \pi_p(a|s) \overset{(i)}{=} \ln \Big(\frac{\exp[-\widetilde{Q}_\tau(\pi_p, p_t; s, a)/\tau]}{\sum_{a'} \exp[-\widetilde{Q}_\tau(\pi_p, p_t; s, a')/\tau]} \Big) \overset{(ii)}{\geq} \ln \Big(\frac{\exp[-1/\tau(1-\gamma)]}{|\mathcal{A}| \exp[\gamma \ln |\mathcal{A}|/(1-\gamma)]} \Big) = -\frac{\ln |\mathcal{A}| + 1/\tau}{1-\gamma} \tag{35}$$

where (i) uses eq. (34) and (ii) uses eq. (32). □

C. Lipschitz Properties

Lemma 5. *The occupancy measure $d_\rho^{\pi,p}(s) := (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{\pi,p}(s_t = s | s_0 \sim \rho)$ satisfies the following Lipschitz properties for any $\pi, \pi' \in \Pi$ and $p, p' \in \mathcal{P}$.*

$$\|d_\rho^{\pi',p} - d_\rho^{\pi,p}\|_1 \leq \frac{\gamma}{1-\gamma} \max_s \|\pi'(\cdot|s) - \pi(\cdot|s)\|_1 \tag{36}$$

$$\|d_\rho^{\pi,p'} - d_\rho^{\pi,p}\|_1 \leq \frac{\gamma}{1-\gamma} \max_{s,a} \|p'(\cdot|s, a) - p(\cdot|s, a)\|_1 \tag{37}$$

Proof. The occupancy measure $d_\rho^{\pi,p}$ satisfies the following equation based on Theorem 3.2 of (Altman, 2004).

$$d_\rho^{\pi,p}(s') = (1-\gamma)\rho(s') + \gamma \sum_{s,a} d_\rho^{\pi,p}(s)\pi(a|s)p(s'|s,a); \forall \pi, p, s'. \tag{38}$$

Therefore, for any $\pi, \pi' \in \Pi$ and $p \in \mathcal{P}$, we have

$$\begin{aligned}
&\|d_\rho^{\pi',p} - d_\rho^{\pi,p}\|_1 \\
&= \sum_{s'} |d_\rho^{\pi',p}(s') - d_\rho^{\pi,p}(s')| \\
&\overset{(i)}{=} \gamma \sum_{s'} \Big| \sum_{s,a} d_\rho^{\pi',p}(s)\pi'(a|s)p(s'|s,a) - \sum_{s,a} d_\rho^{\pi,p}(s)\pi(a|s)p(s'|s,a) \Big| \\
&= \gamma \sum_{s'} p(s'|s,a) \Big| \sum_{s,a} [d_\rho^{\pi',p}(s) - d_\rho^{\pi,p}(s)]\pi'(a|s) + \sum_{s,a} d_\rho^{\pi,p}(s)[\pi'(a|s) - \pi(a|s)] \Big| \\
&\leq \gamma \sum_{s} |d_\rho^{\pi',p}(s) - d_\rho^{\pi,p}(s)| + \gamma \sum_{s,a} d_\rho^{\pi,p}(s)|\pi'(a|s) - \pi(a|s)| \\
&\leq \gamma \|d_\rho^{\pi',p} - d_\rho^{\pi,p}\|_1 + \gamma \max_s \|\pi'(\cdot|s) - \pi(\cdot|s)\|_1
\end{aligned}$$

where (i) uses eq. (38). Then eq. (36) can be proved by rearranging the above inequality.

Next, we will prove eq. (37). For any $\pi \in \Pi$ and $p, p' \in \mathcal{P}$, we have

$$\begin{aligned}
&\|d_\rho^{\pi,p'} - d_\rho^{\pi,p}\|_1 \\
&= \sum_{s'} |d_\rho^{\pi,p'}(s') - d_\rho^{\pi,p}(s')| \\
&\overset{(i)}{=} \gamma \sum_{s'} \Big| \sum_{s,a} d_\rho^{\pi,p'}(s)\pi(a|s)p'(s'|s,a) - \sum_{s,a} d_\rho^{\pi,p}(s)\pi(a|s)p(s'|s,a) \Big| \\
&= \gamma \sum_{s'} \Big| \sum_{s,a} d_\rho^{\pi,p'}(s)\pi(a|s)[p'(s'|s,a) - p(s'|s,a)] + \sum_{s,a} [d_\rho^{\pi,p'}(s) - d_\rho^{\pi,p}(s)]\pi(a|s)p(s'|s,a) \Big| \\
&\leq \gamma \sum_{s,a,s'} d_\rho^{\pi,p'}(s)\pi(a|s)|p'(s'|s,a) - p(s'|s,a)| + \gamma \sum_{s,a,s'} \pi(a|s)p(s'|s,a)|d_\rho^{\pi,p'}(s) - d_\rho^{\pi,p}(s)| \\
&\leq \gamma \max_{s,a} \|p'(\cdot|s,a) - p(\cdot|s,a)\|_1 + \gamma \|d_\rho^{\pi,p'} - d_\rho^{\pi,p}\|_1,
\end{aligned}$$

where (i) uses eq. (38). Then eq. (37) can be proved by rearranging the above inequality. □

Lemma 6. *The function $J_{\rho,\tau}(\pi, p)$ defined by eq. (3) has the following Lipschitz properties for any $\pi, \pi' \in \Pi$, $p, p' \in \mathcal{P}$.*

$$|J_{\rho,\tau}(\pi', p) - J_{\rho,\tau}(\pi, p)| \leq L_\pi \max_s \|\ln \pi'(\cdot|s) - \ln \pi(\cdot|s)\| \tag{39}$$

$$|J_{\rho,\tau}(\pi, p') - J_{\rho,\tau}(\pi, p)| \leq L_p \|p' - p\| \tag{40}$$

$$\|\nabla_p J_{\rho,\tau}(\pi', p) - \nabla_p J_{\rho,\tau}(\pi, p)\| \leq \ell_\pi \max_s \|\ln \pi'(\cdot|s) - \ln \pi(\cdot|s)\| \tag{41}$$

$$\|\nabla_p J_{\rho,\tau}(\pi, p') - \nabla_p J_{\rho,\tau}(\pi, p)\| \leq \ell_p \|p' - p\|, \tag{42}$$

where $L_\pi := \frac{\sqrt{|\mathcal{A}|}(2-\gamma+\gamma\tau\ln|\mathcal{A}|)}{(1-\gamma)^2}$, $L_p := \frac{\sqrt{|\mathcal{S}|}(1+\tau\ln|\mathcal{A}|)}{(1-\gamma)^2}$, $\ell_\pi := \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}(2+3\gamma\tau\ln|\mathcal{A}|)}{(1-\gamma)^3}$ *and* $\ell_p := \frac{2\gamma|\mathcal{S}|(1+\tau\ln|\mathcal{A}|)}{(1-\gamma)^3}$.

Proof. First, we prove eq. (39).

$$\begin{aligned}
&|J_{\rho,\tau}(\pi', p) - J_{\rho,\tau}(\pi, p)| \\
&= \frac{1}{1-\gamma} \Big| \sum_{s,a,s'} \big(d_\rho^{\pi',p}(s)\pi'(a|s)p(s'|s,a)[c(s,a,s') + \tau \ln \pi'(a|s)] - d_\rho^{\pi,p}(s)\pi(a|s)p(s'|s,a)[c(s,a,s') + \tau \ln \pi(a|s)] \big) \Big|
\end{aligned}$$

$$
\begin{aligned}
&\le \frac{1}{1-\gamma}\sum_{s,a,s'}|d_\rho^{\pi',p}(s)-d_\rho^{\pi,p}(s)|\pi'(a|s)p(s'|s,a)[|c(s,a,s')|+\tau|\ln\pi'(a|s)|]\\
&\quad+\frac{1}{1-\gamma}\sum_{s,a,s'}d_\rho^{\pi,p}(s)p(s'|s,a)\big(|\pi'(a|s)-\pi(a|s)||c(s,a,s')|+\tau|\pi'(a|s)\ln\pi'(a|s)-\pi(a|s)\ln\pi(a|s)|\big)\\
&\overset{(i)}{\le}\frac{1}{1-\gamma}\sum_{s}|d_\rho^{\pi',p}(s)-d_\rho^{\pi,p}(s)|-\frac{\tau}{1-\gamma}\sum_{s}|d_\rho^{\pi',p}(s)-d_\rho^{\pi,p}(s)|\sum_a\pi'(a|s)\ln\pi'(a|s)\\
&\quad+\frac{1+\tau}{1-\gamma}\sum_{s,a}d_\rho^{\pi,p}(s)|\ln\pi'(a|s)-\ln\pi(a|s)|\\
&\overset{(ii)}{\le}\frac{\gamma(1+\tau\ln|\mathcal{A}|)}{(1-\gamma)^2}\max_s\|\pi'(\cdot|s)-\pi(\cdot|s)\|_1+\frac{2}{1-\gamma}\sum_s d_\rho^{\pi,p}(s)\|\ln\pi'(\cdot|s)-\ln\pi(\cdot|s)\|_1\\
&\overset{(iii)}{\le}\frac{2-\gamma+\gamma\tau\ln|\mathcal{A}|}{(1-\gamma)^2}\max_s\|\ln\pi'(\cdot|s)-\ln\pi(\cdot|s)\|_1\\
&\le\frac{\sqrt{|\mathcal{A}|}(2-\gamma+\gamma\tau\ln|\mathcal{A}|)}{(1-\gamma)^2}\max_s\|\ln\pi'(\cdot|s)-\ln\pi(\cdot|s)\|,
\end{aligned}
$$

where (i) uses $c(s,a,s')\in[0,1]$ and Lemma 16, (ii) uses eq. (36), $\tau\in[0,1]$ and $-\sum_a\pi'(a|s)\ln\pi'(a|s)\in[0,\ln|\mathcal{A}|]$, and (iii) uses eq. (67).

Then eq. (40) can be proved as follows.

$$
\begin{aligned}
&|J_{\rho,\tau}(\pi,p')-J_{\rho,\tau}(\pi,p)|\\
&\overset{(i)}{=}\Big|\int_0^1(p'-p)^\top\nabla_pJ_{\rho,\tau}(\pi,p_u)du\Big|\\
&\overset{(ii)}{\le}\frac{1}{1-\gamma}\Big|\int_0^1\sum_{s,a,s'}[p'(s'|s,a)-p(s'|s,a)]d_\rho^{\pi,p_u}(s)\pi(a|s)\big[c(s,a,s')+\tau\ln\pi(a|s)+\gamma V_\tau(\pi,p_u;s')\big]du\Big|\\
&\overset{(iii)}{\le}\frac{1}{1-\gamma}\int_0^1\sum_{s,a,s'}|p'(s'|s,a)-p(s'|s,a)|d_\rho^{\pi,p_u}(s)\pi(a|s)\Big[\frac{\max(1,\gamma\tau\ln|\mathcal{A}|)}{1-\gamma}-\tau\ln\pi(a|s)\Big]du\\
&\le\frac{1}{1-\gamma}\max_{s,a}\|p'(\cdot|s,a)-p(\cdot|s,a)\|_1\int_0^1\sum_{s,a}d_\rho^{\pi,p_u}(s)\pi(a|s)\Big[\frac{\max(1,\gamma\tau\ln|\mathcal{A}|)}{1-\gamma}-\tau\ln\pi(a|s)\Big]du\\
&\overset{(iv)}{\le}\frac{1}{1-\gamma}\max_{s,a}\|p'(\cdot|s,a)-p(\cdot|s,a)\|_1\Big[\frac{1+\gamma\tau\ln|\mathcal{A}|}{1-\gamma}+\tau\ln|\mathcal{A}|\Big]\\
&\le\frac{\sqrt{|\mathcal{S}|}(1+\tau\ln|\mathcal{A}|)}{(1-\gamma)^2}\|p'-p\|,
\end{aligned}
$$

where (i) denotes $p_u:=up'+(1-u)p$ for $u\in[0,1]$, (ii) uses eq. (9), (iii) uses $c(s,a,s')\in[0,1]$ and eq. (31) which imply that $|c(s,a,s')+\tau\ln\pi(a|s)+\gamma V_\tau(\pi,p_u;s')|\le\frac{\max(1,\gamma\tau\ln|\mathcal{A}|)}{1-\gamma}-\tau\ln\pi(a|s)$, and (iv) uses $-\sum_a\pi'(a|s)\ln\pi'(a|s)\in[0,\ln|\mathcal{A}|]$.

Then eq. (41) can be proved as follows.

$$
\begin{aligned}
&\|\nabla_pJ_{\rho,\tau}(\pi',p)-\nabla_pJ_{\rho,\tau}(\pi,p)\|\\
&\le\sqrt{|\mathcal{S}|}\sum_{s,a}\max_{s'}\Big|\nabla_pJ_{\rho,\tau}(\pi',p)(s,a,s')-\nabla_pJ_{\rho,\tau}(\pi,p)(s,a,s')\Big|\\
&\overset{(i)}{=}\frac{\sqrt{|\mathcal{S}|}}{1-\gamma}\sum_{s,a}\max_{s'}\Big|d_\rho^{\pi',p}(s)\pi'(a|s)[c(s,a,s')+\tau\ln\pi'(a|s)+\gamma V_\tau(\pi',p;s')]\\
&\quad-d_\rho^{\pi,p}(s)\pi(a|s)[c(s,a,s')+\tau\ln\pi(a|s)+\gamma V_\tau(\pi,p;s')]\Big|
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\sqrt{|\mathcal{S}|}}{1-\gamma}\sum_{s,a}\max_{s'}\Big|[d_\rho^{\pi',p}(s)-d_\rho^{\pi,p}(s)]\pi'(a|s)[c(s,a,s')+\tau\ln\pi'(a|s)+\gamma V_\tau(\pi',p;s')] \\
&\quad + d_\rho^{\pi,p}(s)[\pi'(a|s)-\pi(a|s)][c(s,a,s')+\gamma V_\tau(\pi',p;s')]+\gamma d_\rho^{\pi,p}(s)\pi(a|s)[V_\tau(\pi',p;s')-V_\tau(\pi,p;s')] \\
&\quad + \tau d_\rho^{\pi,p}(s)[\pi'(a|s)\ln\pi'(a|s)-\pi(a|s)\ln\pi(a|s)]\Big| \\
&\overset{(ii)}{\le} \frac{\sqrt{|\mathcal{S}|}}{1-\gamma}\sum_{s,a}\Big[|d_\rho^{\pi',p}(s)-d_\rho^{\pi,p}(s)|\pi'(a|s)\Big(\frac{\max(1,\gamma\tau\ln|\mathcal{A}|)}{1-\gamma}-\tau\ln\pi'(a|s)\Big) \\
&\quad + \frac{\max(1,\gamma\tau\ln|\mathcal{A}|)}{1-\gamma}d_\rho^{\pi,p}(s)|\ln\pi'(a|s)-\ln\pi(a|s)|+\gamma L_\pi d_\rho^{\pi,p}(s)\pi(a|s)\max_s\|\ln\pi'(\cdot|s)-\ln\pi(\cdot|s)\| \\
&\quad + \tau d_\rho^{\pi,p}(s)|\ln\pi'(a|s)-\ln\pi(a|s)|\Big] \\
&\overset{(iii)}{\le} \frac{\gamma\sqrt{|\mathcal{S}|}}{(1-\gamma)^2}\max_s\|\pi'(\cdot|s)-\pi(\cdot|s)\|_1\Big(\frac{\max(1,\gamma\tau\ln|\mathcal{A}|)}{1-\gamma}+\tau\ln|\mathcal{A}|\Big) \\
&\quad + \frac{\sqrt{|\mathcal{S}|}}{1-\gamma}\Big[\frac{\max(1,\gamma\tau\ln|\mathcal{A}|)}{1-\gamma}\max_s\|\ln\pi'(\cdot|s)-\ln\pi(\cdot|s)\|_1+\frac{\sqrt{|\mathcal{A}|}(2\gamma-\gamma^2+\gamma^2\tau\ln|\mathcal{A}|)}{(1-\gamma)^2}\max_s\|\ln\pi'(\cdot|s)-\ln\pi(\cdot|s)\| \\
&\quad + \tau\max_s\|\ln\pi'(\cdot|s)-\ln\pi(\cdot|s)\|_1\Big] \\
&\overset{(iv)}{\le} \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}(2+3\gamma\tau\ln|\mathcal{A}|)}{(1-\gamma)^3}\max_s\|\ln\pi'(\cdot|s)-\ln\pi(\cdot|s)\|, \qquad (43)
\end{aligned}
$$

where (i) uses eq. (9), (ii) uses $c(s,a,s')\in[0,1]$, eq. (31), Lemma 16 and $|V_\tau(\pi',p;s')-V_\tau(\pi,p;s')|\le L_\pi\max_s\|\ln\pi'(\cdot|s)-\ln\pi(\cdot|s)\|$ (eq. (39) when $\rho(s)=I\{s=s'\}$), (iii) uses eq. (36), $L_\pi:=\frac{\sqrt{|\mathcal{A}|}(2-\gamma+\gamma\tau\ln|\mathcal{A}|)}{(1-\gamma)^2}$ and $-\sum_a\pi'(a|s)\ln\pi'(a|s)\in[0,\ln|\mathcal{A}|]$, and (iv) uses $\gamma,\tau\in[0,1]$, $\max_s\|\pi'(\cdot|s)-\pi(\cdot|s)\|_1\le\max_s\|\ln\pi'(\cdot|s)-\ln\pi(\cdot|s)\|_1\le\sqrt{|\mathcal{A}|}\max_s\|\ln\pi'(\cdot|s)-\ln\pi(\cdot|s)\|$ (the first $\le$ comes from eq. (67)).

Then, eq. (42) can be proved as follows.

$$
\begin{aligned}
&\|\nabla_p J_{\rho,\tau}(\pi,p')-\nabla_p J_{\rho,\tau}(\pi,p)\| \\
&\le \sqrt{|\mathcal{S}|}\sum_{s,a}\max_{s'}\Big|\nabla_p J_{\rho,\tau}(\pi,p')(s,a,s')-\nabla_p J_{\rho,\tau}(\pi,p)(s,a,s')\Big| \\
&\overset{(i)}{=} \frac{\sqrt{|\mathcal{S}|}}{1-\gamma}\sum_{s,a}\max_{s'}\Big|d_\rho^{\pi,p'}(s)\pi(a|s)[c(s,a,s')+\tau\ln\pi(a|s)+\gamma V_\tau(\pi,p';s')] \\
&\quad - d_\rho^{\pi,p}(s)\pi(a|s)[c(s,a,s')+\tau\ln\pi(a|s)+\gamma V_\tau(\pi,p;s')]\Big| \\
&\le \frac{\sqrt{|\mathcal{S}|}}{1-\gamma}\sum_{s,a}\pi(a|s)\max_{s'}\Big|\gamma d_\rho^{\pi,p'}(s)[V_\tau(\pi,p';s')-V_\tau(\pi,p;s')] \\
&\quad + [d_\rho^{\pi,p'}(s)-d_\rho^{\pi,p}(s)][c(s,a,s')+\tau\ln\pi(a|s)+\gamma V_\tau(\pi,p;s')]\Big| \\
&\overset{(ii)}{\le} \frac{\sqrt{|\mathcal{S}|}}{1-\gamma}\sum_{s,a}\pi(a|s)\Big[\gamma L_p d_\rho^{\pi,p'}(s)\|p'-p\|+\Big(\frac{\max(1,\gamma\tau\ln|\mathcal{A}|)}{1-\gamma}-\tau\ln\pi(a|s)\Big)|d_\rho^{\pi,p'}(s)-d_\rho^{\pi,p}(s)|\Big] \\
&\overset{(iii)}{\le} \frac{\gamma L_p\sqrt{|\mathcal{S}|}}{1-\gamma}\|p'-p\|+\frac{\sqrt{|\mathcal{S}|}}{1-\gamma}\Big(\frac{1+\gamma\tau\ln|\mathcal{A}|}{1-\gamma}+\tau\ln|\mathcal{A}|\Big)\frac{\gamma}{1-\gamma}\max_{s,a}\|p'(\cdot|s,a)-p(\cdot|s,a)\|_1 \\
&\overset{(iv)}{\le} \frac{\gamma|\mathcal{S}|(1+\tau\ln|\mathcal{A}|)}{(1-\gamma)^3}\|p'-p\|+\frac{\gamma|\mathcal{S}|}{(1-\gamma)^3}(1+\tau\ln|\mathcal{A}|)\|p'-p\| \\
&\le \frac{2\gamma|\mathcal{S}|(1+\tau\ln|\mathcal{A}|)}{(1-\gamma)^3}\|p'-p\|,
\end{aligned}
$$

where (i) uses eq. (9), (ii) uses $c(s,a,s')\in[0,1]$, eq. (31) and $|V_\tau(\pi,p';s')-V_\tau(\pi,p;s')|\le L_p\|p'-p\|$ (eq. (40) when

$\rho(s) = I\{s = s'\}$), (iii) uses $-\sum_a \pi'(a|s) \ln \pi'(a|s) \in [0, \ln |\mathcal{A}|]$ and eq. (37), and (iv) uses $L_p := \frac{\sqrt{|\mathcal{S}|}(1+\tau \ln |\mathcal{A}|)}{(1-\gamma)^2}$. □

Lemma 7. *$F_{\rho,\tau}(p) := \min_{\pi\in\Pi} J_{\rho,\tau}(\pi, p)$ is differentiable with $\nabla F_{\rho,\tau}(p) := \nabla_2 J_{\rho,\tau}(\pi_p, p)$*[6].

Proof. Note that $p \in \mathcal{P}$ where $\mathcal{P}$ is a subset of the Banach space $(\Delta^{\mathcal{S}})^{\mathcal{S}\times\mathcal{A}}$. Also, we have proved in Lemma 6 that $J_{\rho,\tau}(\pi, p)$ is differentiable and $\nabla_p J_{\rho,\tau}(\pi, p)$ is Lipschitz contnuous. Hence, the conditions of the Danskin's Theorem (Bernhard and Rapaport, 1995) hold, so we can apply the Danskin's Theorem which yields this lemma. □

Lemma 8. *The stochastic gradient $\widehat{\nabla}_p J_{\rho,\tau}(\pi_t, p_t)$ obtained from Algorithm 1 has the following estimation error*

$$\|\widehat{\nabla}_p J_{\rho,\tau}(\pi_t, p_t) - \nabla F_{\rho,\tau}(p_t)\| \le \ell_\pi \sqrt{|\mathcal{A}|} \|\ln \pi_t - \ln \pi_t^*\|_\infty + \epsilon_2 \tag{44}$$

Proof. Based on Lemma 7, $\nabla F_{\rho,\tau}(p_t) = \nabla_p J_\rho(\pi_t^*, p_t)$ where $\pi_t^* := \arg\max_\pi J_{\rho,\tau}(\pi, p_t)$. Hence, eq. (44) can be proved as follows.

$$\begin{aligned}
\|\widehat{\nabla}_p J_{\rho,\tau}(\pi_t, p_t) - \nabla_p J_{\rho,\tau}(\pi_t^*, p_t)\| &\le \|\nabla_p J_{\rho,\tau}(\pi_t, p_t) - \nabla_p J_{\rho,\tau}(\pi_t^*, p_t)\| + \|\widehat{\nabla}_p J_{\rho,\tau}(\pi_t, p_t) - \nabla_p J_{\rho,\tau}(\pi_t, p_t)\| \\
&\overset{(i)}{\le} \ell_\pi \max_s \|\ln \pi_t^*(\cdot|s) - \ln \pi_t(\cdot|s)\| + \epsilon_2 \\
&\le \ell_\pi \sqrt{|\mathcal{A}|} \|\ln \pi_t - \ln \pi_t^*\|_\infty + \epsilon_2,
\end{aligned}$$

where (i) uses eq. (41). □

D. Convergence Results of the Policy Updates

Lemma 9 (Convergence of policy updates for small space)**.** *For the policy optimization problem $\min_{\pi\in\Pi} J_{\rho,\tau}(\pi, p_t)$, perform the NPG step (7) with stepsize $\eta = \frac{1-\gamma}{\tau}$. Suppose the Q function is approximated with error ϵ_1, i.e., $\|\widehat{Q}_{t,k'} - Q_\tau(\pi_{t,k'}, p_t)\|_\infty \le \epsilon_1$ for all $k' = 0, 1, \ldots, k-1$. Then the policy $\pi_{t,k}$ satisfies the following properties.*

$$\|\widetilde{Q}_\tau(\pi_t^*, p_t) - \widetilde{Q}_\tau(\pi_{t,k}, p_t)\|_\infty \le \frac{\gamma^k(1+\gamma\tau \ln |\mathcal{A}|)}{1-\gamma} + \frac{2\gamma\epsilon_1}{(1-\gamma)^2}, \tag{45}$$

$$\|\pi_t^* - \pi_{t,k}\|_\infty \le \|\ln \pi_t^* - \ln \pi_{t,k}\|_\infty \le \frac{2\gamma^{k-1}(1/\tau + \gamma \ln |\mathcal{A}|)}{1-\gamma} + \frac{4\epsilon_1}{\tau(1-\gamma)^2}, \tag{46}$$

$$\ln \pi_{t,k} \ge \ln \pi_{\min} := -\frac{3\ln |\mathcal{A}| + 3/\tau}{1-\gamma} - \frac{4\epsilon_1}{\tau(1-\gamma)^2}. \tag{47}$$

Proof. Note that the NPG step (7) can be rewritten into the following form, as used in (Cen et al., 2022).

$$\pi_{t,k+1}(\cdot|s) \propto \pi_{t,k}(\cdot|s)^{1-\frac{\eta\tau}{1-\gamma}} \exp\Big[-\frac{\eta\big(\widehat{Q}_{t,k}(s,\cdot) - \tau \ln \pi_{t,k}(\cdot|s)\big)}{1-\gamma}\Big],$$

where $\widehat{Q}_{t,k}(s,a) - \tau \ln \pi_{t,k}(a|s) \approx \widetilde{Q}_\tau(\pi_{t,k}, p_t; s, a) = Q_\tau(\pi_{t,k}, p_t; s, a) - \tau \ln \pi_{t,k}(a|s)$ with $\sup_{s,a} \big|[\widehat{Q}_{t,k}(s,a) - \tau \ln \pi_{t,k}(a|s)] - \widetilde{Q}_\tau(\pi_{t,k}, p_t; s, a)\big| \le \epsilon_1$. Therefore, based on Theorem 2 of (Cen et al., 2022), we obtain the convergence rates (45) and (46) with stepsize $\eta = \frac{1-\gamma}{\tau}$ as follows.

$$\begin{aligned}
&\|\widetilde{Q}_\tau(\pi_t^*, p_t) - \widetilde{Q}_\tau(\pi_{t,k}, p_t)\|_\infty \le C_1\gamma^k + \frac{2\gamma\epsilon_1}{(1-\gamma)^2} \overset{(i)}{\le} \frac{\gamma^k(1+\gamma\tau \ln |\mathcal{A}|)}{1-\gamma} + \frac{2\gamma\epsilon_1}{(1-\gamma)^2}, \\
&\|\pi_t^* - \pi_{t,k}\|_\infty \overset{(ii)}{\le} \|\ln \pi_t^* - \ln \pi_{t,k}\|_\infty \le 2C_1\tau^{-1}\gamma^{k-1} + \frac{4\epsilon_1}{\tau(1-\gamma)^2} \overset{(iii)}{\le} \frac{2\gamma^{k-1}(1/\tau + \gamma \ln |\mathcal{A}|)}{1-\gamma} + \frac{4\epsilon_1}{\tau(1-\gamma)^2},
\end{aligned}$$

[6] $\nabla_2 J_{\rho,\tau}(\pi_p, p)$ denotes the gradient with respect to the second argument p. We do not use $\nabla_p J_{\rho,\tau}(\pi_p, p)$ since π_p depends on p.

where (i) and (iii) use $C_1 := \|\widetilde{Q}_\tau(\pi_t^*, p_t) - \widetilde{Q}_\tau(\pi_{t,0}, p_t)\|_\infty \le \frac{1+\gamma\tau\ln|\mathcal{A}|}{1-\gamma}$ ($\le$ is based on eq. (32)), and (ii) uses the following inequality for any $u, v \in (0, 1]$.

$$|\ln u - \ln v| = \ln\max(u,v) - \ln\min(u,v) = \int_{\min(u,v)}^{\max(u,v)} \frac{ds}{s} \ge \max(u,v) - \min(u,v) = |u-v|. \tag{48}$$

Note that $\ln\pi_t^* = \ln\pi_{p_t} \ge -\frac{\ln|\mathcal{A}|+1/\tau}{1-\gamma}$ based on eq. (33). This along with eq. (46) implies that for any $k \ge 1$,

$$\ln\pi_{t,k} \ge -\frac{\ln|\mathcal{A}|+1/\tau}{1-\gamma} - \frac{2\gamma^{k-1}(1/\tau+\gamma\ln|\mathcal{A}|)}{1-\gamma} - \frac{4\epsilon_1}{\tau(1-\gamma)^2} \ge -\frac{3\ln|\mathcal{A}|+3/\tau}{1-\gamma} - \frac{4\epsilon_1}{\tau(1-\gamma)^2},$$

The above bound also holds for $k = 0$ as $\pi_{t,0}(a|s) = 1/|\mathcal{A}|$. This proves eq. (47). □

Lemma 10 (Convergence of policy updates for large space)**.** *For the policy optimization problem* $\min_{\pi\in\Pi} J_{\rho,\tau}(\pi, p_{\xi_t})$*, perform the NPG steps* (22)-(23) *with stepsize* $\eta = \frac{1-\gamma}{\tau}$*. Suppose the Q function is approximated with error* ϵ_1*, i.e.,* $\sup_{s,a}|\phi(s,a)^\top w_{t,k'} - Q_\tau(\pi_{t,k'}, p_t; s, a)| \le \epsilon_1$ *for all* $k' = 0, 1, \ldots, k-1$*. Then the policy* $\pi_{t,k}$ *satisfies the following properties.*

$$\|\widetilde{Q}_\tau(\pi_t^*, p_t) - \widetilde{Q}_\tau(\pi_{t,k}, p_t)\|_\infty \le \frac{\gamma^k(1+\gamma\tau\ln|\mathcal{A}|)}{1-\gamma} + \frac{2\gamma\epsilon_1}{(1-\gamma)^2}, \tag{49}$$

$$\|\pi_t^* - \pi_{t,k}\|_\infty \le \|\ln\pi_t^* - \ln\pi_{t,k}\|_\infty \le \frac{2\gamma^{k-1}(1/\tau+\gamma\ln|\mathcal{A}|)}{1-\gamma} + \frac{4\epsilon_1}{\tau(1-\gamma)^2}, \tag{50}$$

$$\ln\pi_{t,k} \ge \ln\pi_{\min} := -\frac{3\ln|\mathcal{A}|+3/\tau}{1-\gamma} - \frac{4\epsilon_1}{\tau(1-\gamma)^2}. \tag{51}$$

Proof. It suffices to prove that the NPG steps (22)-(23) is equivalent to the NPG step (7) with $\widehat{Q}_{t,k}(s,a) = \phi(s,a)^\top w_{t,k}$, so that we can directly apply Lemma 9.

The NPG steps (22)-(23) imply the NPG step (7) as follows.

$$\begin{aligned}
\pi_{t,k+1} &\propto \exp\left[-\frac{\phi(s,\cdot)^\top u_{t,k+1}}{1-\gamma}\right] \\
&\propto \exp\left[-\frac{\phi(s,\cdot)^\top (u_{t,k}+w_{t,k})}{1-\gamma}\right] \\
&= \exp\left[-\frac{\phi(s,\cdot)^\top u_{t,k}}{1-\gamma}\right]\exp\left[-\frac{\phi(s,\cdot)^\top w_{t,k}}{1-\gamma}\right] \\
&\propto \pi_{t,k}\exp\left[-\frac{\widehat{Q}_{t,k}(s,a)}{1-\gamma}\right].
\end{aligned}$$

Conversely, by iterating the NPG step (7) over $k = 0, 1, \ldots, K-1$, we obtain that

$$\pi_{t,K}(a|s) \propto \exp\left[-\frac{1}{1-\gamma}\sum_{k=0}^{K-1}\widehat{Q}_{t,k}(s,a)\right] = \exp\left[-\frac{1}{1-\gamma}\phi(s,a)^\top\sum_{k=0}^{K-1}u_{t,k}\right].$$

Denote $w_{t,K} := \sum_{k=0}^{K-1} u_{t,k}$ which satisfies eq. (23). Then the above update rule becomes eq. (22). □

E. Stochastic Approximation Errors

In this section, we will analyze the approximation error of estimating Q function and transition gradients in both Algorithm 2 (for small state space) and Algorithm 3 (for large state space).

Lemma 11 (Approximation error of Q_τ for large space)**.** *Fix* $p \in \mathcal{P}$ *and* $\pi \in \Pi$*. Suppose that the regularized cost* $c(s,a,s') + \tau\ln\pi(a|s)$ *is bounded and that* $\sup_{s,a}\|\phi(s,a)\| \le 1$*. For any* $\delta_1 \in (0,1)$ *and* $\epsilon_1 > 2\zeta$ *where* $\zeta := \sup_{\pi\in\Pi, p\in\mathcal{P}, s\in\mathcal{S}, a\in\mathcal{A}, \tau\in[0,1]}|\phi(s,a)^\top w_{\pi,p}^* - Q_\tau(\pi, p; s, a)|^2$ *denotes the linear Q function approximation error, update* $w_n \in \mathbb{R}^d$ *by applying the TD rule* (21) *with* $T_1 \ge \mathcal{O}(\epsilon_1^{-2})$ *iterations and stepsize* $\alpha = \mathcal{O}[\ln^{-1}(\epsilon_1^{-1})]$*. Then* $\overline{w}_{T_1} := \frac{1}{T_1}\sum_{t_1=1}^{T_1} w_n$ *satisfies* $\sup_{s\in\mathcal{S},a\in\mathcal{A}}|\phi(s,a)^\top\overline{w}_{T_1} - Q_\tau(\pi, p; s, a)| \le \epsilon_1$ *with probability at least* $1-\delta_1$*.*

Proof. The optimal parameter $w^*_{\pi,p}$ for estimating $Q_\tau(\pi,p;s,a) \approx \phi(s,a)^\top w$ has the following expression. (Li et al., 2023a)

$$w^*_{\pi,p} := \mathbb{E}_{\pi,p}\big[\phi(s,a)\big(\phi(s,a)-\gamma\phi(s',a')\big)^\top\big]^{-1}\mathbb{E}\big[\phi(s,a)\big(c(s,a,s')+\tau\ln\pi(a|s)\big)\big], \tag{52}$$

where the expectation $\mathbb{E}_{\pi,p}$ is taken over $s\sim\mu_{\pi,p}, a\sim\pi(\cdot|s), s'\sim p(\cdot|s,a), a'\sim\pi(\cdot|s')$ where $\mu_{\pi,p}$ is the stationary distribution under policy π and transition kernel p.

Based on Theorem 1 of (Li et al., 2023a), $\|\overline{w}_{T_1}-w^*_{\pi,p}\|\le\mathcal{O}(T^{-1/2})\le\epsilon_1/2$ for $T=\mathcal{O}(\epsilon_1^{-2})$. Therefore,

$$\begin{aligned}
&\max_{s,a}|\phi(s,a)^\top\overline{w}_{T_1}-Q_\tau(\pi,p;s,a)|\\
&\le\max_{s,a}\big[|\phi(s,a)^\top(\overline{w}_{T_1}-w^*_{\pi,p})|+|\phi(s,a)^\top w^*_{\pi,p}-Q_\tau(\pi,p;s,a)|\big]\\
&\overset{(i)}{\le}\|\overline{w}_{T_1}-w^*_{\pi,p}\|+\zeta\le\frac{\epsilon_1}{2}+\zeta\overset{(ii)}{\le}\epsilon_1.
\end{aligned}$$

where (i) uses $\zeta := \sup_{\pi\in\Pi,p\in\mathcal{P},s\in\mathcal{S},a\in\mathcal{A},\tau\in[0,1]}|\phi(s,a)^\top w^*_{\pi,p}-Q_\tau(\pi,p;s,a)|^2$ and the assumption that $\sup_{s,a}\|\phi(s,a)\|\le 1$ and (ii) uses $\epsilon_1>2\zeta$. □

Lemma 12 (Approximation error of Q_τ for small space)**.** *Fix $\pi\in\Pi$ and $p\in\mathcal{P}$. Suppose that the regularized cost $c(s,a,s')+\tau\ln\pi(a|s)$ is bounded. For any $\delta_1\in(0,1)$ and $\epsilon_1>0$, update q_n by applying the TD rule (15) with $T_1\ge\mathcal{O}(\epsilon_1^{-2})$ iterations and stepsize $\alpha=\mathcal{O}[\ln^{-1}(\epsilon_1^{-1})]$. Then $\overline{q}_{T_1}:=\frac{1}{T_1}\sum_{t_1=1}^{T_1}q_n$ satisfies $\sup_{s\in\mathcal{S},a\in\mathcal{A}}|\overline{q}_{T_1}(s,a)-Q_\tau(\pi,p;s,a)|\le\epsilon_1$ with probability at least $1-\delta_1$.*

Proof. In Lemma 11, let $\phi(s,a)\in\{0,1\}^d$ ($d=|\mathcal{S}||\mathcal{A}|$) be a one-hot vector with the (s,a)-th entry being 1. Then this Lemma becomes a special case of Lemma 11 in the following aspects:

(1) The TD rule (21) becomes the TD rule (15) with $q_n=w_n$.

(2) $\overline{q}_{T_1}=\overline{w}_{T_1}$.

(3) $Q_\tau(\pi,p)=w^*_{\pi,p}$ and thus ζ becomes 0.

(4) The condition of Lemma 11 that $\sup_{s,a}\|\phi(s,a)\|\le 1$ is satisfied. □

Lemma 13 (Approximation error of $\nabla_\xi J_{\rho,\tau}(\pi,p_\xi)$ for large space)**.** *Fix $\pi\in\Pi$ and $p\in\mathcal{P}$. Use linear parameterization $p_\xi=\Psi\xi$ and assume it satisfies $\inf_{s,a,s'}p_\xi(s'|s,a)>p_{\min}$ for a constant $p_{\min}>0$. Suppose that the regularized cost $c(s,a,s')+\tau\ln\pi(a|s)$ is bounded, and that the Q function estimation $Q_\tau(\pi,p_\xi;s,a)\approx\phi(s,a)^\top\overline{w}_{T_1}$ satisfies $\sup_{s\in\mathcal{S},a\in\mathcal{A}}|\phi(s,a)^\top\overline{w}_{T_1}-Q_\tau(\pi,p_\xi;s,a)|\le\epsilon_1$ for $\epsilon_1>2\zeta$. Then for any $\delta_2\in(0,1)$ and $\epsilon_2\ge\frac{3\gamma\|\Psi\|\epsilon_1}{p_{\min}}$, the stochastic transition gradient (19) with $N\ge\mathcal{O}(\epsilon_2^{-2})$ and $H\ge\mathcal{O}[\ln(\epsilon_2^{-1})]$ has approximation error $\|\widehat{\nabla}_\xi J_{\rho,\tau}(\pi,p_\xi)-\nabla_\xi J_{\rho,\tau}(\pi,p_\xi)\|\le\epsilon_2$ with probability at least $1-\delta_2$, which requires $NH=\mathcal{O}[\epsilon_2^{-2}\ln(\epsilon_2^{-1})]$ samples.*

Proof. The stochastic gradient (19) can be rewritten as $\widehat{\nabla}_\xi J_{\rho,\tau}(\pi,p_\xi)=\frac{1}{N}\sum_{i=1}^N g(s_{i,H_i},a_{i,H_i},s_{i,H_i+1},a_{i,H_i+1})$ with

$$g(s,a,s',a'):=\frac{\psi(s,a,s')}{(1-\gamma)p_\xi(s'|s,a)}[c(s,a,s')+\tau\ln\pi(a|s)+\gamma\phi(s',a')^\top\overline{w}_{T_1}] \tag{53}$$

Since $\mathbb{P}(H_i=h)\propto\gamma^h(h=0,1,\dots,H-1)$, $s_{i,H_i}\sim d^{\pi,p}_{\rho,H}(s):=\frac{1-\gamma}{1-\gamma^H}\sum_{h=0}^{H-1}\gamma^h\mathbb{P}_{\pi,p}(s_h=s|s_0\sim\rho)$. Therefore,

$$\mathbb{E}g(s_{i,H_i},a_{i,H_i},s_{i,H_i+1},a_{i,H_i+1})=\mathbb{E}_{d^{\pi,p_\xi}_{\rho,H}}g(s,a,s',a'), \tag{54}$$

where the expectation $\mathbb{E}_{d^{\pi,p_\xi}_{\rho,H}}$ is taken over $s\sim d^{\pi,p_\xi}_{\rho,H}, a\sim\pi(\cdot|s), s'\sim p_\xi(\cdot|s,a), a'\sim\pi(\cdot|s')$.

Note that

$$\|g(s,a,s',a')\|$$

$$\begin{aligned}
&\overset{(i)}{\leq} \frac{\|\psi(s,a,s')\|}{(1-\gamma)p_{\min}}\big[|c(s,a,s')+\tau\ln\pi(a|s)|+\gamma|\phi(s',a')^\top\overline{w}_{T_1}-Q_\tau(\pi,p;s,a)|+\gamma|\widetilde{Q}_\tau(\pi,p;s,a)|+\gamma|\tau\ln\pi(a'|s')|\big]\\
&\overset{(ii)}{\leq} \frac{\|\Psi\|}{(1-\gamma)p_{\min}}\Big[\mathcal{O}(1)+\gamma\Big(\epsilon_1+\frac{1+\gamma\tau\ln|\mathcal{A}|}{1-\gamma}\Big)\Big]\\
&\leq \mathcal{O}(1),
\end{aligned} \tag{55}$$

where (i) uses $\widetilde{Q}_\tau(\pi,p;s,a)=Q_\tau(\pi,p;s,a)-\tau\ln\pi(a|s)$ (based on eqs. (5) and (30)) and the assumption that $p_\xi(s'|s,a)\geq p_{\min}$, and (ii) uses eq. (32), $\sup_{s\in\mathcal{S},a\in\mathcal{A}}|\phi(s,a)^\top\overline{w}_{T_1}-Q_\tau(\pi,p_\xi;s,a)|\leq\epsilon_1$, $|c(s,a,s')+\tau\ln\pi(a|s)|\leq\mathcal{O}(1)$ and $|\tau\ln\pi(a'|s')|\leq\mathcal{O}(1)$ (since $|c(s,a,s')|\leq\mathcal{O}(1)$).

Therefore, applying Hoeffding's ineqaulity to the i.i.d. variables $\{g(s_{i,H_i},a_{i,H_i},s_{i,H_i+1},a_{i,H_i+1})\}_{i=1}^N$ with bound (55), the following bound holds with probability at least $1-\delta_2$.

$$\|\widehat{\nabla}_\xi J_{\rho,\tau}(\pi,p_\xi)-\mathbb{E}_{d_{\rho,H}^{\pi,p_\xi}}g_1\|\leq\mathcal{O}\Big[\frac{1}{\sqrt{N}}\ln\Big(\frac{2}{\delta_2}\Big)\Big]\overset{(i)}{\leq}\frac{\epsilon_2}{3}, \tag{56}$$

where (i) holds for $N=\mathcal{O}(\epsilon_2^{-2})$. Note that the transition gradient (18) can be rewritten as follows.

$$\nabla_\xi J_{\rho,\tau}(\pi,p_\xi)=\frac{1}{1-\gamma}\mathbb{E}_{d_\rho^{\pi,p_\xi}}\Big[\frac{\psi(s,a,s')}{p_\xi(s'|s,a)}[c(s,a,s')+\tau\ln\pi(a|s)+\gamma Q_\tau(\pi,p_\xi;s',a')]\Big], \tag{57}$$

where the expectation $\mathbb{E}_{d_\rho^{\pi,p_\xi}}$ is taken over $s\sim d_\rho^{\pi,p_\xi},a\sim\pi(\cdot|s),s'\sim p_\xi(\cdot|s,a),a'\sim\pi(\cdot|s')$ and we used $\mathbb{E}_{a'\sim\pi(\cdot|s')}[Q_\tau(\pi,p_\xi;s',a')|s']=V_\tau(\pi,p_\xi;s')$ based on eqs. (4) and (5).

$$\begin{aligned}
&\|\widehat{\nabla}_\xi J_{\rho,\tau}(\pi,p_\xi)-\nabla_\xi J_{\rho,\tau}(\pi,p_\xi)\|\\
&\leq\|\widehat{\nabla}_\xi J_{\rho,\tau}(\pi,p_\xi)-\mathbb{E}_{d_{\rho,H}^{\pi,p_\xi}}g(s,a,s',s'')\|+\|\mathbb{E}_{d_{\rho,H}^{\pi,p_\xi}}g(s,a,s',s'')-\mathbb{E}_{d_\rho^{\pi,p_\xi}}g(s,a,s',s'')\|\\
&\quad+\|\mathbb{E}_{d_\rho^{\pi,p_\xi}}g(s,a,s',s'')-\nabla_\xi J_{\rho,\tau}(\pi,p_\xi)\|\\
&\overset{(i)}{\leq}\frac{\epsilon_2}{3}+\mathcal{O}(1)\sum_s|d_{\rho,H}^{\pi,p}(s)-d_\rho^{\pi,p}(s)|+\frac{\gamma}{1-\gamma}\Big\|\mathbb{E}_{d_\rho^{\pi,p_\xi}}\Big[\frac{\psi(s,a,s')}{p_\xi(s'|s,a)}\big(\phi(s',a')^\top\overline{w}_{T_1}-Q_\tau(\pi,p_\xi;s',a')\big)\Big]\Big\|\\
&\overset{(ii)}{\leq}\frac{\epsilon_2}{3}+\mathcal{O}(1)(1-\gamma)\sum_s\Big|\sum_{t=0}^{H-1}\gamma^t\Big(\frac{1}{1-\gamma^H}-1\Big)\mathbb{P}_{\pi,p}(s_t=s|s_0\sim\rho)-\sum_{t=H}^{+\infty}\gamma^t\mathbb{P}_{\pi,p}(s_t=s|s_0\sim\rho)\Big|+\frac{\gamma\|\Psi\|\epsilon_1}{p_{\min}(1-\gamma)}\\
&\leq\frac{\epsilon_2}{3}+\mathcal{O}(1)(1-\gamma)\sum_s\Big[\sum_{t=0}^{H-1}\frac{\gamma^{H+t}}{1-\gamma^H}\mathbb{P}_{\pi,p}(s_t=s|s_0\sim\rho)+\sum_{t=H}^{+\infty}\gamma^t\mathbb{P}_{\pi,p}(s_t=s|s_0\sim\rho)\Big]+\frac{\gamma\|\Psi\|\epsilon_1}{p_{\min}(1-\gamma)}\\
&=\frac{\epsilon_2}{3}+\mathcal{O}(1)(1-\gamma)\Big[\sum_{t=0}^{H-1}\frac{\gamma^{H+t}}{1-\gamma^H}+\sum_{t=H}^{+\infty}\gamma^t\Big]+\frac{\gamma\|\Psi\|\epsilon_1}{p_{\min}(1-\gamma)}\\
&\leq\frac{\epsilon_2}{3}+\mathcal{O}(\gamma^H)+\frac{\gamma\|\Psi\|\epsilon_1}{p_{\min}(1-\gamma)}\overset{(iii)}{\leq}\epsilon_2,
\end{aligned}$$

where (i) uses eqs. (53), (55), (56) and (57), (ii) uses $d_\rho^{\pi,p}(s):=(1-\gamma)\sum_{t=0}^\infty\gamma^t\mathbb{P}_{\pi,p}(s_t=s|s_0\sim\rho)$ defined in eq. (10), $d_{\rho,H}^{\pi,p}(s):=\frac{1-\gamma}{1-\gamma^H}\sum_{t=0}^{H-1}\gamma^t\mathbb{P}_{\pi,p}(s_t=s|s_0\sim\rho)$, $\inf_{s,a,s'}p_\xi(s'|s,a)>p_{\min}$ and $\sup_{s\in\mathcal{S},a\in\mathcal{A}}|\phi(s,a)^\top\overline{w}_{T_1}-Q_\tau(\pi,p_\xi;s,a)|\leq\epsilon_1$, and (iii) uses $H=\mathcal{O}[\ln(\epsilon_2^{-1})]$ and $\epsilon_1\leq\frac{p_{\min}\epsilon_2(1-\gamma)}{3\gamma\|\Psi\|}$. □

Lemma 14 (Approximation error of $\nabla_p J_{\rho,\tau}(\pi,p)$ for small space)**.** *Fix $\pi\in\Pi$ and $p\in\mathcal{P}$. Suppose that the Q function estimation $\overline{q}_{T_1}\approx Q_\tau(\pi,p)$ satisfies $\|\overline{q}_{T_1}-Q_\tau(\pi,p)\|_\infty\leq\epsilon_1$. Then for any $\delta_2\in(0,1)$ and $\epsilon_2\geq\frac{3\gamma\epsilon_1\sqrt{|\mathcal{S}|}}{1-\gamma}$, the stochastic transition gradient (16) with $N\geq\mathcal{O}(\epsilon_2^{-2})$ and $H\geq\mathcal{O}[\ln(\epsilon_2^{-1})]$ has approximation error $\|\widehat{\nabla}_p J_{\rho,\tau}(\pi,p)-\nabla_p J_{\rho,\tau}(\pi,p)\|\leq\epsilon_2$ with probability at least $1-\delta_2$, which requires $NH=\mathcal{O}(\epsilon_2^{-2}\ln\epsilon_2^{-1})$ samples.*

Proof. The proof logic is the same as that of Lemma 13. We rewrite the stochastic gradient (16) as $\widehat{\nabla}_p J_{\rho,\tau}(\pi,p)(s,a,s') = \frac{1}{N}\sum_{i=1}^N g(s_{i,H_i};s,a,s')$ where

$$g(\widetilde{s};s,a,s') := \frac{\pi(a|s)\mathbb{1}\{\widetilde{s}=s\}}{1-\gamma}\Big[c(s,a,s') + \tau\ln\pi(a|s) + \gamma\sum_{a'}\pi(a'|s')\overline{q}_{T_1}(s',a')\Big]. \tag{58}$$

This function g has the following bound.

$$\begin{aligned}
&\|g(\widetilde{s};\cdot,\cdot,\cdot)\| \le \sum_{s,a}\sqrt{\sum_{s'}|g(\widetilde{s};s,a,s')|^2}\\
&=\sum_{s,a}\frac{\pi(a|s)\mathbb{1}\{\widetilde{s}=s\}}{1-\gamma}\sqrt{\sum_{s'}\Big|c(s,a,s')+\tau\ln\pi(a|s)+\gamma\sum_{a'}\pi(a'|s')[\overline{q}_{T_1}(s',a')-Q_\tau(\pi,p;s',a')+Q_\tau(\pi,p;s',a')]\Big|^2}\\
&\overset{(i)}{\le}\sum_{s,a}\frac{\pi(a|s)\mathbb{1}\{\widetilde{s}=s\}}{1-\gamma}\sqrt{\sum_{s'}\big[1-\tau\ln\pi(a|s)+\gamma\epsilon_1+\gamma|V_\tau(\pi,p;s')|\big]^2}\\
&\overset{(ii)}{\le}\sqrt{|\mathcal{S}|}\sum_{s,a}\frac{\pi(a|s)\mathbb{1}\{\widetilde{s}=s\}}{1-\gamma}\Big[1-\tau\ln\pi(a|s)+\gamma\epsilon_1+\frac{\gamma+\gamma\tau\ln|\mathcal{A}|}{1-\gamma}\Big]\\
&\overset{(iii)}{\le}\frac{\sqrt{|\mathcal{S}|}}{1-\gamma}\Big[1+\tau\ln|\mathcal{A}|+\gamma\epsilon_1+\frac{\gamma+\gamma\tau\ln|\mathcal{A}|}{1-\gamma}\Big]=\mathcal{O}(1),
\end{aligned}\tag{59}$$

where (i) uses $|c(s,a,s')| \le 1$, $|\tau\ln\pi(a|s)| = -\tau\ln\pi(a|s)$, $\|\overline{q}_{T_1} - Q_\tau(\pi,p)\|_\infty \le \epsilon_1$ and $V_\tau(\pi,p;s') = \sum_{a'}\pi(a'|s')Q_\tau(\pi,p;s',a')$ (based on eqs. (4) and (5)), (ii) uses eq. (31), (iii) uses $-\sum_a \pi(a|s)\ln\pi(a|s) \in [0,\ln|\mathcal{A}|]$. Hence, applying Hoeffding's inequality to $\widehat{\nabla}_p J_{\rho,\tau}(\pi,p)(s,a,s') = \frac{1}{N}\sum_{i=1}^N g(s_{i,H_i};s,a,s')$, the following bound holds with probability at least $1-\delta_2$.

$$\big\|\widehat{\nabla}_p J_{\rho,\tau}(\pi,p) - \mathbb{E}_{\widetilde{s}\sim d^{\pi,p}_{\rho,H}} g(\widetilde{s};\cdot,\cdot,\cdot)\big\| \le \mathcal{O}\Big[\frac{1}{\sqrt{N}}\ln\Big(\frac{2}{\delta_2}\Big)\Big] \overset{(i)}{\le} \frac{\epsilon_2}{3}, \tag{60}$$

where $d^{\pi,p}_{\rho,H}(s) := \frac{1-\gamma}{1-\gamma^H}\sum_{t=0}^{H-1}\gamma^t\mathbb{P}_{\pi,p}(s_t=s|s_0\sim\rho)$ is the distribution of s_{i,H_i}, and (i) holds for $N=\mathcal{O}(\epsilon_2^{-2})$. Moreover,

$$\big\|\mathbb{E}_{\widetilde{s}\sim d^{\pi,p}_{\rho,H}} g(\widetilde{s};\cdot,\cdot,\cdot) - \mathbb{E}_{\widetilde{s}\sim d^{\pi,p}_{\rho}} g(\widetilde{s};\cdot,\cdot,\cdot)\big\| \overset{(i)}{\le} \mathcal{O}(1)\sum_s |d^{\pi,p}_{\rho,H}(s) - d^{\pi,p}_{\rho}(s)| \overset{(ii)}{\le} \mathcal{O}(\gamma^H) \overset{(iii)}{\le} \frac{\epsilon_2}{3} \tag{61}$$

where (i) uses eq. (59), (ii) follows the proof of Lemma 13, and (iii) uses $H=\mathcal{O}[\ln(\epsilon_2^{-1})]$.

Note that the transition gradient (9) can be rewritten as

$$\nabla_p J_{\rho,\tau}(\pi,p)(s,a,s') = \frac{d^{\pi,p}_\rho(s)\pi(a|s)}{1-\gamma}\Big[c(s,a,s') + \tau\ln\pi(a|s) + \gamma\sum_{a'}\pi(a'|s')Q_\tau(\pi,p;s',a')\Big] \tag{62}$$

Therefore,

$$\begin{aligned}
\|\mathbb{E}_{\widetilde{s}\sim d^{\pi,p}_\rho} g(\widetilde{s};\cdot,\cdot,\cdot) - \nabla_p J_{\rho,\tau}(\pi,p)\| &\le \sum_{s,a}\sqrt{\sum_{s'}\big|\mathbb{E}_{\widetilde{s}\sim d^{\pi,p}_\rho} g(\widetilde{s};s,a,s') - \nabla_p J_{\rho,\tau}(\pi,p)(s,a,s')\big|^2}\\
&\overset{(i)}{\le}\sum_{s,a}\sqrt{\sum_{s'}\Big|\frac{\gamma d^{\pi,p}_\rho(s)\pi(a|s)}{1-\gamma}\sum_{a'}\pi(a'|s')[\overline{q}_{T_1}(s',a') - Q_\tau(\pi,p;s',a')]\Big|^2}\\
&\overset{(ii)}{\le}\sum_{s,a}\sqrt{\sum_{s'}\Big|\frac{\gamma\epsilon_1 d^{\pi,p}_\rho(s)\pi(a|s)}{1-\gamma}\Big|^2}\\
&=\sqrt{|\mathcal{S}|}\sum_{s,a}\frac{\gamma\epsilon_1 d^{\pi,p}_\rho(s)\pi(a|s)}{1-\gamma}
\end{aligned}$$

$$= \frac{\gamma\epsilon_1\sqrt{|\mathcal{S}|}}{1-\gamma} \overset{(iii)}{\leq} \frac{\epsilon_2}{3}, \tag{63}$$

where (i) uses eqs. (58) and (62), (ii) uses $\|\overline{q}_{T_1} - Q_\tau(\pi,p)\|_\infty \leq \epsilon_1$, and (iii) uses $\epsilon_2 \geq \frac{3\gamma\epsilon_1\sqrt{|\mathcal{S}|}}{1-\gamma}$.

As a result, we conclude that $\|\widehat{\nabla}_p J_{\rho,\tau}(\pi,p) - \nabla_p J_{\rho,\tau}(\pi,p)\| \leq \epsilon_2$ by applying triangular inequality to eqs. (60), (61) and (63). □

F. Supporting Lemmas

Lemma 15. *Suppose $\mathcal{P}$ is a convex set. For any $p' \in \mathcal{P}$, the variable $p_{t+1} = proj_{\mathcal{P}}\big(p_t + \beta\widehat{\nabla}_p J_\rho(\pi_t,p_t)\big)$ generated from Algorithm 1 satisfies*

$$\langle p' - p_{t+1}, p_t + \beta\widehat{\nabla}_p J_\rho(\pi_t,p_t) - p_{t+1}\rangle \leq 0. \tag{64}$$

Similarly, if Ξ is a convex set, then for any $\xi \in \Xi$, the variable $\xi_{t+1} = proj_{\Xi}\big(\xi_t + \beta\widehat{\nabla}_\xi J_\rho(\pi_t,p_{\xi_t})\big)$ generated from Algorithm 3 satisfies

$$\langle \xi' - \xi_{t+1}, \xi_t + \beta\widehat{\nabla}_\xi J_\rho(\pi_t,p_{\xi_t}) - \xi_{t+1}\rangle \leq 0. \tag{65}$$

Proof. We will only prove eq. (64) since eq. (65) can be proved in a similar way.

Define the function $f(u) := \|p_t + \beta\widehat{\nabla}_p J_\rho(\pi_t,p_t) - [up' + (1-u)p_{t+1}]\|^2$.

Note that $p', p_{t+1} \in \mathcal{P}$. Hence, for any $u \in [0,1]$, $up' + (1-u)p_{t+1} \in \mathcal{P}$. Since $p_{t+1} = \text{proj}_{\mathcal{P}}\big(p_t + \beta\widehat{\nabla}_p J_\rho(\pi_t,p_t)\big)$, we have

$$\begin{aligned} f(u) &= \|p_t + \beta\widehat{\nabla}_p J_\rho(\pi_t,p_t) - [up' + (1-u)p_{t+1}]\|^2 \\ &\geq \|p_t + \beta\widehat{\nabla}_p J_\rho(\pi_t,p_t) - p_{t+1}]\|^2 = f(0). \end{aligned}$$

Therefore,

$$f'(0) = -2\langle p' - p_{t+1}, p_t + \beta\widehat{\nabla}_p J_\rho(\pi_t,p_t) - p_{t+1}\rangle \geq 0, \tag{66}$$

which proves eq. (64). □

Lemma 16. *Any $x, y \in (0,1]$ satisfy the following inequalities.*

$$|x - y| \leq |\ln x - \ln y| \tag{67}$$
$$|x\ln x - y\ln y| \leq |\ln x - \ln y| \tag{68}$$

Proof. Denote $a_1 = \ln x \leq 0$, $a_2 = \ln y \leq 0$, $g(a) := e^a$ and $h(a) := ae^a$. Then this lemma can be proved as follows

$$|x-y| = |g(a_1) - g(a_2)| \leq |a_1 - a_2| \sup_{a\leq 0} |g'(a)| = |a_1 - a_2| \sup_{a\leq 0} e^a = |\ln x - \ln y|$$

$$|x\ln x - y\ln y| = |h(a_1) - h(a_2)| \leq |a_1 - a_2| \sup_{a\leq 0} |h'(a)| \overset{(i)}{=} |\ln x - \ln y|,$$

where (i) uses $\sup_{a\leq 0}|h'(a)| = 1$ which will be proved next.

Note that $h'(a) = e^a(a+1)$ and $h''(a) = e^a(a+2)$. Hence, $h'(a)$ is monotonically decreasing in $(-\infty,-2]$ and increasing in $[-2,0]$. Since $\lim_{a\to-\infty} h'(a) = 0$, $h'(-2) = -e^{-2}$ and $h'(0) = 1$, we have $\sup_{a\leq 0}|h'(a)| = 1$. □

Lemma 17. *The diameter of $\mathcal{P}$ defined as $D_{\mathcal{P}} := \sup_{p,\widetilde{p}\in\mathcal{P}} \|\widetilde{p} - p\|$ ranges in $[0, \sqrt{2|\mathcal{S}||\mathcal{A}|}]$.*

Proof. For any $p, \widetilde{p} \in \mathcal{P}$,

$$\begin{aligned}
0 &\le \|\widetilde{p} - p\|^2 \\
&= \sum_{s,a,s'} [\widetilde{p}(s'|s,a) - p(s'|s,a)]^2 \\
&\le |\mathcal{S}||\mathcal{A}| \max_{s,a} \sum_{s'} [\widetilde{p}(s'|s,a) - p(s'|s,a)]^2 \\
&= |\mathcal{S}||\mathcal{A}| \max_{s,a} \sum_{s'} [\widetilde{p}(s'|s,a)^2 + p(s'|s,a)^2 - 2p(s'|s,a)\widetilde{p}(s'|s,a)] \\
&\overset{(i)}{\le} |\mathcal{S}||\mathcal{A}| \max_{s,a} \sum_{s'} [\widetilde{p}(s'|s,a) + p(s'|s,a)] = 2|\mathcal{S}||\mathcal{A}|,
\end{aligned}$$

where (i) uses $p(s'|s,a), \widetilde{p}(s'|s,a) \in [0,1]$. □

G. Proof of Proposition 1

Proposition 1. *Under Assumption 1, for any $\epsilon \ge 0$ and $\tau > 0$, (ϵ, τ)-Nash equilibrium exists. If $(\pi, p) \in \Pi \times \mathcal{P}$ is an (ϵ, τ)-Nash equilibrium, then π is a $\left(2\epsilon + \frac{\tau \ln |\mathcal{A}|}{1-\gamma}\right)$-optimal robust policy to the optimization problem (2).*

Proof. **Proof of (ϵ, τ)-Nash equilibrium existence:**

Fix any $\tau > 0$. Based on (Cen et al., 2022), for any $p \in \mathcal{P}$, there exists a unique optimal policy $\pi_p := \arg\min_{\pi \in \Pi} J_{\rho,\tau}(\pi, p)$. Then based on the Danskin's Theorem (Bernhard and Rapaport, 1995), $F_{\rho,\tau}(p) := J_{\rho,\tau}(\pi_p, p)$ is differentiable with $\nabla F_{\rho,\tau}(p) := \nabla_2 J_{\rho,\tau}(\pi_p, p)$. Such a differentiable function $F_{\rho,\tau}(p)$ has minimum in the compact set $\mathcal{P}$, so there exists $p^* \in \arg\min_{p \in \mathcal{P}} F_{\rho,\tau}(p)$.

Note that $J_{\rho,\tau}(\pi_{p^*}, p^*) = \min_{\pi' \in \Pi} J_{\rho,\tau}(\pi', p^*)$ based on the definition of π_p. Then it suffices to prove that $J_{\rho,\tau}(\pi_{p^*}, p^*) = \max_{p' \in \mathcal{P}} J_{\rho,\tau}(\pi_{p^*}, p')$, which along with $J_{\rho,\tau}(\pi_{p^*}, p^*) = \min_{\pi' \in \Pi} J_{\rho,\tau}(\pi', p^*)$ implies that (π', p^*) is a $(0,\tau)$-Nash equilibrium and thus also an (ϵ, τ)-Nash equilibrium for any $\epsilon \ge 0$.

Note that the proof of Proposition 3 does not rely on the existence of (ϵ, τ)-Nash equilibrium, so we can apply Proposition 3 and obtain that

$$0 \le \max_{p' \in \mathcal{P}} J_{\rho,\tau}(\pi_{p^*}, p') - J_{\rho,\tau}(\pi_{p^*}, p^*) \le \frac{D}{1-\gamma} \max_{p' \in \mathcal{P}} (p' - p)^\top \nabla_2 J_{\rho,\tau}(\pi_{p^*}, p^*) \overset{(i)}{=} 0, \tag{69}$$

where (i) uses $\nabla_2 J_{\rho,\tau}(\pi_{p^*}, p^*) = \nabla F_{\rho,\tau}(p^*) = 0$ since $p^* \in \arg\min_{p \in \mathcal{P}} F_{\rho,\tau}(p)$. Hence, $J_{\rho,\tau}(\pi_{p^*}, p^*) = \max_{p' \in \mathcal{P}} J_{\rho,\tau}(\pi_{p^*}, p')$.

Proof of optimal robust policy: Note that (π, p) satisfy the following (ϵ, τ)-Nash equilibrium conditions

$$J_{\rho,\tau}(\pi, p) - \min_{\pi' \in \Pi} J_{\rho,\tau}(\pi', p) \le \epsilon, \quad \max_{p' \in \mathcal{P}} J_{\rho,\tau}(\pi, p') - J_{\rho,\tau}(\pi, p) \le \epsilon. \tag{70}$$

Therefore,

$$\begin{aligned}
&\Phi_\rho(\pi) - \Phi_\rho(\pi^*) \\
&= \max_{p' \in \mathcal{P}} J_\rho(\pi, p') - \min_{\pi' \in \Pi} \max_{p'' \in \mathcal{P}} J_\rho(\pi', p'') \\
&\le \max_{p' \in \mathcal{P}} J_\rho(\pi, p') - \min_{\pi' \in \Pi} J_\rho(\pi', p) \\
&\overset{(i)}{\le} \max_{p' \in \mathcal{P}} J_{\rho,\tau}(\pi, p') - \min_{\pi' \in \Pi} J_{\rho,\tau}(\pi', p) + \frac{\tau \ln |\mathcal{A}|}{1-\gamma} \\
&\overset{(ii)}{\le} 2\epsilon + \frac{\tau \ln |\mathcal{A}|}{1-\gamma}
\end{aligned} \tag{71}$$

where (i) uses $J_\rho(\pi, p) := J_{\rho,\tau}(\pi, p) + \tau \mathcal{H}_{\rho,p}(\pi)$ with entropy regularizer $\mathcal{H}_{\rho,p}(\pi) := -\mathbb{E}_{\pi,p}[\sum_{t=0}^{\infty} \gamma^t \ln \pi(a_t \mid s_t) \,|\, s_0 \sim \rho] \in \left[0, \frac{\tau \ln |\mathcal{A}|}{1-\gamma}\right]$ and (ii) uses the conditions (70). The above inequality means π is a $\left(2\epsilon + \frac{\tau \ln |\mathcal{A}|}{1-\gamma}\right)$-optimal policy by Definition 1. □

H. Proof of Proposition 2

Proposition 2. *Under Assumption 1, $F_{\rho,\tau}(p)$ is Lipschitz smooth with parameter $\ell_F := \frac{8|\mathcal{S}||\mathcal{A}|(1+\gamma\tau\ln|\mathcal{A}|)^2}{\tau(1-\gamma)^5}$, i.e., for any $p, p' \in \mathcal{P}$,*

$$\|\nabla F_{\rho,\tau}(p') - \nabla F_{\rho,\tau}(p)\| \leq \ell_F \|p' - p\|. \tag{11}$$

Proof. Based on Lemma 2 of (Cen et al., 2022), $\widetilde{Q}_\tau(\pi_p, p)$ defined by eq. (30) is the unique fixed point of the following Bellman operator T_p.

$$T_p Q(s,a) := \min_{\pi\in\Pi} \sum_{s'} p(s'|s,a)\Big(c(s,a,s') + \gamma \sum_{a'} \pi(a'|s')[Q(s',a') + \tau \ln \pi(a'|s')]\Big) \tag{72}$$

The Bellman operator T_p above has the following two properties.

1. Based on Lemma 2 of (Cen et al., 2022), T_p is a γ-contraction under ℓ_∞-norm, i.e.,

$$\|T_p Q' - T_p Q\|_\infty \leq \gamma \|Q' - Q\|_\infty; \forall p \in \mathcal{P}, Q, Q' \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}. \tag{73}$$

2. For any $p, p' \in \mathcal{P}$, $\pi \in \Pi$, $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, we have

$$\begin{aligned}
&\Big|\sum_{s'} [p'(s'|s,a) - p(s'|s,a)]\Big(c(s,a,s') + \gamma \sum_{a'} \pi(a'|s')[Q(s',a') + \tau \ln \pi(a'|s')]\Big)\Big| \\
&\overset{(i)}{\leq} [1 + \gamma(\|Q\|_\infty + \tau \ln|\mathcal{A}|)] \sum_{s'} |p'(s'|s,a) - p(s'|s,a)|,
\end{aligned}$$

where (i) uses $c(s,a,s') \in [0,1]$ and $\sum_{a'} \pi(a'|s') \ln \pi(a'|s') \in [-\ln|\mathcal{A}|, 0]$. Hence,

$$\|T_{p'} Q - T_p Q\|_\infty \leq [1 + \gamma(\|Q\|_\infty + \tau \ln|\mathcal{A}|)] \max_{s,a} \|p'(\cdot|s,a) - p(\cdot|s,a)\|_1. \tag{74}$$

Based on the above two properties, for any $p, p' \in \mathcal{P}$, we have

$$\begin{aligned}
&\|\widetilde{Q}_\tau(\pi_{p'}, p') - \widetilde{Q}_\tau(\pi_p, p)\|_\infty \\
&= \|T_{p'} \widetilde{Q}_\tau(\pi_{p'}, p') - T_p \widetilde{Q}_\tau(\pi_p, p)\|_\infty \\
&\leq \|T_{p'} \widetilde{Q}_\tau(\pi_{p'}, p') - T_p \widetilde{Q}_\tau(\pi_{p'}, p')\|_\infty + \|T_p \widetilde{Q}_\tau(\pi_{p'}, p') - T_p \widetilde{Q}_\tau(\pi_p, p)\|_\infty \\
&\overset{(i)}{\leq} [1 + \gamma(\|\widetilde{Q}_\tau(\pi_{p'}, p')\|_\infty + \tau \ln|\mathcal{A}|)] \max_{s,a} \|p'(\cdot|s,a) - p(\cdot|s,a)\|_1 + \gamma \|\widetilde{Q}_\tau(\pi_{p'}, p') - \widetilde{Q}_\tau(\pi_p, p)\|_\infty \\
&\overset{(ii)}{\leq} \Big(1 + \frac{\gamma \max(1, \gamma\tau \ln|\mathcal{A}|)}{1-\gamma} + \gamma\tau \ln|\mathcal{A}|\Big) \max_{s,a} \|p'(\cdot|s,a) - p(\cdot|s,a)\|_1 + \gamma \|\widetilde{Q}_\tau(\pi_{p'}, p') - \widetilde{Q}_\tau(\pi_p, p)\|_\infty \\
&\leq \frac{1 + \gamma\tau \ln|\mathcal{A}|}{1-\gamma} \max_{s,a} \|p'(\cdot|s,a) - p(\cdot|s,a)\|_1 + \gamma \|\widetilde{Q}_\tau(\pi_{p'}, p') - \widetilde{Q}_\tau(\pi_p, p)\|_\infty,
\end{aligned}$$

where (i) uses eqs. (73)-(74) and (ii) uses eq. (32). Rearranging the above inequality yields that

$$\|\widetilde{Q}_\tau(\pi_{p'}, p') - \widetilde{Q}_\tau(\pi_p, p)\|_\infty \leq \frac{1 + \gamma\tau \ln|\mathcal{A}|}{(1-\gamma)^2} \max_{s,a} \|p'(\cdot|s,a) - p(\cdot|s,a)\|_1. \tag{75}$$

Therefore,

$$\begin{aligned}
&|\ln \pi_{p'}(a|s) - \ln \pi_p(a|s)| \\
&\overset{(i)}{=} \frac{1}{\tau} |\widetilde{Q}_\tau(\pi_p, p; s, a) - \widetilde{Q}_\tau(\pi_{p'}, p'; s, a)| + \Big|\ln \frac{\sum_{a'} \exp[-\widetilde{Q}_\tau(\pi_p, p; s, a')/\tau]}{\sum_{a'} \exp[-\widetilde{Q}_\tau(\pi_{p'}, p'; s, a')/\tau]}\Big|
\end{aligned}$$

$$
\begin{aligned}
&\overset{(ii)}{\leq} \frac{1}{\tau}\|\widetilde{Q}_\tau(\pi_{p'},p') - \widetilde{Q}_\tau(\pi_p,p)\|_\infty + \Big|\ln \frac{\sum_{a'}\exp[-\widetilde{Q}_\tau(\pi_p,p;s,a')/\tau]}{\sum_{a'}\exp[-\widetilde{Q}_\tau(\pi_p,p;s,a')/\tau - \|\widetilde{Q}_\tau(\pi_{p'},p') - \widetilde{Q}_\tau(\pi_p,p)\|_\infty/\tau]}\Big| \\
&= \frac{2}{\tau}\|\widetilde{Q}_\tau(\pi_{p'},p') - \widetilde{Q}_\tau(\pi_p,p)\|_\infty \\
&\overset{(iii)}{\leq} \frac{2+2\gamma\tau\ln|\mathcal{A}|}{\tau(1-\gamma)^2}\max_{s,a}\|p'(\cdot|s,a) - p(\cdot|s,a)\|_1 \\
&\leq \frac{2\sqrt{|\mathcal{S}|}(1+\gamma\tau\ln|\mathcal{A}|)}{\tau(1-\gamma)^2}\|p'-p\|, \qquad (76)
\end{aligned}
$$

where (i) uses eq. (34), (ii) uses $\widetilde{Q}_\tau(\pi_{p'},p';s,a') \leq \widetilde{Q}_\tau(\pi_p,p;s,a') + \|\widetilde{Q}_\tau(\pi_{p'},p') - \widetilde{Q}_\tau(\pi_p,p)\|_\infty$ and (iii) uses eq. (75). Therefore, eq. (11) can be proved as follows.

$$
\begin{aligned}
&\|\nabla F_{\rho,\tau}(p') - \nabla F_{\rho,\tau}(p)\| \overset{(i)}{=} \|\nabla_2 J_{\rho,\tau}(\pi_{p'},p') - \nabla_2 J_{\rho,\tau}(\pi_p,p)\| \\
&\leq \|\nabla_2 J_{\rho,\tau}(\pi_{p'},p') - \nabla_2 J_{\rho,\tau}(\pi_p,p')\| + \|\nabla_2 J_{\rho,\tau}(\pi_p,p') - \nabla_2 J_{\rho,\tau}(\pi_p,p)\| \\
&\overset{(ii)}{\leq} \ell_\pi \max_s \|\ln\pi_{p'}(\cdot|s) - \ln\pi_p(\cdot|s)\| + \ell_p\|p'-p\| \\
&\leq \ell_\pi\sqrt{|\mathcal{A}|}\max_{s,a}|\ln\pi_{p'}(a|s) - \ln\pi_p(a|s)| + \ell_p\|p'-p\| \\
&\overset{(iii)}{\leq} \Big(\frac{|\mathcal{A}|\sqrt{|\mathcal{S}|}(2+3\gamma\tau\ln|\mathcal{A}|)}{(1-\gamma)^3}\frac{2\sqrt{|\mathcal{S}|}(1+\gamma\tau\ln|\mathcal{A}|)}{\tau(1-\gamma)^2} + \frac{2\gamma|\mathcal{S}|(1+\tau\ln|\mathcal{A}|)}{(1-\gamma)^3}\Big)\|p'-p\| \\
&\leq \frac{8|\mathcal{S}||\mathcal{A}|(1+\gamma\tau\ln|\mathcal{A}|)^2}{\tau(1-\gamma)^5}\|p'-p\|,
\end{aligned}
$$

where (i) uses $\nabla F_{\rho,\tau}(p) = \nabla_2 J_{\rho,\tau}(\pi_p,p)$ based on the Danskin's Theorem (Bernhard and Rapaport, 1995), (ii) uses eqs. (41) and (42), and (iii) uses $\ell_\pi := \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}(2+3\gamma\tau\ln|\mathcal{A}|)}{(1-\gamma)^3}$, $\ell_p := \frac{2\gamma|\mathcal{S}|(1+\tau\ln|\mathcal{A}|)}{(1-\gamma)^3}$ and eq. (76). □

I. Proof of Proposition 3

Proposition 3 (Gradient dominance). *Under Assumption 1, the function $J_{\rho,\tau}$ satisfies the following gradient dominance property for any $\pi \in \Pi$ and $p \in \mathcal{P}$,*

$$
\begin{aligned}
&\max_{p'\in\mathcal{P}} J_{\rho,\tau}(\pi,p') - J_{\rho,\tau}(\pi,p) \\
&\leq \frac{D}{1-\gamma}\max_{p'\in\mathcal{P}}(p'-p)^\top\nabla_p J_{\rho,\tau}(\pi,p). \qquad (12)
\end{aligned}
$$

Proof. Based on Lemma 4.3 of (Wang et al., 2023), the gradient dominance property (12) holds for J_ρ, i.e.,

$$
\max_{p'\in\mathcal{P}} J_\rho(\pi,p') - J_\rho(\pi,p) \leq \frac{D}{1-\gamma}\max_{p'\in\mathcal{P}}(p'-p)^\top\nabla_p J_\rho(\pi,p).
$$

Note that for any fixed policy π, the function $J_\rho(\pi,p) = \mathbb{E}\big[\sum_{t=0}^\infty \gamma^t c_t \big| s_0 = s\big]$ becomes $J_{\rho,\tau}(\pi,p) = \mathbb{E}\big[\sum_{t=0}^\infty \gamma^t[c_t + \tau\ln\pi(a_t|s_t)]\big|s_0=s\big]$ after replacing the cost $c_t = c(s_t,a_t,s_{t+1})$ with $c_t + \tau\ln\pi(a_t|s_t)$. Therefore, the gradient dominance property (12) also holds for $J_{\rho,\tau}$.

If $p\in\mathcal{P}$ satisfies $\|\nabla_p F_{\rho,\tau}(p)\| \leq \frac{\epsilon(1-\gamma)}{DD_\mathcal{P}}$, then we prove below that (π_p,p) is an (ϵ,τ)-Nash equilibrium.

$$
\begin{aligned}
&J_{\rho,\tau}(\pi_p,p) - \min_{\pi'\in\Pi} J_{\rho,\tau}(\pi',p) = 0 \leq \epsilon, \\
&\max_{p'\in\mathcal{P}} J_{\rho,\tau}(\pi,p') - J_{\rho,\tau}(\pi_p,p) \leq \frac{D}{1-\gamma}\max_{p'\in\mathcal{P}}(p'-p)^\top\nabla_2 J_{\rho,\tau}(\pi_p,p) \\
&\qquad\qquad \leq \frac{D}{1-\gamma}\max_{p'\in\mathcal{P}}\|p'-p\|\|\nabla F_{\rho,\tau}(p)\| \leq \frac{DD_\mathcal{P}}{1-\gamma}\frac{\epsilon(1-\gamma)}{DD_\mathcal{P}} = \epsilon.
\end{aligned}
$$

□

J. Proof of Proposition 4

Proposition 4. *Under Assumption 1, the function $J_{\rho,\tau}$ satisfies the following gradient dominance property for any $\pi \in \Pi$, $\xi \in \Xi$.*

$$\begin{aligned}&\max_{p' \in \mathcal{P}} J_{\rho,\tau}(\pi, p') - J_{\rho,\tau}(\pi, p_\xi)\\ &\leq \frac{D}{1-\gamma} \max_{\xi' \in \Xi} (\xi' - \xi)^\top \nabla_\xi J_{\rho,\tau}(\pi, p_\xi).\end{aligned} \tag{24}$$

Proof. Proposition 4 can be proved as follows.

$$\begin{aligned}\max_{p' \in \mathcal{P}} J_\rho(\pi, p') - J_\rho(\pi, p_\xi) &\overset{(i)}{\leq} \frac{D}{1-\gamma} \max_{p' \in \mathcal{P}} (p' - p_\xi)^\top \nabla_p J_\rho(\pi, p_\xi)\\ &\overset{(ii)}{=} \frac{D}{1-\gamma} \max_{\xi' \in \Xi} (p_{\xi'} - p_\xi)^\top \nabla_p J_\rho(\pi, p_\xi)\\ &\overset{(iii)}{=} \frac{D}{1-\gamma} \max_{\xi' \in \Xi} (\xi' - \xi)^\top \Psi^\top \nabla_p J_\rho(\pi, p_\xi)\\ &\overset{(iv)}{=} \frac{D}{1-\gamma} \max_{\xi' \in \Xi} (\xi' - \xi)^\top \nabla_\xi J_\rho(\pi, p_\xi),\end{aligned} \tag{77}$$

where (i) uses Proposition 3, (ii) uses $\mathcal{P} := \{p_\xi : \xi \in \Xi\}$ and (iii)-(iv) use $p_\xi = \Psi \xi$. □

K. Proof of Theorem 1

Theorem 1. *Implement Algorithm 1 with $\beta \leq \frac{1}{2\ell_F}$, $\eta = \frac{1-\gamma}{\tau}$. Then the output $(\pi_{\widetilde{T}}, p_{\widetilde{T}})$ satisfies the following rates under Assumption 1.*

$$J_{\rho,\tau}(\pi_{\widetilde{T}}, p_{\widetilde{T}}) - \min_{\pi \in \Pi} J_{\rho,\tau}(\pi, p_{\widetilde{T}}) \leq \mathcal{O}\Big(\frac{\gamma^{T'} + \epsilon_1}{\tau}\Big), \tag{13}$$

$$\begin{aligned}&\max_{p \in \mathcal{P}} J_{\rho,\tau}(\pi_{\widetilde{T}}, p) - J_{\rho,\tau}(\pi_{\widetilde{T}}, p_{\widetilde{T}})\\ &\leq \mathcal{O}\Big[(1 + \tau\epsilon_2)\Big(\frac{\gamma^{T'} + \epsilon_1}{\tau} + \epsilon_2 + \frac{1}{\sqrt{T\beta}}\Big)\Big].\end{aligned} \tag{14}$$

Proof. Based on Lemma 9, the output $\pi_t := \pi_{t,T'}$ of the NPG step (7) with stepsize $\eta = \frac{1-\gamma}{\tau}$ has the following convergence rates.

$$\|Q_\tau(\pi_t, p_t) - Q_\tau(\pi_t^*, p_t)\|_\infty \leq \frac{\gamma^{T'+1}(1 + \gamma\tau \ln|\mathcal{A}|)}{1-\gamma} + \frac{2\gamma\epsilon_1}{(1-\gamma)^2}, \tag{78}$$

$$\|\pi_t^* - \pi_t\|_\infty \leq \|\ln \pi_t^* - \ln \pi_t\|_\infty \leq \frac{2\gamma^{T'}(1 + \gamma\tau \ln|\mathcal{A}|)}{\tau(1-\gamma)} + \frac{4\epsilon_1}{\tau(1-\gamma)^2}. \tag{79}$$

Hence, the convergence rate (13) can be proved as follows.

$$\begin{aligned}&J_{\rho,\tau}(\pi_{\widetilde{T}}, p_{\widetilde{T}}) - \min_{\pi \in \Pi} J_{\rho,\tau}(\pi, p_{\widetilde{T}}) = J_{\rho,\tau}(\pi_{\widetilde{T}}, p_{\widetilde{T}}) - J_{\rho,\tau}(\pi_{\widetilde{T}}^*, p_{\widetilde{T}})\\ &\leq \mathbb{E}_{s \sim \rho}[V_\tau(\pi_{\widetilde{T}}, p_{\widetilde{T}}; s) - V_\tau(\pi_{\widetilde{T}}^*, p_{\widetilde{T}}; s)]\\ &\overset{(i)}{=} \mathbb{E}_{s \sim \rho} \sum_a \Big[\pi_{\widetilde{T}}(a|s)[Q_\tau(\pi_{\widetilde{T}}, p_{\widetilde{T}}; s, a) - \tau \ln \pi_{\widetilde{T}}(a|s)] - \pi_{\widetilde{T}}^*(a|s)[Q_\tau(\pi_{\widetilde{T}}^*, p_{\widetilde{T}}; s, a) - \tau \ln \pi_{\widetilde{T}}^*(a|s)]\Big]\\ &= \mathbb{E}_{s \sim \rho} \sum_a \Big[[\pi_{\widetilde{T}}(a|s) - \pi_{\widetilde{T}}^*(a|s)][Q_\tau(\pi_{\widetilde{T}}^*, p_{\widetilde{T}}; s, a) - \tau \ln \pi_{\widetilde{T}}^*(a|s)]\\ &\quad + \pi_{\widetilde{T}}(a|s)[Q_\tau(\pi_{\widetilde{T}}, p_{\widetilde{T}}; s, a) - Q_\tau(\pi_{\widetilde{T}}^*, p_{\widetilde{T}}; s, a) - \tau \ln \pi_{\widetilde{T}}(a|s) + \tau \ln \pi_{\widetilde{T}}^*(a|s)]\Big]\end{aligned}$$

$$
\begin{aligned}
&\overset{(ii)}{\leq} \mathbb{E}_{s\sim\rho}\sum_a\Big[\Big(\frac{2\gamma^{T'}(1+\gamma\tau\ln|\mathcal{A}|)}{\tau(1-\gamma)}+\frac{4\epsilon_1}{\tau(1-\gamma)^2}\Big)\Big(\frac{2+\tau\ln|\mathcal{A}|}{1-\gamma}\Big)\\
&\quad+\pi_{\widetilde{T}}(a|s)\Big(\frac{\gamma^{T'+1}(1+\gamma\tau\ln|\mathcal{A}|)}{1-\gamma}+\frac{2\gamma\epsilon_1}{(1-\gamma)^2}+\tau\Big(\frac{2\gamma^{T'}(1+\gamma\tau\ln|\mathcal{A}|)}{\tau(1-\gamma)}+\frac{4\epsilon_1}{\tau(1-\gamma)^2}\Big)\Big)\Big]\\
&\leq \frac{3|\mathcal{A}|}{1-\gamma}\Big(\frac{2\gamma^{T'}(1+\gamma\tau\ln|\mathcal{A}|)}{\tau(1-\gamma)}+\frac{4\epsilon_1}{\tau(1-\gamma)^2}\Big)+\frac{12\gamma^{T'}}{1-\gamma}+\frac{6\epsilon_1}{(1-\gamma)^2}\\
&=\frac{2+3\tau\ln|\mathcal{A}|}{\tau(1-\gamma)^2}\Big(2\gamma^{T'}(1+\gamma\tau\ln|\mathcal{A}|)+\frac{4\epsilon_1}{1-\gamma}\Big)\leq\mathcal{O}\Big(\frac{\gamma^{T'}+\epsilon_1}{\tau}\Big),
\end{aligned}
$$

where (i) uses $V_\tau(\pi,p;s)=\mathbb{E}_{a\sim\pi(\cdot|s)}\big[Q_\tau(\pi,p;s,a)-\tau\ln\pi(a|s)\big]$ based on eqs. (4) and (5), (ii) uses eqs. (32), (33), (78) and (79).

Next, we will prove the convergence rate (14). Note that

$$
\begin{aligned}
\|\widehat{\nabla}_p J_{\rho,\tau}(\pi_t,p_t)\| &= \frac{1}{\beta}\big\|\big(p_t+\beta\widehat{\nabla}_p J_{\rho,\tau}(\pi_t,p_t)\big)-p_t\big\|\\
&\overset{(i)}{\geq}\frac{1}{\beta}\big\|\big(p_t+\beta\widehat{\nabla}_p J_{\rho,\tau}(\pi_t,p_t)\big)-\mathrm{proj}_{\mathcal{P}}\big(p_t+\beta\widehat{\nabla}_p J_{\rho,\tau}(\pi_t,p_t)\big)\big\|\\
&\overset{(ii)}{=}\|\widehat{\nabla}_p J_{\rho,\tau}(\pi_t,p_t)-G_t\|,
\end{aligned}
$$

where (i) uses $p_t\in\mathcal{P}$ and (ii) denotes $G_t:=\frac{1}{\beta}\big(\mathrm{proj}_{\mathcal{P}}[p_t+\beta\widehat{\nabla}_p J_{\rho,\tau}(\pi_t,p_t)]-p_t\big)$. The above inequality implies that

$$
G_t^\top\widehat{\nabla}_p J_{\rho,\tau}(\pi_t,p_t)\geq\frac{1}{2}\|G_t\|^2. \tag{80}
$$

Since $F_{\rho,\tau}(p):=\max_{\pi\in\Pi}J_{\rho,\tau}(\pi,p)$ is ℓ_F-smooth as shown in Proposition 2, we have

$$
\begin{aligned}
F_{\rho,\tau}(p_{t+1})-F_{\rho,\tau}(p_t)&\geq\nabla F_{\rho,\tau}(p_t)^\top(p_{t+1}-p_t)-\frac{\ell_F}{2}\|p_{t+1}-p_t\|^2\\
&\overset{(i)}{=}\beta G_t^\top[\nabla F_{\rho,\tau}(p_t)-\widehat{\nabla}_p J_{\rho,\tau}(\pi_t,p_t)]+\beta G_t^\top\widehat{\nabla}_p J_{\rho,\tau}(\pi_t,p_t)-\frac{\ell_F\beta^2}{2}\|G_t\|^2\\
&\overset{(ii)}{\geq}-\beta\|G_t\|\big(\ell_\pi\sqrt{|\mathcal{A}|}\|\ln\pi_t-\ln\pi_t^*\|_\infty+\epsilon_2\big)+\frac{\beta}{2}\|G_t\|^2-\frac{\beta}{4}\|G_t\|^2\\
&\overset{(iii)}{\geq}\frac{\beta}{8}\|G_t\|^2-2\beta\big(\ell_\pi\sqrt{|\mathcal{A}|}\|\ln\pi_t-\ln\pi_t^*\|_\infty+\epsilon_2\big)^2,
\end{aligned} \tag{81}
$$

where (i) uses $p_{t+1}-p_t=\beta G_t$, (ii) uses $\beta\leq\frac{1}{2\ell_F}$ and eqs. (44) and (80), and (iii) uses $c\|G_t\|\leq 2c^2+\frac{\|G_t\|^2}{8}$ for $c:=\ell_\pi\sqrt{|\mathcal{A}|}\|\ln\pi_t-\ln\pi_t^*\|_\infty+\epsilon_2$. Averaging the above inequality over $t=0,1,\ldots,T-1$, we obtain that

$$
\begin{aligned}
\|G_{\widetilde{T}}\|=\min_{0\leq t\leq T-1}\|G_t\|&\leq\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\|G_t\|^2}\\
&\leq\sqrt{\frac{16}{T}\sum_{t=0}^{T-1}\big(\ell_\pi\sqrt{|\mathcal{A}|}\|\ln\pi_t-\ln\pi_t^*\|_\infty+\epsilon_2\big)^2+\frac{8[F_{\rho,\tau}(p_T)-F_{\rho,\tau}(p_0)]}{T\beta}}\\
&\overset{(i)}{\leq}\sqrt{16\Big[\frac{|\mathcal{A}|\sqrt{|\mathcal{S}|}(2+3\gamma\tau\ln|\mathcal{A}|)}{(1-\gamma)^3}\Big(\frac{2\gamma^{T'}(1+\gamma\tau\ln|\mathcal{A}|)}{\tau(1-\gamma)}+\frac{4\epsilon_1}{\tau(1-\gamma)^2}\Big)+\epsilon_2\Big]^2+\frac{8(1+\tau\ln|\mathcal{A}|)}{T\beta(1-\gamma)}}\\
&\leq\frac{4|\mathcal{A}|\sqrt{|\mathcal{S}|}(2+3\gamma\tau\ln|\mathcal{A}|)}{\tau(1-\gamma)^4}\Big(2\gamma^{T'}(1+\gamma\tau\ln|\mathcal{A}|)+\frac{4\epsilon_1}{1-\gamma}\Big)+4\epsilon_2+\sqrt{\frac{8(1+\tau\ln|\mathcal{A}|)}{T\beta(1-\gamma)}}
\end{aligned} \tag{82}
$$

where (i) uses eq. (31), eq. (79), $\ell_\pi:=\frac{\sqrt{|\mathcal{S}||\mathcal{A}|}(2+3\gamma\tau\ln|\mathcal{A}|)}{(1-\gamma)^3}$.

Then, the convergence rate (14) can be proved as follows.

$$
\begin{aligned}
&\max_{p\in\mathcal{P}} J_{\rho,\tau}(\pi_{\widetilde{T}},p) - J_{\rho,\tau}(\pi_{\widetilde{T}},p_{\widetilde{T}}) \\
&\overset{(i)}{\leq} \frac{D}{1-\gamma}\max_{p'\in\mathcal{P}}\langle p'-p_{\widetilde{T}}, \nabla_p J_{\rho,\tau}(\pi_{\widetilde{T}},p_{\widetilde{T}})\rangle \\
&\leq \frac{D}{1-\gamma}\max_{p'\in\mathcal{P}}\langle p'-p_{\widetilde{T}}, \widehat{\nabla}_p J_{\rho,\tau}(\pi_{\widetilde{T}},p_{\widetilde{T}})\rangle + \frac{D}{1-\gamma}\max_{p'\in\mathcal{P}}\|p'-p_{\widetilde{T}}\|\|\widehat{\nabla}_p J_{\rho,\tau}(\pi_{\widetilde{T}},p_{\widetilde{T}}) - \nabla_p J_{\rho,\tau}(\pi_{\widetilde{T}},p_{\widetilde{T}})\| \\
&\overset{(ii)}{\leq} \frac{D}{1-\gamma}\max_{p'\in\mathcal{P}}\left[\langle p'-p_{\widetilde{T}+1}, \widehat{\nabla}_p J_{\rho,\tau}(\pi_{\widetilde{T}},p_{\widetilde{T}})\rangle + \langle p_{\widetilde{T}+1}-p_{\widetilde{T}}, \widehat{\nabla}_p J_{\rho,\tau}(\pi_{\widetilde{T}},p_{\widetilde{T}})\rangle\right] + \frac{DD_{\mathcal{P}}\epsilon_2}{1-\gamma} \\
&\overset{(iii)}{\leq} \frac{D}{1-\gamma}\max_{p'\in\mathcal{P}}\left[\frac{1}{\beta}\langle p'-p_{\widetilde{T}+1}, p_{\widetilde{T}}+\beta\widehat{\nabla}_p J_{\rho,\tau}(\pi_{\widetilde{T}},p_{\widetilde{T}}) - p_{\widetilde{T}+1}\rangle - \frac{1}{\beta}\langle p'-p_{\widetilde{T}+1}, p_{\widetilde{T}}-p_{\widetilde{T}+1}\rangle + \|\beta G_{\widetilde{T}}\|(L_p+\epsilon_2)\right] \\
&\quad + \frac{DD_{\mathcal{P}}\epsilon_2}{1-\gamma} \\
&\overset{(iv)}{\leq} \frac{D}{1-\gamma}\left[0 + \frac{1}{\beta}\max_{p'\in\mathcal{P}}\|p'-p_{\widetilde{T}+1}\|\|\beta G_{\widetilde{T}}\| + \beta(L_p+\epsilon_2)\|G_{\widetilde{T}}\| + D_{\mathcal{P}}\epsilon_2\right] \\
&\overset{(v)}{\leq} \frac{D}{1-\gamma}\left[[D_{\mathcal{P}}+\beta(L_p+\epsilon_2)]\left(\frac{4|\mathcal{A}|\sqrt{|\mathcal{S}|}(2+3\gamma\tau\ln|\mathcal{A}|)}{\tau(1-\gamma)^4}\left(2\gamma^{T'}(1+\gamma\tau\ln|\mathcal{A}|)+\frac{4\epsilon_1}{1-\gamma}\right)+4\epsilon_2+\sqrt{\frac{8(1+\tau\ln|\mathcal{A}|)}{T\beta(1-\gamma)}}\right)\right. \\
&\quad \left. + D_{\mathcal{P}}\epsilon_2\right] \\
&\overset{(vi)}{\leq} \mathcal{O}\left[(1+\tau\epsilon_2)\left(\frac{\gamma^{T'}+\epsilon_1}{\tau}+\epsilon_2+\frac{1}{\sqrt{T\beta}}\right)\right],
\end{aligned}
$$

where (i) uses Proposition 3, (ii) uses $\|p'-p_{\widetilde{T}}\| \leq D_{\mathcal{P}}$ for $p', p_{\widetilde{T}} \in \mathcal{P}$ where $D_{\mathcal{P}} := \sup_{p,\widetilde{p}\in\mathcal{P}}\|\widetilde{p}-p\|$ denotes the diameter of $\mathcal{P}$ ($D_{\mathcal{P}} \leq \sqrt{2|\mathcal{S}||\mathcal{A}|}$ as shown in Lemma 17) and $\|\widehat{\nabla}_p J_{\rho,\tau}(\pi_{\widetilde{T}},p_{\widetilde{T}}) - \nabla_p J_{\rho,\tau}(\pi_{\widetilde{T}},p_{\widetilde{T}})\| \leq \epsilon_2$, (iii) uses $p_{\widetilde{T}+1}-p_{\widetilde{T}} = \beta G_{\widetilde{T}}$ and $\|\widehat{\nabla}_p J_{\rho,\tau}(\pi_{\widetilde{T}},p_{\widetilde{T}})\| \leq \|\nabla_p J_{\rho,\tau}(\pi_{\widetilde{T}},p_{\widetilde{T}})\| + \|\widehat{\nabla}_p J_{\rho,\tau}(\pi_{\widetilde{T}},p_{\widetilde{T}}) - \nabla_p J_{\rho,\tau}(\pi_{\widetilde{T}},p_{\widetilde{T}})\| \leq L_p+\epsilon_2$ (The second $\leq$ uses eq. (40)), (iv) uses $p_{\widetilde{T}+1}-p_{\widetilde{T}} = \beta G_{\widetilde{T}}$ and eq. (64), (v) uses eq. (82) and $\|p'-p_{\widetilde{T}+1}\| \leq D_{\mathcal{P}}$, and (vi) uses $\beta \leq \frac{1}{2\ell_F} = \frac{\tau(1-\gamma)^5}{16|\mathcal{S}||\mathcal{A}|(1+\gamma\tau\ln|\mathcal{A}|)^2} = \mathcal{O}(\tau)$ and $L_p = \frac{\sqrt{|\mathcal{S}|}(1+\tau\ln|\mathcal{A}|)}{(1-\gamma)^2} = \mathcal{O}(1)$. □

L. Proof of Corollary 1

Corollary 1 (Iteration Complexity of Algorithm 1)**.** *Implement Algorithm 1 under deterministic setting ($\epsilon_1 = \epsilon_2 = 0$). For any $\epsilon > 0$, select hyperparameters $\tau = \min\left(\frac{\epsilon(1-\gamma)}{3\ln|\mathcal{A}|}, 1\right)$, $T = \mathcal{O}(\epsilon^{-3})$, $T' = \mathcal{O}[\ln(\epsilon^{-1})]$, $\eta = \frac{1-\gamma}{\tau}$, $\beta = \frac{1}{2\ell_F}$. Then the output $(\pi_{\widetilde{T}}, p_{\widetilde{T}})$ is both ϵ-optimal robust policy and (ϵ,τ)-Nash equilibrium under Assumption 1. This requires $T = \mathcal{O}(\epsilon^{-3})$ transition kernel updates, $TT' = \mathcal{O}(\epsilon^{-3}\ln\epsilon^{-1})$ policy updates and iteration complexity $T + TT' = \mathcal{O}(\epsilon^{-3}\ln\epsilon^{-1})$.*

Proof. Select the following hyperparameters for Algorithm 1 which satisfies the conditions of Theorem 1.

$$
\begin{aligned}
\epsilon_1 &= \epsilon_2 = 0 && (83)\\
\tau &= \min\left(\frac{\epsilon(1-\gamma)}{3\ln|\mathcal{A}|}, 1\right) && (84)\\
\beta &= \frac{1}{2\ell_F} = \frac{\tau(1-\gamma)^5}{16|\mathcal{S}||\mathcal{A}|(1+\gamma\tau\ln|\mathcal{A}|)^2} && (85)\\
\eta &= \frac{1-\gamma}{\tau} && (86)\\
T &= \mathcal{O}(\epsilon^{-3}) && (87)\\
T' &= \frac{\mathcal{O}[\ln(\tau^{-1}\epsilon^{-1})]}{\ln(\gamma^{-1})} = \mathcal{O}[\ln(\epsilon^{-1})]. && (88)
\end{aligned}
$$

Therefore, the convergence rates (13) and (14) along with the above hyperparameter choices imply that

$$J_{\rho,\tau}(\pi_{\widetilde{T}}, p_{\widetilde{T}}) - \min_{\pi\in\Pi} J_{\rho,\tau}(\pi, p_{\widetilde{T}}) \le \mathcal{O}\Big(\frac{\gamma^{T'}+\epsilon_1}{\tau}\Big) \le \frac{\epsilon}{3}, \tag{89}$$

$$\max_{p\in\mathcal{P}} J_{\rho,\tau}(\pi_{\widetilde{T}}, p) - J_{\rho,\tau}(\pi_{\widetilde{T}}, p_{\widetilde{T}}) \le \mathcal{O}(1+\tau\epsilon_2)\Big(\frac{\gamma^{T'}+\epsilon_1}{\tau} + \epsilon_2 + \frac{1}{\sqrt{T\beta}}\Big) \le \frac{\epsilon}{3}, \tag{90}$$

which means $(\pi_{\widetilde{T}}, p_{\widetilde{T}})$ is $(\epsilon/3, \tau)$-Nash equilibrium and thus ϵ-optimal robust policy by Proposition 1. □

M. Proof of Corollary 2

Corollary 2 (Sample Complexity of Algorithm 2)**.** *For any $\epsilon > 0$ and $\delta \in (0,1)$, implement Algorithm 2 with hyperparameters $\tau = \min\big(\frac{\epsilon(1-\gamma)}{3\ln|\mathcal{A}|}, 1\big)$, $T = \mathcal{O}(\epsilon^{-3})$, $T' = \mathcal{O}[\ln(\epsilon^{-1})]$, $T_1 = \mathcal{O}(\epsilon^{-4})$, $\alpha = \mathcal{O}[\ln^{-1}(\epsilon^{-1})]$, $\eta = \frac{1-\gamma}{\tau}$, $\beta = \frac{1}{2\ell_F}$, $N = \mathcal{O}(\epsilon^{-2})$, $H = \mathcal{O}[\ln(\epsilon^{-1})]$. The output $(\pi_{\widetilde{T}}, p_{\widetilde{T}})$ is both ϵ-optimal robust policy and (ϵ,τ)-Nash equilibrium with probability at least $1-\delta$ under Assumption 1. Furthermore, the sample complexity is $T(T'T_1 + NH) = \mathcal{O}(\epsilon^{-7}\ln\epsilon^{-1})$.*

Proof. Select the following hyperparameters.

$$\tau = \min\Big(\frac{\epsilon(1-\gamma)}{3\ln|\mathcal{A}|}, 1\Big) = \mathcal{O}(\epsilon) \tag{91}$$

$$\epsilon_1 = \mathcal{O}(\tau\epsilon) = \mathcal{O}(\epsilon^2) \tag{92}$$

$$\epsilon_2 = \mathcal{O}(\epsilon) \tag{93}$$

$$\beta = \frac{1}{2\ell_F} = \mathcal{O}(\tau) = \mathcal{O}(\epsilon) \tag{94}$$

$$\eta = \frac{1-\gamma}{\tau} \tag{95}$$

$$T = \mathcal{O}(\epsilon^{-3}) \tag{96}$$

$$T' = \frac{\mathcal{O}[\ln(\tau^{-1}\epsilon^{-1})]}{\ln(\gamma^{-1})} = \mathcal{O}[\ln(\epsilon^{-1})] \tag{97}$$

$$T_1 = \mathcal{O}(\epsilon_1^{-2}) = \mathcal{O}(\epsilon^{-4}) \tag{98}$$

$$\alpha = \mathcal{O}[\ln^{-1}(\epsilon_1^{-1})] = \mathcal{O}[\ln^{-1}(\epsilon^{-1})] \tag{99}$$

$$\delta_1 = \frac{\delta}{2TT'} \tag{100}$$

$$N = \mathcal{O}(\epsilon_2^{-2}) = \mathcal{O}(\epsilon^{-2}), \tag{101}$$

$$H = \mathcal{O}[\ln(\epsilon_2^{-1})] = \mathcal{O}[\ln(\epsilon^{-1})], \tag{102}$$

$$\delta_2 = \frac{\delta}{2T}, \tag{103}$$

Based on the conditions of Lemmas 9 and 12, for all $t = 0, 1, \ldots, T-1$ and $k = 0, 1, \ldots, T'-1$, eq. (47) of Lemmas 9 and the conclusion of Lemma 12 below hold with probability at least $1 - TT'\delta_1 = 1 - \delta/2$.

$$\|\widehat{Q}_{t,k'} - Q_\tau(\pi_{t,k'}, p_t)\|_\infty \le \epsilon_1; \forall k' = 0, 1, \ldots, k-1 \Rightarrow \inf_{s,a} \pi_{t,k}(a|s) \ge \pi_{\min},$$

$$\inf_{s,a} \ln \pi_{t,k}(a|s) \ge \ln \pi_{\min} = -\mathcal{O}(\tau^{-1}) \Rightarrow |c(s,a,s') + \tau \ln \pi_{t,k}(a|s)| \le \mathcal{O}(1) \Rightarrow \|\widehat{Q}_{t,k} - Q_\tau(\pi_{t,k}, p_t)\|_\infty \le \epsilon_1.$$

Note that $\pi_{t,0}(a|s) \equiv 1/|\mathcal{A}| \ge \pi_{\min}$. Hence, by induction over k, the above statements imply that $\|\widehat{Q}_{t,k} - Q_\tau(\pi_{t,k}, p_t)\|_\infty \le \epsilon_1$ and $\inf_{s,a} \pi_{t,k}(a|s) \ge \pi_{\min}$ for all $t = 0, 1, \ldots, T-1$ and $k = 0, 1, \ldots, T'-1$.

Note that $\epsilon_1 = \mathcal{O}(\epsilon^2)$ and $\epsilon_2 = \mathcal{O}(\epsilon)$ for sufficiently small $\epsilon > 0$ can satisfy the condition of Lemma 14 that $\epsilon_2 \ge \frac{3\gamma\epsilon_1\sqrt{|\mathcal{S}|}}{1-\gamma}$. Hence, based on Lemma 14, the stochastic transition gradients $\widehat{\nabla}_p J_{\rho,\tau}(\pi_t, p_t)$ obtained by eq. (16) for all $t = 0, 1, \ldots, T-1$ satisfy $\|\widehat{\nabla}_p J_{\rho,\tau}(\pi_t, p_t) - \nabla_p J_{\rho,\tau}(\pi_t, p_t)\| \le \epsilon_2$ with probability at least $1 - T\delta_2 = 1 - \delta/2$.

Hence, we proved that $\|\widehat{\nabla}_p J_{\rho,\tau}(\pi_t, p_t) - \nabla_p J_{\rho,\tau}(\pi_t, p_t)\| \leq \epsilon_2$ and $\|\widehat{Q}_{t,k} - Q_\tau(\pi_{t,k}, p_t)\|_\infty \leq \epsilon_1$ hold for all $t = 0, \ldots, T-1$ and $k = 0, \ldots, T'-1$ with probability at least $1-\delta$. Therefore, Algorithm 2 with the above hyperparameter choices can be seen as a special case of Algorithm 1, so the convergence rates (13) and (14) in Theorem 1 hold which imply

$$J_\rho(\pi_{\widetilde{T}}, p_{\widetilde{T}}) - \min_{\pi\in\Pi} J_\rho(\pi, p_{\widetilde{T}}) \leq \mathcal{O}\Big(\frac{\gamma^{T'} + \epsilon_1}{\tau}\Big) \leq \frac{\epsilon}{3},$$
$$\max_{p\in\mathcal{P}} J_\rho(\pi_{\widetilde{T}}, p) - J_\rho(\pi_{\widetilde{T}}, p_{\widetilde{T}}) \leq \mathcal{O}(1+\tau\epsilon_2)\Big(\frac{\gamma^{T'} + \epsilon_1}{\tau} + \epsilon_2 + \frac{1}{\sqrt{T\beta}}\Big) \leq \frac{\epsilon}{3}.$$

Therefore, with probability at least $1-\delta$, $(\pi_{\widetilde{T}}, p_{\widetilde{T}})$ is $(\epsilon/3, \tau)$-Nash equilibrium and thus ϵ-optimal robust policy by Proposition 1. The required total sample complexity is $T(T'T_1 + NH) = \mathcal{O}(\epsilon^{-7}\ln\epsilon^{-1})$. □

N. Proof of Theorem 2

Theorem 2 (Sample Complexity of Algorithm 3)**.** *For any $\epsilon > 0$ and $\delta \in (0,1)$, implement Algorithm 3 with hyperparameters $\tau = \min\big[\mathcal{O}(\sqrt{\zeta}+\epsilon), 1\big]$, $T = \mathcal{O}(\epsilon^{-3})$, $T' = \mathcal{O}[\ln(\epsilon^{-1})]$, $T_1 := \mathcal{O}(\epsilon^{-4})$, $\alpha = \mathcal{O}[\ln^{-1}(\zeta+\epsilon^2)^{-1}]$, $\eta = \frac{1-\gamma}{\tau}$, $\beta = \frac{1}{2\ell_F\|\Psi\|}$, $N = \mathcal{O}(\epsilon^{-4})$, $H = \mathcal{O}[\ln(\epsilon^{-1})]$. Then under Assumption 1 and the assumption that $\inf_{s,a,s'} p_\xi(s'|s,a) > p_{\min}$ for a constant $p_{\min} > 0$, $(\pi_{\widetilde{T}}, p_{\widetilde{T}})$ is both $\big(\mathcal{O}(\sqrt{\zeta}+\zeta+\epsilon), \tau\big)$-Nash equilibrium and $\mathcal{O}(\sqrt{\zeta}+\zeta+\epsilon)$-optimal robust policy with probability at least $1-\delta$. The required sample complexity is $T(T'T_1 + NH) = \mathcal{O}(\epsilon^{-7}\ln\epsilon^{-1})$.*

Proof. Select the following hyperparameters for Algorithm 3.

$$\epsilon_1 = 2\zeta + \epsilon^2 \tag{104}$$
$$\epsilon_2 = \frac{3\gamma\epsilon_1\sqrt{|\mathcal{S}|}}{1-\gamma} = \frac{3\gamma\sqrt{|\mathcal{S}|}(2\zeta+\epsilon^2)}{1-\gamma} \tag{105}$$
$$\delta_1 = \frac{\delta}{2TT'} \tag{106}$$
$$\delta_2 = \frac{\delta}{2T} \tag{107}$$
$$\tau = \min\big[\mathcal{O}(\sqrt{\zeta}+\epsilon), 1\big] \tag{108}$$
$$\beta = \frac{1}{2\ell_F\|\Psi\|} = \mathcal{O}(\tau) \geq \mathcal{O}(\epsilon) \tag{109}$$
$$\eta = \frac{1-\gamma}{\tau} \tag{110}$$
$$T = \mathcal{O}(\epsilon^{-3}) \tag{111}$$
$$T' = \frac{\mathcal{O}[\ln(\epsilon^{-2})]}{\ln(\gamma^{-1})} = \mathcal{O}[\ln(\epsilon^{-1})] \tag{112}$$
$$T_1 = \mathcal{O}(\epsilon^{-4}) \geq \mathcal{O}(\epsilon_1^{-2}) \tag{113}$$
$$\alpha = \mathcal{O}(\ln^{-1}\epsilon_1^{-1}) = \mathcal{O}[\ln^{-1}(\zeta+\epsilon^2)^{-1}] \tag{114}$$
$$N = \mathcal{O}(\epsilon^{-4}) \geq \mathcal{O}(\epsilon_2^{-2}) \tag{115}$$
$$H = \mathcal{O}[\ln(\epsilon^{-1})] \geq \mathcal{O}[\ln(\epsilon_2^{-1})] \tag{116}$$

Based on the conditions of Lemmas 10 and 11, select the following hyperparameters for the TD update rule (21).

Then for all $t = 0, 1, \ldots, T-1$ and $k = 0, 1, \ldots, T'-1$, eq. (51) of Lemmas 10 and the conclusion of Lemma 11 below hold with probability at least $1 - TT'\delta_1' = 1 - \delta/2$.

$$\sup_{s,a} |\phi(s,a)^\top w_{t,k'} - Q_\tau(\pi_{t,k'}, p_t; s, a)| \leq \epsilon_1; \forall k' = 0, 1, \ldots, k-1 \Rightarrow \inf_{s,a} \ln \pi_{t,k}(a|s) \geq \ln \pi_{\min},$$
$$\inf_{s,a} \ln \pi_{t,k}(a|s) \geq \ln \pi_{\min} = -\mathcal{O}(\tau^{-1}) \Rightarrow |c(s,a,s') + \tau \ln \pi_{t,k}(a|s)| \leq \mathcal{O}(1) \Rightarrow \sup_{s,a} |\phi(s,a)^\top w_{t,k} - Q_\tau(\pi_{t,k}, p_t; s, a)| \leq \epsilon_1.$$

Note that $u_{t,0} = 0 \Rightarrow \pi_{t,0}(a|s) \equiv 1/|\mathcal{A}| \geq \pi_{\min}$ based on eq. (22). Hence, by induction over k, the above statements imply that $\sup_{s,a} |\phi(s,a)^\top w_{t,k} - Q_\tau(\pi_{t,k}, p_t; s, a)| \leq \epsilon_1$ and $\inf_{s,a} \pi_{t,k}(a|s) \geq \pi_{\min}$ for all $t = 0, 1, \ldots, T-1$ and $k = 0, 1, \ldots, T'-1$.

Then based on Lemma 13, the stochastic transition gradients $\widehat{\nabla}_\xi J_{\rho,\tau}(\pi_t, p_{\xi_t})$ obtained by eq. (19) for all $t = 0, 1, \ldots, T-1$ satisfy $\|\widehat{\nabla}_\xi J_{\rho,\tau}(\pi_t, p_{\xi_t}) - \nabla_\xi J_{\rho,\tau}(\pi_t, p_{\xi_t})\| \leq \epsilon_2$ with probability at least $1 - T\delta_2 = 1 - \delta/2$.

Hence, we have proved that $\|\widehat{\nabla}_\xi J_{\rho,\tau}(\pi_t, p_{\xi_t}) - \nabla_\xi J_{\rho,\tau}(\pi_t, p_{\xi_t})\| \leq \epsilon_2$ and $\sup_{s,a} |\phi(s,a)^\top w_{t,k} - Q_\tau(\pi_{t,k}, p_t; s, a)| \leq \epsilon_1$ holds for all $t = 0, 1, \ldots, T-1$ and $k = 0, 1, \ldots, T'-1$ with probability at least $1 - \delta$. Therefore, we can prove that the convergence rates in Theorem 1 also hold for Algorithm 3 with probability at least $1 - \delta$. The proof logic is the same as that of Theorem 1. The major difference is that we replace the transition kernel $p \in \mathcal{P}$ with its corresponding parameter ξ. Note that the proof of Theorem 2 uses the gradient dominance property (Proposition 12) about $\nabla_p J_{\rho,\tau}(\pi, p)$ to obtain global convergence, and $\nabla_\xi F_{\rho,\tau}(p_\xi)$ also satisfies gradient dominance property (Proposition 24) of the same form. Hence, we can use the latter here. In addition, since $p_\xi = \Psi\xi$, we have $\nabla_\xi J_{\rho,\tau}(\pi, p_\xi) = \Psi^\top \nabla_p J_{\rho,\tau}(\pi, p)$ and $\nabla_\xi F_{\rho,\tau}(p_\xi) = \Psi^\top \nabla F_{\rho,\tau}(p)$, so the Lipschitz constants L_p, ℓ_p and ℓ_F will be changed to $L_p\|\Psi\|$, $\ell_p\|\Psi\|$ and $\ell_F\|\Psi\|$ respectively, which does not change the order of the convergence rate as $\|\Psi\| = \mathcal{O}(1)$.

Substituting the hyperparameters (104)-(116) into the convergence rates in Theorem 1, we obtain that

$$J_{\rho,\tau}(\pi_{\widetilde{T}}, p_{\widetilde{T}}) - \min_{\pi\in\Pi} J_{\rho,\tau}(\pi, p_{\widetilde{T}}) \leq \mathcal{O}\Big(\frac{\gamma^{T'} + \epsilon_1}{\tau}\Big) \leq \mathcal{O}\Big(\frac{\epsilon^2 + 2\zeta + \epsilon^2}{\min\big[\mathcal{O}(\sqrt{\zeta} + \epsilon), 1\big]}\Big) \leq \mathcal{O}(\zeta + \sqrt{\zeta} + \epsilon),$$

$$\max_{p\in\mathcal{P}} J_{\rho,\tau}(\pi_{\widetilde{T}}, p) - J_{\rho,\tau}(\pi_{\widetilde{T}}, p_{\widetilde{T}}) \leq \mathcal{O}(1 + \tau\epsilon_2)\Big(\frac{\gamma^{T'} + \epsilon_1}{\tau} + \epsilon_2 + \frac{1}{\sqrt{T\beta}}\Big) \leq \mathcal{O}(\zeta + \sqrt{\zeta} + \epsilon).$$

Therefore, with probability at least $1-\delta$, $(\pi_{\widetilde{T}}, p_{\widetilde{T}})$ is $\big(\mathcal{O}(\sqrt{\zeta}+\zeta+\epsilon), \tau\big)$-Nash equilibrium and thus $\mathcal{O}(\sqrt{\zeta}+\zeta+\epsilon)$-optimal robust policy by Proposition 1. The required total sample complexity is

$$T(T'T_1 + NH) = \mathcal{O}\big[\epsilon^{-3}\big(\epsilon^{-4}\ln(\epsilon^{-1}) + \epsilon^{-4}\ln(\epsilon^{-1})\big)\big] = \mathcal{O}\big(\epsilon^{-7}\ln(\epsilon^{-1})\big).$$

□

O. Experiments

The experiments are implemented on Python 3.9 in a MacBook Pro laptop with 500 GB Storage and 8-core CPU (16 GB Memory). The code can be downloaded from https://github.com/changy12/ICML2024-Accelerated-Policy-Gradient-for-s-rectangular-Robust-MDPs-with-Large-State-Spaces.

O.1. Experiments on Small State Space under Deterministic Setting

We compare our Algorithm 1 with the existing double-loop robust policy gradient (DRPG) algorithm (Wang et al., 2023) and actor-critic algorithm (Li et al., 2023b) under deterministic setting (i.e., when exact values of some quantities are available, including gradients, Q functions, V functions, etc.) on the Garnet problem (Archibald et al., 1995; Wang and Zou, 2022) with spaces $\mathcal{S} = \{0, 1, 2, 3, 4\}$ of 5 states and $\mathcal{A} = \{0, 1, 2\}$ of 3 actions. The agent gets cost 0 if it takes action 0 at state 0 or action 1 at other states, and gets cost 1 otherwise. We use s-rectangular L_2-norm ambiguity set $\mathcal{P} := \{p \in (\Delta^{\mathcal{S}})^{\mathcal{S}\times\mathcal{A}} : \|p(s,\cdot,\cdot) - p_0(s,\cdot,\cdot)\| \leq 0.03\}$ where $p_0(s,a,s') \equiv 0.2$ is the nominal transition kernel. The initial state distribution ρ is uniform with $\rho(s) \equiv \frac{1}{5}$. The discount factor is $\gamma = 0.95$.

We implement an exact version of Algorithm 1 (i.e., $\epsilon_1 = \epsilon_2 = 0$) using $T_p = 5$ outer transition kernel updates with stepsize $\beta = 0.001$, and $T' = 1$ inner policy update with stepsize $\eta = \frac{1-\gamma}{\tau} = 50$ per outer update. For DRPG algorithm, we use $T = 5$ outer policy updates (Algorithm 1 of (Wang et al., 2023)) with stepsize $\alpha_t = 10$ and $T_k = 1$ inner transition kernel update (Algorithm 2 of (Wang et al., 2023)) with stepsize $\beta_t = 0.001$ per outer update. For actor-critic algorithm (Algorithm 4.1 of (Li et al., 2023b)), we use $K = 5$ outer iterations, where the actor step (policy update) uses stepsize $\eta = 500$, and the critic step (transition kernel update) uses only 1 iteration of Algorithm 3.2 of (Li et al., 2023b) with $\alpha_m = 1$ as well as P_ϵ obtained by exactly solving the direction-finding subproblem in eq. (3.4) of (Li et al., 2023b). We plot learning curves of the objective function $\Phi_\rho(\pi_t) := \max_{p\in\mathcal{P}} J_\rho(\pi_t, p)$ at each t-th outer iteration on the left of Figure 1. The x-axis is iteration complexity defined as the total number of policy updates and transition kernel updates up to each iteration t. Figure 1 shows

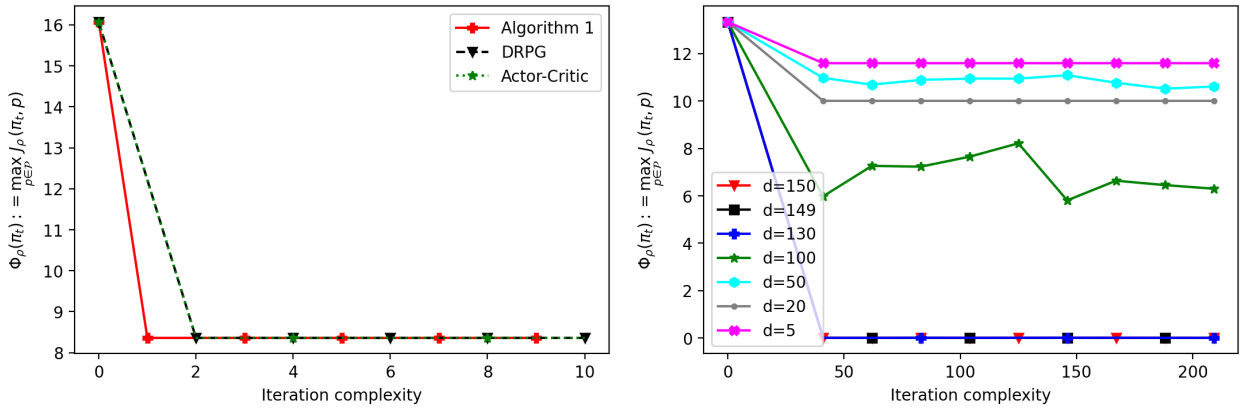

Figure 1: Experimental Results on Small State Space (Left) and Large State Space (Right).

that our Algorithm 1 converges faster to the optimal robust value $\min_{\pi\in\Pi}\Phi_\rho(\pi)=0$ than DRPG algorithm (Wang et al., 2023) and actor-critic algorithm (Li et al., 2023b).

O.2. Experiments on Large State Space

We test Algorithm 3 on the Garnet problem (Archibald et al., 1995; Wang and Zou, 2022) with spaces $\mathcal{S}=\{0,1,\dots,49\}$ of 50 states and $\mathcal{A}=\{0,1,2\}$ of 3 actions. The agent gets cost 0 if it takes action 0 at state 0 or action 1 at other states, and gets cost 1 otherwise. We use transition kernel parameterization $p_\xi(s'|s,a)=\psi(a,s')\xi(s)$ with parameter $\xi(s)\in\mathbb{R}^{d_p}$ ($d_p=10$) and randomly generated feature vectors $\psi(a,s')\in\mathbb{R}^{d_p}$. This parameterization is both s-rectangular and a special case of the linear kernel parameterization $p_{\widetilde{\xi}}(s'|s,a)=\widetilde{\psi}(s,a,s')\widetilde{\xi}$ introduced in Section 4.1 with parameter $\widetilde{\xi}=[\xi(1),\xi(2),\dots,\xi(|\mathcal{S}|)]\in\mathbb{R}^{d_p|\mathcal{S}|}$ and the following feature vector

$$\widetilde{\psi}(s,a,s')=[\underbrace{0,\dots,0}_{(s-1)d_p\text{ elements}},\underbrace{\psi(a,s')}_{d_p\text{ elements}},\underbrace{0,\dots,0}_{(|\mathcal{S}|-s)d_p\text{ elements}}]\in\mathbb{R}^{d_p|\mathcal{S}|}.$$

We first generate $\psi_{\text{pre}}^{(j)}(a,s')\in\mathbb{R}$ from uniform distribution $U(1,2)$ for all the entries $j=1,\dots,d_p$ and for all a,s'. Then we obtain $\psi(a,s')=[\psi^{(1)}(a,s'),\dots,\psi^{(d_p)}(a,s')]\in\mathbb{R}^{d_p}$ by normalization as follows.

$$\psi^{(j)}(a,s')=\frac{\psi_{\text{pre}}^{(j)}(a,s')}{\sum_{s''}\psi_{\text{pre}}^{(j)}(a,s'')}.$$

In this way, we obtain $\psi(a,s')=[\psi^{(1)}(a,s'),\dots,\psi^{(d_p)}(a,s')]\in\mathbb{R}^{d_p}$ where $p_\xi(s'|s,a)=\psi(a,s')\xi(s)$ is a distribution of $s'\in\mathcal{S}$ for any $\xi(s)\in\mathbb{R}_+^{d_p}$ satisfying $\|\xi(s)\|_1=1$. We use s-rectangular L_2-norm ambiguity set $\Xi:=\{[\xi(1),\xi(2),\dots,\xi(|\mathcal{S}|)]\in\mathbb{R}^{d_p|\mathcal{S}|}:\|\xi(s)-\xi_0(s)\|\le 0.03,\forall s\in\mathcal{S}\}$ where $\xi_0(s)=[0.1,0.1,\dots,0.1]\in\mathbb{R}^{d_p}$ is the nominal kernel parameter. We also adopt the linear Q function approximation $Q_\tau(\pi,p;s,a)\approx\phi(s,a)^\top w$ with parameter $w\in\mathbb{R}^d$ as well as feature vectors $\phi(s,a)\in\mathbb{R}^d$ generated entrywise from uniform distribution $U(0,1)$. The initial state distribution ρ is uniform with $\rho(s)\equiv\frac{1}{50},\forall s\in\mathcal{S}$. The discount factor is $\gamma=0.95$.

In the above robust MDP setting with varying $d\in\{5,20,50,100,130,140,150\}$, we implement Algorithm 3 with $\tau=0.1$, $T=10$, $T'=20$, $T_1=10^5$, $\eta=1$, $\beta=0.001$, $\alpha=0.001$, $N=1$, $H=500$, $N=10^4$. The learning curves of the objective function $\Phi_\rho(\pi_t):=\max_{p\in\mathcal{P}}J_\rho(\pi_t,p)$ for each d is plotted on the right of Figure 1, which shows that our Algorithm 3 converges to the optimal robust value $\min_{\pi\in\Pi}\Phi_\rho(\pi)=0$ with sufficiently large d, and the convergence gap gets larger with smaller d, due to larger transition kernel parameterization error. Hence, a proper value of d is important to trade off between performance and the amount of computation.