ACHIEVE PERFORMATIVELY OPTIMAL POLICY FOR PERFORMATIVE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Performative reinforcement learning is an emerging dynamical decision making framework, which extends reinforcement learning to the common applications where the agent's policy can change the environmental dynamics. Existing works on performative reinforcement learning only aim at a performatively stable (PS) policy that maximizes an approximate value function. However, there is a provably positive constant gap between the PS policy and the desired performatively optimal (PO) policy that maximizes the original value function. In contrast, this work proposes a zeroth-order Frank-Wolfe algorithm (0-FW) algorithm with a zeroth-order approximation of the performative policy gradient in the Frank-Wolfe framework, and obtains **the first polynomial-time convergence to the desired PO** policy under the standard regularizer dominance condition. For the convergence analysis, we prove two important properties of the nonconvex value function. First, when the policy regularizer dominates the environmental shift, the value function satisfies a certain gradient dominance property, so that any stationary point (not PS) of the value function is a desired PO. Second, though the value function has unbounded gradient, we prove that all the sufficiently stationary points lie in a convex and compact policy subspace Π_Δ, where the policy value has a constant lower bound $\Delta > 0$ and thus the gradient becomes bounded and Lipschitz continuous. Experimental results also demonstrate that our 0-FW algorithm is more effective than the existing algorithms in finding the desired PO policy.

1 INTRODUCTION

Reinforcement learning is a useful dynamic decision making framework with many successes in AI, such as AlphaGo (Silver et al., 2017), AlphaStar (Vinyals et al., 2019), Pluribus (Brown & Sandholm, 2019), large language model alignment (Bai et al., 2022) and reasoning (Havrilla et al., 2024). However, most reinforcement learning works ignore the effect of the deployed policy on the environmental dynamics, including transition kernel and reward function. This effect is significant in multi-agent systems, particularly the Stackelberg game, where leaders' policy change triggers the followers' policy change, which in turn affects the environmental dynamics faced by the leader (Mandal et al., 2023). For example, a recommender system (leader) affects the users' (followers) demographics and their interaction strategy with the system (Chaney et al., 2018; Mansoury et al., 2020). Autonomous vehicles (leaders) affect the strategies of the pedestrians and the other vehicles (followers) (Nikolaidis et al., 2017).

To account for such effect of deployed policy on environmental dynamics, performative reinforcement learning has been proposed by (Mandal et al., 2023) where the transition kernel p_π and reward function r_π are modeled as functions of the deployed policy π. The ultimate goal is to find the *performatively optimal (PO)* policy that maximizes the *performative value function*, defined as the accumulated discounted reward when deploying a policy π to its corresponding environment (p_π, r_π). However, the policy-dependent environmental dynamics pose significant challenges to achieve PO. Hence, (Mandal et al., 2023) pursues a suboptimal *performatively stable (PS)* policy using repeated retraining method with environmental dynamics fixed for the current policy at each policy optimization step. However, (Mandal et al., 2023) shows that PS can have a positive constant distance to PO.

Extensions of the basic performative reinforcement learning problem (Mandal et al., 2023) have been proposed and all of them focus on the suboptimal PS policy. For example, Rank et al. (2024) allows the environmental dynamics to gradually adjust to the currently deployed policy, and proposes a mixed delayed repeated retraining algorithm with accelerated convergence to a PS policy. Mandal & Radanovic (2024) extends (Mandal et al., 2023) from tabular setting to linear Markov decision processes with large number of states, and also obtains the convergence rate of the repeated retraining algorithm to a PS policy. Pollatos et al. (2025) obtains a PS policy that is robust to data contamination. Sahitaj et al. (2025) obtains a performatively stable equilibrium as an extension of PS policy to performative Markov potential games with multiple competitive agents.

In sum, all these existing performative reinforcement learning works pursue a suboptimal PS policy by repeated retraining algorithms. Therefore, we want to ask the following basic research question:

***Q:** Is there an algorithm that converges to the desired performatively optimal (PO) policy?*

1.1 Our Contributions

We will answer affirmatively to the research question above in the following steps. Each step yields a novel contribution.

- We study an entropy regularized performative reinforcement learning problem, compatible with the basic performative reinforcement learning problem in (Mandal et al., 2023). We prove that the objective function satisfies a certain gradient dominance condition, which implies that an approximate stationary point (not the suboptimal PS) is the desired approximate PO policy, under a standard regularizer dominance condition similar to that used by (Mandal et al., 2023; Rank et al., 2024; Mandal & Radanovic, 2024; Pollatos et al., 2025) to ensure convergence to a suboptimal PS policy. The proof adopts novel techniques such as recursion for p_π-related error term and frequent switch among various necessary and sufficient conditions of smoothness and strong concavity like properties for various variables (see Section 3.2).

- We obtain a policy lower bound as a decreasing function of a stationary measure. This bound not only implies the unbounded *performative policy gradient* (a challenge to find a stationary policy and thus PO), but also inspires us to find a stationary policy in the policy subspace Π_Δ with a constant policy lower bound $\Delta > 0$ where we prove the objective function to be Lipschitz continuous and Lipschitz smooth (a solution to this challenge). The lower bound Δ is obtained using a novel technique which simplifies a complicated inequality of the minimum policy value $\pi[a_{\min}(s)|s]$ in two cases (see Section 3.3).

- We construct a zeroth-order estimation of the *performative policy gradient* and obtains its estimation error. This is more challenging than the existing zeroth-order estimation methods since our objective function is only well-defined on the policy space, a compact subset of a linear subspace of the Euclidean space $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. To solve this puzzle, we adjust a two-point estimation to the linear subspace $\mathcal{L}_0$ of policy difference, and simplify the estimation error analysis by mapping policies onto the Euclidean space $\mathbb{R}^{|\mathcal{S}|(|\mathcal{A}|-1)}$ via orthogonal transformation (see Section 4.1).

- We propose a zeroth-order Frank-Wolfe (0-FW) algorithm (see Algorithm 1) by combining the *performative policy gradient* estimation above with the Frank-Wolfe algorithm. Then we obtain a polynomial computation complexity of our 0-FW algorithm to converge to a stationary policy, which is also the desired PO policy under the regularizer dominance condition above. The convergence analysis uses a policy averaging technique to show that an approximate stationary policy on Π_Δ is also approximately stationary on the whole policy space Π (see Section 4.2).

Finally, we briefly show that the results above, including gradient dominance, Lipschitz properties and the finite-time convergence of 0-FW algorithm to the desired PO, can be adjusted to the performative reinforcement learning problem with the quadratic regularizer used by (Mandal et al., 2023; Rank et al., 2024; Pollatos et al., 2025) (see Appendix M).

2 Preliminary: Performative Reinforcement Learning

2.1 Problem Formulation

Performative reinforcement learning is characterized by a Markov decision process (MDP) $\mathcal{M}_\pi = (\mathcal{S}, \mathcal{A}, p_\pi, r_\pi, \rho)$ that depends on a certain policy π. Here, $\mathcal{S}$ and $\mathcal{A}$ denote the finite state and action spaces respectively. The policy $\pi \in [0,1]^{|\mathcal{S}||\mathcal{A}|}$, transition kernel $p_\pi \in [0,1]^{|\mathcal{S}|^2|\mathcal{A}|}$, reward $r_\pi \in [0,1]^{|\mathcal{S}||\mathcal{A}|}$, and initial state distribution $\rho \in [0,1]^{|\mathcal{S}|}$ are vectors that represent distributions. Specifically, the policy $\pi \in [0,1]^{|\mathcal{S}||\mathcal{A}|}$, with entries $\pi(a|s)$ for any state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, lies in the policy space $\Pi \overset{\text{def}}{=} \big\{\pi \in [0,1]^{|\mathcal{S}|^2|\mathcal{A}|} : \sum_{a\in\mathcal{A}} \pi(a|s) = 1, \forall s \in \mathcal{S}\big\}$, such that $\pi(\cdot|s)$ for any state s can be seen as a distribution over $\mathcal{A}$. The transition kernel $p_\pi \in [0,1]^{|\mathcal{S}|^2|\mathcal{A}|}$ dependent on policy $\pi \in \Pi$, with entries $p_\pi(s'|s,a)$ for any $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$, lies in the transition kernel space $\mathcal{P} \overset{\text{def}}{=} \big\{p \in [0,1]^{|\mathcal{S}|^2|\mathcal{A}|} : \sum_{s'\in\mathcal{S}} p(s'|s,a) = 1, \forall s \in \mathcal{S}, a \in \mathcal{A}\big\}$ such that $p_\pi(\cdot|s,a)$ can be seen as a state distribution on $\mathcal{S}$. $r_\pi \in \mathcal{R} \overset{\text{def}}{=} [0,1]^{|\mathcal{S}||\mathcal{A}|}$ is the reward function with entries $r_\pi(s,a) \in [0,1]$ for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$. $\rho \in [0,1]^{|\mathcal{S}|}$ is the initial state distribution such that $\sum_{s\in\mathcal{S}} \rho(s) = 1$. Note that we consider p_π, r_π, ρ, π as Euclidean vectors, so that we can conveniently define their Euclidean norm. For example, we define $\|p_\pi\|_q = \big[\sum_{s,a,s'} |p_\pi(s'|s,a)|^q\big]^{1/q}$ for any $q > 1$ and $\|p_\pi\|_\infty = \max_{s,a,s'} |p_\pi(s'|s,a)|$. Such norms can be similarly defined over r_π, ρ, π by summing or maximizing over all the entries. Specifically, denote $\|\cdot\| = \|\cdot\|_2$ by convention.

When an agent applies its policy $\pi \in \Pi$ to MDP $\mathcal{M}_{\pi'} = (\mathcal{S}, \mathcal{A}, p_{\pi'}, r_{\pi'}, \rho)$, the initial environmental state $s_0 \in \mathcal{S}$ is generated from the distribution ρ. Then at each time $t = 0, 1, 2, \ldots$, the agent takes a random action $a_t \sim \pi(\cdot|s_t)$ based on the current state $s_t \in \mathcal{S}$, the environment transitions to the next state $s_{t+1} \sim p_{\pi'}(\cdot|s_t, a_t)$ and provides reward $r_t = r_{\pi'}(s_t, a_t) \in [0,1]$ to the agent. The value of applying policy π to $\mathcal{M}_{\pi'}$ can be characterized by the following *value function*:

$$V^{\pi}_{\lambda,\pi'} \overset{\text{def}}{=} \mathbb{E}_{\pi,p_{\pi'},\rho}\Big[\sum_{t=0}^{\infty} \gamma^t r_{\pi'}(s_t, a_t)\Big] - \lambda \mathcal{H}_{\pi'}(\pi). \tag{1}$$

Here, $\mathbb{E}_{\pi,p_{\pi'},\rho}$ is the expectation under policy π, transition kernel $p_{\pi'}$ and initial state distribution ρ. $\gamma \in (0,1)$ is the discount factor. $\mathcal{H}_{\pi'}(\pi)$ is a regularizer with coefficient $\lambda \geq 0$ to ensure or accelerate algorithm convergence. Existing works use the quadratic regularizers such as $\mathcal{H}_{\pi'}(\pi) = \frac{1}{2}\|d_{\pi,p_{\pi'}}\|^2$ (Mandal et al., 2023; Rank et al., 2024; Pollatos et al., 2025) and $\mathcal{H}_{\pi'}(\pi) = \frac{1}{2}\|\Phi^\top d_{\pi,p_{\pi'}}\|^2$ (Mandal & Radanovic, 2024) with a feature matrix Φ, where the occupancy measure $d_{\pi,p} \in [0,1]^{|\mathcal{S}||\mathcal{A}|}$ for any policy π and transition kernel p is defined as the following distribution on $\mathcal{S} \times \mathcal{A}$.

$$d_{\pi,p}(s,a) \overset{\text{def}}{=} (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{\pi,p,\rho}\{s_t = s, a_t = a\}, \tag{2}$$

Then the state occupancy measure defined as $d_{\pi,p}(s) \overset{\text{def}}{=} \sum_a d_{\pi,p}(s,a)$ satisfies the following well-known Bellman equation for any state $s' \in \mathcal{S}$.

$$d_{\pi,p}(s') = (1-\gamma)\rho(s') + \gamma \sum_{s,a} d_{\pi,p}(s)\pi(a|s)p(s'|s,a). \tag{3}$$

The goal of performative reinforcement learning is to find the *performatively optimal (PO)* policy π that maximizes the *performative value function* $V^{\pi}_{\lambda,\pi}$ (with $\pi' = \pi$ in Eq. (1)), as defined below.

Definition 1 (Ultimate Goal: PO). *For any $\epsilon \geq 0$, a policy $\pi \in \Pi$ is defined as ϵ-performatively optimal (ϵ-PO) if $\max_{\pi'\in\Pi} V^{\pi'}_{\lambda,\pi'} - V^{\pi}_{\lambda,\pi} \leq \epsilon$. Specifically, we call a 0-PO policy as a PO policy.*

Conventional reinforcement learning can be seen as a special case of performative reinforcement learning with fixed environmental dynamics, namely, fixed transition kernel $p_\pi \equiv p$ and fixed reward function $r_\pi \equiv r$. However, this may fail on applications with policy-dependent environmental dynamics, such as recommender system and autonomous driving as explained in Section 1.

2.2 Existing Repeated Retraining Methods for Performatively Stable (PS) Policy

Achieving an ϵ-PO policy (defined by Definition 1) is challenging, due to the policy-dependent environmental dynamics p_π and r_π. To alleviate the challenge, all the existing works (Mandal et al., 2023; Rank et al., 2024; Mandal & Radanovic, 2024; Pollatos et al., 2025; Sahitaj et al., 2025) aim at a *performatively stable (PS)* policy π_{PS} defined as follows, as an approximation to a *PO policy*.

$$\pi_{\text{PS}} \in \arg\max_{\pi\in\Pi} V^{\pi}_{\lambda,\pi_{\text{PS}}}. \tag{4}$$

In other words, a PS policy $\pi_{\rm PS}$ has the optimal value on the fixed environment $\mathcal{M}_{\pi_{\rm PS}}$. However, Mandal et al. (2023) shows that a PS policy can be suboptimal.

Nevertheless, we will briefly introduce the suboptimal repeated retraining algorithms in their works, to later partially inspire our method that converges to the global optimal PO policy. All these repeated retraining algorithms share the fundamental idea that in each iteration t, the next policy $\pi_{t+1}\approx\arg\max_{\pi\in\Pi}V_{\lambda,\pi_t}^{\pi}$ is obtained by solving the conventional reinforcement learning problem under fixed dynamics p_{π_t} and r_{π_t}. This strategy highly relies on conventional reinforcement learning but fail to make full use of the policy-dependent dynamics, which leads to the suboptimal PS policy. Next, we will propose our significantly different strategies to achieve the desired PO policy.

3 Entropy Regularized Performative Reinforcement Learning

In this section, we obtain critical properties of an entropy regularized performative reinforcement learning problem for achieving the desired PO policy.

3.1 Negative Entropy Regularizer

We consider the following negative entropy regularizer of the policy π, which is widely used in reinforcement learning to encourage environment exploration and accelerate convergence (Mnih et al., 2016; Mankowitz et al., 2019; Cen et al., 2022; Chen & Huang, 2024).

$$\mathcal{H}_{\pi'}(\pi)=\mathbb{E}_{\pi,p_{\pi'},\rho}\Big[\sum_{t=0}^{\infty}\gamma^t\log\pi(a_t|s_t)\Big]. \tag{5}$$

In addition, this negative entropy regularizer can be seen as a strongly convex function of the occupancy measure $d_{\pi,p_{\pi'}}$ (proved in Appendix D), which is critical to develop algorithms convergent to a PO (see Theorem 1 later) or PS policy (Mandal et al., 2023). For optimization problem on a probability simplex variable (policy π or occupancy measure d), negative entropy regularizer is more natural and yields faster theoretical convergence than the quadratic regularizers used in the existing performative reinforcment learning works (Mandal et al., 2023; Rank et al., 2024; Pollatos et al., 2025) (see pages 43-45 of (Chen, 2020) for explanation).

Therefore, we will mainly focus on the following entropy-regularized value function, which is obtained by substituting the negative entropy regularizer (5) into the general value function (1).

$$V_{\lambda,\pi'}^{\pi}\overset{\rm def}{=}\mathbb{E}_{\pi,p_{\pi'},\rho}\Big[\sum_{t=0}^{\infty}\gamma^t[r_{\pi'}(s_t,a_t)-\lambda\log\pi(a_t|s_t)]\Big]. \tag{6}$$

Specifically, we will study the critical properties of the entropy-regularized value function (6) (Section 4) to develop algorithm that converges to PO (Sections 4.1-4.2). Then we will briefly discuss about how to adjust these results to the existing quadratic regularizers (Appendix M).

We make the following standard assumptions to study the properties of the value function (6).

Assumption 1 (Sensitivity). *There exist constants $\epsilon_p,\epsilon_r>0$ such that for any $\pi,\pi'\in\Pi$,*

$$\|p_{\pi'}-p_\pi\|\le\epsilon_p\|\pi'-\pi\|,\quad \|r_{\pi'}-r_\pi\|\le\epsilon_r\|\pi'-\pi\| \tag{7}$$

Assumption 2 (Smoothness). *p_π and r_π are Lipschitz smooth with modulus $S_p,S_r>0$ respectively, that is, for any $\pi\in\Pi$, $s,s'\in\mathcal{S}$, $a\in\mathcal{A}$, we have*

$$\|\nabla_\pi p_{\pi'}(s'|s,a)-\nabla_\pi p_\pi(s'|s,a)\|\le S_p\|\pi'-\pi\|, \tag{8}$$

$$\|\nabla_\pi r_{\pi'}(s,a)-\nabla_\pi r_\pi(s,a)\|\le S_r\|\pi'-\pi\|. \tag{9}$$

Assumption 3. *There exists a constant $D>0$ such that $\inf_{\pi\in\Pi,p\in\mathcal{P},s\in\mathcal{S}}d_{\pi,p}(s)\ge D$.*

Assumptions 1-2 ensure that the environmental dynamics p_π and r_π adjust continuously and smoothly to policy π, and thus the *performative value function* $V_{\lambda,\pi}^{\pi}$ is differentiable with *performative policy gradient* $\nabla_\pi V_{\lambda,\pi}^{\pi}$. Similar versions of Assumption 1 on environmental sensitivity have also been used for performative reinforcement learning (Mandal et al., 2023; Rank et al., 2024; Mandal & Radanovic, 2024; Pollatos et al., 2025; Sahitaj et al., 2025). Assumption 3 has been used (Zhang et al., 2021;

Sahitaj et al., 2025) or implied by stronger assumptions (Wei et al., 2021; Chen et al., 2022; Agarwal et al., 2021; Leonardos et al., 2022; Wang et al., 2023; Chen & Huang, 2024; Bhandari & Russo, 2024) in conventional reinforcement learning (see Appendix E for the proof), which guarantees that each state is visited sufficiently often.

3.2 GRADIENT DOMINANCE

For the nonconvex policy optimization problem $\max_{\pi\in\Pi} V_{\lambda,\pi}^{\pi}$ in Eq. (6) on the convex policy space Π, it is natural to consider its approximate stationary solution as defined below.

Definition 2 (Stationary Policy). *For any $\epsilon \geq 0$, a policy $\pi \in \Pi$ is ϵ-stationary if $\max_{\tilde{\pi}\in\Pi} \langle \nabla_{\pi} V_{\lambda,\pi}^{\pi}, \tilde{\pi}-\pi \rangle \leq \epsilon$. We call a 0-stationary policy as a stationary policy.*

Note that for a policy to be the desired PO, it is necessary to be stationary, while the PS policy targeted by existing works is neither necessary nor sufficient. Furthermore, we will show that stationary policy can also be a sufficient condition of the desired PO under mild conditions. As a preliminary step, we show the important gradient dominance property of the objective function as follows.

Theorem 1 (Gradient Dominance). *Under Assumptions 1-3, the entropy regularized value function (6) satisfies the following gradient dominance property for any $\pi_0, \pi_1 \in \Pi$.*

$$V_{\lambda,\pi_1}^{\pi_1} \leq V_{\lambda,\pi_0}^{\pi_0} + D^{-1} \max_{\pi\in\Pi} \langle \nabla_{\pi_0} V_{\lambda,\pi_0}^{\pi_0}, \pi-\pi_0 \rangle - \frac{\mu}{2}\|\pi_1-\pi_0\|^2, \tag{10}$$

where

$$\mu \overset{\text{def}}{=} \frac{D\lambda}{1-\gamma} - \frac{6\gamma|\mathcal{S}|(1+\lambda\log|\mathcal{A}|)}{D(1-\gamma)^3}\big[\epsilon_p\big(\sqrt{|\mathcal{A}|}+\gamma\epsilon_p\sqrt{|\mathcal{S}|}\big)+S_p(1-\gamma)\big] - \frac{S_r(1-\gamma)+4\epsilon_r(\sqrt{|\mathcal{A}|}+\epsilon_p\sqrt{|\mathcal{S}|})}{D^2(1-\gamma)^2}, \tag{11}$$

The gradient dominance property above generalizes that used in the conventional unregularized reinforcement learning (see Lemma 4 of (Agarwal et al., 2021)), which implies that stationary policy is close to a PO policy as shown in the corollary below.

Corollary 1. *Under Assumptions 1-3, any $D\epsilon$-stationary policy is an $(\epsilon+|\mu||\mathcal{S}|)$-PO policy. Furthermore, this is also the desired ϵ-PO policy if $\mu \geq 0$. The PO policy is unique if $\mu > 0$.*

Remark: Corollary 1 implies that a $D\epsilon$-stationary policy is always $(\epsilon+|\mu||\mathcal{S}|)$-close to the desired PO policy with $|\mu|$ proportional to the environmental sensitivity $\mathcal{O}(\epsilon_p+\epsilon_r+S_p+S_r)$. Furthermore, since $\mu = [\mathcal{O}(1)-\mathcal{O}(\epsilon_p+S_p)]\lambda - \mathcal{O}(\epsilon_p+\epsilon_r+S_p+S_r)$ by Eq. (11), when $\mathcal{O}(\epsilon_p+S_p) < \mathcal{O}(1)$ and the regularizer strength dominates the environmental shift ($\lambda \geq \frac{\mathcal{O}(\epsilon_p+\epsilon_r+S_p+S_r)}{\mathcal{O}(1)-\mathcal{O}(\epsilon_p+S_p)}$), we have $\mu \geq 0$ so that the $D\epsilon$-stationary policy is also the desired ϵ-PO policy. Note that similar regularizer dominance condition has also been used to guarantee convergence to a suboptimal PS policy (Mandal et al., 2023; Rank et al., 2024; Mandal & Radanovic, 2024; Pollatos et al., 2025).

3.3 POLICY LOWER BOUND AND LIPSCHITZ PROPERTIES

Policy Lower Bound: Based on Section 3.2, we can focus on achieving an ϵ-stationary policy. A major challenge is the unbounded *performative policy gradient* $\nabla_{\pi} V_{\lambda,\pi}^{\pi}$ on Π. Specifically, we will show that as $\pi(a|s)\to 0$ for any state s and action a, $\|\nabla_{\pi} V_{\lambda,\pi}^{\pi}\| \to +\infty$. To tackle this challenge, we prove the following policy lower bound.

Theorem 2. *If Assumptions 1 and 3 hold, and p_π, r_π are differentiable functions of π, then there exists a constant $\pi_{\min} > 0$ (see its value in Eq. (96) in Appendix H) such that the following policy lower bound holds for any $\pi \in \Pi$, $s \in \mathcal{S}$, $a \in \mathcal{A}$.*

$$\pi(a|s) \geq \pi_{\min} \exp\Big[-\frac{2|\mathcal{A}|}{\lambda}(1-\gamma)\langle \nabla_{\pi} V_{\lambda,\pi}^{\pi}, \pi'-\pi \rangle\Big], \tag{12}$$

Here, the policy π' is defined as follows depending on π:

$$\pi'(a|s) = \begin{cases} \pi[a_{\min}(s)|s], & a = a_{\max}(s) \\ \pi[a_{\max}(s)|s], & a = a_{\min}(s) \\ \pi(a|s), & \text{Otherwise} \end{cases}, \tag{13}$$

where $a_{\max}(s) \in \arg\max_a \pi(a|s)$ *and* $a_{\min}(s) \in \arg\min_a \pi(a|s)$.

Implications of Theorem 2: First, as $\pi(a|s) \to 0$, we have $\langle \nabla_\pi V_{\lambda,\pi}^{\pi}, \pi' - \pi \rangle \to +\infty$, so $\|\nabla_\pi V_{\lambda,\pi}^{\pi}\| \to +\infty$ as aforementioned. Second, any stationary policy π satisfies $\langle \nabla_\pi V_{\lambda,\pi}^{\pi}, \pi' - \pi \rangle \le 0$, so $\pi(a|s) \ge \pi_{\min}$. Therefore, we can search ϵ-stationary policy on the convex and compact policy subspace $\Pi_\Delta \overset{\text{def}}{=} \{\pi \in \Pi : \pi(a|s) \ge \Delta\}$ with lower bound $\Delta \in (0, \pi_{\min}]$.

Lipschitz Properties: Theorem 2 inspires us to find an ϵ-stationary policy in the policy subspace Π_Δ, where the *performative value function* $V_{\lambda,\pi}^{\pi}$ is Lipschitz continuous and Lipschitz smooth as follows.

Theorem 3. *Under Assumptions 1-2, there exist constants* $L_\lambda, \ell_\lambda > 0$ *(see the values in Eqs. (98) and (100) in Appendix I) such that the following Lipschitz propreties hold for any* $\Delta > 0$ *and* $\pi, \pi' \in \Pi_\Delta$.

$$|V_{\lambda,\pi'}^{\pi'} - V_{\lambda,\pi}^{\pi}| \le \frac{L_\lambda}{\Delta}\|\pi' - \pi\|, \qquad \|\nabla_{\pi'} V_{\lambda,\pi'}^{\pi'} - \nabla_\pi V_{\lambda,\pi}^{\pi}\| \le \frac{\ell_\lambda}{\Delta}\|\pi' - \pi\|. \tag{14}$$

4 Zeroth-Order Frank-Wolfe (0-FW) Algorithm

4.1 Performative Policy Gradient Estimation

In Section 3, we have obtained important properties of the entropy regularized *performative value function* $V_{\lambda,\pi}^{\pi}$ (defined by Eq. (6)), which indicates that it suffices to find an ϵ-stationary policy in the subspace Π_Δ for $\Delta \in (0, \pi_{\min}]$. To achieve this goal, an accurate estimation of the *performative policy gradient* $\nabla_\pi V_{\lambda,\pi}^{\pi}$ is important but also challenging, since the performative policy gradient involves the unknown gradients $\nabla_\pi p_\pi(s'|s,a)$ and $\nabla_\pi r_\pi(s,a)$.

Despite these challenges in estimating $\nabla_\pi V_{\lambda,\pi}^{\pi}$, note that $V_{\lambda,\pi}^{\pi}$ for any policy π can be evaluated by policy evaluation in conventional reinforcement learning under fixed environment p_π and r_π (for fixed π). Furthermore, for any $\epsilon_V > 0$ and $\eta \in (0,1)$, many existing policy evaluation algorithms such as temporal difference (Bhandari et al., 2018; Li et al., 2023; Samsonov et al., 2023), can obtain $\hat{V}_{\lambda,\pi}^{\pi} \approx V_{\lambda,\pi}^{\pi}$ with small error bound $|\hat{V}_{\lambda,\pi}^{\pi} - V_{\lambda,\pi}^{\pi}| \le \epsilon_V$ with probability at least $1 - \eta$.

As a result, we will consider a zeroth-order estimation of $\nabla_\pi V_{\lambda,\pi}^{\pi}$ using policy evaluation. However, this has another challenge that $V_{\lambda,\pi}^{\pi}$ is only well-defined on $\pi \in \Pi$, so we cannot directly apply the existing zeroth-order estimation methods (Agarwal et al., 2010; Shamir, 2017; Malik et al., 2020) which require the objective function to be well-defined on a sphere. Fortunately, for any $\pi, \pi' \in \Pi$, the policy difference $\pi' - \pi$ lies in the following linear subspace of dimensionality $|\mathcal{S}|(|\mathcal{A}| - 1)$.

$$\mathcal{L}_0 \overset{\text{def}}{=} \Big\{u \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} : \sum_a u(a|s) = 0, \forall s \in \mathcal{S}\Big\}. \tag{15}$$

Therefore, inspired by the popular two-point zeroth-order estimations, we estimate $\nabla_\pi V_{\lambda,\pi}^{\pi}$ as follows.

$$\hat{g}_{\lambda,\delta}(\pi) = \frac{|\mathcal{S}|(|\mathcal{A}|-1)}{2N\delta} \sum_{i=1}^{N} \big(\hat{V}_{\lambda,\pi+\delta u_i}^{\pi+\delta u_i} - \hat{V}_{\lambda,\pi-\delta u_i}^{\pi-\delta u_i}\big) u_i, \tag{16}$$

where $\{u_i\}_{i=1}^N$ are i.i.d. samples uniformly from $U_1 \cap \mathcal{L}_0$ with $U_1 \overset{\text{def}}{=} \{u \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} : \|u\| = 1\}$. Our estimation (16) above is more tricky than the existing two-point zeroth-order estimations (Agarwal et al., 2010; Shamir, 2017; Malik et al., 2020) where u_i is uniformly distributed on U_1. To elaborate, we replace their U_1 with $U_1 \cap \mathcal{L}_0$, a unit sphere on the linear subspace $\mathcal{L}_0$, and further require $\pi \in \Pi_\Delta$ and $\delta < \Delta$, to guarantee that $\pi + \delta u_i, \pi - \delta u_i \in \Pi$ for any $u_i \in U_1 \cap \mathcal{L}_0$ and thus the gradient estimation (16) is well-defined (see Appendix J for the proof). Moreover, we use the following three steps to obtain u_i uniformly from $U_1 \cap \mathcal{L}_0$: (1) Obtain v_i uniformly from U_1; (2) Project v_i onto $\mathcal{L}_0$ as Eq. (17) below; (3) Normalize this projection by $u_i = \text{proj}_{\mathcal{L}_0}(v_i)/\|\text{proj}_{\mathcal{L}_0}(v_i)\|$.

$$\text{proj}_{\mathcal{L}_0}(v_i)(a|s) = v_i(a|s) - \frac{1}{|\mathcal{A}|}\sum_{a'} v_i(a'|s). \tag{17}$$

The gradient estimation (16) has the following provable error bound.

Proposition 1. *For any $\Delta > \delta > 0$, $\eta \in (0,1)$ and $\pi \in \Pi_\Delta$, the stochastic gradient (16) is well-defined (i.e., $\pi + \delta u_i$ and $\pi - \delta u_i$ therein are valid policies defined by Π) and approximates the projected performative policy gradient $\mathrm{proj}_{\mathcal{L}_0}(\nabla_\pi V_{\lambda,\pi}^{\pi})$ with the following error bound (see its full expression in Eq. (110) in Appendix J), with probability at least $1-\eta$.*

$$\|\hat{g}_{\lambda,\delta}(\pi) - \mathrm{proj}_{\mathcal{L}_0}(\nabla_\pi V_{\lambda,\pi}^{\pi})\| \le \mathcal{O}\big(\frac{\epsilon_V}{\delta} + \frac{\log(N/\eta)}{\sqrt{N}} + \delta\big). \tag{18}$$

Remark: Proposition 1 above aims to approximate $\mathrm{proj}_{\mathcal{L}_0}(\nabla_\pi V_{\lambda,\pi}^{\pi})$ instead of $\nabla_\pi V_{\lambda,\pi}^{\pi}$. This is sufficient to find an ϵ-stationary policy, because for any policies π, π', the stationarity measure only involves $\langle \nabla_\pi V_{\lambda,\pi}^{\pi}, \pi' - \pi\rangle = \langle \mathrm{proj}_{\mathcal{L}_0}(\nabla_\pi V_{\lambda,\pi}^{\pi}), \pi' - \pi\rangle$ as $\pi' - \pi \in \mathcal{L}_0$. Therefore, we only care about $\mathrm{proj}_{\mathcal{L}_0}(\nabla_\pi V_{\lambda,\pi}^{\pi})$. The estimation error (18) above can be arbitrarily small with sufficiently large batchsize N (to reduce the variance), small δ (to reduce the bias), and policy evaluation error $\epsilon_V \ll \delta$.

Algorithm 1 Zeroth-order Frank-Wolfe (0-FW) Algorithm

Inputs: $T, N, \Delta > \delta > 0, \epsilon_V \ge 0, \beta > 0$.
Initialize: policy $\pi_0 \in \Pi_\Delta$.
for Iterations $t = 0, 1, \ldots, T-1$ **do**
    Obtain i.i.d. vectors $\{v_i\}_{i=1}^N$ uniformly from the unit sphere $U_1 \overset{\mathrm{def}}{=} \{u \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} : \|u\| = 1\}$.
    Obtain $\{\mathrm{proj}_{\mathcal{L}_0}(v_i)\}_{i=1}^N$ from Eq. (17).
    Obtain $\{u_i\}_{i=1}^N$ where $u_i = \mathrm{proj}_{\mathcal{L}_0}(v_i)/\|\mathrm{proj}_{\mathcal{L}_0}(v_i)\|$.
    Obtain stochastic policy evaluation $\hat{V}_{\lambda,\pi}^{\pi} \approx V_{\lambda,\pi}^{\pi}$ which satisfies $|\hat{V}_{\lambda,\pi}^{\pi} - V_{\lambda,\pi}^{\pi}| \le \epsilon_V$ for $\pi \in \{\pi_t \pm \delta u_i\}_{i=1}^N$.
    Obtain stochastic performative policy gradient estimation $\hat{g}_{\lambda,\delta}(\pi_t)$ using Eq. (16).
    Obtain $\tilde{\pi}_t$ by Eq. (21).
    Update π_{t+1} by Eq. (20).
end for
Output: $\pi_{\widetilde{T}}$ where $\widetilde{T} \in \arg\min_{0 \le t \le T-1} \langle \hat{g}_{\lambda,\delta}(\pi_t), \tilde{\pi}_t - \pi_t\rangle$.

4.2 Zeroth-Order Frank-Wolfe (0-FW) Algorithm

With the estimated gradient $\hat{g}_{\lambda,\delta}(\pi_t)$ defined by Eq. (16), we consider the following Frank-Wolfe algorithm to find an ϵ-stationary policy.

$$\tilde{\pi}_t = \arg\max_{\pi \in \Pi_\Delta} \langle \pi, \hat{g}_{\lambda,\delta}(\pi_t)\rangle, \tag{19}$$

$$\pi_{t+1} = \pi_t + \beta(\tilde{\pi}_t - \pi_t). \tag{20}$$

Lemma 1. *The step (19) has the analytical solution below.*

$$\tilde{\pi}_t(a|s) = \begin{cases} \Delta; a \ne \tilde{a}_t(s) \\ 1 - \Delta(|\mathcal{A}| - 1); a = \tilde{a}_t(s) \end{cases}, \tag{21}$$

where $\tilde{a}_t(s) \in \arg\max_a \hat{g}_{\lambda,\delta}(\pi_t)(a|s)$.

See the proof of Lemma 1 in Section C.1. Then combining the *performative policy gradient* estimation (see Section 3.1) with the Frank-Wolfe algorithm, we propose our zeroth-order Frank-Wolfe (0-FW) algorithm (see Algorithm 1).

We obtain the following convergence result of Algorithm 1 in Theorem 4, the main theoretical result of this work, as follows.

Theorem 4. *Suppose Assumptions 1-3 hold. For any $\eta \in (0,1)$ and precision $0 < \epsilon \le \min\big[24\sqrt{2|\mathcal{S}|}\frac{\ell_\lambda}{D}, \frac{2\lambda}{5|\mathcal{A}|D^2(1-\gamma)}, \frac{288 L_\lambda |\mathcal{S}|^{1.5}|\mathcal{A}|}{D\pi_{\min}}\big]$, select the following hyperparameters for Algorithm 1: $\Delta = \frac{\pi_{\min}}{3}$, $\beta = \frac{D\pi_{\min}\epsilon}{36\ell_\lambda|\mathcal{S}|}$, $\delta = \mathcal{O}(\epsilon)$, $\epsilon_V = \mathcal{O}(\epsilon^2)$, $N = \mathcal{O}[\epsilon^{-2}\log(\eta^{-1}\epsilon^{-1})]$, and the number of iterations $T = \mathcal{O}(\epsilon^{-2})$ (see Eqs. (116)-(121) in Appendix L for detailed expression of these hyperparameters). Then with probability at least $1-\eta$, the output policy $\tilde{\pi}_{\widetilde{T}}$ of Algorithm 1 is a $D\epsilon$-stationary policy. Furthermore, if $\mu \ge 0$, $\tilde{\pi}_{\widetilde{T}}$ is also an ϵ-PO policy. The total number of policy evaluations is $2NT = \mathcal{O}[\epsilon^{-4}\log(\eta^{-1}\epsilon^{-1})]$.*

Comparison with Existing Works: Theorem 4 indicates that our 0-FW algorithm for the first time converges to the desire PO policy with arbitrarily small precision ϵ in polynomial computation complexity, under the regularizer dominance condition that $\mu \ge 0$. In contrast, existing works only converge to a suboptimal PS policy under a similar regularizer dominance condition (Mandal et al., 2023; Rank et al., 2024; Mandal & Radanovic, 2024; Pollatos et al., 2025). Our preferable convergence result is due to the main algorithmic difference that existing works use repeated retraining algorithms with iteration $\pi_{t+1} \approx \arg\max_{\pi\in\Pi} V_{\lambda,\pi}^{\pi_t}$ where the policy π is deployed in a fixed environment $\mathcal{M}_{\pi_t}$ with $\pi \neq \pi_t$, while our 0-FW algorithm evaluates $V_{\lambda,\pi}^{\pi}$ where π is always deployed at its corresponding environment $\mathcal{M}_{\pi}$.

Proposition 2. *If $\Delta \leq \pi_{\min}/3$ and a policy π satisfies $\max_{\tilde{\pi}\in\Pi_\Delta}\langle \nabla_\pi V_{\lambda,\pi}^{\pi}, \tilde{\pi}-\pi\rangle \leq \frac{D\lambda}{5|\mathcal{A}|(1-\gamma)}$, then the stationary measures on Π_Δ and Π bound each other as follows.*

$$\max_{\tilde{\pi}\in\Pi}\langle \nabla_\pi V_{\lambda,\pi}^{\pi}, \tilde{\pi}-\pi\rangle \leq 2\max_{\tilde{\pi}\in\Pi_\Delta}\langle \nabla_\pi V_{\lambda,\pi}^{\pi}, \tilde{\pi}-\pi\rangle \tag{22}$$

To prove Proposition 2, note that π' defined by Eq. (13) also belongs to Π_Δ, so Theorem 2 implies $\pi(a|s) \geq 2\Delta$. Then for any $\pi_2 \in \Pi$, we have $\frac{\pi_2+\pi}{2} \in \Pi_\Delta$ and thus

$$\max_{\pi_2\in\Pi}\langle \nabla_\pi V_{\lambda,\pi}^{\pi}, \pi_2-\pi\rangle = 2\max_{\pi_2\in\Pi}\Big\langle \nabla_\pi V_{\lambda,\pi}^{\pi}, \frac{\pi_2+\pi}{2}-\pi\Big\rangle \leq 2\max_{\tilde{\pi}\in\Pi_\Delta}\langle \nabla_\pi V_{\lambda,\pi}^{\pi}, \tilde{\pi}-\pi\rangle.$$

5 Proof Sketch and Novelty

Intuition and Novelty for Proving Theorem 1: Define the following more refined value function

$$J_\lambda(\pi,\pi',p,r) \overset{\text{def}}{=} \mathbb{E}_{\pi,p}\Big[\sum_{t=0}^{\infty}\gamma^t[r(s_t,a_t)-\lambda\log\pi'(a_t|s_t)]\Big|s_0\sim\rho\Big]. \tag{23}$$

To get the intuition, we will first prove the bound (10) in the special case with fixed $p_\pi \equiv p$ and $r_\pi \equiv r$. Then we allow non-constant p_π to inspect the perturbation on the bound (10), and finally see the effect of non-constant r_π on the bound (10).

(Step 1): For conventional reinforcement learning with fixed $p_\pi \equiv p$ and $r_\pi \equiv r$, denote $d_\alpha = \alpha d_{\pi_1,p} + (1-\alpha)d_{\pi_0,p}$ ($\alpha \in [0,1]$). Based on the Bellman equation (3), $d_\alpha = d_{\pi_\alpha,p}$ is the occupancy measure of the policy $\pi_\alpha(a|s) = \frac{d_\alpha(s,a)}{d_\alpha(s)}$. Therefore, $V_{\lambda,\pi_\alpha}^{\pi_\alpha}$ can be rewritten as $J_\lambda(\pi_\alpha,\pi_\alpha,p,r) = \sum_{s,a} d_\alpha(s,a)[r(s,a)-\lambda\log\pi_\alpha(a|s)]$, which has the following strong concavity like property by Pinsker's inequality.

$$\begin{aligned}
&J_\lambda(\pi_\alpha,\pi_\alpha,p,r) - \alpha J_\lambda(\pi_1,\pi_1,p,r) - (1-\alpha)J_\lambda(\pi_0,\pi_0,p,r)\\
=&\frac{1}{1-\gamma}\sum_s\big[\alpha d_1(s)\mathrm{KL}[\pi_1(\cdot|s)\|\pi_\alpha(a|s)] + (1-\alpha)d_0(s)\mathrm{KL}[\pi_0(\cdot|s)\|\pi_\alpha(a|s)]\big]\\
\geq&\frac{D\lambda\alpha(1-\alpha)}{2(1-\gamma)}\|\pi_1-\pi_0\|^2.
\end{aligned} \tag{24}$$

(Step 2): Consider a harder case with non-constant p_π and constant reward $r_\pi \equiv r$. Similarly, denote $d_\alpha = \alpha d_{\pi_1,p_{\pi_1}} + (1-\alpha)d_{\pi_0,p_{\pi_0}}$ and $\pi_\alpha(a|s) = \frac{d_\alpha(s,a)}{d_\alpha(s)}$. The non-constant p_π brings a major challenge that $d_\alpha = d_{\pi_\alpha,p_{\pi_\alpha}}$ required by Step 1 above no longer holds. To solve this challenge, we need to bound the error term $e_\alpha(s) = d_{\pi_\alpha,p_\alpha}(s) - d_\alpha(s)$ which we prove to satisfy the following novel recursion.

$$e_\alpha(s') = \gamma\sum_{s,a}\big[e_\alpha(s)\pi_\alpha(a|s)p_{\pi_\alpha}(s'|s,a) + h_\alpha(s,a,s')\big],$$

where $h_\alpha(s,a,s') = d_\alpha(s,a)p_{\pi_\alpha}(s'|s,a) - \alpha d_1(s,a)p_{\pi_1}(s'|s,a) - (1-\alpha)d_0(s,a)p_{\pi_0}(s'|s,a)$. Since $d_\alpha(s,a)p_{\pi_\alpha}(s'|s,a)$ is a Lipschitz smooth function of α, we can upper bound $|h_\alpha(s,a,s')|$ and substitute this bound to the recursion above, which yields the following novel error bound.

$$\sum_s|e_\alpha(s)| \leq \frac{3\gamma|\mathcal{S}|\alpha(1-\alpha)}{D(1-\gamma)^2}\|\pi_1-\pi_0\|^2\big[\epsilon_p\big(\sqrt{|\mathcal{A}|}+\gamma\epsilon_p\sqrt{|\mathcal{S}|}\big) + S_p(1-\gamma)\big],$$

The bound above reflects the effect of non-constant p_π, which perturbs the bound (24) into

$$J_\lambda(\pi_\alpha,\pi_\alpha,p_\alpha,r)-\alpha J_\lambda(\pi_1,\pi_1,p_1,r)-(1-\alpha)J_\lambda(\pi_0,\pi_0,p_0,r)\ge\frac{\alpha(1-\alpha)\mu_1}{2}\|\pi_1-\pi_0\|^2,\quad(25)$$

where $\mu_1\overset{\rm def}{=}\frac{D\lambda}{1-\gamma}-\frac{6\gamma|\mathcal{S}|(1+\lambda\log|\mathcal{A}|)}{D(1-\gamma)^3}\big[\epsilon_p\big(\sqrt{|\mathcal{A}|}+\gamma\epsilon_p\sqrt{|\mathcal{S}|}\big)+S_p(1-\gamma)\big]$ equals μ in Eq. (11) when $\epsilon_r=S_r=0$.

(Step 3): Now we consider performative reinforcement learning with non-constant p_π and r_π. The policy π_α and its occupancy measure d_α are the same as in Case II above. Then the function $w(\alpha)=\alpha J_\lambda(\pi_1,\pi_1,p_1,r_\alpha)+(1-\alpha)J_\lambda(\pi_0,\pi_0,p_0,r_\alpha)$ can be proved $\mu_2\|\pi_1-\pi_0\|^2$-Lipschitz smooth with parameter $\mu_2=\mu-\mu_1\ge 0$. Using $r=r_\alpha$ in Eq. (25), we obtain the following strong concavity like property with $\mu=\mu_1-\mu_2$.

$$\begin{aligned}&V_{\lambda,\pi_\alpha}^{\pi_\alpha}-\alpha V_{\lambda,\pi_1}^{\pi_1}-(1-\alpha)V_{\lambda,\pi_0}^{\pi_0}\\=&J_\lambda(\pi_\alpha,\pi_\alpha,p_\alpha,r_\alpha)-\alpha J_\lambda(\pi_1,\pi_1,p_1,r_1)-(1-\alpha)J_\lambda(\pi_0,\pi_0,p_0,r_0)\\\ge&\frac{\alpha(1-\alpha)\mu_1}{2}\|\pi_1-\pi_0\|^2+w(\alpha)-\alpha w(1)-(1-\alpha)w(0)\ge\frac{\alpha(1-\alpha)\mu}{2}\|\pi_1-\pi_0\|^2.\end{aligned}$$

Finally, the dominance property (10) follows from the inequality above as $\alpha\to+0$.

Intuition and Novelty for Proving Theorem 2: At first, consider conventional reinforcement learning with fixed environmental dynamics $p_\pi\equiv p$ and $r_\pi\equiv r$. In this case, $\nabla_\pi V_{\lambda,\pi}^{\pi}$ has analytical form (see Eq. (90)), so by direct computation we obtain the following inequality with constant $C=1+\frac{\gamma(1+\lambda\log|\mathcal{A}|)}{1-\gamma}$ (see Eq. (91) for detail)

$$\langle\nabla_\pi J_\lambda(\pi,\pi,p,r),\pi'-\pi\rangle\ge\frac{1}{1-\gamma}\max_s\Big\{\big(\pi[a_{\max}(s)|s]-\pi[a_{\min}(s)|s]\big)\Big[\lambda\log\frac{\pi[a_{\max}(s)|s]}{\pi[a_{\min}(s)|s]}-C\Big]\Big\}.$$

To obtain a lower bound of $\pi[a_{\min}(s)|s]$, we simplify the inequality above by considering two cases, $\pi[a_{\min}(s)|s]\ge\frac{1}{2}\pi[a_{\max}(s)|s]\ge\frac{1}{2|\mathcal{A}|}$ and $\pi[a_{\min}(s)|s]<\frac{1}{2}\pi[a_{\max}(s)|s]$. In the second case, we replace $\pi[a_{\max}(s)|s]$ and $\pi[a_{\max}(s)|s]-\pi[a_{\min}(s)|s]$ above with their lower bounds $\frac{1}{|\mathcal{A}|}$ and $\frac{1}{2|\mathcal{A}|}$ respectively. Then combining the two cases proves the lower bound (12) at the special case of $\epsilon_p=\epsilon_r=0$. Then we extend from conventional reinforcement learning to performative reinforcement learning which involves a gradient perturbation with magnitude of at most $\mathcal{O}(\epsilon_p+\epsilon_r)$ (see Eq. (94) for detail) based on the chain rule and leads to the lower bound (12) for any $\epsilon_p,\epsilon_r\ge 0$.

Intuition and Novelty for Proving Proposition 1: Unlike existing zeroth-order estimations on the whole Euclidean space, our estimation (16) is made on the policy space Π, which lies in the linear manifold $\mathcal{L}_0+|\mathcal{A}|^{-1}\subset\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. The key to our proof is to find an orthogonal transformation $T:\mathbb{R}^{|\mathcal{S}|(|\mathcal{A}|-1)}\to\mathcal{L}_0$, so that the goal is simplified to analyze the gradient estimation of $f_\lambda(x)\overset{\rm def}{=}V_{\lambda,T(x)+|\mathcal{A}|^{-1}}^{T(x)+|\mathcal{A}|^{-1}}$ on any $x\in\mathbb{R}^{|\mathcal{S}|(|\mathcal{A}|-1)}$.

Intuition and Novelty for Proving Theorem 4: Standard convergence analysis of Frank-Wolfe algorithm yields that $\max_{\tilde\pi\in\Pi_\Delta}\langle\nabla_\pi V_{\lambda,\pi_{\tilde T}}^{\pi_{\tilde T}},\tilde\pi-\pi_{\tilde T}\rangle\le\frac{D\epsilon}{2}$ on Π_Δ. However, it requires a trick to prove the following Proposition 2 which implies that $\pi_{\tilde T}$ is $D\epsilon$-stationary on Π.

6 Experiments

We compare our Algorithm 1 with the existing repeated retraining algorithm in a simulation environment. See Appendix B for the implementation details. Then for the policies π_t obtained by each algorithm, we plot the training curves of the performative value function $V_{\lambda,\pi_t}^{\pi_t}$ ($\lambda=0.5$) and the unregularized performative value function $V_{0,\pi_t}^{\pi_t}$ in Figure 1 in Appendix B, which show that our Algorithm 1 converges better than the existing repeated retraining algorithm on both regularized and unregularized performative value functions.

7 Conclusion

We have studied an entropy-regularized performative reinforcement learning problem, obtained its important properties including gradient dominance, policy lower bound, Lipschitz continuity and smoothness. Based on these properties, we have proposed a zeroth-order Frank-Wolfe (0-FW) algorithm only using sample-based policy evaluation, which for the first time converges to a *performatively optimal (PO)* policy with polynomial number of policy evaluations under the regularizer dominance condition. These theoretical results also holds for the quadratice regularizers used in the existing works on performative reinforcement learning (see Appendix M for discussion).

REFERENCES

Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Colt*, pp. 28–40. Citeseer, 2010.

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

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv:2204.05862*, 2022.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *Operations Research*, 2024.

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

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *International conference on artificial intelligence and statistics*, pp. 6045–6061, 2022.

Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456): 885–890, 2019.

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

Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM conference on recommender systems*, pp. 224–232, 2018.

Yuxin Chen. Mirror descent. `https://yuxinchen2020.github.io/ele522_optimization/lectures/mirror_descent.pdf`, 2020.

Ziyi Chen and Heng Huang. Accelerated policy gradient for s-rectangular robust mdps with large state spaces. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Ziyi Chen, Shaocong Ma, and Yi Zhou. Sample efficient stochastic policy extragradient algorithm for zero-sum markov game. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism. In *International conference on machine learning*, pp. 1843–1854, 2020.

Ofer Dekel and Elad Hazan. Better rates for any adversarial deterministic mdp. In *International Conference on Machine Learning*, pp. 675–683, 2013.

Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. A kernel-based approach to non-stationary reinforcement learning in metric spaces. In *International Conference on Artificial Intelligence and Statistics*, pp. 3538–3546, 2021.

Eyal Even-Dar and Yishay Mansour. Experts in a markov decision process. In *Proceedings of the International Conference on Neural Information Processing Systems (Neurips)*, volume 17, pp. 401. MIT Press, 2004.

Yingjie Fei, Zhuoran Yang, Zhaoran Wang, and Qiaomin Xie. Dynamic regret of policy optimization in non-stationary environments. In *Proceedings of the International Conference on Neural Information Processing Systems (Neurips)*, pp. 6743–6754, 2020.

Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 385–394, 2005.

Pratik Gajane, Ronald Ortner, and Peter Auer. A sliding-window algorithm for markov decision processes with arbitrarily changing rewards and transitions. *ArXiv:1805.10066*, 2018.

LIU Haitong, LI Qiang, and Hoi To Wai. Two-timescale derivative free optimization for performative prediction with markovian data. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Moritz Hardt and Celestine Mendler-Dünner. Performative prediction: Past and future. *ArXiv:2310.16608*, 2023.

Alexander Havrilla, Yuqing Du, Sharath Chandra Raparthy, Christoforos Nalmpantis, Jane Dwivedi-Yu, Eric Hambro, Sainbayar Sukhbaatar, and Roberta Raileanu. Teaching large language models to reason with reinforcement learning. In *AI for Math Workshop@ ICML 2024*, 2024.

Zachary Izzo, Lexing Ying, and James Zou. How to learn when data reacts to your model: performative gradient descent. In *International Conference on Machine Learning*, pp. 4641–4650, 2021.

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games. In *ICLR 2022 Workshop on Gamification and Multiagent Solutions*, 2022.

Gen Li, Weichen Wu, Yuejie Chi, Cong Ma, Alessandro Rinaldo, and Yuting Wei. Sharp high-probability sample complexities for policy evaluation with linear function approximation. *ArXiv:2305.19001*, 2023.

Qiang Li and Hoi-To Wai. State dependent performative prediction with stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3164–3186, 2022.

Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter L Bartlett, and Martin J Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. *Journal of Machine Learning Research*, 21(21):1–51, 2020.

Debmalya Mandal and Goran Radanovic. Performative reinforcement learning with linear markov decision process. *ArXiv:2411.05234*, 2024.

Debmalya Mandal, Stelios Triantafyllou, and Goran Radanovic. Performative reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 23642–23680, 2023.

Daniel J Mankowitz, Nir Levine, Rae Jeong, Abbas Abdolmaleki, Jost Tobias Springenberg, Yuanyuan Shi, Jackie Kay, Todd Hester, Timothy Mann, and Martin Riedmiller. Robust reinforcement learning for continuous control with model misspecification. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 2145–2148, 2020.

Celestine Mendler-Dünner, Juan C Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. In *Proceedings of the International Conference on Neural Information Processing Systems*, pp. 4929–4939, 2020.

John P Miller, Juan C Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, pp. 7710–7720, 2021.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 48, pp. 1928–1937, 2016.

Stefanos Nikolaidis, Swaprava Nath, Ariel D Procaccia, and Siddhartha Srinivasa. Game-theoretic modeling of human adaptation in human-robot collaboration. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, pp. 323–331, 2017.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609, 2020.

Vasilis Pollatos, Debmalya Mandal, and Goran Radanovic. On corruption-robustness in performative reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19939–19947, 2025.

Ben Rank, Stelios Triantafyllou, Debmalya Mandal, and Goran Radanovic. Performative reinforcement learning in gradually shifting environments. In *The 40th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2024.

Mitas Ray, Lillian J Ratliff, Dmitriy Drusvyatskiy, and Maryam Fazel. Decision-dependent risk minimization in geometrically decaying dynamic environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8081–8088, 2022.

Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning (ICML)*, pp. 5478–5486, 2019.

Abhishek Roy, Krishnakumar Balasubramanian, and Saeed Ghadimi. Constrained stochastic nonconvex optimization with state-dependent markov data. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 23256–23270, 2022.

Rilind Sahitaj, Paulius Sasnauskas, Yiğit Yalın, Debmalya Mandal, and Goran Radanović. Independent learning in performative markov potential games. *ArXiv:2504.20593*, 2025.

Sergey Samsonov, Daniil Tiapkin, Alexey Naumov, and Eric Moulines. Finite-sample analysis of the temporal difference learning. *ArXiv:2310.14286*, 2023.

Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52):1–11, 2017.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *nature*, 575(7782):350–354, 2019.

Qiuhao Wang, Chin Pang Ho, and Marek Petrik. Policy gradient in robust mdps with global convergence guarantee. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 202, pp. 35763–35797, 23–29 Jul 2023.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Conference on learning theory (COLT)*, pp. 4300–4354, 2021.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In *Proceedings of the Conference on Learning Theory (COLT)*, 2021.

Junyu Zhang, Amrit Singh Bedi, Mengdi Wang, and Alec Koppel. Beyond cumulative returns via reinforcement learning over state-action occupancy measures. In *2021 American Control Conference (ACC)*, pp. 894–901. IEEE, 2021.

Appendix

Table of Contents

A Related Works

Non-stationary Reinforcement Learning: The performative reinforcement learning studied in this work relates to some non-stationary reinforcement learning. For example, Gajane et al. (2018); Fei et al. (2020); Cheung et al. (2020); Wei & Luo (2021); Domingues et al. (2021) provide theoretical results assuming that the non-stationary environment (rewards and transitions) change in a bounded amount or number, and Even-Dar & Mansour (2004); Dekel & Hazan (2013); Rosenberg & Mansour (2019) study reinforcement learning with adversarial reward functions.

Performative Prediction: Performative prediction proposed by (Perdomo et al., 2020) is a stochastic optimization framework where the data distribution depends on the decision policy. Compared with performative prediction, performative reinforcement learning is similar but more complex due to the policy-dependent transition dynamics.

Various algorithms have been obtained with finite-time convergence to various solutions of performative prediction. For example, Mendler-Dünner et al. (2020); Brown et al. (2022); Li & Wai (2022) converge to a performatively stable solution that approximates the performatively optimal solution (the primary goal). Izzo et al. (2021); Roy et al. (2022); Haitong et al. (2024) converge to a stationary point of the nonconvex performative prediction objective. Miller et al. (2021); Ray et al. (2022) converge to the performatively optimal solution (the primary goal), which relies on the strong assumptions that the loss function is strongly convex with degree dominating the distribution shift and that the data distribution satisfies mixture dominance condition or belongs to a location-scale family, such that the objective function becomes convex as proved by (Miller et al., 2021). In contrast, we have proved an analogous result that the objective of performative reinforcement learning (harder than performative prediction) is gradient dominant (see our Theorem 1) without these strong assumptions. In particular, our condition of regularizer dominating the environmental shift is analogous to their condition of strong convexity dominating the distribution shift, but our value function still remains nonconvex which is more challenging than their strongly convex losses.

A survey of performative prediction can be seen in (Hardt & Mendler-Dünner, 2023).

B Experimental Details and Results

We compare our Algorithm 1 with the existing repeated retraining algorithm in a simulation environment with 5 states, 4 actions, discount factor $\gamma = 0.95$, entropy regularizer coefficient $\lambda = 0.5$, as well as transition kernel $p_\pi(s'|s,a) = \frac{\pi(a|s)+\pi(a|s')+1}{\sum_{s''}[\pi(a|s)+\pi(a|s'')+1]}$ and reward $r_\pi(s,a) = \pi(a|s)$ that depend on the policy π. We implement our Algorithm 1 for 401 iterations with $N = 1000$, $\beta = 0.01$, $\Delta = 10^{-3}$, $\delta = 10^{-4}$, the uniform policy initialization (i.e. $\pi_0(a|s) \equiv 1/4$) and the performative value functions are evaluated by value iteration.

Recall that the repeated retraining algorithm is a general framework which obtains the next policy $\pi_{t+1} \approx \arg\max_{\pi\in\Pi} V_{\lambda,\pi_t}^{\pi}; t = 0, 1, \ldots, T-1$ by solving the conventional entropy-regularized reinforcement learning problem under the fixed dynamics p_{π_t} and r_{π_t}. To solve this conventional entropy-regularized reinforcement learning problem, we select the following natural policy gradient algorithm because its output $\pi_{t+1} := \pi_{t,K}$ has been proved to converge linearly to the optimal solution of $\arg\max_{\pi\in\Pi} V_{\lambda,\pi_t}^{\pi}$ as we increase the number K of natural policy gradient steps (Cen et al., 2022).

$$\pi_{t,k+1}(a|s) = \frac{1}{Z_{t,k}(s)} \pi_{t,k}(a|s)^{1-\frac{\eta\lambda}{1-\gamma}} \exp\Big[\frac{\eta Q_\lambda(s,a;\pi_{t,k})}{1-\gamma}\Big], k = 0, 1, \ldots, K-1. \tag{26}$$

where

$$Z_{t,k}(s) \overset{\text{def}}{=} \sum_{a'\in\mathcal{A}} \pi_{t,k}(a'|s)^{1-\frac{\eta\lambda}{1-\gamma}} \exp\Big[\frac{\eta Q_\lambda(s,a';\pi_{t,k})}{1-\gamma}\Big],$$

$$Q_\lambda(s,a;\pi) \overset{\text{def}}{=} \mathbb{E}_{\pi,p_\pi,\rho}\Big[\sum_{t=0}^{\infty} \gamma^t [r_\pi(s_t,a_t) - \lambda \log \pi(a_t|s_t)] \Big| s_0 = s, a_0 = a\Big].$$

Here, we also implement $T = 401$ outer iterations of the repeated retraining algorithm, and for the inner loop we apply $K = 1000$ natural policy gradient steps with stepsize $\eta = 0.01$.

The experiment is implemented on Python 3.9, using Apple M1 Pro with 8 cores and 16 GB memory, which costs about 110 minutes in total. Then for the policies $\{\pi_t\}_{t=0}^{400}$ obtained by each algorithm, we plot the training curves of the performative value function $V_{\lambda,\pi_t}^{\pi_t}$ (defined by Eq. (6) with $\lambda = 0.5$) and the unregularized performative value function $V_{0,\pi_t}^{\pi_t}$ (defined by Eq. (6) with $\lambda = 0$) on the left and right side of Figure 1 respectively, which show that the existing repeated retraining algorithm stucks at the initial uniform policy π_0 since π_0 is a performatively stable (PS) policy, while our Algorithm 1 converges well on both regularized and unregularized performative value functions in a similar pattern.

C Supporting Lemmas

C.1 Frank-Wolfe Step

We repeat Lemma 1 as follows.

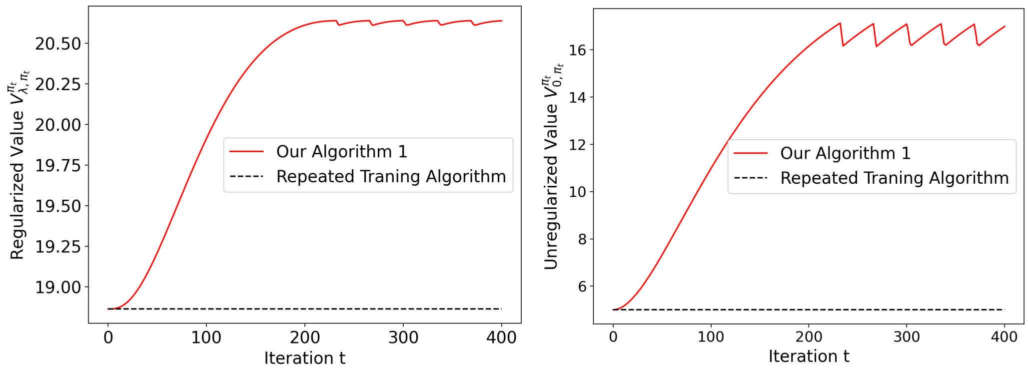

Figure 1: Experimental Results.

Lemma 2. *The step (19) has the following analytical solution.*

$$\tilde{\pi}_t(a|s) = \begin{cases} \Delta; a \neq \tilde{a}_t(s) \\ 1 - \Delta(|\mathcal{A}| - 1); a = \tilde{a}_t(s) \end{cases}, \tag{27}$$

where $\tilde{a}_t(s) \in \arg\max_a \hat{g}_{\lambda,\delta}(\pi_t)(a|s)$.

Proof. For $\tilde{\pi}_t$ defined by Eq. (27) and for any $\pi \in \Pi_\Delta$, we have

$$\begin{aligned}
&\langle \tilde{\pi}_t - \pi, \hat{g}_{\lambda,\delta}(\pi_t) \rangle \\
=& \sum_{s,a} \hat{g}_{\lambda,\delta}(\pi_t)(a|s)[\tilde{\pi}_t(a|s) - \pi(a|s)] \\
=& \sum_s \Big\{ \hat{g}_{\lambda,\delta}(\pi_t)[\tilde{a}_t(s)|s]\big[1 - \Delta(|\mathcal{A}| - 1) - \pi[\tilde{a}_t(s)|s]\big] - \sum_{a \neq \tilde{a}_t(s)} \hat{g}_{\lambda,\delta}(\pi_t)(a|s)[\pi(a|s) - \Delta] \Big\} \\
\overset{(a)}{\geq}& \sum_s \Big\{ \hat{g}_{\lambda,\delta}(\pi_t)[\tilde{a}_t(s)|s]\big[1 - \Delta(|\mathcal{A}| - 1) - \pi[\tilde{a}_t(s)|s]\big] \\
&- \sum_{a \neq \tilde{a}_t(s)} \hat{g}_{\lambda,\delta}(\pi_t)[\tilde{a}_t(s)|s][\pi(a|s) - \Delta] \Big\} \\
=& \sum_s \Big\{ \hat{g}_{\lambda,\delta}(\pi_t)[\tilde{a}_t(s)|s]\big[1 - \Delta(|\mathcal{A}| - 1) - \pi[\tilde{a}_t(s)|s]\big] \\
&- \hat{g}_{\lambda,\delta}(\pi_t)[\tilde{a}_t(s)|s]\big[1 - \pi[\tilde{a}_t(s)|s] - \Delta(|\mathcal{A}| - 1)\big] \Big\} \\
=&0,
\end{aligned}$$

where (a) uses $\pi(a|s) - \Delta \geq 0$ and $\hat{g}_{\lambda,\delta}(\pi_t)(a|s) \leq \hat{g}_{\lambda,\delta}(\pi_t)[\tilde{a}_t(s)|s]$. Therefore, Eq. (19) holds, that is, $\tilde{\pi}_t = \arg\max_{\pi \in \Pi_\Delta} \langle \pi, \hat{g}_{\lambda,\delta}(\pi_t) \rangle$. □

C.2 Lipschitz Property of Occupany Measure

Lemma 3. *The occupancy measure $d_{\pi,p}$ defined by Eq. (2) has the following Lipschitz properties for any $\pi, \pi' \in \Pi$, $p, p' \in \mathcal{P}$ and $\tilde{s} \in \mathcal{S}$.*

$$\sum_s |d_{\pi',p}(s) - d_{\pi,p}(s)| \leq \frac{\gamma}{1-\gamma} \max_s \|\pi'(\cdot|s) - \pi(\cdot|s)\|_1 \leq \frac{\gamma\sqrt{|\mathcal{A}|}}{1-\gamma} \|\pi' - \pi\| \tag{28}$$

$$\sum_s |d_{\pi,p'}(s) - d_{\pi,p}(s)| \leq \frac{\gamma}{1-\gamma} \max_{s,a} \|p'(\cdot|s,a) - p(\cdot|s,a)\|_1 \leq \frac{\gamma\sqrt{|\mathcal{S}|}}{1-\gamma} \|p' - p\| \tag{29}$$

$$\sum_{s,a} |d_{\pi',p'}(s,a) - d_{\pi,p}(s,a)| \leq \frac{1}{1-\gamma} \max_s \|\pi'(\cdot|s) - \pi(\cdot|s)\|_1 + \frac{\gamma}{1-\gamma} \max_{s,a} \|p'(\cdot|s,a) - p(\cdot|s,a)\|_1$$

$$\leq \frac{\sqrt{|\mathcal{A}|}}{1-\gamma}\|\pi' - \pi\| + \frac{\gamma\sqrt{|\mathcal{S}|}}{1-\gamma}\|p' - p\| \tag{30}$$

Proof. The first $\leq$ of Eqs. (28) and (29) follows from Lemma 5 of (Chen & Huang, 2024). The second $\leq$ of Eqs. (28) and (29) uses $\|x\|_1 \leq \sqrt{d}\|x\|$ for any $x \in \mathbb{R}^d$.

Eq. (30) can be proved as follows.

$$\begin{aligned}
&\sum_{s,a} |d_{\pi',p'}(s,a) - d_{\pi,p}(s,a)| \\
=&\sum_{s,a} |d_{\pi',p'}(s)\pi'(a|s) - d_{\pi,p}(s)\pi(a|s)| \\
\leq&\sum_{s,a} d_{\pi',p'}(s)|\pi'(a|s) - \pi(a|s)| + \pi(a|s)|d_{\pi',p'}(s) - d_{\pi,p}(s)| \\
\leq&\sum_{s}[d_{\pi',p'}(s) \max_{s'} \|\pi'(\cdot|s') - \pi(\cdot|s')\|_1] + \sum_s |d_{\pi',p'}(s) - d_{\pi,p}(s)| \\
\overset{(a)}{\leq}& \max_{s'} \|\pi'(\cdot|s') - \pi(\cdot|s')\|_1 + \frac{\gamma}{1-\gamma} \max_s \|\pi'(\cdot|s) - \pi(\cdot|s)\|_1 + \frac{\gamma}{1-\gamma} \max_{s,a} \|p'(\cdot|s,a) - p(\cdot|s,a)\|_1 \\
\leq&\frac{1}{1-\gamma} \max_s \|\pi'(\cdot|s) - \pi(\cdot|s)\|_1 + \frac{\gamma}{1-\gamma} \max_{s,a} \|p'(\cdot|s,a) - p(\cdot|s,a)\|_1 \\
\leq&\frac{\sqrt{|\mathcal{A}|}}{1-\gamma}\|\pi' - \pi\| + \frac{\gamma\sqrt{|\mathcal{S}|}}{1-\gamma}\|p' - p\|,
\end{aligned}$$

where (a) uses Eqs. (28) and (29). □

C.3 VARIOUS VALUE FUNCTIONS

Define the following value functions.

$$\begin{aligned}
J_\lambda(\pi, \pi', p, r) \overset{\text{def}}{=}& \mathbb{E}_{\pi,p}\Big[\sum_{t=0}^{\infty} \gamma^t [r(s_t, a_t) - \lambda \log \pi'(a_t|s_t)]\Big| s_0 \sim \rho\Big] \\
=& \frac{1}{1-\gamma} \sum_{s,a} d_{\pi,p}(s,a)[r(s,a) - \lambda \log \pi'(a|s)], \qquad (31)
\end{aligned}$$

$$V_\lambda(\pi, \pi', p, r; s) \overset{\text{def}}{=} \mathbb{E}_{\pi,p}\Big[\sum_{t=0}^{\infty} \gamma^t [r(s_t, a_t) - \lambda \log \pi'(a_t|s_t)]\Big| s_0 = s\Big], \tag{32}$$

$$\begin{aligned}
Q_\lambda(\pi, \pi', p, r; s, a) \overset{\text{def}}{=}& \mathbb{E}_{\pi,p}\Big[\sum_{t=0}^{\infty} \gamma^t [r(s_t, a_t) - \lambda \log \pi'(a_t|s_t)]\Big| s_0 = s, a_0 = a\Big] \\
=& r(s,a) - \lambda \log \pi'(a|s) + \gamma \sum_{s'} p(s'|s,a) V_\lambda(\pi, \pi', p, r; s'). \qquad (33)
\end{aligned}$$

Note that the value function (6) of interest can be rewritten into the above functions as follows.

$$\begin{aligned}
V_{\lambda,\pi'}^{\pi} =& J_\lambda(\pi, \pi, p_{\pi'}, r_{\pi'}) \\
=& \sum_s \rho(s) V_\lambda(\pi, \pi, p_{\pi'}, r_{\pi'}; s) \\
=& \sum_{s,a} \rho(s)\pi(a|s) Q_\lambda(\pi, \pi, p_{\pi'}, r_{\pi'}; s, a).
\end{aligned}$$

Hence, we will investigate the properties of the value functions (31)-(33) as follows.

Lemma 4. *For any* $\pi \in \Pi$, $p \in \mathcal{P}$, $r \in \mathcal{R}$, *we have* $V_{\lambda,\pi}^{\pi}, J_\lambda(\pi, \pi, p, r), V_\lambda(\pi, \pi, p, r; s), Q_\lambda(\pi, \pi, p, r; s, a) \in \Big[0, \frac{1+\lambda \log |\mathcal{A}|}{1-\gamma}\Big]$.

Proof. We will prove the range of $J_\lambda(\pi,\pi,p,r)$ as follows using $r(s,a)\in[0,1]$. The proof for the other value functions follow the same way.

$$\begin{aligned}0\le J_\lambda(\pi,\pi,p,r)=&\mathbb{E}_{\pi,p,\rho}\Big[\sum_{t=0}^{\infty}\gamma^t[r(s_t,a_t)-\lambda\log\pi(a_t|s_t)]\Big]\\ \le&\sum_{t=0}^{\infty}\gamma^t+\lambda\mathbb{E}_{\pi,p,\rho}\Big[\sum_{t=0}^{\infty}\gamma^t\sum_a[-\pi(a|s_t)\log\pi(a|s_t)]\Big]\\ \le&\frac{1}{1-\gamma}+\lambda\sum_{t=0}^{\infty}\gamma^t\log|\mathcal{A}|\\ \le&\frac{1+\lambda\log|\mathcal{A}|}{1-\gamma}.\end{aligned}$$

□

Lemma 5. *The gradients of $J_\lambda(\pi,\pi',p,r)$ defined by Eq. (31) have the following expressions.*

$$\frac{\partial J_\lambda(\pi,\pi',p,r)}{\partial\pi(a|s)}=\frac{d_{\pi,p}(s)Q_\lambda(\pi,\pi',p,r;s,a)}{1-\gamma},\tag{34}$$

$$\frac{\partial J_\lambda(\pi,\pi',p,r)}{\partial\pi'(a|s)}=-\frac{\lambda d_{\pi,p}(s,a)}{(1-\gamma)\pi'(a|s)},\tag{35}$$

$$\frac{\partial J_\lambda(\pi,\pi',p,r)}{\partial p(s'|s,a)}=\frac{d_{\pi,p}(s,a)}{1-\gamma}\big[r(s,a)-\lambda\log\pi'(a|s)+\gamma V_\lambda(\pi,\pi',p,r;s')\big],\tag{36}$$

$$\frac{\partial J_\lambda(\pi,\pi',p,r)}{\partial r(s,a)}=\frac{d_{\pi,p}(s,a)}{1-\gamma},\tag{37}$$

$$\frac{\partial J_\lambda(\pi,\pi,p,r)}{\partial\pi(a|s)}=\frac{d_{\pi,p}(s)[Q_\lambda(\pi,\pi,p,r;s,a)-\lambda]}{1-\gamma}.\tag{38}$$

Proof. Eq. (34) follows from the policy gradient expression in Eq. (7) of (Agarwal et al., 2021), with reward function $r(s,a)$ replaced by $r(s,a)-\lambda\log\pi'(a|s)$.

Eq. (36) can be proved as follows.

$$\begin{aligned}p(s'|s,a)\overset{(a)}{=}&\frac{d_{\pi,p}(s)\pi(a|s)}{1-\gamma}\big[r(s,a)-\lambda\log\pi(a|s)+\gamma V_\lambda(\pi,\pi',p,r;s')\big]\\ =&\frac{d_{\pi,p}(s,a)}{1-\gamma}\big[r(s,a)-\lambda\log\pi(a|s)+\gamma V_\lambda(\pi,\pi',p,r;s')\big],\end{aligned}$$

where (a) uses Eq. (9) in (Chen & Huang, 2024).

Eqs. (35) and (37) can be proved by taking derivatives of Eq. (31).

Based on the chain rule, Eq. (38) can be proved as follows by adding Eqs. (34) and (35) with $\pi'=\pi$.

$$\begin{aligned}\frac{\partial J_\lambda(\pi,\pi,p,r)}{\partial\pi(a|s)}=&\Big[\frac{\partial J_\lambda(\pi,\pi',p,r)}{\partial\pi(a|s)}+\frac{\partial J_\lambda(\pi,\pi',p,r)}{\partial\pi'(a|s)}\Big]\Big|_{\pi'=\pi}\\ =&\frac{d_{\pi,p}(s)Q_\lambda(\pi,\pi,p,r;s,a)}{1-\gamma}-\frac{\lambda d_{\pi,p}(s,a)}{(1-\gamma)\pi(a|s)}\\ =&\frac{d_{\pi,p}(s)[Q_\lambda(\pi,\pi,p,r;s,a)-\lambda]}{1-\gamma},\end{aligned}$$

where the final = uses $d_{\pi,p}(s,a)=d_{\pi,p}(s)\pi(a|s)$. □

Lemma 6. *The function J_λ defined by Eq. (31) has the following Lipschitz properties for any $\pi,\pi'\in\Pi$, $p,p'\in\mathcal{P}$ and $r,r'\in\mathcal{R}$.*

$$|J_\lambda(\pi',\pi',p,r)-J_\lambda(\pi,\pi,p,r)|\le L_\pi\max_s\|\log\pi'(\cdot|s)-\log\pi(\cdot|s)\|\tag{39}$$

$$|J_\lambda(\pi,\pi,p',r)-J_\lambda(\pi,\pi,p,r)|\le L_p\|p'-p\|\tag{40}$$

$$|J_\lambda(\pi,\pi,p,r')-J_\lambda(\pi,\pi,p,r)|\le\frac{\|r'-r\|_\infty}{1-\gamma}\le\frac{\|r'-r\|}{1-\gamma} \tag{41}$$

$$\|\nabla_p J_\lambda(\pi',\pi',p,r)-\nabla_p J_\lambda(\pi,\pi,p,r)\|\le\ell_\pi\max_s\|\log\pi'(\cdot|s)-\log\pi(\cdot|s)\| \tag{42}$$

$$\|\nabla_p J_\lambda(\pi,\pi,p',r)-\nabla_p J_\lambda(\pi,\pi,p,r)\|\le\ell_p\|p'-p\| \tag{43}$$

$$\begin{aligned}&\|\nabla_p J_\lambda(\pi',\pi',p',r')-\nabla_p J_\lambda(\pi,\pi,p,r)\|\\ \le&\ell_\pi\max_s\|\log\pi'(\cdot|s)-\log\pi(\cdot|s)\|+\ell_p\|p'-p\|+\frac{\sqrt{|\mathcal{S}|}}{(1-\gamma)^2}\|r'-r\|_\infty\end{aligned} \tag{44}$$

$$\begin{aligned}&\|\nabla_r J_\lambda(\pi',\pi',p',r')-\nabla_r J_\lambda(\pi,\pi,p,r)\|\\ \le&\frac{\max_s\|\pi'(\cdot|s)-\pi(\cdot|s)\|_1+\gamma\max_{s,a}\|p'(\cdot|s,a)-p(\cdot|s,a)\|_1}{(1-\gamma)^2}\end{aligned} \tag{45}$$

$$\begin{aligned}&\|\nabla_\pi J_\lambda(\pi',\pi',p',r')-\nabla_\pi J_\lambda(\pi,\pi,p,r)\|\\ \le&\Big(\frac{|\mathcal{A}|(1+2\lambda\log|\mathcal{A}|)}{(1-\gamma)^2}+\gamma L_\pi\Big)\max_s\|\log\pi'(\cdot|s)-\log\pi(\cdot|s)\|\\ &+\gamma\sqrt{|\mathcal{A}|}\Big[\frac{2\sqrt{|\mathcal{S}|}(1+\lambda\log|\mathcal{A}|)}{(1-\gamma)^2}+L_p\Big]\|p'-p\|+\frac{\sqrt{|\mathcal{A}|}\|r'-r\|_\infty}{1-\gamma},\end{aligned} \tag{46}$$

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

Proof. Eqs. (39), (40), (42) and (43) directly follow from Lemma 6 of (Chen & Huang, 2024). Eq. (41) can be proved as follows.

$$\begin{aligned}|J_\lambda(\pi,p,r')-J_\lambda(\pi,p,r)|=&\Big|\frac{1}{1-\gamma}\sum_{s,a}d_{\pi,p}(s,a)[r'(s,a)-r(s,a)]\Big|\\ \le&\frac{1}{1-\gamma}\sum_{s,a}d_{\pi,p}(s,a)|r'(s,a)-r(s,a)|\\ =&\frac{1}{1-\gamma}\sum_{s,a}d_{\pi,p}(s,a)\|r'-r\|_\infty\\ =&\frac{1}{1-\gamma}\|r'-r\|_\infty\le\frac{1}{1-\gamma}\|r'-r\|.\end{aligned}$$

To prove Eq. (44), note that

$$\begin{aligned}&\Big|\frac{\partial J_\lambda(\pi,\pi,p,r')}{\partial p(s'|s,a)}-\frac{\partial J_\lambda(\pi,\pi,p,r)}{\partial p(s'|s,a)}\Big|\\ \overset{(a)}{=}&\frac{d_{\pi,p}(s,a)}{1-\gamma}\big|r'(s,a)-r(s,a)+\gamma[V_\lambda(\pi,\pi',p,r';s')-V_\lambda(\pi,\pi',p,r;s')]\big|\\ \overset{(b)}{\le}&\frac{d_{\pi,p}(s,a)}{1-\gamma}\Big[\|r'-r\|_\infty+\gamma\sum_{t=0}^{\infty}\gamma^t\|r'-r\|_\infty\Big]\\ \le&\frac{d_{\pi,p}(s,a)}{(1-\gamma)^2}\|r'-r\|_\infty\end{aligned} \tag{47}$$

where (a) uses Eq. (36) and (b) uses Eq. (32). Therefore, we can prove Eq. (44) as follows.

$$\begin{aligned}&\|\nabla_p J_\lambda(\pi',\pi',p',r')-\nabla_p J_\lambda(\pi,\pi,p,r)\|\\ \le&\|\nabla_p J_\lambda(\pi',\pi',p',r')-\nabla_p J_\lambda(\pi,\pi,p',r')\|+\|\nabla_p J_\lambda(\pi,\pi,p',r')-\nabla_p J_\lambda(\pi,\pi,p,r')\|\\ &+\|\nabla_p J_\lambda(\pi,\pi,p,r')-\nabla_p J_\lambda(\pi,\pi,p,r)\|\\ \overset{(a)}{\le}&\ell_\pi\max_s\|\log\pi'(\cdot|s)-\log\pi(\cdot|s)\|+\ell_p\|p'-p\|+\sqrt{\sum_{s,a,s'}\Big|\frac{\partial J_\lambda(\pi,\pi,p,r')}{\partial p(s'|s,a)}-\frac{\partial J_\lambda(\pi,\pi,p,r)}{\partial p(s'|s,a)}\Big|^2}\end{aligned}$$

$$\overset{(b)}{\leq}\ell_\pi \max_s \|\log\pi'(\cdot|s)-\log\pi(\cdot|s)\| + \ell_p\|p'-p\| + \sqrt{\frac{\|r'-r\|_\infty^2}{(1-\gamma)^4}\sum_{s,a,s'} d_{\pi,p}^2(s,a)}$$

$$\leq\ell_\pi \max_s \|\log\pi'(\cdot|s)-\log\pi(\cdot|s)\| + \ell_p\|p'-p\| + \frac{\sqrt{|\mathcal{S}|}}{(1-\gamma)^2}\|r'-r\|_\infty,$$

where (a) uses Eqs. (42) and (43) and (b) uses Eq. (47).

Then, we prove Eq. (45) as follows.

$$\begin{aligned}
&\|\nabla_r J_\lambda(\pi',\pi',p',r') - \nabla_r J_\lambda(\pi,\pi,p,r)\| \\
\overset{(a)}{=}&\frac{\|d_{\pi',p'}-d_{\pi,p}\|}{1-\gamma} \\
\leq&\frac{\|d_{\pi',p'}-d_{\pi,p}\|_1}{1-\gamma} \\
\overset{(b)}{\leq}&\frac{1}{(1-\gamma)^2}\max_s\|\pi'(\cdot|s)-\pi(\cdot|s)\|_1 + \frac{\gamma}{(1-\gamma)^2}\max_{s,a}\|p'(\cdot|s,a)-p(\cdot|s,a)\|_1,
\end{aligned}$$

where (a) uses Eq. (37), (b) uses Eq. (30).

To prove Eq. (46), we will first prove the following auxiliary bounds.

$$Q_\lambda(\pi,\pi,p,r;s,a)-\lambda \overset{(a)}{\in} \Big[-\lambda, \frac{1+\lambda\log|\mathcal{A}|}{1-\gamma}-\lambda\Big] \Rightarrow \big|Q_\lambda(\pi,\pi,p,r;s,a)-\lambda\big| \leq \frac{1+\lambda\log|\mathcal{A}|}{1-\gamma}, \quad (48)$$

where (a) uses Lemma 4.

$$\begin{aligned}
&|V_\lambda(\pi',\pi',p',r';s) - V_\lambda(\pi,\pi,p,r;s)| \\
\leq&|V_\lambda(\pi',\pi',p',r';s)-V_\lambda(\pi,\pi,p',r';s)|+|V_\lambda(\pi,\pi,p',r';s)-V_\lambda(\pi,\pi,p,r';s)| \\
&+|V_\lambda(\pi,\pi,p,r';s)-V_\lambda(\pi,\pi,p,r;s)| \\
\overset{(a)}{\leq}&L_\pi\max_s\|\log\pi'(\cdot|s)-\log\pi(\cdot|s)\| + L_p\|p'-p\| + \frac{\|r'-r\|_\infty}{1-\gamma}, \quad (49)
\end{aligned}$$

where (a) applies Eqs. (39)-(41) to the case where the initial state distribution ρ is probability 1 at s (so $J_\lambda(\pi,\pi,p,r)$ becomes $V_\lambda(\pi,\pi,p,r;s)$).

$$\begin{aligned}
&|Q_\lambda(\pi,\pi,p,r';s,a) - Q_\lambda(\pi,\pi,p,r;s,a)| \\
\overset{(a)}{=}&\Big|\mathbb{E}_{\pi,p}\Big[\sum_{t=0}^\infty \gamma^t[r'(s_t,a_t)-r(s_t,a_t)]\Big|s_0=s,a_0=a\Big]\Big| \\
\leq&\mathbb{E}_{\pi,p}\Big[\sum_{t=0}^\infty \gamma^t|r'(s_t,a_t)-r(s_t,a_t)|\Big|s_0=s,a_0=a\Big] \\
\leq&\mathbb{E}_{\pi,p}\Big[\sum_{t=0}^\infty \gamma^t\|r'-r\|_\infty\Big|s_0=s,a_0=a\Big] \\
\leq&\frac{\|r'-r\|_\infty}{1-\gamma}, \quad (50)
\end{aligned}$$

where (a) uses Eq. (33).

$$\begin{aligned}
&|Q_\lambda(\pi',\pi',p',r;s,a) - Q_\lambda(\pi,\pi,p,r;s,a)| \\
\overset{(a)}{\leq}&\lambda|\log\pi'(a|s)-\log\pi(a|s)| + \gamma\Big|\sum_{s'}[p'(s'|s,a)V_\lambda(\pi',\pi',p',r;s) - p(s'|s,a)V_\lambda(\pi,\pi,p,r;s)]\Big| \\
\leq&\lambda|\log\pi'(a|s)-\log\pi(a|s)| + \gamma\sum_{s'}p'(s'|s,a)|V_\lambda(\pi',\pi',p',r;s) - V_\lambda(\pi,\pi,p,r;s)| \\
&+\gamma\sum_{s'}|p'(s'|s,a)-p(s'|s,a)||V_\lambda(\pi,\pi,p,r;s)|
\end{aligned}$$

$$\overset{(b)}{\leq}\lambda|\log\pi'(a|s)-\log\pi(a|s)|+\gamma L_\pi\max_{s'}\|\log\pi'(\cdot|s')-\log\pi(\cdot|s')\|+\gamma L_p\|p'-p\|$$
$$+\frac{\gamma(1+\lambda\log|\mathcal{A}|)}{1-\gamma}\|p'(\cdot|s,a)-p(\cdot|s,a)\|_1, \tag{51}$$

where (a) uses Eq. (33), and (b) uses Eq. (49) and Lemma 4.

Note that

$$\begin{aligned}
&(1-\gamma)\Big|\frac{\partial J_\lambda(\pi',\pi',p',r')}{\partial\pi'(a|s)}-\frac{\partial J_\lambda(\pi,\pi,p,r)}{\partial\pi(a|s)}\Big|\\
\overset{(a)}{=}&\big|d_{\pi',p'}(s)[Q_\lambda(\pi',\pi',p',r';s,a)-\lambda]-d_{\pi,p}(s)[Q_\lambda(\pi,\pi,p,r;s,a)-\lambda]\big|\\
\leq&\big|[d_{\pi',p'}(s)-d_{\pi,p}(s)][Q_\lambda(\pi',\pi',p',r';s,a)-\lambda]\\
&+d_{\pi,p}(s)[Q_\lambda(\pi',\pi',p',r';s,a)-Q_\lambda(\pi',\pi',p',r;s,a)]\\
&+d_{\pi,p}(s)[Q_\lambda(\pi',\pi',p',r;s,a)-Q_\lambda(\pi,\pi,p,r;s,a)]\big|\\
\leq&\big|d_{\pi',p'}(s)-d_{\pi,p}(s)\big|\cdot\big|Q_\lambda(\pi',\pi',p',r';s,a)-\lambda\big|\\
&+d_{\pi,p}(s)\big|Q_\lambda(\pi',\pi',p',r';s,a)-Q_\lambda(\pi',\pi',p',r;s,a)\big|\\
&+d_{\pi,p}(s)\big|Q_\lambda(\pi',\pi',p',r;s,a)-Q_\lambda(\pi,\pi,p,r;s,a)\big|\\
\overset{(b)}{\leq}&\frac{1+\lambda\log|\mathcal{A}|}{1-\gamma}\big|d_{\pi',p'}(s)-d_{\pi,p}(s)\big|+\frac{d_{\pi,p}(s)\|r'-r\|_\infty}{1-\gamma}\\
&+d_{\pi,p}(s)\Big[\lambda|\log\pi'(a|s)-\log\pi(a|s)|+\gamma L_\pi\max_{s'}\|\log\pi'(\cdot|s')-\log\pi(\cdot|s')\|\\
&+\gamma L_p\|p'-p\|+\frac{\gamma(1+\lambda\log|\mathcal{A}|)}{1-\gamma}\|p'(\cdot|s,a)-p(\cdot|s,a)\|_1\Big],
\end{aligned}$$

where (a) uses Eq. (38), (b) uses Eqs. (48), (50) and (51). Applying triangular inequality to the bound above, we can prove Eq. (46) as follows.

$$\begin{aligned}
&(1-\gamma)\big\|\nabla_{\pi'}J_\lambda(\pi',\pi',p',r')-\nabla_\pi J_\lambda(\pi,\pi,p,r)\big\|\\
\leq&\frac{1+\lambda\log|\mathcal{A}|}{1-\gamma}\sqrt{\sum_{s,a}\big|d_{\pi',p'}(s)-d_{\pi,p}(s)\big|^2}+\frac{\|r'-r\|_\infty}{1-\gamma}\sqrt{\sum_{s,a}d_{\pi,p}(s)^2}\\
&+\lambda\sqrt{\sum_{s,a}d_{\pi,p}(s)^2|\log\pi'(a|s)-\log\pi(a|s)|^2}\\
&+\big[\gamma L_\pi\max_{s'}\|\log\pi'(\cdot|s')-\log\pi(\cdot|s')\|+\gamma L_p\|p'-p\|\big]\sqrt{\sum_{s,a}d_{\pi,p}(s)^2}\\
&+\frac{\gamma(1+\lambda\log|\mathcal{A}|)}{1-\gamma}\sqrt{\sum_{s,a}d_{\pi,p}(s)^2\|p'(\cdot|s,a)-p(\cdot|s,a)\|_1^2}\\
\leq&\frac{\sqrt{|\mathcal{A}|}(1+\lambda\log|\mathcal{A}|)}{1-\gamma}\sum_s|d_{\pi',p'}(s)-d_{\pi,p}(s)|+\frac{\sqrt{|\mathcal{A}|}\|r'-r\|_\infty}{1-\gamma}\\
&+\lambda\sqrt{\sum_s d_{\pi,p}(s)\|\log\pi'(\cdot|s)-\log\pi(\cdot|s)\|^2}\\
&+\big[\gamma L_\pi\max_{s'}\|\log\pi'(\cdot|s')-\log\pi(\cdot|s')\|+\gamma L_p\|p'-p\|\big]\sqrt{|\mathcal{A}|}\\
&+\frac{\gamma(1+\lambda\log|\mathcal{A}|)}{1-\gamma}\sqrt{|\mathcal{S}|\sum_{s,a}\|p'(\cdot|s,a)-p(\cdot|s,a)\|^2}\\
\overset{(a)}{\leq}&\frac{\gamma\sqrt{|\mathcal{A}|}(1+\lambda\log|\mathcal{A}|)}{(1-\gamma)^2}\big[\max_s\|\pi'(\cdot|s)-\pi(\cdot|s)\|_1+\max_{s,a}\|p'(\cdot|s,a)-p(\cdot|s,a)\|_1\big]\\
&+\frac{\sqrt{|\mathcal{A}|}\|r'-r\|_\infty}{1-\gamma}+\lambda\max_{s'}\|\log\pi'(\cdot|s')-\log\pi(\cdot|s')\|
\end{aligned}$$

$$
\begin{aligned}
&+ \sqrt{|\mathcal{A}|}\big[\gamma L_\pi \max_{s'} \|\log \pi'(\cdot|s') - \log \pi(\cdot|s')\| + \gamma L_p \|p' - p\|\big] \\
&+ \frac{\gamma\sqrt{|\mathcal{S}|}(1+\lambda \log|\mathcal{A}|)}{1-\gamma}\|p'-p\| \\
&\overset{(b)}{\le} \Big[\frac{|\mathcal{A}|(\gamma + 2\lambda \log|\mathcal{A}|)}{(1-\gamma)^2} + \gamma L_\pi\Big] \max_{s'} \|\log \pi'(\cdot|s') - \log \pi(\cdot|s')\| \\
&+ \gamma\sqrt{|\mathcal{A}|}\Big[\frac{2\sqrt{|\mathcal{S}|}(1+\lambda\log|\mathcal{A}|)}{(1-\gamma)^2} + L_p\Big]\|p'-p\| + \frac{\sqrt{|\mathcal{A}|}\|r'-r\|_\infty}{1-\gamma},
\end{aligned}
$$

where (a) uses Lemma 3, (b) uses $\|\pi'(\cdot|s) - \pi(\cdot|s)\|_1 \le \|\log \pi'(\cdot|s) - \log \pi(\cdot|s)\|_1$, $\|p'(\cdot|s,a) - p(\cdot|s,a)\|_1 \le \sqrt{|\mathcal{S}|}\|p'(\cdot|s,a) - p(\cdot|s,a)\| \le \sqrt{|\mathcal{S}|}\|p'-p\|$, $\frac{\gamma\sqrt{|\mathcal{S}|}(1+\lambda\log|\mathcal{A}|)}{1-\gamma} \le \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}(1+\lambda\log|\mathcal{A}|)}{(1-\gamma)^2}$ and $\lambda \le \frac{\lambda|\mathcal{A}|\log|\mathcal{A}|}{(1-\gamma)^2}$. □

C.4 Zeroth-order Gradient Estimation Error

We import Theorem 1.6.2 of (Tropp et al., 2015) as follows.

Lemma 7 (Matrix Bernstein Inequality). *Suppose complex-valued matrices $S_1, \ldots, S_N \in \mathbb{C}^{d_1\times d_2}$ are independently distributed with $\mathbb{E}S_k = 0$ and $\|S_k\| \le C$ for each $k = 1, \ldots, N$. Denote the sum $Z_N = \sum_{k=1}^N S_k$ its variance statistic as follows*

$$
v(Z_N) = \max\Big[\Big\|\sum_{k=1}^N \mathbb{E}(S_k S_k^*)\Big\|, \Big\|\sum_{k=1}^N \mathbb{E}(S_k^* S_k)\Big\|\Big], \tag{52}
$$

where S_k^* denotes the conjugate transpose of S_k. Then for any $\epsilon \ge 0$, we have

$$
\mathbb{P}\{\|Z_N\| \ge \epsilon\} \le (d_1 + d_2)\exp\Big[\frac{-\epsilon^2/2}{v(Z_N) + C\epsilon/3}\Big]. \tag{53}
$$

Applying the above lemma to vectors, we obtain the following vector Bernstein inequality.

Lemma 8 (Vector Bernstein Inequality). *Suppose independently distributed vectors $x_1, \ldots, x_N \in \mathbb{C}^d$ satisfies $\|x_k\| \le c$ for each $k = 1, \ldots, N$. Then for any $\eta \in (0,1)$, with probability at least $1-\eta$, we have*

$$
\Big\|\frac{1}{N}\sum_{k=1}^N (x_k - \mathbb{E}x_k)\Big\| < \frac{4c}{3N}\log\Big(\frac{d+1}{\eta}\Big) + 2c\sqrt{\frac{2}{N}\log\Big(\frac{d+1}{\eta}\Big)}. \tag{54}
$$

Proof. Note that $S_k = x_k - \mathbb{E}x_k$ satisfies the conditions of Lemma 7 with $d_1 = d$, $d_2 = 1$ and C replaced by $2c$. In addition, $v(Z_N)$ defined by Eq. (52) satisfies $v(Z_N) \le 4Nc^2$ since

$$
\max[\|S_k S_k^*\|, \|S_k^* S_k\|^2] \le \|S_k^*\|^2\|S_k\|^2 \le 4c^2.
$$

For any $\eta \in (0,1)$, let

$$
\epsilon = \frac{4c}{3}\log\Big(\frac{d+1}{\eta}\Big) + c\sqrt{2N\log\Big(\frac{d+1}{\eta}\Big)}.
$$

Therefore, Lemma 7 implies that

$$
\mathbb{P}\Big\{\frac{1}{N}\Big\|\sum_{k=1}^N (x_k - \mathbb{E}x_k)\Big\| \ge \frac{\epsilon}{N}\Big\} \le (d+1)\exp\Big[\frac{-\epsilon^2/2}{4Nc^2 + 2c\epsilon/3}\Big] \le \eta,
$$

which implies that with probability at least $1-\eta$, we have

$$
\frac{1}{N}\Big\|\sum_{k=1}^N (x_k - \mathbb{E}x_k)\Big\| < \frac{\epsilon}{N} = \frac{4c}{3N}\log\Big(\frac{d+1}{\eta}\Big) + 2c\sqrt{\frac{2}{N}\log\Big(\frac{d+1}{\eta}\Big)}.
$$

□

For any function $f : \mathbb{R}^d \to \mathbb{R}$, obtain the following zeroth-order stochastic estimator of the gradient ∇f.

$$g_\delta(x) = \frac{d}{2N\delta} \sum_{i=1}^{N} [f(x + \delta u_i) - f(x - \delta u_i)]u_i \approx \nabla f(x) \tag{55}$$

where $\delta > 0$ and $\{u_i\}_{i=1}^N$ are i.i.d. samples of the uniform distribution on the sphere $\mathbb{S}_d = \{u \in \mathbb{R}^d : \|u\| = 1\}$.

Lemma 9. *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is an L_f-Lipschitz continuous and ℓ_f-smooth function. Then for any $\eta \in (0, 1)$, with probability at least $1 - \eta$, the gradient estimator g_δ defined by Eq. (55) has the following error bound.*

$$\|g_\delta(x) - \nabla f(x)\| \leq \frac{4L_f d}{3N} \log\Big(\frac{d+1}{\eta}\Big) + 2L_f d\sqrt{\frac{2}{N} \log\Big(\frac{d+1}{\eta}\Big)} + \delta\ell_f. \tag{56}$$

Proof. Note that $g_{\delta,i}(x) \stackrel{\text{def}}{=} \frac{d}{2\delta}[f(x + \delta u_i) - f(x - \delta u_i)]u_i$ has the following norm bound

$$\|g_{\delta,i}(x)\| \leq \frac{d}{2\delta}\big|f(x + \delta u_i) - f(x - \delta u_i)\big| \cdot \|u_i\| \leq \frac{d}{2\delta} \cdot L_f\|2\delta u_i\| = L_f d. \tag{57}$$

Define the following smoothed approximation of f as follows.

$$f_\delta(x) \stackrel{\text{def}}{=} \mathbb{E}_{v\sim\text{Unif}(\mathbb{B}_d)}[f(x + \delta v)], \tag{58}$$

where $\text{Unif}(\mathbb{B}_d)$ denotes the uniform distribution on the ball $\mathbb{B}_d \stackrel{\text{def}}{=} \{u \in \mathbb{R}^d : \|u\| \leq 1\}$. Then based on Lemma 1 of (Flaxman et al., 2005), we have

$$\mathbb{E}[g_{\delta,i}(x)] = \nabla f_\delta(x) = \mathbb{E}_{v\sim\text{Unif}(\mathbb{B}_d)}[\nabla f(x + \delta v)]. \tag{59}$$

Therefore, applying Lemma 8 to $g_{\delta,i}(x)$, the following bound holds with probability at least $1 - \eta$.

$$\frac{1}{N}\Big\|\sum_{i=1}^{N}[g_{\delta,i}(x) - \nabla f_\delta(x)]\Big\| < \frac{4L_f d}{3N} \log\Big(\frac{d+1}{\eta}\Big) + 2L_f d\sqrt{\frac{2}{N} \log\Big(\frac{d+1}{\eta}\Big)}. \tag{60}$$

Note that

$$\|\nabla f_\delta(x) - \nabla f(x)\| = \big\|\mathbb{E}_{v\sim\text{Unif}(\mathbb{B}_d)}[\nabla f(x + \delta v) - \nabla f(x)]\big\| \leq \delta\ell_f. \tag{61}$$

As a result, we can prove the conclusion as follows by using Eqs. (60) and (61) above.

$$\begin{aligned}
\|g_\delta(x) - \nabla f(x)\| =& \Big\|\Big[\frac{1}{N}\sum_{i=1}^{N} g_{\delta,i}(x)\Big] - \nabla f(x)\Big\| \\
\leq& \Big\|\Big[\frac{1}{N}\sum_{i=1}^{N} g_{\delta,i}(x)\Big] - \nabla f_\delta(x)\Big\| + \|\nabla f_\delta(x) - \nabla f(x)\| \\
<& \frac{4L_f d}{3N} \log\Big(\frac{d+1}{\eta}\Big) + 2L_f d\sqrt{\frac{2}{N} \log\Big(\frac{d+1}{\eta}\Big)} + \delta\ell_f.
\end{aligned}$$

□

C.5 Orthogonal Transformation

Lemma 10. *There exists an orthogonal transformation $\mathcal{T}$ from the space $\mathbb{R}^{d-1}$ to $\mathcal{Z}_d = \{z = [z_1, \ldots, z_d] \in \mathbb{R}^d : \sum_i z_i = 0\}$, that is, $\mathcal{T}$ is invertible and satisfies the following properties for any $x, y \in \mathcal{Z}_d$ and $\alpha, \beta \in \mathbb{R}$.*

$$\mathcal{T}(\alpha x + \beta y) = \alpha\mathcal{T}(x) + \beta\mathcal{T}(y), \tag{62}$$

$$\langle\mathcal{T}(x), \mathcal{T}(y)\rangle = \langle x, y\rangle. \tag{63}$$

Proof. It can be verified that $\mathbb{R}^d$ admits the following orthonormal basis with $\langle e_i, e_j\rangle = 0$ for any $i \neq j$ and $\|e_i\| = 1$.

$$e_k = \frac{1}{\sqrt{k(k+1)}}[\underbrace{1, 1, \ldots, 1}_{k\ 1's}, -k, \underbrace{0, 0, \ldots, 0}_{(d-k-1)\ 0's}] \in \mathbb{R}^d; k = 1, 2, \ldots, d-1.$$

$$e_d = \frac{1}{\sqrt{d}}[\underbrace{1, 1, \ldots, 1}_{d\ 1's}] \in \mathbb{R}^d.$$

Define the transformation $\mathcal{T}$ at $x = [x_1, x_2, \ldots, x_{d-1}] \in \mathbb{R}^{d-1}$ as follows.

$$\mathcal{T}(x) = \sum_{i=1}^{d-1} x_i e_i. \tag{64}$$

Since $\mathcal{Z}_d$ is a linear subspace of $\mathbb{R}^d$ orthogonal to e_d, $\mathcal{Z}_d$ admits the orthonormal basis $\{e_i\}_{i=1}^{d-1}$. Hence, $\mathcal{T}(x) \in \mathcal{Z}_d$. Conversely, for any $y \in \mathcal{Z}_d$, there exists unique $x \in \mathbb{R}^{d-1}$ such that $y = \sum_{i=1}^{d-1} x_i e_i$. Hence, $\mathcal{T} : \mathbb{R}^{d-1} \to \mathcal{Z}^d$ is invertible.

For any $x = [x_1, \ldots, x_{d-1}], y = [y_1, \ldots, y_{d-1}] \in \mathbb{R}^{d-1}$ and $\alpha, \beta \in \mathbb{R}$, we can prove Eqs. (62) and (63) respectively as follows.

$$\begin{aligned}\mathcal{T}(\alpha x + \beta y) =& \sum_{i=1}^{d-1} (\alpha x_i + \beta y_i) e_i \\ =& \alpha \sum_{i=1}^{d-1} x_i e_i + \beta \sum_{i=1}^{d-1} y_i e_i \\ =& \alpha \mathcal{T}(x) + \beta \mathcal{T}(y).\end{aligned}$$

$$\begin{aligned}\langle \mathcal{T}(x), \mathcal{T}(y)\rangle =& \Big\langle \sum_{i=1}^{d-1} x_i e_i, \sum_{j=1}^{d-1} y_j e_j \Big\rangle \\ =& \sum_{i=1}^{d-1}\sum_{j=1}^{d-1} x_i y_j \langle e_i, e_j\rangle \\ =& \sum_{i=1}^{d-1} x_i y_i = \langle x, y\rangle.\end{aligned}$$

□

C.6 Basic Inequalities

Lemma 11. *For any $\epsilon \in (0, 0.5]$ and $x \geq 4\epsilon^{-1}\log(\epsilon^{-1})$, the following inequality holds.*

$$0 < \frac{\log x}{x} \leq \epsilon \tag{65}$$

Specifically, any $x \geq 3$ satisfies $\frac{\log x}{x} \leq \frac{1}{2}$.

Proof. As $\epsilon^{-1} \geq 2$, we have $x \geq 4\epsilon^{-1}\log(\epsilon^{-1}) \geq (4)(2)\log(2) > 5.54$, so $\log x > \log 5.54 > 1.71$, which proves the first $<$ of Eq. (65).

Note that the function $f(x) = \frac{\log x}{x}$ has the following derivative

$$f'(x) = \frac{1 - \log x}{x^2} < 0,$$

where $<$ uses $\log x > 1.71$. Hence, f is monotonic decreasing in $x \geq 4\epsilon^{-1}\log(\epsilon^{-1}) > 5.54$, Therefore, we prove the second $\leq$ of Eq. (65) as follows.

$$\frac{\log x}{x\epsilon} \leq \frac{\log[4\epsilon^{-1}\log(\epsilon^{-1})]}{\epsilon[4\epsilon^{-1}\log(\epsilon^{-1})]}$$

$$
\begin{aligned}
&= \frac{\log 4 + \log(\epsilon^{-1}) + \log[\log(\epsilon^{-1})]}{4\log(\epsilon^{-1})} \\
&\overset{(a)}{\leq} \frac{\log 4}{4\log(2)} + \frac{\log(\epsilon^{-1}) + \log(\epsilon^{-1})}{4\log(\epsilon^{-1})} = 1,
\end{aligned} \tag{66}
$$

where (a) uses $\epsilon^{-1} \geq 2$ and $\log u \leq u$ for $u = \log(\epsilon^{-1})$.

When $x \geq 3$, $f'(x) = \frac{1-\log x}{x^2} < 0$, so $f(x) \leq f(3) = \frac{\log 3}{3} < \frac{1}{2}$. □

Lemma 12. *For any $\pi, \pi' \in \Pi$, we have $\|\pi' - \pi\| \leq \sqrt{2|\mathcal{S}|}$.*

Proof.

$$
\|\pi' - \pi\|^2 = \sum_{s,a} |\pi'(a|s) - \pi(a|s)|^2 \leq \sum_{s,a} [\pi'^2(a|s) + \pi^2(a|s)] \leq \sum_{s,a} [\pi'(a|s) + \pi(a|s)] = 2|\mathcal{S}|.
$$

□

D NEGATIVE ENTROPY REGULARIZER AS A STRONGLY CONVEX FUNCTION OF OCCUPANCY MEASURE

The negative entropy regularizer (5) can be rewritten as follows

$$
\mathcal{H}_{\pi'}(\pi) = \mathbb{E}_{\pi, p_{\pi'}, \rho}\Big[\sum_{t=0}^{\infty} \gamma^t \log \pi(a_t|s_t)\Big] = \frac{1}{1-\gamma} \sum_{s,a} d_{\pi, p_{\pi'}}(s,a) \log \frac{d_{\pi, p_{\pi'}}(s,a)}{d_{\pi, p_{\pi'}}(s)}, \tag{67}
$$

where $d_{\pi, p_{\pi'}}(s) = \sum_{a'} d_{\pi, p_{\pi'}}(s, a')$. Hence, it suffices to prove that the following function of occupancy measure d is strongly convex.

$$
H(d) = \sum_{s,a} d(s,a) \log \frac{d(s,a)}{d(s)}, \tag{68}
$$

where $d(s) = \sum_{a'} d(s, a')$. For any $\alpha \in [0,1]$ and occupancy measures d_1, d_0, denote $d_\alpha = \alpha d_1 + (1-\alpha) d_0$ and the corresponding policy as $\pi_\alpha(a|s) = \frac{d_\alpha(s,a)}{d_\alpha(s)}$. Then we have

$$
\begin{aligned}
&\alpha H(d_1) + (1-\alpha) H(d_0) - H(d_\alpha) \\
=& \sum_{s,a} \Big[\alpha d_1(s,a) \log \pi_1(a|s) + (1-\alpha) d_0(s,a) \log \pi_0(a|s) \\
&- [\alpha d_1(s,a) + (1-\alpha) d_0(s,a)] \log \pi_\alpha(a|s)\Big] \\
=& \sum_{s,a} \Big[\alpha d_1(s,a) \log \frac{\pi_1(a|s)}{\pi_\alpha(a|s)} + (1-\alpha) d_0(s,a) \log \frac{\pi_0(a|s)}{\pi_\alpha(a|s)}\Big] \\
=& \sum_{s,a} \Big[\alpha d_1(s) \pi_1(a|s) \log \frac{\pi_1(a|s)}{\pi_\alpha(a|s)} + (1-\alpha) d_0(s) \pi_0(a|s) \log \frac{\pi_0(a|s)}{\pi_\alpha(a|s)}\Big] \\
=& \sum_{s} \Big[\alpha d_1(s) \mathrm{KL}[\pi_1(\cdot|s) \| \pi_\alpha(a|s)] + (1-\alpha) d_0(s) \mathrm{KL}[\pi_0(\cdot|s) \| \pi_\alpha(a|s)]\Big] \\
\overset{(a)}{\geq}& \frac{1}{2} \sum_{s} \Big[\alpha d_1(s) \|\pi_1(\cdot|s) - \pi_\alpha(\cdot|s)\|_1^2 + (1-\alpha) d_0(s) \|\pi_0(\cdot|s) - \pi_\alpha(\cdot|s)\|_1^2\Big] \\
\overset{(b)}{\geq}& \frac{D}{2} \sum_{s} \Big[\alpha \|\pi_1(\cdot|s) - \pi_\alpha(\cdot|s)\|_1^2 + (1-\alpha) \|\pi_0(\cdot|s) - \pi_\alpha(\cdot|s)\|_1^2\Big] \\
\geq& \frac{D}{2} \Big[\alpha \max_s \|\pi_1(\cdot|s) - \pi_\alpha(\cdot|s)\|_1^2 + (1-\alpha) \max_s \|\pi_0(\cdot|s) - \pi_\alpha(\cdot|s)\|_1^2\Big] \\
\overset{(c)}{\geq}& \frac{D(1-\gamma)}{2} \Big[\alpha \|d_1 - d_\alpha\|_1^2 + (1-\alpha) \max_s \|d_0 - d_\alpha\|_1^2\Big]
\end{aligned}
$$

$$=\frac{D(1-\gamma)}{2}\Big[\alpha(1-\alpha)^2\|d_1-d_0\|_1^2+(1-\alpha)\alpha^2\|d_1-d_0\|_1^2\Big]$$
$$=\frac{\alpha(1-\alpha)}{2}\cdot D(1-\gamma)\|d_1-d_0\|_1^2. \tag{69}$$

where (a) uses Pinsker's inequality, (b) uses Assumption 3, (c) uses Eq. (30) with $p'=p$. The inequality above implies that $H(d)$ is $D(1-\gamma)$-strongly convex, so the negative entropy regularizer (67) can be seen as a D-strongly convex function of the occupancy measure $d_{\pi,p_{\pi'}}$.

E Existing Assumptions That Implies Assumption 3

The following assumptions have been used in the reinforcement learning literature. We will show that each of these assumptions implies Assumption 3.

Assumption 4. *(Bhandari & Russo, 2024) $\rho(s)>0$ for any $s\in\mathcal{S}$.*

Assumption 5. *(Agarwal et al., 2021; Leonardos et al., 2022; Wang et al., 2023; Chen & Huang, 2024) $D_\rho:=\sup_{\pi\in\Pi,p\in\mathcal{P}}\|d_{\pi,p}/\rho\|_\infty<\infty$.*

Assumption 6. *(Wei et al., 2021; Chen et al., 2022) There exists a constant $\mu_{\min}>0$ and mixing time $t_{mix}\in\mathbb{N}$ such that under any policy $\pi\in\Pi$ and transition kernel $p\in\mathcal{P}$, the stationary state distribution $\mu_{\pi,p}(s)$ has uniform lower bound $\min_{s\in\mathcal{S}}\mu_{\pi,p}(s)\ge\mu_{\min}$, and*

$$d_{\rm TV}\big[\mathbb{P}_{\pi,p,\rho}(s_{t_{mix}}=\cdot),\mu_{\pi,p}\big]\le\frac{1}{4}, \tag{70}$$

where $\mathbb{P}_{\pi,p,\rho}(s_{t_{mix}}=\cdot)$ denotes the state distribution at time t_{mix}, under the policy π, transition kernel p and initial state distribution ρ, and $d_{\rm TV}$ denotes the total variation distance between two probability distributions.

Proof of Assumption 4⇒Assumption 3: For any policy $\pi\in\Pi$, transition kernel $p\in\mathcal{P}$ and state $s\in\mathcal{S}$, we have

$$\begin{aligned}
d_{\pi,p}(s)=&\sum_a d_{\pi,p}(s,a)\\
\overset{(a)}{=}&\sum_a(1-\gamma)\sum_{t=0}^\infty\gamma^t\mathbb{P}_{\pi,p,\rho}\{s_t=s,a_t=a\}\\
=&(1-\gamma)\sum_{t=0}^\infty\gamma^t\mathbb{P}_{\pi,p,\rho}\{s_t=s\}\\
\ge&(1-\gamma)\mathbb{P}_{\pi,p,\rho}\{s_0=s\}\\
=&(1-\gamma)\rho(s)\\
\ge&(1-\gamma)\min_{s\in\mathcal{S}}\rho(s).
\end{aligned}$$

As $\mathcal{S}$ is a finite state space, $\rho(s)>0,\forall s\in\mathcal{S}$ implies that $\min_{s\in\mathcal{S}}\rho(s)>0$. Hence, Assumption 3 holds with $D=(1-\gamma)\min_{s\in\mathcal{S}}\rho(s)>0$.

Proof of Assumption 5⇒Assumption 3: If $\rho(s)=0$ for a state s, then Assumption 5 implies that $d_{\pi,p}(s)=(1-\gamma)\sum_{t=0}^\infty\gamma^t\mathbb{P}_{\pi,p,\rho}\{s_t=s\}=0$ for any $\pi\in\Pi$ and $p\in\mathcal{P}$, which means the state s will never be visited. Therefore, we can exclude all such states s from $\mathcal{S}$ such that Assumption 4 holds, which implies Assumption 3 as proved above.

Proof of Assumption 6⇒Assumption 3: Eq. (70) implies that for any $n\in\mathbb{N}_+$, we have

$$d_{\rm TV}\big[\mathbb{P}_{\pi,p,\rho}(s_{nt_{\rm mix}}=\cdot),\mu_{\pi,p}\big]=\frac{1}{2}\sum_s|\mathbb{P}_{\pi,p,\rho}\{s_{nt_{\rm mix}}=s\}-\mu_{\pi,p}(s)|\le\frac{1}{4^n}.$$

Select $n=\lceil\log(\mu_{\min}^{-1})/\log 4\rceil$. Then the bound above implies $|\mathbb{P}_{\pi,p,\rho}\{s_{nt_{\rm mix}}=s\}-\mu_{\pi,p}(s)|\le\mu_{\min}/2$ for any state s, which along with $\mu_{\pi,p}(s)\ge\mu_{\min}$ implies that $\mathbb{P}_{\pi,p,\rho}\{s_{nt_{\rm mix}}=s\}\ge\mu_{\min}/2$. Therefore, we can prove Assumption 3 as follows.

$$d_{\pi,p}(s)=(1-\gamma)\sum_{t=0}^\infty\gamma^t\mathbb{P}_{\pi,p,\rho}\{s_t=s\}\ge(1-\gamma)\gamma^{nt_{\rm mix}}\mathbb{P}_{\pi,p,\rho}\{s_{nt_{\rm mix}}=s\}\ge\frac{\mu_{\min}}{2}\gamma^{nt_{\rm mix}}(1-\gamma).$$

F Proof of Theorem 1

Fix any $\pi_0, \pi_1 \in \Pi$. For any $\alpha \in [0,1]$, denote $d_\alpha = \alpha d_{\pi_1, p_{\pi_1}} + (1-\alpha) d_{\pi_0, p_{\pi_0}}$, $\pi_\alpha(a|s) = \frac{d_\alpha(s,a)}{d_\alpha(s)}$ where $d_\alpha(s) = \sum_{a'} d_\alpha(s,a')$, and $p_\alpha = p_{\pi_\alpha}$. It can be easily verified that $d_0 = d_{\pi_0,p_0}$, $d_1 = d_{\pi_1,p_1}$ and $d_\alpha = \alpha d_0 + (1-\alpha) d_1$. Then we can obtain the following derivatives and their bounds about π_α, d_α in Eqs. (71)-(77).

$$\begin{aligned}
&\frac{d_\alpha(s)[d_1(s,a) - d_0(s,a)] - d_\alpha(s,a)[d_1(s) - d_0(s)]}{d_\alpha^2(s)} \\
=&\frac{[\alpha d_1(s) + (1-\alpha) d_0(s)][d_1(s,a) - d_0(s,a)] - [\alpha d_1(s,a) + (1-\alpha) d_0(s,a)][d_1(s) - d_0(s)]}{d_\alpha^2(s)} \\
=&\frac{d_0(s) d_1(s,a) - d_0(s,a) d_1(s)}{d_\alpha^2(s)} \\
=&\frac{d_0(s) d_1(s)[\pi_1(a|s) - \pi_0(a|s)]}{d_\alpha^2(s)}. \qquad (71)
\end{aligned}$$

Hence,

$$\begin{aligned}
\Big\|\frac{d\pi_\alpha}{d\alpha}\Big\|^2 =& \sum_{s,a} \Big|\frac{d_0(s) d_1(s)[\pi_1(a|s) - \pi_0(a|s)]}{d_\alpha^2(s)}\Big|^2 \\
\overset{(a)}{\le}& \sum_{s,a} \Big[\frac{\max[d_0(s), d_1(s)] \min[d_0(s), d_1(s)]}{\min^2[d_0(s), d_1(s)]}\Big]^2 [\pi_1(a|s) - \pi_0(a|s)]^2 \\
\overset{(b)}{\le}& D^{-2} \sum_{s,a} [\pi_1(a|s) - \pi_0(a|s)]^2 \le D^{-2} \|\pi_1 - \pi_0\|^2, \qquad (72)
\end{aligned}$$

where (a) uses $d_\alpha(s) = \alpha d_1(s) + (1-\alpha) d_0(s) \ge \min[d_0(s), d_1(s)]$ and (b) uses Assumption 3. Then by taking derivative of Eq. (71), we have

$$\frac{d^2}{d\alpha^2} \pi_\alpha(a|s) = -\frac{2 d_0(s) d_1(s)[\pi_1(a|s) - \pi_0(a|s)][d_1(s) - d_0(s)]}{d_\alpha^3(s)}. \qquad (73)$$

Hence,

$$\begin{aligned}
\Big\|\frac{d^2\pi_\alpha}{d\alpha^2}\Big\|^2 =& \sum_{s,a} \Big|\frac{2 d_0(s) d_1(s)[\pi_1(a|s) - \pi_0(a|s)][d_1(s) - d_0(s)]}{[\alpha d_1(s) + (1-\alpha) d_0(s)]^3}\Big|^2 \\
\overset{(a)}{\le}& \sum_{s,a} \Big[\frac{2 \max[d_0(s), d_1(s)] \min[d_0(s), d_1(s)] \big|d_1(s) - d_0(s)\big|}{D^2 \min[d_0(s), d_1(s)]}\Big]^2 [\pi_1(a|s) - \pi_0(a|s)]^2 \\
\le& (2D^{-2})^2 \max_s \big[|d_1(s) - d_0(s)|^2\big] \sum_{s,a} [\pi_1(a|s) - \pi_0(a|s)]^2 \\
\le& (2D^{-2})^2 \|\pi_1 - \pi_0\|^2 \Big[\sum_s |d_1(s) - d_0(s)|\Big]^2 \\
\overset{(b)}{\le}& (2D^{-2})^2 \|\pi_1 - \pi_0\|^2 \Big[\frac{\gamma\sqrt{|\mathcal{A}|}}{1-\gamma} \|\pi_1 - \pi_0\| + \frac{\gamma\sqrt{|\mathcal{S}|}}{1-\gamma} \|p_{\pi_1} - p_{\pi_0}\|\Big]^2 \\
\overset{(c)}{\le}& (2D^{-2})^2 \|\pi_1 - \pi_0\|^2 \Big[\frac{\gamma\sqrt{|\mathcal{A}|}}{1-\gamma} \|\pi_1 - \pi_0\| + \frac{\gamma\epsilon_p\sqrt{|\mathcal{S}|}}{1-\gamma} \|\pi_1 - \pi_0\|\Big]^2 \\
\le& (2D^{-2})^2 \|\pi_1 - \pi_0\|^4 \Big[\frac{\gamma(\epsilon_p\sqrt{|\mathcal{S}|} + \sqrt{|\mathcal{A}|})}{1-\gamma}\Big]^2, \qquad (74)
\end{aligned}$$

where (a) uses $d_\alpha(s) = \alpha d_1(s) + (1-\alpha) d_0(s) \ge \min[d_0(s), d_1(s)] \ge D$, (b) uses Lemma 3, and (c) uses Assumption 1.

$$d_0(s) d_1(s) \Big| \frac{d}{d\alpha} \Big[\frac{d_\alpha(s,a)}{d_\alpha^2(s)}\Big]\Big|$$

$$
\begin{aligned}
=&\Big|\frac{d_0(s)d_1(s)}{d_\alpha^2(s)}[d_1(s,a)-d_0(s,a)]-\frac{2d_0(s)d_1(s)d_\alpha(s,a)}{d_\alpha^3(s)}[d_1(s)-d_0(s)]\Big|\\
\le&\frac{d_0(s)d_1(s)}{d_\alpha^2(s)}\Big[|d_1(s,a)-d_0(s,a)|+\frac{2d_\alpha(s,a)}{d_\alpha(s)}|d_1(s)-d_0(s)|\Big]\\
\le&\frac{\max[d_0(s),d_1(s)]\min[d_0(s),d_1(s)]}{\min^2[d_0(s),d_1(s)]}\big[|d_1(s,a)-d_0(s,a)|+2\pi_\alpha(a|s)|d_1(s)-d_0(s)|\big]\\
\le&D^{-1}\big[|d_1(s,a)-d_0(s,a)|+2\pi_\alpha(a|s)|d_1(s)-d_0(s)|\big].
\end{aligned}\tag{75}
$$

$$
\begin{aligned}
&\frac{d}{d\alpha}[d_\alpha(s,a)p_\alpha(s'|s,a)]\\
=&p_\alpha(s'|s,a)[d_1(s,a)-d_0(s,a)]+d_\alpha(s,a)\cdot\frac{d}{d\alpha}\pi_\alpha(a|s)\cdot\nabla_\pi p_{\pi_\alpha}(s'|s,a)\\
=&p_\alpha(s'|s,a)[d_1(s,a)-d_0(s,a)]+\frac{d_\alpha(s,a)d_0(s)d_1(s)[\pi_1(a|s)-\pi_0(a|s)]}{d_\alpha^2(s)}\cdot\nabla_\pi p_{\pi_\alpha}(s'|s,a)
\end{aligned}\tag{76}
$$

Then for any $\alpha,\alpha'\in[0,1]$, we have

$$
\begin{aligned}
&\Big|\frac{d}{d\alpha}[d_{\alpha'}(s,a)p_{\alpha'}(s'|s,a)]-\frac{d}{d\alpha}[d_\alpha(s,a)p_\alpha(s'|s,a)]\Big|\\
\overset{(a)}{\le}&|p_{\alpha'}(s'|s,a)-p_\alpha(s'|s,a)|\cdot|d_1(s,a)-d_0(s,a)|+d_0(s)d_1(s)|\pi_1(a|s)-\pi_0(a|s)|\cdot\\
&\Big[\Big|\frac{d_{\alpha'}(s,a)}{d_{\alpha'}^2(s)}\Big|\|\nabla_\pi p_{\pi_{\alpha'}}(s'|s,a)-\nabla_\pi p_{\pi_\alpha}(s'|s,a)\|+\Big|\frac{d_{\alpha'}(s,a)}{d_{\alpha'}^2(s)}-\frac{d_\alpha(s,a)}{d_\alpha^2(s)}\Big|\|\nabla_\pi p_{\pi_\alpha}(s'|s,a)\|\Big]\\
\overset{(b)}{\le}&\epsilon_p\|\pi_{\alpha'}-\pi_\alpha\||d_1(s,a)-d_0(s,a)|\\
&+\pi_{\alpha'}(a|s)|\pi_1(a|s)-\pi_0(a|s)|\cdot\frac{\max[d_0(s),d_1(s)]\min[d_0(s),d_1(s)]}{\min[d_0(s),d_1(s)]}\cdot S_p\|\pi_{\alpha'}-\pi_\alpha\|\\
&+D^{-1}\epsilon_p|\pi_1(a|s)-\pi_0(a|s)|\cdot\big[|d_1(s,a)-d_0(s,a)|+2\pi_\alpha(a|s)|d_1(s)-d_0(s)|\big]\cdot|\alpha'-\alpha|\\
\overset{(c)}{\le}&\epsilon_pD^{-1}\|\pi_1-\pi_0\|\cdot|\alpha'-\alpha|\cdot|d_1(s,a)-d_0(s,a)|\\
&+S_p\pi_{\alpha'}(a|s)\cdot|\pi_1(a|s)-\pi_0(a|s)|\cdot[d_0(s)+d_1(s)]\cdot D^{-1}\|\pi_1-\pi_0\|\cdot|\alpha'-\alpha|\\
&+D^{-1}\epsilon_p|\pi_1(a|s)-\pi_0(a|s)|\cdot\big[|d_1(s,a)-d_0(s,a)|+2\pi_\alpha(a|s)|d_1(s)-d_0(s)|\big]\cdot|\alpha'-\alpha|\\
\overset{(d)}{\le}&\ell_{dp}(s,a)|\alpha'-\alpha|,
\end{aligned}\tag{77}
$$

where (a) uses Eq. (76), (b) uses Assumptions 1-2, $d_{\alpha'}(s,a)=d_{\alpha'}(s)\pi_{\alpha'}(a|s)$, $d_{\alpha'}(s)=\alpha'd_1(s)+(1-\alpha')d_0(s)\ge\min[d_0(s),d_1(s)]$ and Eq. (75), (c) uses Assumption 3 as well as Eq. (72), (d) defines $\ell_{dp}(s,a)$ as the following Eq. (78) and uses $\pi_\alpha(a|s)=\frac{\alpha d_1(s)\pi_1(a|s)+(1-\alpha)d_0(s)\pi_0(a|s)}{\alpha d_1(s)+(1-\alpha)d_0(s)}\le\pi_0(a|s)+\pi_1(a|s)$.

$$
\begin{aligned}
\ell_{dp}(s,a)=&2D^{-1}\epsilon_p\|\pi_1-\pi_0\||d_1(s,a)-d_0(s,a)|\\
&+2D^{-1}\epsilon_p[\pi_1(a|s)+\pi_0(a|s)]\cdot|\pi_1(a|s)-\pi_0(a|s)|\cdot|d_1(s)-d_0(s)|\\
&+D^{-1}S_p[\pi_1(a|s)+\pi_0(a|s)]\cdot|\pi_1(a|s)-\pi_0(a|s)|\cdot\|\pi_1-\pi_0\|\cdot[d_0(s)+d_1(s)].
\end{aligned}\tag{78}
$$

Denote $e_\alpha(s)=d_{\pi_\alpha,p_\alpha}(s)-d_\alpha(s)$ as the error term due to the policy-dependent transition kernel $p_\alpha=p_{\pi_\alpha}$[1]. Note that the occupancy measure (2) satisfies that the Bellman equation (3) repeated as follows.

$$
d_{\pi,p}(s')=(1-\gamma)\rho(s')+\gamma\sum_{s,a}d_{\pi,p}(s)\pi(a|s)p(s'|s,a),\quad s'\in\mathcal{S}.\tag{79}
$$

Therefore, the error term $e_\alpha(s)$ satisfies the following recursion.

$$
e_\alpha(s')
$$

[1] If $p_{\pi_\alpha}\equiv p$ does not depend on the policy π_α, it can be easily verified that $e_\alpha(s)=0$ for all $s\in\mathcal{S}$.

$$
\begin{aligned}
=&d_{\pi_\alpha,p_\alpha}(s')-\alpha d_1(s')-(1-\alpha)d_0(s')\\
=&\gamma\sum_{s,a}[d_{\pi_\alpha,p_\alpha}(s)\pi_\alpha(a|s)p_\alpha(s'|s,a)-\alpha d_{\pi_1,p_1}(s)\pi_1(a|s)p_1(s'|s,a)\\
&-(1-\alpha)d_{\pi_0,p_0}(s)\pi_0(a|s)p_0(s'|s,a)]\\
=&\gamma\sum_{s,a}[e_\alpha(s)\pi_\alpha(a|s)p_\alpha(s'|s,a)+d_\alpha(s,a)p_\alpha(s'|s,a)-\alpha d_1(s,a)p_1(s'|s,a)\\
&-(1-\alpha)d_0(s,a)p_0(s'|s,a)].
\end{aligned}
\tag{80}
$$

The above inequality implies that

$$
\begin{aligned}
&\sum_{s'}|e_\alpha(s')|\\
\le&\gamma\sum_{s,a,s'}\big[|e_\alpha(s)|\pi_\alpha(a|s)p_\alpha(s'|s,a)\\
&+|d_\alpha(s,a)p_\alpha(s'|s,a)-\alpha d_1(s,a)p_1(s'|s,a)-(1-\alpha)d_0(s,a)p_0(s'|s,a)|\big]\\
\overset{(a)}{\le}&\gamma\sum_s|e_\alpha(s)|+\frac{\gamma\alpha(1-\alpha)}{2}\sum_{s,a,s'}\ell_{dp}(s,a)\\
\overset{(b)}{\le}&\gamma\sum_s|e_\alpha(s)|+\frac{\gamma|\mathcal{S}|\alpha(1-\alpha)}{2}\Big[2D^{-1}\epsilon_p\|\pi_1-\pi_0\|\sum_{s,a}|d_1(s,a)-d_0(s,a)|\\
&+4D^{-1}\epsilon_p\|\pi_1-\pi_0\|_\infty\sum_s|d_1(s)-d_0(s)|+4D^{-1}S_p\|\pi_1-\pi_0\|_\infty\cdot\|\pi_1-\pi_0\|\Big]\\
\overset{(c)}{\le}&\gamma\sum_s|e_\alpha(s)|+\frac{\gamma|\mathcal{S}|\alpha(1-\alpha)}{2}\Big[6D^{-1}\epsilon_p\|\pi_1-\pi_0\|\cdot\frac{1}{1-\gamma}\Big(\sqrt{|\mathcal{A}|}\|\pi_1-\pi_0\|+\gamma\sqrt{|\mathcal{S}|}\|p_{\pi_1}-p_{\pi_0}\|\Big)\\
&+4D^{-1}S_p\|\pi_1-\pi_0\|^2\Big]\\
\overset{(d)}{\le}&\gamma\sum_s|e_\alpha(s)|+3D^{-1}\gamma|\mathcal{S}|\alpha(1-\alpha)\|\pi_1-\pi_0\|^2\Big[\frac{\epsilon_p}{1-\gamma}(\sqrt{|\mathcal{A}|}+\gamma\epsilon_p\sqrt{|\mathcal{S}|})+S_p\Big],
\end{aligned}
$$

where (a) uses Eq. (77) which implies that $d_\alpha(s,a)p_\alpha(s'|s,a)$ is a Lipschitz smooth function with Lipschitz constant $\ell_{dp}(s,a)$ defined by Eq. (78), (b) uses Eq. (78), (c) uses $\|\pi_1-\pi_0\|_\infty\le\|\pi_1-\pi_0\|$ and Lemma 3, and (d) uses Assumption 1. Rearranging the above inequality, we get

$$
\sum_s|e_\alpha(s)|\le\frac{3\gamma|\mathcal{S}|\alpha(1-\alpha)}{D(1-\gamma)^2}\|\pi_1-\pi_0\|^2\big[\epsilon_p\big(\sqrt{|\mathcal{A}|}+\gamma\epsilon_p\sqrt{|\mathcal{S}|}\big)+S_p(1-\gamma)\big].
\tag{81}
$$

Therefore, for any reward function r, we have

$$
\begin{aligned}
&J_\lambda(\pi_\alpha,\pi_\alpha,p_\alpha,r)-\alpha J_\lambda(\pi_1,\pi_1,p_1,r)-(1-\alpha)J_\lambda(\pi_0,\pi_0,p_0,r)\\
\overset{(a)}{=}&\frac{1}{1-\gamma}\sum_{s,a}\Big[d_{\pi_\alpha,p_\alpha}(s,a)[r(s,a)-\lambda\log\pi_\alpha(a|s)]-\alpha d_1(s,a)[r(s,a)-\lambda\log\pi_1(a|s)]\\
&-(1-\alpha)d_0(s,a)[r(s,a)-\lambda\log\pi_0(a|s)]\Big]\\
=&\frac{1}{1-\gamma}\sum_{s,a}\Big[[d_{\pi_\alpha,p_\alpha}(s,a)-d_\alpha(s,a)][r(s,a)-\lambda\log\pi_\alpha(a|s)]+d_\alpha(s,a)[r(s,a)-\lambda\log\pi_\alpha(a|s)]\\
&-\alpha d_1(s,a)[r(s,a)-\lambda\log\pi_1(a|s)]-(1-\alpha)d_0(s,a)[r(s,a)-\lambda\log\pi_0(a|s)]\Big]\\
\overset{(b)}{=}&\frac{1}{1-\gamma}\sum_{s,a}[d_{\pi_\alpha,p_\alpha}(s)-d_\alpha(s)]\pi_\alpha(a|s)[r(s,a)-\lambda\log\pi_\alpha(a|s)]\\
&+\frac{\lambda}{1-\gamma}\sum_{s,a}\Big[\alpha d_1(s,a)\log\frac{\pi_1(a|s)}{\pi_\alpha(a|s)}+(1-\alpha)d_0(s,a)\log\frac{\pi_0(a|s)}{\pi_\alpha(a|s)}\Big]
\end{aligned}
$$

$$\overset{(c)}{\geq} -\frac{1+\lambda\log|\mathcal{A}|}{1-\gamma}\sum_s|e_\alpha(s)|$$
$$+\frac{\lambda}{1-\gamma}\sum_s\Big[\alpha d_1(s)\sum_a\Big(\pi_1(a|s)\log\frac{\pi_1(a|s)}{\pi_\alpha(a|s)}\Big)+(1-\alpha)d_0(s)\sum_a\Big(\pi_0(a|s)\log\frac{\pi_0(a|s)}{\pi_\alpha(a|s)}\Big)\Big]$$
$$\overset{(d)}{\geq} -\frac{1+\lambda\log|\mathcal{A}|}{1-\gamma}\frac{3\gamma|\mathcal{S}|\alpha(1-\alpha)}{D(1-\gamma)^2}\|\pi_1-\pi_0\|^2\big[\epsilon_p\big(\sqrt{|\mathcal{A}|}+\gamma\epsilon_p\sqrt{|\mathcal{S}|}\big)+S_p(1-\gamma)\big]$$
$$+\frac{\lambda}{1-\gamma}\sum_s\Big[\alpha d_1(s)\mathrm{KL}[\pi_1(\cdot|s)\|\pi_\alpha(\cdot|s)]+(1-\alpha)d_0(s)\mathrm{KL}[\pi_0(\cdot|s)\|\pi_\alpha(\cdot|s)]\Big]$$
$$\overset{(e)}{\geq} -\frac{3\gamma|\mathcal{S}|\alpha(1-\alpha)(1+\lambda\log|\mathcal{A}|)}{D(1-\gamma)^3}\|\pi_1-\pi_0\|^2\big[\epsilon_p\big(\sqrt{|\mathcal{A}|}+\gamma\epsilon_p\sqrt{|\mathcal{S}|}\big)+S_p(1-\gamma)\big]$$
$$+\frac{\lambda}{2(1-\gamma)}\sum_s\Big[\alpha d_1(s)\|\pi_1(\cdot|s)-\pi_\alpha(\cdot|s)\|_1^2+(1-\alpha)d_0(s)\|\pi_0(\cdot|s)-\pi_\alpha(\cdot|s)\|_1^2\Big]$$
$$\overset{(f)}{=} -\frac{3\gamma|\mathcal{S}|\alpha(1-\alpha)(1+\lambda\log|\mathcal{A}|)}{D(1-\gamma)^3}\|\pi_1-\pi_0\|^2\big[\epsilon_p\big(\sqrt{|\mathcal{A}|}+\gamma\epsilon_p\sqrt{|\mathcal{S}|}\big)+S_p(1-\gamma)\big]$$
$$+\frac{\lambda}{2(1-\gamma)}\sum_s\Big[\alpha d_1(s)\Big\|\frac{(1-\alpha)d_0(s)}{d_\alpha(s)}[\pi_1(\cdot|s)-\pi_0(\cdot|s)]\Big\|_1^2$$
$$+(1-\alpha)d_0(s)\Big\|\frac{\alpha d_1(s)}{d_\alpha(s)}[\pi_1(\cdot|s)-\pi_0(\cdot|s)]\Big\|_1^2\Big]$$
$$\overset{(g)}{=}\frac{\lambda\alpha(1-\alpha)}{2(1-\gamma)}\sum_s\frac{d_0(s)d_1(s)}{d_\alpha(s)}\|\pi_1(\cdot|s)-\pi_0(\cdot|s)\|_1^2$$
$$-\frac{3\gamma|\mathcal{S}|\alpha(1-\alpha)(1+\lambda\log|\mathcal{A}|)}{D(1-\gamma)^3}\|\pi_1-\pi_0\|^2\big[\epsilon_p\big(\sqrt{|\mathcal{A}|}+\gamma\epsilon_p\sqrt{|\mathcal{S}|}\big)+S_p(1-\gamma)\big]$$
$$\overset{(h)}{\geq}\frac{D\lambda\alpha(1-\alpha)}{2(1-\gamma)}\|\pi_1-\pi_0\|^2$$
$$-\frac{3\gamma|\mathcal{S}|\alpha(1-\alpha)(1+\lambda\log|\mathcal{A}|)}{D(1-\gamma)^3}\|\pi_1-\pi_0\|^2\big[\epsilon_p\big(\sqrt{|\mathcal{A}|}+\gamma\epsilon_p\sqrt{|\mathcal{S}|}\big)+S_p(1-\gamma)\big]$$
$$\overset{(i)}{=}\frac{\mu_1\alpha(1-\alpha)}{2}\|\pi_1-\pi_0\|^2, \tag{82}$$

where (a) uses Eq. (31), (b) uses $d_{\pi_\alpha,p_\alpha}(s,a)=d_{\pi_\alpha,p_\alpha}(s)\pi_\alpha(a|s)$, $d_\alpha(s,a)=d_\alpha(s)\pi_\alpha(a|s)$ and $d_\alpha=\alpha d_1+(1-\alpha)d_0$, (c) uses $r(s,a)\in[0,1]$, $-\sum_a\pi_\alpha(a|s)\log\pi_\alpha(a|s)\in[0,\log|\mathcal{A}|]$ and $e_\alpha(s)=d_{\pi_\alpha,p_\alpha}(s)-d_\alpha(s)$, (d) uses Eq. (22), (e) uses Pinsker's inequality, (f) uses $\pi_\alpha(a|s)=\frac{d_\alpha(s,a)}{d_\alpha(s)}=\frac{\alpha d_1(s)}{d_\alpha(s)}\pi_1(a|s)+\frac{(1-\alpha)d_0(s)}{d_\alpha(s)}\pi_0(a|s)$, (g) uses $d_\alpha(s)=\alpha d_1(s)+(1-\alpha)d_0(s)$, (h) uses Assumption 3 and $d_\alpha(s)\leq\max[d_0(s),d_1(s)]$, and (i) defines the constant μ_1 below.

$$\mu_1\overset{\text{def}}{=}\frac{D\lambda}{1-\gamma}-\frac{6\gamma|\mathcal{S}|(1+\lambda\log|\mathcal{A}|)}{D(1-\gamma)^3}\big[\epsilon_p\big(\sqrt{|\mathcal{A}|}+\gamma\epsilon_p\sqrt{|\mathcal{S}|}\big)+S_p(1-\gamma)\big]. \tag{83}$$

Next, we begin to consider the policy-dependent reward $r_\alpha=r_{\pi_\alpha}$. Define the function $w(\alpha)=\alpha J_\lambda(\pi_1,\pi_1,p_1,r_\alpha)+(1-\alpha)J_\lambda(\pi_0,\pi_0,p_0,r_\alpha)$, which has the following derivative

$$w'(\alpha)=J_\lambda(\pi_1,\pi_1,p_1,r_\alpha)-J_\lambda(\pi_0,\pi_0,p_0,r_\alpha)$$
$$+[\alpha\nabla_rJ_\lambda(\pi_1,\pi_1,p_1,r_\alpha)+(1-\alpha)\nabla_rJ_\lambda(\pi_0,\pi_0,p_0,r_\alpha)](\nabla_\pi r_{\pi_\alpha})\frac{d\pi_\alpha}{d\alpha} \tag{84}$$

For any $0\leq\alpha\leq\alpha'\leq1$, we prove the smoothness of $w(\alpha)$ as follows.

$$|w'(\alpha')-w'(\alpha)|$$
$$=\Big|\int_\alpha^{\alpha'}\nabla_r[J_\lambda(\pi_1,\pi_1,p_1,r_{\tilde\alpha})-J_\lambda(\pi_0,\pi_0,p_0,r_{\tilde\alpha})](\nabla_\pi r_{\pi_{\tilde\alpha}})\frac{d\pi_{\tilde\alpha}}{d\tilde\alpha}d\tilde\alpha$$
$$+[\alpha'\nabla_rJ_\lambda(\pi_1,\pi_1,p_1,r_{\alpha'})+(1-\alpha')\nabla_rJ_\lambda(\pi_0,\pi_0,p_0,r_{\alpha'})](\nabla_\pi r_{\pi_{\alpha'}})\Big(\frac{d\pi_{\alpha'}}{d\alpha'}-\frac{d\pi_\alpha}{d\alpha}\Big)$$

$$
\begin{aligned}
&+[\alpha'\nabla_r J_\lambda(\pi_1,\pi_1,p_1,r_{\alpha'})+(1-\alpha')\nabla_r J_\lambda(\pi_0,\pi_0,p_0,r_{\alpha'})](\nabla_\pi r_{\pi_{\alpha'}}-\nabla_\pi r_{\pi_\alpha})\frac{d\pi_\alpha}{d\alpha}\\
&+\big\{\alpha'[\nabla_r J_\lambda(\pi_1,\pi_1,p_1,r_{\alpha'})-\nabla_r J_\lambda(\pi_1,\pi_1,p_1,r_\alpha)]\\
&+(1-\alpha')[\nabla_r J_\lambda(\pi_0,\pi_0,p_0,r_{\alpha'})-\nabla_r J_\lambda(\pi_0,\pi_0,p_0,r_\alpha)]\big\}(\nabla_\pi r_{\pi_\alpha})\frac{d\pi_\alpha}{d\alpha}\\
&+(\alpha'-\alpha)[\nabla_r J_\lambda(\pi_1,\pi_1,p_1,r_\alpha)-\nabla_r J_\lambda(\pi_0,\pi_0,p_0,r_\alpha)](\nabla_\pi r_{\pi_\alpha})\frac{d\pi_\alpha}{d\alpha}\Big|\\
&\overset{(a)}{\le}\int_\alpha^{\alpha'}\frac{\epsilon_r\|\pi_1-\pi_0\|}{D(1-\gamma)^2}\big(\max_s\|\pi_1(\cdot|s)-\pi_0(\cdot|s)\|_1+\gamma\max_{s,a}\|p_1(\cdot|s,a)-p_0(\cdot|s,a)\|_1\big)d\tilde{\alpha}\\
&+\frac{\epsilon_r}{1-\gamma}\cdot 2D^{-2}\|\pi_1-\pi_0\|^2\Big[\frac{\gamma(\epsilon_p\sqrt{|\mathcal{S}|}+\sqrt{|\mathcal{A}|})}{1-\gamma}\Big]|\alpha'-\alpha|+\frac{S_r\|\pi_{\alpha'}-\pi_\alpha\|}{1-\gamma}\cdot D^{-1}\|\pi_1-\pi_0\|\\
&+0+|\alpha'-\alpha|\cdot\frac{\epsilon_r\|\pi_1-\pi_0\|}{D(1-\gamma)^2}\big(\max_s\|\pi_1(\cdot|s)-\pi_0(\cdot|s)\|_1+\gamma\max_{s,a}\|p_1(\cdot|s,a)-p_0(\cdot|s,a)\|_1\big)\\
&\overset{(b)}{\le}2|\alpha'-\alpha|\cdot\frac{\epsilon_r\|\pi_1-\pi_0\|}{D(1-\gamma)^2}\big(\sqrt{|\mathcal{A}|}\|\pi_1-\pi_0\|+\gamma\sqrt{|\mathcal{S}|}\|p_1-p_0\|\big)\\
&+\frac{2\epsilon_r\|\pi_1-\pi_0\|^2}{D^2(1-\gamma)}\Big[\frac{\gamma(\epsilon_p\sqrt{|\mathcal{S}|}+\sqrt{|\mathcal{A}|})}{1-\gamma}\Big]|\alpha'-\alpha|+\frac{S_r\|\pi_1-\pi_0\|^2}{D^2(1-\gamma)}|\alpha'-\alpha|\\
&\overset{(c)}{\le}\frac{2\epsilon_r\|\pi_1-\pi_0\|}{D(1-\gamma)^2}\big(\sqrt{|\mathcal{A}|}\|\pi_1-\pi_0\|+\gamma\epsilon_p\sqrt{|\mathcal{S}|}\|\pi_1-\pi_0\|\big)|\alpha'-\alpha|\\
&+\frac{2\gamma\epsilon_r\|\pi_1-\pi_0\|^2}{D^2(1-\gamma)^2}\big(\sqrt{|\mathcal{A}|}+\epsilon_p\sqrt{|\mathcal{S}|}\big)|\alpha'-\alpha|+\frac{S_r(1-\gamma)\|\pi_1-\pi_0\|^2}{D^2(1-\gamma)^2}|\alpha'-\alpha|\\
&\overset{(d)}{\le}\frac{4\epsilon_r(\sqrt{|\mathcal{A}|}+\gamma\epsilon_p\sqrt{|\mathcal{S}|})+S_r(1-\gamma)}{D^2(1-\gamma)^2}\|\pi_1-\pi_0\|^2|\alpha'-\alpha|,
\end{aligned}
$$

where (a) uses Assumptions 1-2, $\|\nabla_r J_\lambda(\cdot,\cdot,\cdot,\cdot)\|\le\frac{1}{1-\gamma}$ (implied by Eq. (41)) as well as Eqs. (45), (72) and (74), (b) uses Eq. (72) and $\|x\|_1\le\sqrt{d}\|x\|$ for any $x\in\mathbb{R}^d$, (c) uses Assumption 1, and (d) uses $D,\gamma\in[0,1]$. The inequality above implies that $w(\alpha)$ is $\mu_2\|\pi_1-\pi_0\|^2$-Lipschitz smooth with the constant μ_2 defined as follows.

$$\mu_2=\frac{4\epsilon_r(\sqrt{|\mathcal{A}|}+\epsilon_p\sqrt{|\mathcal{S}|})+S_r(1-\gamma)}{D^2(1-\gamma)^2} \tag{85}$$

Therefore,

$$
\begin{aligned}
&V_{\lambda,\pi_\alpha}^{\pi_\alpha}-\alpha V_{\lambda,\pi_1}^{\pi_1}-(1-\alpha)V_{\lambda,\pi_0}^{\pi_0}\\
=&J_\lambda(\pi_\alpha,\pi_\alpha,p_\alpha,r_\alpha)-\alpha J_\lambda(\pi_1,\pi_1,p_1,r_1)-(1-\alpha)J_\lambda(\pi_0,\pi_0,p_0,r_0)\\
\overset{(a)}{\ge}&\alpha J_\lambda(\pi_1,\pi_1,p_1,r_\alpha)+(1-\alpha)J_\lambda(\pi_0,\pi_0,p_0,r_\alpha)+\frac{\mu_1\alpha(1-\alpha)}{2}\|\pi_1-\pi_0\|^2\\
&-\alpha J_\lambda(\pi_1,\pi_1,p_1,r_1)-(1-\alpha)J_\lambda(\pi_0,\pi_0,p_0,r_0)\\
=&w(\alpha)-\alpha w(1)-(1-\alpha)w(0)+\frac{\mu_1\alpha(1-\alpha)}{2}\|\pi_1-\pi_0\|^2\\
\overset{(b)}{\ge}&\frac{(\mu_1-\mu_2)\alpha(1-\alpha)}{2}\|\pi_1-\pi_0\|^2\\
\overset{(c)}{=}&\frac{\mu\alpha(1-\alpha)}{2}\|\pi_1-\pi_0\|^2,
\end{aligned} \tag{86}
$$

where (a) uses Eq. (82) with r replaced by r_α, (b) uses the fact proved above that $w(\alpha)$ is $\mu_2\|\pi_1-\pi_0\|^2$-Lipschitz smooth, and (c) defines the following constant μ.

$$
\begin{aligned}
\mu\overset{\text{def}}{=}&\mu_1-\mu_2\\
\overset{(a)}{=}&\frac{D\lambda}{1-\gamma}-\frac{6\gamma|\mathcal{S}|(1+\lambda\log|\mathcal{A}|)}{D(1-\gamma)^3}\big[\epsilon_p\big(\sqrt{|\mathcal{A}|}+\gamma\epsilon_p\sqrt{|\mathcal{S}|}\big)+S_p(1-\gamma)\big]
\end{aligned}
$$

$$-\frac{S_r(1-\gamma)+4\epsilon_r(\sqrt{|\mathcal{A}|}+\epsilon_p\sqrt{|\mathcal{S}|})}{D^2(1-\gamma)^2}, \tag{87}$$

where (a) uses Eqs. (83) and (85). Rearranging Eq. (86), we obtain that

$$\frac{V_{\lambda,\pi_\alpha}^{\pi_\alpha}-V_{\lambda,\pi_0}^{\pi_0}}{\alpha}\ge V_{\lambda,\pi_1}^{\pi_1}-V_{\lambda,\pi_0}^{\pi_0}+\frac{\mu(1-\alpha)}{2}\|\pi_1-\pi_0\|^2.$$

Letting $\alpha\to+0$ above, we can prove the conclusion as follows.

$$\begin{aligned}
&V_{\lambda,\pi_1}^{\pi_1}-V_{\lambda,\pi_0}^{\pi_0}+\frac{\mu}{2}\|\pi_1-\pi_0\|^2\\
\le&\Big[\frac{d}{d\alpha}V_{\lambda,\pi_\alpha}^{\pi_\alpha}\Big]\Big|_{\alpha=0}\\
\le&\sum_{s,a}\frac{\partial V_{\lambda,\pi_0}^{\pi_0}}{\partial\pi_0(s,a)}\Big[\frac{d}{d\alpha}\pi_\alpha(a|s)\Big]\Big|_{\alpha=0}\\
\overset{(a)}{=}&\sum_s\frac{d_1(s)}{d_0(s)}\sum_a\frac{\partial V_{\lambda,\pi_0}^{\pi_0}}{\partial\pi_0(s,a)}[\pi_1(a|s)-\pi_0(a|s)]\\
\le&\sum_s\frac{d_1(s)}{d_0(s)}\Big[\max_{a'}\frac{\partial V_{\lambda,\pi_0}^{\pi_0}}{\partial\pi_0(s,a')}-\sum_a\pi_0(a|s)\frac{\partial V_{\lambda,\pi_0}^{\pi_0}}{\partial\pi_0(s,a)}\Big]\\
\overset{(b)}{\le}&D^{-1}\sum_{s,a}\frac{\partial V_{\lambda,\pi_0}^{\pi_0}}{\partial\pi_0(s,a)}[\pi_0^*(a|s)-\pi_0(a|s)]\\
\le&D^{-1}\max_{\pi\in\Pi}\big\langle\nabla_{\pi_0}V_{\lambda,\pi_0}^{\pi_0},\pi-\pi_0\big\rangle,
\end{aligned}$$

where (a) uses Eq. (71), and (b) uses Assumption 3 as well as the following Eq. (88) where $\pi_0^*\in\Pi$ is defined as $\pi_0^*(a^*|s)=1$ for a certain $a^*\in\arg\max_{a'}\frac{\partial V_{\lambda,\pi_0}^{\pi_0}}{\partial\pi_0(s,a')}$ and $\pi_0^*(a'|s)=0$ for $a'\neq a^*$.

$$\sum_a\pi_0^*(a|s)\frac{\partial V_{\lambda,\pi_0}^{\pi_0}}{\partial\pi_0(s,a)}=\max_{a'}\frac{\partial V_{\lambda,\pi_0}^{\pi_0}}{\partial\pi_0(s,a')}\ge\sum_a\pi_0(a|s)\frac{\partial V_{\lambda,\pi_0}^{\pi_0}}{\partial\pi_0(s,a)}. \tag{88}$$

G PROOF OF COROLLARY 1

Based on Theorem 1, Eq. (87) holds for any $\pi_0,\pi_1\in\Pi$ as repeated below.

$$V_{\lambda,\pi_1}^{\pi_1}\le V_{\lambda,\pi_0}^{\pi_0}+D^{-1}\max_{\pi\in\Pi}\big\langle\nabla_{\pi_0}V_{\lambda,\pi_0}^{\pi_0},\pi-\pi_0\big\rangle-\frac{\mu}{2}\|\pi_1-\pi_0\|^2, \tag{89}$$

In the above inequality, let $\pi_1\in\arg\max_{\pi\in\Pi}V_{\lambda,\pi}^{\pi}$ and $\pi_0=\pi$ is any a $D\epsilon$-stationary policy of interest. Then the inequality above becomes

$$\max_{\tilde\pi\in\Pi}V_{\lambda,\tilde\pi}^{\tilde\pi}\le V_{\lambda,\pi}^{\pi}+D^{-1}\cdot D\epsilon-\frac{\mu}{2}\|\pi_1-\pi\|^2\overset{\text{(a)}}{\le}V_{\lambda,\pi}^{\pi}+\epsilon+|\mu||\mathcal{S}|,$$

where (a) uses Lemma 12. This implies that $\max_{\tilde\pi\in\Pi}V_{\lambda,\tilde\pi}^{\tilde\pi}-V_{\lambda,\pi}^{\pi}\le\epsilon+|\mu||\mathcal{S}|$, that is, the $D\epsilon$-stationary policy π is also an $(\epsilon+|\mu||\mathcal{S}|)$-PO policy.

If $\mu\ge0$, the inequality above further implies that $\max_{\tilde\pi\in\Pi}V_{\lambda,\tilde\pi}^{\tilde\pi}-V_{\lambda,\pi}^{\pi}\le\epsilon$, that is, the $D\epsilon$-stationary policy π is also an ϵ-PO policy.

Furthermore, suppose $\mu>0$ and there are two PO policies $\pi_0,\pi_1\in\Pi$, which should satisfy

$$V_{\lambda,\pi_1}^{\pi_1}=V_{\lambda,\pi_0}^{\pi_0}=\max_{\pi\in\Pi}V_{\lambda,\pi}^{\pi},$$
$$\max_{\pi\in\Pi}\big\langle\nabla_{\pi_0}V_{\lambda,\pi_0}^{\pi_0},\pi-\pi_0\big\rangle=0.$$

Substituting the two equalities above into Eq. (10), we obtain that $\frac{\mu}{2}\|\pi_1-\pi_0\|^2\le0$, which along with $\mu>0$ implies $\pi_1=\pi_0$, that is, the PO policy is unique.

H PROOF OF THEOREM 2

For any $\pi \in \Pi$, $p \in \mathcal{P}$, $r \in \mathcal{R}$, we have

$$\begin{aligned}\frac{\partial J_\lambda(\pi,\pi,p,r)}{\partial \pi(a|s)} &\overset{(a)}{=} \frac{d_{\pi,p}(s)[Q_\lambda(\pi,\pi,p,r;s,a)-\lambda]}{1-\gamma}\\ &\overset{(b)}{=} \frac{d_{\pi,p}(s)}{1-\gamma}\Big[r(s,a)-\lambda-\lambda\log\pi(a|s)+\gamma\sum_{s'}p(s'|s,a)V_\lambda(\pi,p,r;s')\Big],\end{aligned} \tag{90}$$

where (a) uses Eqs. (38), and (b) uses Eq. (33).

Then we have

$$\begin{aligned}&\nabla_\pi J_\lambda(\pi,\pi,p,r)^\top(\pi'-\pi)\\ =&\sum_s\Big[\frac{\partial J_\lambda(\pi,\pi,p,r)}{\partial\pi[a_{\max}(s)|s]}\big(\pi'[a_{\max}(s)|s]-\pi[a_{\max}(s)|s]\big)\\ &+\frac{\partial J_\lambda(\pi,\pi,p,r)}{\partial\pi[a_{\min}(s)|s]}\big(\pi'[a_{\min}(s)|s]-\pi[a_{\min}(s)|s]\big)\Big]\\ =&\sum_s\Big\{\frac{d_{\pi,p}(s)}{1-\gamma}\big(\pi[a_{\max}(s)|s]-\pi[a_{\min}(s)|s]\big)\Big[r[s,a_{\min}(s)]-r[s,a_{\max}(s)]\\ &+\lambda\log\frac{\pi[a_{\max}(s)|s]}{\pi[a_{\min}(s)|s]}+\gamma\sum_{s'}[p(s'|s,a_{\min}(s))-p(s'|s,a_{\max}(s))]V_\lambda(\pi,p,r;s')\Big]\Big\}\\ \overset{(a)}{\ge}&\frac{1}{1-\gamma}\max_s\Big\{\big(\pi[a_{\max}(s)|s]-\pi[a_{\min}(s)|s]\big)\Big[\lambda\log\frac{\pi[a_{\max}(s)|s]}{\pi[a_{\min}(s)|s]}-1-\frac{\gamma(1+\lambda\log|\mathcal{A}|)}{1-\gamma}\Big]\Big\},\end{aligned} \tag{91}$$

where (a) uses $\pi[a_{\max}(s)|s]-\pi[a_{\min}(s)|s]\ge 0$, $r(a|s)\in[0,1]$, $p(s'|s,a)\in[0,1]$ for any s,a,s' and Lemma 4.

Consider the following two cases.

(Case I) If $\pi[a_{\min}(s)|s]\ge\frac{1}{2}\pi[a_{\max}(s)|s]$, then as $\pi[a_{\max}(s)|s]\ge\frac{1}{|\mathcal{A}|}$, we have $\pi[a_{\min}(s)|s]\ge\frac{1}{2|\mathcal{A}|}$.

(Case II) $\pi[a_{\min}(s)|s]<\frac{1}{2}\pi[a_{\max}(s)|s]$, then as $\pi[a_{\max}(s)|s]\ge\frac{1}{|\mathcal{A}|}$, Eq. (91) implies that

$$\begin{aligned}&\nabla_\pi J_\lambda(\pi,\pi,p,r)^\top(\pi'-\pi)\\ \ge&\max_s\Big\{\frac{\pi[a_{\max}(s)|s]}{2(1-\gamma)}\Big[\lambda\log\frac{1}{|\mathcal{A}|\pi[a_{\min}(s)|s]}-\frac{1+\gamma\lambda\log|\mathcal{A}|}{1-\gamma}\Big]\Big\}\\ \ge&-\frac{1}{2|\mathcal{A}|(1-\gamma)}\Big[\lambda\log\big(|\mathcal{A}|\min_s\pi[a_{\min}(s)|s]\big)+\frac{1+\gamma\lambda\log|\mathcal{A}|}{1-\gamma}\Big],\end{aligned} \tag{92}$$

which further implies that for any $s\in\mathcal{S}$ and $a\in\mathcal{A}$, we have

$$\begin{aligned}\pi(a|s)\ge&\pi[a_{\min}(s)|s]\\ \ge&\frac{1}{|\mathcal{A}|}\exp\Big[-\frac{1/\lambda+\gamma\log|\mathcal{A}|}{1-\gamma}-\frac{2|\mathcal{A}|}{\lambda}(1-\gamma)\nabla_\pi J_\lambda(\pi,\pi,p,r)^\top(\pi'-\pi)\Big]\\ \ge&\frac{1}{2|\mathcal{A}|^{1/(1-\gamma)}}\exp\Big[-\frac{1}{\lambda(1-\gamma)}-\frac{2|\mathcal{A}|}{\lambda}(1-\gamma)\nabla_\pi J_\lambda(\pi,\pi,p,r)^\top(\pi'-\pi)\Big],\end{aligned} \tag{93}$$

Note that in the two cases above, Eq. (93) always holds.

Furthermore, if Assumption 1 holds and p_π, r_π are differentiable functions of π, then we have

$$\begin{aligned}&\big\|\nabla_\pi J_\lambda(\pi,\pi,p_\pi,r_\pi)-\nabla_\pi J_\lambda(\pi,\pi,p_{\tilde\pi},r_{\tilde\pi})|_{\tilde\pi=\pi}\big\|\\ =&\big\|\nabla_p J_\lambda(\pi,\pi,p_\pi,r_\pi)\nabla_\pi p_\pi+\nabla_r J_\lambda(\pi,\pi,p_\pi,r_\pi)\nabla_\pi r_\pi\big\|\\ \le&\big\|\nabla_p J_\lambda(\pi,\pi,p_\pi,r_\pi)\big\|\big\|\nabla_\pi p_\pi\big\|+\big\|\nabla_r J_\lambda(\pi,\pi,p_\pi,r_\pi)\big\|\big\|\nabla_\pi r_\pi\big\|\\ \overset{(a)}{\le}&\frac{\epsilon_p\sqrt{|\mathcal{S}|}(1+\lambda\log|\mathcal{A}|)}{(1-\gamma)^2}+\frac{\epsilon_r}{1-\gamma},\end{aligned} \tag{94}$$

where (a) uses Assumption 1 as well as Eqs. (40) and (41). Therefore,

$$\begin{aligned}
&\big[\nabla_\pi J_\lambda(\pi,\pi,p_{\tilde{\pi}},r_{\tilde{\pi}})|_{\tilde{\pi}=\pi}\big]^\top(\pi'-\pi)\\
=&\nabla_\pi J_\lambda(\pi,\pi,p_\pi,r_\pi)^\top(\pi'-\pi)-\big[\nabla_\pi J_\lambda(\pi,\pi,p_\pi,r_\pi)-\nabla_\pi J_\lambda(\pi,\pi,p_{\tilde{\pi}},r_{\tilde{\pi}})|_{\tilde{\pi}=\pi}\big]^\top(\pi'-\pi)\\
\le&\nabla_\pi J_\lambda(\pi,\pi,p_\pi,r_\pi)^\top(\pi'-\pi)+\big\|\nabla_\pi J_\lambda(\pi,\pi,p_\pi,r_\pi)-\nabla_\pi J_\lambda(\pi,\pi,p_{\tilde{\pi}},r_{\tilde{\pi}})|_{\tilde{\pi}=\pi}\big\|\|\pi'-\pi\|\\
\overset{(a)}{\le}&\nabla_\pi J_\lambda(\pi,\pi,p_\pi,r_\pi)^\top(\pi'-\pi)+\sqrt{2|\mathcal{S}|}\Big(\frac{\epsilon_p\sqrt{|\mathcal{S}|}(1+\lambda\log|\mathcal{A}|)}{(1-\gamma)^2}+\frac{\epsilon_r}{1-\gamma}\Big),
\end{aligned}\tag{95}$$

where (a) uses Eq. (94) and Lemma 12. Substituting $p=p_\pi$, $r=r_\pi$ and then Eq. (95) into Eq. (93), we can prove Eq. (12) as follows.

$$\begin{aligned}
\pi(a|s)\ge&\frac{1}{2|\mathcal{A}|^{1/(1-\gamma)}}\exp\Big\{-\frac{1}{\lambda(1-\gamma)}-\frac{2|\mathcal{A}|}{\lambda}(1-\gamma)\cdot\\
&\Big[\nabla_\pi J_\lambda(\pi,\pi,p_\pi,r_\pi)^\top(\pi'-\pi)+\sqrt{2|\mathcal{S}|}\Big(\frac{\epsilon_p\sqrt{|\mathcal{S}|}(1+\lambda\log|\mathcal{A}|)}{(1-\gamma)^2}+\frac{\epsilon_r}{1-\gamma}\Big)\Big]\Big\}\\
=&\pi_{\min}\exp\Big[-\frac{2|\mathcal{A}|}{\lambda}(1-\gamma)\langle\nabla_\pi V_{\lambda,\pi}^\pi,\pi'-\pi\rangle\Big],
\end{aligned}$$

where the $=$ uses $V_{\lambda,\pi}^\pi=J_\lambda(\pi,\pi,p_\pi,r_\pi)$ and $\pi_{\min}$ defined as follows.

$$\pi_{\min}\overset{\text{def}}{=}\frac{1}{2|\mathcal{A}|^{1/(1-\gamma)}}\exp\Big\{-\frac{1}{\lambda(1-\gamma)}-\frac{2|\mathcal{A}|\sqrt{2|\mathcal{S}|}}{\lambda}\Big[\frac{\epsilon_p\sqrt{|\mathcal{S}|}(1+\lambda\log|\mathcal{A}|)}{1-\gamma}+\epsilon_r\Big]\Big\},\tag{96}$$

I PROOF OF THEOREM 3

For any policies π,π', we have

$$\begin{aligned}
&|V_{\lambda,\pi'}^{\pi'}-V_{\lambda,\pi}^\pi|\\
\le&|J_\lambda(\pi',p_{\pi'},r_{\pi'})-J_\lambda(\pi,p_\pi,r_\pi)|\\
\le&|J_\lambda(\pi',p_{\pi'},r_{\pi'})-J_\lambda(\pi',p_{\pi'},r_\pi)|+|J_\lambda(\pi',p_{\pi'},r_\pi)-J_\lambda(\pi',p_\pi,r_\pi)|\\
&+|J_\lambda(\pi',p_\pi,r_\pi)-J_\lambda(\pi,p_\pi,r_\pi)|\\
\overset{(a)}{\le}&\frac{\|r_{\pi'}-r_\pi\|}{1-\gamma}+L_p\|p_{\pi'}-p_\pi\|+L_\pi\max_s\|\log\pi'(\cdot|s)-\log\pi(\cdot|s)\|\\
\overset{(b)}{\le}&\Big(L_p\epsilon_p+\frac{\epsilon_r}{1-\gamma}\Big)\|\pi'-\pi\|+L_\pi\sqrt{\sum_s\|\log\pi'(\cdot|s)-\log\pi(\cdot|s)\|^2}\\
\overset{(c)}{\le}&\Big(L_p\epsilon_p+\frac{\epsilon_r}{1-\gamma}\Big)\|\log\pi'-\log\pi\|+L_\pi\|\log\pi'-\log\pi\|\\
\overset{(d)}{=}&L_\lambda\|\log\pi'-\log\pi\|,
\end{aligned}\tag{97}$$

where (a) uses Eqs. (39), (40) and (41), (b) uses Assumption 7, (c) uses $|\log y-\log x|\le|y-x|$ for any $x,y\in\mathbb{R}$, and (d) defines the following constant.

$$L_\lambda=L_p\epsilon_p+\frac{\epsilon_r}{1-\gamma}+L_\pi=\frac{\sqrt{|\mathcal{A}|}(2-\gamma+\gamma\lambda\log|\mathcal{A}|)+\epsilon_p\sqrt{|\mathcal{S}|}(1+\lambda\log|\mathcal{A}|)+\epsilon_r(1-\gamma)}{(1-\gamma)^2}.$$

$$L_\lambda\overset{\text{def}}{=}L_p\epsilon_p+\frac{\epsilon_r}{1-\gamma}+L_\pi=\frac{\sqrt{|\mathcal{A}|}(2-\gamma+\gamma\lambda\log|\mathcal{A}|)+\epsilon_p\sqrt{|\mathcal{S}|}(1+\lambda\log|\mathcal{A}|)}{(1-\gamma)^2}+\frac{\epsilon_r}{1-\gamma}\tag{98}$$

Note that for any $u,v\ge\Delta>0$,

$$\begin{aligned}
|\log u-\log v|=&\log\max(u,v)-\log\min(u,v)\\
=&\int_{\min(u,v)}^{\max(u,v)}\frac{1}{x}dx\le\frac{1}{\Delta}[\max(u,v)-\min(u,v)]=\frac{|u-v|}{\Delta}.
\end{aligned}$$

Therefore, for any $\pi, \pi' \in \Pi_\Delta \overset{\text{def}}{=} \{\pi \in \Pi : \pi(a|s) \geq \Delta\}$, we have

$$\begin{aligned}\|\log\pi' - \log\pi\|^2 =& \sum_{s,a} |\log\pi'(a|s) - \log\pi(a|s)|^2 \\ \leq& \Delta^{-2}\sum_{s,a}|\pi'(a|s) - \pi(a|s)|^2 = \Delta^{-2}\|\pi' - \pi\|^2.\end{aligned}$$

Substituting the above inequality into Eq. (97) proves the first inequality of Eq. (97).

Next, we will prove the second inequality of Eq. (97) about the Lipschitz continuity of the following performative policy gradient.

$$\begin{aligned}\nabla_\pi V_{\lambda,\pi}^\pi =& \nabla_\pi J_\lambda(\pi,\pi,p_\pi,r_\pi) \\ =& \nabla_\pi J_\lambda(\pi,\pi,p_{\tilde\pi},r_{\tilde\pi})|_{\tilde\pi=\pi} + (\nabla_\pi p_\pi)\nabla_{p_\pi}J_\lambda(\pi,\pi,p_\pi,r_\pi) + (\nabla_\pi r_\pi)\nabla_{r_\pi}J_\lambda(\pi,\pi,p_\pi,r_\pi).\end{aligned}$$

For any $\pi, \pi' \in \Pi_\Delta$, we have

$$\begin{aligned}
&\|\nabla_{\pi'}V_{\lambda,\pi'}^{\pi'} - \nabla_\pi V_{\lambda,\pi}^\pi\| \\
\leq& \big\|\nabla_{\pi'}J_\lambda(\pi',\pi',p_{\tilde\pi},r_{\tilde\pi})|_{\tilde\pi=\pi'} - \nabla_\pi J_\lambda(\pi,\pi,p_{\tilde\pi},r_{\tilde\pi})|_{\tilde\pi=\pi}\big\| \\
&+ \|\nabla_{\pi'}p_{\pi'}\|\cdot\|\nabla_{p_{\pi'}}J_\lambda(\pi',\pi',p_{\pi'},r_{\pi'}) - \nabla_{p_\pi}J_\lambda(\pi,\pi,p_\pi,r_\pi)\| \\
&+ \|\nabla_{p_\pi}J_\lambda(\pi,\pi,p_\pi,r_\pi)\|\cdot\|\nabla_{\pi'}p_{\pi'} - \nabla_\pi p_\pi\| \\
&+ \|\nabla_{\pi'}r_{\pi'}\|\cdot\|\nabla_{r_{\pi'}}J_\lambda(\pi',\pi',p_{\pi'},r_{\pi'}) - \nabla_{r_\pi}J_\lambda(\pi,\pi,p_\pi,r_\pi)\| \\
&+ \|\nabla_{r_\pi}J_\lambda(\pi,\pi,p_\pi,r_\pi)\|\cdot\|\nabla_{\pi'}r_{\pi'} - \nabla_\pi r_\pi\| \\
\overset{(a)}{\leq}& \Big(\frac{|\mathcal{A}|(1+2\lambda\log|\mathcal{A}|)}{(1-\gamma)^2} + \gamma L_\pi\Big)\max_s\|\log\pi'(\cdot|s) - \log\pi(\cdot|s)\| \\
&+ \Big[\frac{2(1+\lambda\log|\mathcal{A}|)}{(1-\gamma)^2} + \gamma L_p\Big]\sqrt{|\mathcal{S}||\mathcal{A}|}\|p_{\pi'} - p_\pi\| + \frac{\sqrt{|\mathcal{A}|}\|r_{\pi'} - r_\pi\|_\infty}{1-\gamma} \\
&+ \epsilon_p\Big[\ell_\pi\max_s\|\log\pi'(\cdot|s) - \log\pi(\cdot|s)\| + \ell_p\|p_{\pi'} - p_\pi\| + \frac{2-\gamma}{1-\gamma}\sqrt{|\mathcal{S}|}\|r_{\pi'} - r_\pi\|_\infty\Big] \\
&+ L_pS_p\|\pi' - \pi\| + \frac{\gamma\epsilon_r}{(1-\gamma)^2}\big(\max_s\|\pi'(\cdot|s) - \pi(\cdot|s)\|_1 + \max_{s,a}\|p_{\pi'}(\cdot|s,a) - p_\pi(\cdot|s,a)\|_1\big) \\
&+ \frac{S_r}{1-\gamma}\|\pi' - \pi\| \\
\overset{(b)}{\leq}& \Big(\frac{|\mathcal{A}|(1+2\lambda\log|\mathcal{A}|)}{\Delta(1-\gamma)^2} + \frac{\gamma L_\pi}{\Delta}\Big)\|\pi' - \pi\| + \epsilon_p\sqrt{|\mathcal{S}||\mathcal{A}|}\Big[\frac{2(1+\lambda\log|\mathcal{A}|)}{(1-\gamma)^2} + \gamma L_p\Big]\|\pi' - \pi\| \\
&+ \frac{\epsilon_r\sqrt{|\mathcal{A}|}\|\pi' - \pi\|}{1-\gamma} + \epsilon_p\Big[\frac{\ell_\pi}{\Delta}\|\pi' - \pi\| + \ell_p\epsilon_p\|\pi' - \pi\| + \frac{2-\gamma}{1-\gamma}\epsilon_r\sqrt{|\mathcal{S}|}\|\pi' - \pi\|\Big] \\
&+ L_pS_p\|\pi' - \pi\| + \frac{\gamma\epsilon_r}{(1-\gamma)^2}\big(\sqrt{|\mathcal{S}|}\|\pi' - \pi\| + \epsilon_p\sqrt{|\mathcal{S}|}\|\pi' - \pi\|\big) + \frac{S_r}{1-\gamma}\|\pi' - \pi\| \\
\overset{(c)}{\leq}& \Big(\frac{|\mathcal{A}|(1+2\lambda\log|\mathcal{A}|)}{\Delta(1-\gamma)^2} + \frac{\gamma L_\pi}{\Delta}\Big)\|\pi' - \pi\| + \frac{\epsilon_p}{\Delta}\sqrt{\frac{|\mathcal{S}|}{|\mathcal{A}|}}\Big[\frac{2(1+\lambda\log|\mathcal{A}|)}{(1-\gamma)^2} + \gamma L_p\Big]\|\pi' - \pi\| \\
&+ \frac{\epsilon_r\|\pi' - \pi\|}{\Delta\sqrt{|\mathcal{A}|}(1-\gamma)} + \frac{\epsilon_p}{\Delta}\Big[\ell_\pi + \frac{\ell_p\epsilon_p}{|\mathcal{A}|} + \frac{2-\gamma}{|\mathcal{A}|(1-\gamma)}\epsilon_r\sqrt{|\mathcal{S}|}\Big]\|\pi' - \pi\| \\
&+ \frac{\gamma\epsilon_r\sqrt{|\mathcal{S}|}(1+\epsilon_p)}{\Delta|\mathcal{A}|(1-\gamma)^2}\|\pi' - \pi\| + \frac{L_pS_p + S_r/(1-\gamma)}{\Delta|\mathcal{A}|}\|\pi' - \pi\| \\
\overset{(d)}{\leq}& \Big(\frac{|\mathcal{A}|(1+2\lambda\log|\mathcal{A}|)}{\Delta(1-\gamma)^2} + \frac{\gamma\sqrt{|\mathcal{A}|}(2-\gamma+\gamma\lambda\log|\mathcal{A}|)}{\Delta(1-\gamma)^2}\Big)\|\pi' - \pi\| \\
&+ \frac{\epsilon_p}{\Delta}\sqrt{\frac{|\mathcal{S}|}{|\mathcal{A}|}}\Big[\frac{2(1+\lambda\log|\mathcal{A}|)}{(1-\gamma)^2} + \frac{\gamma\sqrt{|\mathcal{S}|}(1+\lambda\log|\mathcal{A}|)}{(1-\gamma)^2}\Big]\|\pi' - \pi\|
\end{aligned}$$

$$
\begin{aligned}
&+\frac{\epsilon_p}{\Delta}\Big[\frac{\sqrt{|\mathcal{S}||\mathcal{A}|}(2+3\gamma\lambda\log|\mathcal{A}|)}{(1-\gamma)^3}+\frac{2\epsilon_p\gamma|\mathcal{S}|(1+\lambda\log|\mathcal{A}|)}{|\mathcal{A}|(1-\gamma)^3}+\frac{2-\gamma}{|\mathcal{A}|(1-\gamma)}\epsilon_r\sqrt{|\mathcal{S}|}\Big]\|\pi'-\pi\| \\
&+\frac{\epsilon_r\sqrt{|\mathcal{A}|}(1-\gamma)+\gamma\epsilon_r\sqrt{|\mathcal{S}|}(1+\epsilon_p)}{\Delta|\mathcal{A}|(1-\gamma)^2}\|\pi'-\pi\| \\
&+\frac{S_p\sqrt{|\mathcal{S}|}(1+\lambda\log|\mathcal{A}|)+S_r(1-\gamma)}{\Delta|\mathcal{A}|(1-\gamma)^2}\|\pi'-\pi\| \\
\le&\frac{3|\mathcal{A}|(1+\lambda\log|\mathcal{A}|)}{\Delta(1-\gamma)^2}\|\pi'-\pi\|+\frac{\epsilon_p\sqrt{|\mathcal{S}||\mathcal{A}|}(5+6\lambda\log|\mathcal{A}|)}{\Delta(1-\gamma)^3}\|\pi'-\pi\| \\
&+\frac{\epsilon_r\big[\sqrt{|\mathcal{A}|}(1-\gamma)+\sqrt{|\mathcal{S}|}(\gamma+2\epsilon_p)\big]+S_p\sqrt{|\mathcal{S}|}(1+\lambda\log|\mathcal{A}|)+S_r(1-\gamma)}{\Delta|\mathcal{A}|(1-\gamma)^2}\|\pi'-\pi\|, \quad (99)
\end{aligned}
$$

where (a) uses Eqs. (40), (41) and (44)-(46) as well as Assumptions 1-2, and (b) uses the following bounds for any $\pi,\pi'\in\Delta$, (c) uses $\Delta\le|\mathcal{A}|^{-1}$ (since for any $\pi\in\Pi_\Delta$, $1=\sum_a\pi(a|s)\ge\Delta|\mathcal{A}|$), (d) uses $L_\pi:=\frac{\sqrt{|\mathcal{A}|}(2-\gamma+\gamma\lambda\log|\mathcal{A}|)}{(1-\gamma)^2}$, $L_p:=\frac{\sqrt{|\mathcal{S}|}(1+\lambda\log|\mathcal{A}|)}{(1-\gamma)^2}$, $\ell_\pi:=\frac{\sqrt{|\mathcal{S}||\mathcal{A}|}(2+3\gamma\lambda\log|\mathcal{A}|)}{(1-\gamma)^3}$ and $\ell_p:=\frac{2\gamma|\mathcal{S}|(1+\lambda\log|\mathcal{A}|)}{(1-\gamma)^3}$ defined in Lemma 6, (e) uses ℓ_λ defined by Eq. (100).

$$
\begin{aligned}
\max_s\|\log\pi'(\cdot|s)-\log\pi(\cdot|s)\|\le&\Delta^{-1}\max_s\|\pi'(\cdot|s)-\pi(\cdot|s)\|\le\Delta^{-1}\|\pi'-\pi\|, \\
\|p_{\pi'}-p_\pi\|\overset{(a)}{\le}&\epsilon_p\|\pi'-\pi\|, \\
\|r_{\pi'}-r_\pi\|_\infty\le&\|r_{\pi'}-r_\pi\|\overset{(a)}{\le}\epsilon_r\|\pi'-\pi\|, \\
\max_s\|\pi'(\cdot|s)-\pi(\cdot|s)\|_1\le&\sqrt{|\mathcal{S}|}\max_s\|\pi'(\cdot|s)-\pi(\cdot|s)\|\le\sqrt{|\mathcal{S}|}\|\pi'-\pi\|, \\
\max_{s,a}\|p_{\pi'}(\cdot|s,a)-p_\pi(\cdot|s,a)\|_1\le&\sqrt{|\mathcal{S}|}\max_{s,a}\|p_{\pi'}(\cdot|s,a)-p_\pi(\cdot|s,a)\| \\
\le&\sqrt{|\mathcal{S}|}\|p_{\pi'}-p_\pi\|\overset{(a)}{\le}\epsilon_p\sqrt{|\mathcal{S}|}\|\pi'-\pi\|.
\end{aligned}
$$

Here, (a) uses Assumption 1. Finally, define the Lipschitz constant ℓ_λ as follows and thus Eq. (99) implies the second inequality of Eq. (97) that $\|\nabla_{\pi'}V_{\lambda,\pi'}^{\pi'}-\nabla_\pi V_{\lambda,\pi}^{\pi}\|\le\frac{\ell_\lambda}{\Delta}\|\pi'-\pi\|$.

$$
\begin{aligned}
\ell_\lambda\overset{\text{def}}{=}&\frac{3|\mathcal{A}|(1+\lambda\log|\mathcal{A}|)}{(1-\gamma)^2}+\frac{\epsilon_p\sqrt{|\mathcal{S}||\mathcal{A}|}(5+6\lambda\log|\mathcal{A}|)}{(1-\gamma)^3} \\
&+\frac{\epsilon_r\big[\sqrt{|\mathcal{A}|}(1-\gamma)+\sqrt{|\mathcal{S}|}(\gamma+2\epsilon_p)\big]}{|\mathcal{A}|(1-\gamma)^2}+\frac{S_p\sqrt{|\mathcal{S}|}(1+\lambda\log|\mathcal{A}|)+S_r(1-\gamma)}{|\mathcal{A}|(1-\gamma)^2}. \quad (100)
\end{aligned}
$$

J Proof of Proposition 1

We prove the validity of the stochastic gradient (16) first. For any $\pi\in\Pi_\Delta$, $s\in\mathcal{S}$ and $a\in\mathcal{A}$, we have $\pi(a|s)\ge\Delta$, so $\pi(a|s)\le1-\Delta$ (since $\sum_{a'}\pi(a'|s)=1$). For any $u_i\in U_1$, we have $|u_i(a|s)|\le1$. Therefore,

$$(\pi\pm\delta u_i)(a|s)\ge\pi(a|s)-\delta|u_i(a|s)|\ge\Delta-\delta>0, \quad (101)$$

which means $\pi\pm\delta u_i\in\Pi$. Hence, $V_{\lambda,\pi'}^{\pi'}$ is well defined for $\pi'\in\{\pi+\delta u_i,\pi-\delta u_i\}$.

Then we will prove the estimation error bound (18). Based on Lemma 10, there exists an orthogonal transformation $\mathcal{T}:\mathbb{R}^{|\mathcal{A}|}\to\mathcal{Z}_{|\mathcal{A}|-1}=\{z=[z_1,\ldots,z_{|\mathcal{A}|}]\in\mathbb{R}^{|\mathcal{A}|}:\sum_i z_i=0\}$.

Note that any $x\in\mathbb{R}^{|\mathcal{S}|(|\mathcal{A}|-1)}$ can be written as $x=[x_s]_{s\in\mathcal{S}}$, a concatenation of $|\mathcal{S}|$ vectors $x_s\in\mathbb{R}^{|\mathcal{A}|}$. Therefore, we can define the transformation $T:\mathbb{R}^{|\mathcal{S}|(|\mathcal{A}|-1)}\to\mathcal{L}_0\overset{\text{def}}{=}\big\{u\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}:u(\cdot|s)\in\mathcal{Z}_{|\mathcal{A}|-1},\forall s\in\mathcal{S}\big\}$ as follows

$$[T(x)](\cdot|s)=\mathcal{T}(x_s),\forall s\in\mathcal{S} \quad (102)$$

where $x_s \in \mathbb{R}^{|\mathcal{A}|}$ are extracted from $|\mathcal{A}|$ entries of $x = [x_s]_{s\in\mathcal{S}}$. For any $x = [x_s]_{s\in\mathcal{S}}, y = [y_s]_{s\in\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|(|\mathcal{A}|-1)}$ and $\alpha, \beta \in \mathbb{R}$, we can prove that T is an orthogonal transformation as follows.

$$\begin{aligned}&[T(\alpha x + \beta y)](\cdot|s) = \mathcal{T}(\alpha x_s + \beta y_s) = \alpha\mathcal{T}(x_s) + \beta\mathcal{T}(y_s) = \alpha[T(x)](\cdot|s) + \beta[T(x)](\cdot|s)\\ \Rightarrow &T(\alpha x + \beta y) = \alpha T(x) + \beta T(y).\end{aligned}$$

$$\langle T(x), T(y)\rangle = \sum_s \big\langle [T(x)](\cdot|s), [T(y)](\cdot|s)\big\rangle = \sum_s \langle \mathcal{T}(x_s), \mathcal{T}(y_s)\rangle = \sum_s \langle x_s, y_s\rangle = \langle x, y\rangle.$$

Define the following set.

$$T^{-1}(\Pi_\Delta - |\mathcal{A}|^{-1}) \overset{\text{def}}{=} \{\pi \in \Pi_\Delta : T^{-1}(\pi - |\mathcal{A}|^{-1})\}, \tag{103}$$

where $\pi - |\mathcal{A}|^{-1} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ has entries $(\pi - |\mathcal{A}|^{-1})(a|s) = \pi(a|s) - |\mathcal{A}|^{-1}$, so $\pi - |\mathcal{A}|^{-1} \in \mathcal{L}_0$. Furthermore, since Π_Δ is a convex and compact set and T^{-1} is an orthogonal transformation, $T^{-1}(\Pi_\Delta - |\mathcal{A}|^{-1})$ is a convex and compact subset of $\mathcal{L}_0$.

Then for any $x \in T^{-1}(\Pi_\Delta - |\mathcal{A}|^{-1})$, we have $T(x) + |\mathcal{A}|^{-1} \in \Pi_\Delta$, so we can define the function $f_\lambda(x) \overset{\text{def}}{=} V_{\lambda,T(x)+|\mathcal{A}|^{-1}}^{T(x)+|\mathcal{A}|^{-1}}$.

Note that as $V_{\lambda,\pi}^{\pi}$ is a differentiable function of π, so for any $\pi' \in \Pi$ and fixed $\pi \in \Pi$ we have

$$\begin{aligned}\frac{V_{\lambda,\pi'}^{\pi'} - V_{\lambda,\pi}^{\pi} - \langle \nabla_\pi V_{\lambda,\pi}^{\pi}, \pi' - \pi\rangle}{\|\pi' - \pi\|} =& \frac{V_{\lambda,\pi'}^{\pi'} - V_{\lambda,\pi}^{\pi} - \langle \mathrm{proj}_{\mathcal{L}_0}(\nabla_\pi V_{\lambda,\pi}^{\pi}), \pi' - \pi\rangle}{\|\pi' - \pi\|}\\ \to& 0 \quad (\text{ as } \pi' \in \Pi \text{ and } \pi' \to \pi),\end{aligned} \tag{104}$$

where the above $=$ uses $\pi' - \pi \in \mathcal{L}_0$. Then, we can prove that f_λ is differentiable with gradient $\nabla f_\lambda(x) = T^{-1}\big(\mathrm{proj}_{\mathcal{L}_0} \nabla_\pi V_{\lambda,\pi}^{\pi}\big|_{\pi=T(x)+|\mathcal{A}|^{-1}}\big)$, since for any $x' \in T^{-1}(\Pi_\Delta - |\mathcal{A}|^{-1})$ and fixed $x \in T^{-1}(\Pi_\Delta - |\mathcal{A}|^{-1})$ we have

$$\begin{aligned}&\frac{f_\lambda(x') - f_\lambda(x) - \big\langle T^{-1}\big[\mathrm{proj}_{\mathcal{L}_0}\big(\nabla_\pi V_{\lambda,\pi}^{\pi}\big|_{\pi=T(x)+|\mathcal{A}|^{-1}}\big)\big], x' - x\big\rangle}{\|x' - x\|}\\ \overset{(a)}{=}&\frac{1}{\big\|[T(x') + |\mathcal{A}|^{-1}] - [T(x) + |\mathcal{A}|^{-1}]\big\|}\Big[V_{\lambda,T(x')+|\mathcal{A}|^{-1}}^{T(x')+|\mathcal{A}|^{-1}} - V_{\lambda,T(x)+|\mathcal{A}|^{-1}}^{T(x)+|\mathcal{A}|^{-1}}\\ &- \big\langle \mathrm{proj}_{\mathcal{L}_0}\big(\nabla_\pi V_{\lambda,\pi}^{\pi}\big|_{\pi=T(x)+|\mathcal{A}|^{-1}}\big), [T(x') + |\mathcal{A}|^{-1}] - [T(x) + |\mathcal{A}|^{-1}]\big\rangle\Big]\\ \overset{(b)}{\to}& 0 \text{ as } x' \in T^{-1}(\Pi_\Delta - |\mathcal{A}|^{-1}) \text{ and } x' \to x,\end{aligned} \tag{105}$$

where (a) uses the property of the orthogonal transformation T, and (b) uses Eq. (104) and the fact that $x' \to x$ means $\big\|[T(x') + |\mathcal{A}|^{-1}] - [T(x) + |\mathcal{A}|^{-1}]\big\| = \|x' - x\| \to 0$.

Furthermore, we will show that $f_\lambda(x)$ is a Lipscthiz continuous and Lipschitz smooth function of $x \in \Pi_\Delta$. For any $x, x' \in T^{-1}(\Pi_\Delta - |\mathcal{A}|^{-1})$, we have

$$|f_\lambda(x') - f_\lambda(x)| = \big|V_{\lambda,T(x')+|\mathcal{A}|^{-1}}^{T(x')+|\mathcal{A}|^{-1}} - V_{\lambda,T(x)+|\mathcal{A}|^{-1}}^{T(x)+|\mathcal{A}|^{-1}}\big| \overset{(a)}{\le} \frac{L_\lambda}{\Delta}\|T(x') - T(x)\| \overset{(b)}{=} \frac{L_\lambda}{\Delta}\|x' - x\|,$$

$$\begin{aligned}\|\nabla f_\lambda(x') - \nabla f_\lambda(x)\| =& \big\|T^{-1}\big[\mathrm{proj}_{\mathcal{L}_0}\big(\nabla_\pi V_{\lambda,\pi}^{\pi}\big|_{\pi=T(x')}\big)\big] - T^{-1}\big[\mathrm{proj}_{\mathcal{L}_0}\big(\nabla_\pi V_{\lambda,\pi}^{\pi}\big|_{\pi=T(x)}\big)\big]\big\|\\ \overset{(b)}{=}& \big\|\mathrm{proj}_{\mathcal{L}_0}\big(\nabla_\pi V_{\lambda,\pi}^{\pi}\big|_{\pi=T(x')+|\mathcal{A}|^{-1}}\big) - \mathrm{proj}_{\mathcal{L}_0}\big(\nabla_\pi V_{\lambda,\pi}^{\pi}\big|_{\pi=T(x)+|\mathcal{A}|^{-1}}\big)\big\|\\ \le& \big\|\big(\nabla_\pi V_{\lambda,\pi}^{\pi}\big|_{\pi=T(x')+|\mathcal{A}|^{-1}}\big) - \big(\nabla_\pi V_{\lambda,\pi}^{\pi}\big|_{\pi=T(x)+|\mathcal{A}|^{-1}}\big)\big\|\\ \overset{(a)}{\le}& \frac{\ell_\lambda}{\Delta}\|T(x') - T(x)\| \overset{(b)}{=} \frac{\ell_\lambda}{\Delta}\|x' - x\|,\end{aligned}$$

In both the inequalities above, (a) applies Theorem 3 to $T(x) + |\mathcal{A}|^{-1}, T(x') + |\mathcal{A}|^{-1} \in \Pi_\Delta$ and (b) uses the property of the orthogonal transformation T. The two inequalities above implies that f_λ is an $\frac{L_\lambda}{\Delta}$-Lipschitz continuous and $\frac{\ell_\lambda}{\Delta}$-Lipschitz smooth function on $T^{-1}(\Pi_\Delta - |\mathcal{A}|^{-1})$.

Denote

$$g_{\lambda,\delta}(\pi)=\frac{|\mathcal{S}|(|\mathcal{A}|-1)}{2N\delta}\sum_{i=1}^{N}\big(V_{\lambda,\pi+\delta u_i}^{\pi+\delta u_i}-V_{\lambda,\pi-\delta u_i}^{\pi-\delta u_i}\big)u_i, \tag{106}$$

which replaces $\hat{V}_{\lambda,\pi'}^{\pi'}$ with $V_{\lambda,\pi'}^{\pi'}$ in Eq. (16). The estimation error of the performative policy gradient estimator above can be rewritten as follows for any $\pi\in\Pi_\Delta$.

$$\begin{aligned}
&g_{\lambda,\delta}(\pi)-\mathrm{proj}_{\mathcal{L}_0}(\nabla_\pi V_{\lambda,\pi}^{\pi})\\
\overset{(a)}{=}&\Big(\frac{|\mathcal{S}|(|\mathcal{A}|-1)}{2N\delta}\sum_{i=1}^{N}\big(V_{\lambda,\pi+\delta u_i}^{\pi+\delta u_i}-V_{\lambda,\pi-\delta u_i}^{\pi-\delta u_i}\big)u_i\Big)-\mathrm{proj}_{\mathcal{L}_0}(\nabla_\pi V_{\lambda,\pi}^{\pi})\\
\overset{(b)}{=}&\Big(\frac{|\mathcal{S}|(|\mathcal{A}|-1)}{2N\delta}\sum_{i=1}^{N}\big(f_\lambda\big[T^{-1}(\pi-|\mathcal{A}|^{-1})+\delta T^{-1}(u_i)\big]-f_\lambda\big[T^{-1}(\pi-|\mathcal{A}|^{-1}])-\delta T^{-1}(u_i)\big]\big)\cdot\\
&T^{-1}(u_i)\Big)-T^{-1}[\mathrm{proj}_{\mathcal{L}_0}(\nabla_\pi V_{\lambda,\pi}^{\pi})]\\
\overset{(c)}{=}&\Big(\frac{|\mathcal{S}|(|\mathcal{A}|-1)}{2N\delta}\sum_{i=1}^{N}\big(f_\lambda\big[T^{-1}(\pi-|\mathcal{A}|^{-1})+\delta T^{-1}(u_i)\big]-f_\lambda\big[T^{-1}(\pi-|\mathcal{A}|^{-1}])-\delta T^{-1}(u_i)\big]\big)\cdot\\
&T^{-1}(u_i)\Big)-\nabla f_\lambda[T^{-1}(\pi-|\mathcal{A}|^{-1})],
\end{aligned} \tag{107}$$

where (a) uses Eq. (16), (b) uses $f_\lambda(x)\overset{\text{def}}{=}V_{\lambda,T(x)+|\mathcal{A}|^{-1}}^{T(x)+|\mathcal{A}|^{-1}}$ and the property of the orthogonal transformation T^{-1}, (c) uses $\nabla f_\lambda(x)=T^{-1}\big(\mathrm{proj}_{\mathcal{L}_0}\nabla_\pi V_{\lambda,\pi}^{\pi}\big|_{\pi=T(x)+|\mathcal{A}|^{-1}}\big)$. Note that in the above Eq. (107), $\pi\in\Pi_\Delta$ and u_i is uniformly distributed on the sphere $U_1\cap\mathcal{L}_0$ with $U_1\overset{\text{def}}{=}\{u\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}:\|u\|=1\}$.

Hence, $\pi\pm\delta u_i\in\Pi_{\Delta-\delta}$ which implies $T^{-1}(\pi-|\mathcal{A}|^{-1})\pm\delta T^{-1}(u_i)=T^{-1}(\pi\pm\delta u_i-|\mathcal{A}|^{-1})\in T^{-1}(\Pi_{\Delta-\delta}-|\mathcal{A}|^{-1})$. Also, $T^{-1}(u_i)$ is uniformly distributed on the sphere $T^{-1}(U_{1,0})=\mathbb{S}_{|\mathcal{S}|(|\mathcal{A}|-1)}=\{u\in\mathbb{R}^{|\mathcal{S}|(|\mathcal{A}|-1)}:\|u\|=1\}$. Therefore, we can apply Lemma 9 to the above Eq. (107) where the function f_λ is an $\frac{L_\lambda}{\Delta-\delta}$-Lipschitz continuous and $\frac{\ell_\lambda}{\Delta-\delta}$-Lipschitz smooth function on $T^{-1}(\Pi_{\Delta-\delta}-|\mathcal{A}|^{-1})$, and obtain the following bound which holds with probability at least $1-\eta$.

$$\begin{aligned}
&\|g_{\lambda,\delta}(\pi)-\mathrm{proj}_{\mathcal{L}_0}(\nabla_\pi V_{\lambda,\pi}^{\pi})\|\\
\le&\frac{4L_\lambda|\mathcal{S}|(|\mathcal{A}|-1)}{3N(\Delta-\delta)}\log\Big(\frac{|\mathcal{S}|(|\mathcal{A}|-1)+1}{\eta}\Big)+\frac{L_\lambda|\mathcal{S}|(|\mathcal{A}|-1)}{\Delta-\delta}\sqrt{\frac{2}{N}\log\Big(\frac{|\mathcal{S}|(|\mathcal{A}|-1)+1}{\eta}\Big)}+\frac{\delta\ell_\lambda}{\Delta-\delta}\\
\le&\frac{4L_\lambda|\mathcal{S}||\mathcal{A}|}{3N(\Delta-\delta)}\log\Big(\frac{|\mathcal{S}||\mathcal{A}|}{\eta}\Big)+\frac{L_\lambda|\mathcal{S}||\mathcal{A}|}{\Delta-\delta}\sqrt{\frac{2}{N}\log\Big(\frac{|\mathcal{S}||\mathcal{A}|}{\eta}\Big)}+\frac{\delta\ell_\lambda}{\Delta-\delta}.
\end{aligned} \tag{108}$$

Note that $|\hat{V}_{\lambda,\pi}^{\pi}-V_{\lambda,\pi}^{\pi}|\le\epsilon_V$ holds for any a certain policy π with probability at least $1-\eta$. Therefore, with probability at least $1-2N\eta$, we have

$$|\hat{V}_{\lambda,\pi'}^{\pi'}-V_{\lambda,\pi'}^{\pi'}|\le\epsilon_V,\forall\pi'\in\{\pi\pm\delta u_i\}_{i=1}^{N} \tag{109}$$

Therefore, with probability at least $1-(2N+1)\eta$, Eqs. (108) and (109) hold and thus we have

$$\begin{aligned}
&\|\hat{g}_{\lambda,\delta}(\pi)-\mathrm{proj}_{\mathcal{L}_0}(\nabla_\pi V_{\lambda,\pi}^{\pi})\|\\
\le&\|\hat{g}_{\lambda,\delta}(\pi)-g_{\lambda,\delta}(\pi)\|+\|g_{\lambda,\delta}(\pi)-\mathrm{proj}_{\mathcal{L}_0}(\nabla_\pi V_{\lambda,\pi}^{\pi})\|\\
\overset{(a)}{\le}&\Big\|\frac{|\mathcal{S}|(|\mathcal{A}|-1)}{2N\delta}\sum_{i=1}^{N}\big(\hat{V}_{\lambda,\pi+\delta u_i}^{\pi+\delta u_i}-V_{\lambda,\pi+\delta u_i}^{\pi+\delta u_i}-\hat{V}_{\lambda,\pi-\delta u_i}^{\pi-\delta u_i}+V_{\lambda,\pi-\delta u_i}^{\pi-\delta u_i}\big)u_i\Big\|\\
&+\frac{4L_\lambda|\mathcal{S}||\mathcal{A}|}{3N(\Delta-\delta)}\log\Big(\frac{|\mathcal{S}||\mathcal{A}|}{\eta}\Big)+\frac{L_\lambda|\mathcal{S}||\mathcal{A}|}{\Delta-\delta}\sqrt{\frac{2}{N}\log\Big(\frac{|\mathcal{S}||\mathcal{A}|}{\eta}\Big)}+\frac{\delta\ell_\lambda}{\Delta-\delta}
\end{aligned}$$

$$\overset{(b)}{\le}\frac{|\mathcal{S}||\mathcal{A}|}{N\delta}\sum_{i=1}^N\big\|\big(\hat{V}_{\lambda,\pi+\delta u_i}^{\pi+\delta u_i}-V_{\lambda,\pi+\delta u_i}^{\pi+\delta u_i}-\hat{V}_{\lambda,\pi-\delta u_i}^{\pi-\delta u_i}+V_{\lambda,\pi-\delta u_i}^{\pi-\delta u_i}\big)u_i\big\|$$
$$+\frac{4L_\lambda|\mathcal{S}||\mathcal{A}|}{3N(\Delta-\delta)}\log\Big(\frac{|\mathcal{S}||\mathcal{A}|}{\eta}\Big)+\frac{L_\lambda|\mathcal{S}||\mathcal{A}|}{\Delta-\delta}\sqrt{\frac{2}{N}\log\Big(\frac{|\mathcal{S}||\mathcal{A}|}{\eta}\Big)}+\frac{\delta\ell_\lambda}{\Delta-\delta}$$
$$\le\frac{|\mathcal{S}||\mathcal{A}|}{N\delta}\sum_{i=1}^N\big(|\hat{V}_{\lambda,\pi+\delta u_i}^{\pi+\delta u_i}-V_{\lambda,\pi+\delta u_i}^{\pi+\delta u_i}|+|\hat{V}_{\lambda,\pi-\delta u_i}^{\pi-\delta u_i}+V_{\lambda,\pi-\delta u_i}^{\pi-\delta u_i}|\big)$$
$$+\frac{4L_\lambda|\mathcal{S}||\mathcal{A}|}{3N(\Delta-\delta)}\log\Big(\frac{|\mathcal{S}||\mathcal{A}|}{\eta}\Big)+\frac{L_\lambda|\mathcal{S}||\mathcal{A}|}{\Delta-\delta}\sqrt{\frac{2}{N}\log\Big(\frac{|\mathcal{S}||\mathcal{A}|}{\eta}\Big)}+\frac{\delta\ell_\lambda}{\Delta-\delta}$$
$$\overset{(c)}{\le}\frac{2|\mathcal{S}||\mathcal{A}|\epsilon_V}{\delta}+\frac{4L_\lambda|\mathcal{S}||\mathcal{A}|}{3N(\Delta-\delta)}\log\Big(\frac{|\mathcal{S}||\mathcal{A}|}{\eta}\Big)+\frac{L_\lambda|\mathcal{S}||\mathcal{A}|}{\Delta-\delta}\sqrt{\frac{2}{N}\log\Big(\frac{|\mathcal{S}||\mathcal{A}|}{\eta}\Big)}+\frac{\delta\ell_\lambda}{\Delta-\delta},$$

where (a) uses Eqs. (16), (55) and (108), (b) uses Jensen's inequality that $\|\frac{1}{N}\sum_{i=1}^N x_i\|^2\le\frac{1}{N}\sum_{i=1}^N\|x_i\|^2$ for any vectors $\{x_i\}_{i=1}^N$ of the same dimensionality, (c) uses $|\hat{V}_{\lambda,\pi}^{\pi'}-V_{\lambda,\pi'}^{\pi'}|\le\epsilon_V$ for any policy π'. By replacing η with $\frac{\eta}{3N}$ in the inequality above, we prove the error bound (18) as follows which holds with probability at least $1-\eta$.

$$\|\hat{g}_{\lambda,\delta}(\pi)-\mathrm{proj}_{\mathcal{L}_0}(\nabla_\pi V_{\lambda,\pi}^\pi)\|$$
$$\le\frac{2|\mathcal{S}||\mathcal{A}|\epsilon_V}{\delta}+\frac{4L_\lambda|\mathcal{S}||\mathcal{A}|}{3N(\Delta-\delta)}\log\Big(\frac{3N|\mathcal{S}||\mathcal{A}|}{\eta}\Big)+\frac{L_\lambda|\mathcal{S}||\mathcal{A}|}{\Delta-\delta}\sqrt{\frac{2}{N}\log\Big(\frac{3N|\mathcal{S}||\mathcal{A}|}{\eta}\Big)}+\frac{\delta\ell_\lambda}{\Delta-\delta}\quad(110)$$
$$=\mathcal{O}\Big(\frac{\epsilon_V}{\delta}+\frac{\log(N/\eta)}{\sqrt{N}}+\delta\Big)$$

K PROOF OF PROPOSITION 2

For any $\pi\in\Pi_\Delta$, it is easily seen that the corresponding π' defined by Eq. (13) also belongs to Π_Δ. Therefore,

$$\langle\nabla_\pi V_{\lambda,\pi}^\pi,\pi'-\pi\rangle\le\max_{\tilde{\pi}\in\Pi_\Delta}\langle\nabla_\pi V_{\lambda,\pi}^\pi,\tilde{\pi}-\pi\rangle\le\frac{D\lambda}{5|\mathcal{A}|(1-\gamma)}.$$

Substituting the above inequality into Eq. (12), we obtain that

$$\pi(a|s)\ge\pi_{\min}\exp\Big[-\frac{2|\mathcal{A}|}{D\lambda}(1-\gamma)\langle\nabla_\pi V_{\lambda,\pi}^\pi,\pi'-\pi\rangle\Big]\ge\frac{2\pi_{\min}}{3}\ge2\Delta.$$

Therefore, for any $\pi_2\in\Pi$, we can prove that $\frac{\pi_2+\pi}{2}\in\Pi_\Delta$ as follows.

$$\frac{\pi_2(a|s)+\pi(a|s)}{2}\ge\frac{0+2\Delta}{2}=\Delta.$$

Therefore, we can prove Eq. (22) as follows.

$$\max_{\pi_2\in\Pi}\langle\nabla_\pi V_{\lambda,\pi}^\pi,\pi_2-\pi\rangle=2\max_{\pi_2\in\Pi}\Big\langle\nabla_\pi V_{\lambda,\pi}^\pi,\frac{\pi_2+\pi}{2}-\pi\Big\rangle\overset{(a)}{\le}2\max_{\tilde{\pi}\in\Pi_\Delta}\langle\nabla_\pi V_{\lambda,\pi}^\pi,\tilde{\pi}-\pi\rangle.$$

where (a) uses $\frac{\pi_2+\pi}{2}\in\Pi_\Delta$.

L PROOF OF THEOREM 4

If $\pi_t\in\Pi_\Delta$, then $\pi_{t+1}\in\Pi_\Delta$, since Π_Δ is a convex set and π_{t+1} obtained by Eq. (20) is a convex combination of $\pi_t,\tilde{\pi}_t\in\Pi_\Delta$. Since $\pi_0\in\Pi_\Delta$, we have $\pi_t\in\Pi_\Delta$ for all t by induction. Therefore, Proposition 1 implies that the following bound holds simultaneously for all $\{\pi_t\}_{t=1}^T\subseteq\Pi_\Delta$ with probability at least $1-\eta$.

$$\|\hat{g}_{\lambda,\delta}(\pi_t)-\mathrm{proj}_{\mathcal{L}_0}(\nabla_\pi V_{\lambda,\pi_t}^{\pi_t})\|$$

$$\leq \frac{2|\mathcal{S}||\mathcal{A}|\epsilon_V}{\delta} + \frac{4L_\lambda|\mathcal{S}||\mathcal{A}|}{3TN(\Delta-\delta)}\log\Big(\frac{3TN|\mathcal{S}||\mathcal{A}|}{\eta}\Big) + \frac{L_\lambda|\mathcal{S}||\mathcal{A}|}{\Delta-\delta}\sqrt{\frac{2}{N}\log\Big(\frac{3TN|\mathcal{S}||\mathcal{A}|}{\eta}\Big)} + \frac{\delta\ell_\lambda}{\Delta-\delta}. \tag{111}$$

The bound above further implies that for any $\pi\in\Pi$, we have

$$\begin{aligned}
&\big|\big\langle \hat{g}_{\lambda,\delta}(\pi_t) - \nabla_\pi V_{\lambda,\pi_t}^{\pi_t}, \pi-\pi_t\big\rangle\big| \\
\overset{(a)}{=}&\big|\big\langle \hat{g}_{\lambda,\delta}(\pi_t) - \mathrm{proj}_{\mathcal{L}_0}(\nabla_\pi V_{\lambda,\pi_t}^{\pi_t}), \pi-\pi_t\big\rangle\big| \\
\leq&\|\hat{g}_{\lambda,\delta}(\pi_t) - \mathrm{proj}_{\mathcal{L}_0}(\nabla_\pi V_{\lambda,\pi_t}^{\pi_t})\|\cdot\|\pi-\pi_t\| \\
\overset{(b)}{\leq}&\sqrt{2|\mathcal{S}|}\Big[\frac{2|\mathcal{S}||\mathcal{A}|\epsilon_V}{\delta} + \frac{4L_\lambda|\mathcal{S}||\mathcal{A}|}{3TN(\Delta-\delta)}\log\Big(\frac{3TN|\mathcal{S}||\mathcal{A}|}{\eta}\Big) \\
&\quad + \frac{L_\lambda|\mathcal{S}||\mathcal{A}|}{\Delta-\delta}\sqrt{\frac{2}{N}\log\Big(\frac{3TN|\mathcal{S}||\mathcal{A}|}{\eta}\Big)} + \frac{\delta\ell_\lambda}{\Delta-\delta}\Big],
\end{aligned} \tag{112}$$

where (a) uses $\tilde{\pi}_t-\pi_t, \tilde{\pi}-\pi_t\in\mathcal{L}_0$ for $\tilde{\pi}_t,\tilde{\pi}\in\Pi_\Delta$, and (b) uses Eq. (111) and Lemma 12.

Under the conditions above, we have

$$\begin{aligned}
&V_{\lambda,\pi_{t+1}}^{\pi_{t+1}} \\
\overset{(a)}{\geq}&V_{\lambda,\pi_t}^{\pi_t} + \langle\nabla_\pi V_{\lambda,\pi_t}^{\pi_t}, \pi_{t+1}-\pi_t\rangle - \frac{\ell_\lambda}{2\Delta}\|\pi_{t+1}-\pi_t\|^2 \\
\overset{(b)}{=}&V_{\lambda,\pi_t}^{\pi_t} + \beta\langle\nabla_\pi V_{\lambda,\pi_t}^{\pi_t}, \tilde{\pi}_t-\pi_t\rangle - \frac{\ell_\lambda\beta^2}{2\Delta}\|\tilde{\pi}_t-\pi_t\|^2 \\
=&V_{\lambda,\pi_t}^{\pi_t} + \beta\langle\hat{g}_{\lambda,\delta}(\pi_t), \tilde{\pi}_t-\pi_t\rangle + \beta\langle\nabla_\pi V_{\lambda,\pi_t}^{\pi_t} - \hat{g}_{\lambda,\delta}(\pi_t), \tilde{\pi}_t-\pi_t\rangle - \frac{\ell_\lambda\beta^2}{2\Delta}\|\tilde{\pi}_t-\pi_t\|^2 \\
\overset{(c)}{\geq}&V_{\lambda,\pi_t}^{\pi_t} + \beta\langle\hat{g}_{\lambda,\delta}(\pi_t), \tilde{\pi}_t-\pi_t\rangle - \frac{\ell_\lambda|\mathcal{S}|\beta^2}{\Delta} - \beta\sqrt{2|\mathcal{S}|}\Big[\frac{2|\mathcal{S}||\mathcal{A}|\epsilon_V}{\delta} \\
&\quad + \frac{4L_\lambda|\mathcal{S}||\mathcal{A}|}{3TN(\Delta-\delta)}\log\Big(\frac{3TN|\mathcal{S}||\mathcal{A}|}{\eta}\Big) + \frac{L_\lambda|\mathcal{S}||\mathcal{A}|}{\Delta-\delta}\sqrt{\frac{2}{N}\log\Big(\frac{3TN|\mathcal{S}||\mathcal{A}|}{\eta}\Big)} + \frac{\delta\ell_\lambda}{\Delta-\delta}\Big],
\end{aligned} \tag{113}$$

where (a) uses the $\frac{\ell_\lambda}{\Delta}$-Lipschitz smoothness of $V_{\lambda,\pi}^{\pi}$ on Π_Δ, (b) uses Eq. (20), (c) uses Eq. (112) and Lemma 12.

Rearranging and averaging Eq. (113) over $t=0,1,\ldots,T-1$, we obtain that

$$\begin{aligned}
&\max_{\tilde{\pi}\in\Pi_\Delta}\langle\hat{g}_{\lambda,\delta}(\pi_{\widetilde{T}}), \tilde{\pi}-\pi_{\widetilde{T}}\rangle \\
\overset{(a)}{=}&\langle\hat{g}_{\lambda,\delta}(\pi_{\widetilde{T}}), \tilde{\pi}_{\widetilde{T}}-\pi_{\widetilde{T}}\rangle \\
\overset{(b)}{\leq}&\frac{1}{T}\sum_{t=0}^{T-1}\langle\hat{g}_{\lambda,\delta}(\pi_t), \tilde{\pi}_t-\pi_t\rangle \\
\leq&\frac{V_{\lambda,\pi_T}^{\pi_T}-V_{\lambda,\pi_0}^{\pi_0}}{T\beta} + \frac{\ell_\lambda|\mathcal{S}|\beta}{\Delta} + \sqrt{2|\mathcal{S}|}\Big[\frac{2|\mathcal{S}||\mathcal{A}|\epsilon_V}{\delta} \\
&\quad + \frac{4L_\lambda|\mathcal{S}||\mathcal{A}|}{3TN(\Delta-\delta)}\log\Big(\frac{3TN|\mathcal{S}||\mathcal{A}|}{\eta}\Big) + \frac{L_\lambda|\mathcal{S}||\mathcal{A}|}{\Delta-\delta}\sqrt{\frac{2}{N}\log\Big(\frac{3TN|\mathcal{S}||\mathcal{A}|}{\eta}\Big)} + \frac{\delta\ell_\lambda}{\Delta-\delta}\Big] \\
\leq&\frac{1+\lambda\log|\mathcal{A}|}{T\beta(1-\gamma)} + \frac{\ell_\lambda|\mathcal{S}|\beta}{\Delta} + \sqrt{2|\mathcal{S}|}\Big[\frac{2|\mathcal{S}||\mathcal{A}|\epsilon_V}{\delta} \\
&\quad + \frac{4L_\lambda|\mathcal{S}||\mathcal{A}|}{3TN(\Delta-\delta)}\log\Big(\frac{3TN|\mathcal{S}||\mathcal{A}|}{\eta}\Big) + \frac{L_\lambda|\mathcal{S}||\mathcal{A}|}{\Delta-\delta}\sqrt{\frac{2}{N}\log\Big(\frac{3TN|\mathcal{S}||\mathcal{A}|}{\eta}\Big)} + \frac{\delta\ell_\lambda}{\Delta-\delta}\Big],
\end{aligned} \tag{114}$$

where (a) uses Lemma 1 which means $\tilde{\pi}_t$ satisfies Eq. (19) and (b) uses the output rule of Algorithm 1 that $\widetilde{T}\in\arg\min_{0\leq t\leq T-1}\langle\hat{g}_{\lambda,\delta}(\pi_t), \tilde{\pi}_t-\pi_t\rangle$. Therefore,

$$\max_{\tilde{\pi}\in\Pi_\Delta}\big\langle\nabla_\pi V_{\lambda,\pi_{\widetilde{T}}}^{\pi_{\widetilde{T}}}, \tilde{\pi}-\pi_{\widetilde{T}}\big\rangle$$

$$
\begin{aligned}
&= \max_{\tilde{\pi}\in\Pi_\Delta} \left[\left\langle \nabla_\pi V_{\lambda,\pi_{\widetilde{T}}}^{\pi_{\widetilde{T}}} - \hat{g}_{\lambda,\delta}(\pi_{\pi_{\widetilde{T}}}), \tilde{\pi}-\pi_{\widetilde{T}}\right\rangle + \left\langle \hat{g}_{\lambda,\delta}(\pi_{\pi_{\widetilde{T}}}), \tilde{\pi}-\pi_{\widetilde{T}}\right\rangle\right] \\
&\overset{(a)}{\le} \frac{1+\lambda\log|\mathcal{A}|}{T\beta(1-\gamma)} + \frac{\ell_\lambda|\mathcal{S}|\beta}{\Delta} + 2\sqrt{2|\mathcal{S}|}\Big[\frac{2|\mathcal{S}||\mathcal{A}|\epsilon_V}{\delta} \\
&\quad + \frac{4L_\lambda|\mathcal{S}||\mathcal{A}|}{3TN(\Delta-\delta)}\log\Big(\frac{3TN|\mathcal{S}||\mathcal{A}|}{\eta}\Big) + \frac{L_\lambda|\mathcal{S}||\mathcal{A}|}{\Delta-\delta}\sqrt{\frac{2}{N}\log\Big(\frac{3TN|\mathcal{S}||\mathcal{A}|}{\eta}\Big)} + \frac{\delta\ell_\lambda}{\Delta-\delta}\Big], \quad (115)
\end{aligned}
$$

where (a) uses Eqs. (112) and (114).

Use the following hyperparameter choices for Algorithm 1.

$$
\begin{aligned}
\Delta =& \frac{\pi_{\min}}{3}, && (116)\\
\beta =& \frac{D\Delta\epsilon}{12\ell_\lambda|\mathcal{S}|} = \frac{D\pi_{\min}\epsilon}{36\ell_\lambda|\mathcal{S}|} = \mathcal{O}(\epsilon), && (117)\\
T =& \frac{12(1+\lambda\log|\mathcal{A}|)}{D\epsilon\beta(1-\gamma)} = \frac{432\ell_\lambda|\mathcal{S}|(1+\lambda\log|\mathcal{A}|)}{\pi_{\min}D^2(1-\gamma)\epsilon^2} = \mathcal{O}(\epsilon^{-2}) && (118)\\
\delta =& \frac{D\Delta\epsilon}{48\sqrt{2|\mathcal{S}|}\ell_\lambda} = \frac{D\pi_{\min}\epsilon}{144\sqrt{2|\mathcal{S}|}\ell_\lambda} = \mathcal{O}(\epsilon) \overset{(a)}{\le} \frac{\Delta}{2}, && (119)\\
\epsilon_V =& \frac{D\delta\epsilon}{48|\mathcal{S}||\mathcal{A}|\sqrt{2|\mathcal{S}|}} = \frac{\pi_{\min}D^2\epsilon^2}{13824\ell_\lambda|\mathcal{S}|^2|\mathcal{A}|} = \mathcal{O}(\epsilon^2) && (120)\\
N =& \frac{663552L_\lambda^2|\mathcal{S}|^3|\mathcal{A}|^2}{D^2\pi_{\min}^2\epsilon^2}\log\max\Big(\frac{165888L_\lambda^2|\mathcal{S}|^3|\mathcal{A}|^2}{D^2\pi_{\min}^2\epsilon^2}, \frac{1296\ell_\lambda|\mathcal{S}|^2|\mathcal{A}|(1+\lambda\log|\mathcal{A}|)}{D^2\eta\pi_{\min}(1-\gamma)\epsilon^2}\Big) \\
&+ 2\log\Big(\frac{3|\mathcal{S}||\mathcal{A}|}{\eta}\Big) + 3 \\
=& \mathcal{O}[\epsilon^{-2}\log(\eta^{-1}\epsilon^{-1})] && (121)
\end{aligned}
$$

where (a) uses $\epsilon \le 24\sqrt{2|\mathcal{S}|}\ell_\lambda/D$. With the hyperparameter choices above, we obtain the following inequalities (122)-(124).

$$
\begin{aligned}
&2\sqrt{2|\mathcal{S}|}\cdot\frac{L_\lambda|\mathcal{S}||\mathcal{A}|}{\Delta-\delta}\sqrt{\frac{2}{N}\log\Big(\frac{3TN|\mathcal{S}||\mathcal{A}|}{\eta}\Big)} \\
\overset{(a)}{\le}& \frac{24L_\lambda|\mathcal{S}|^{1.5}|\mathcal{A}|}{\pi_{\min}}\sqrt{\frac{\log N}{N} + \frac{1}{N}\log\Big(\frac{1296\ell_\lambda|\mathcal{S}|^2|\mathcal{A}|(1+\lambda\log|\mathcal{A}|)}{\eta\pi_{\min}D^2(1-\gamma)\epsilon^2}\Big)} \\
\overset{(b)}{\le}& \frac{24L_\lambda|\mathcal{S}|^{1.5}|\mathcal{A}|}{\pi_{\min}}\sqrt{\tilde{\epsilon}+\frac{\tilde{\epsilon}}{4}} \\
=& \frac{12\sqrt{5}L_\lambda|\mathcal{S}|^{1.5}|\mathcal{A}|}{\pi_{\min}}\cdot\frac{D\pi_{\min}\epsilon}{\sqrt{165888}L_\lambda|\mathcal{S}|^{1.5}|\mathcal{A}|} \le \frac{D\epsilon}{12}, \quad (122)
\end{aligned}
$$

where (a) uses Eq. (118) and $\delta \le \Delta/2 = \pi_{\min}/6$ implied by Eqs. (116) and (119), (b) uses Eq. (121) and its implication that $N \ge 4\tilde{\epsilon}^{-1}\log(\tilde{\epsilon}^{-1})$ with $\tilde{\epsilon} = \frac{\pi_{\min}^2\epsilon^2}{165888D^2L_\lambda^2|\mathcal{S}|^3|\mathcal{A}|^2} \le 0.5$ (since $\epsilon \le \frac{288DL_\lambda|\mathcal{S}|^{1.5}|\mathcal{A}|}{\pi_{\min}}$), which implies $\frac{\log N}{N} \le \tilde{\epsilon}$ based on Lemma 11.

$$
\frac{1}{TN}\log\Big(\frac{3TN|\mathcal{S}||\mathcal{A}|}{\eta}\Big) = \frac{\log(TN)}{TN} + \frac{1}{TN}\log\Big(\frac{3|\mathcal{S}||\mathcal{A}|}{\eta}\Big) \overset{(a)}{\le} \frac{1}{2}+\frac{1}{2} = 1, \quad (123)
$$

where (a) uses $NT \ge N \ge \max\Big[3, 2\log\Big(\frac{3|\mathcal{S}||\mathcal{A}|}{\eta}\Big)\Big]$ and Lemma 11.

$$
\begin{aligned}
2\sqrt{2|\mathcal{S}|}\cdot\frac{4L_\lambda|\mathcal{S}||\mathcal{A}|}{3TN(\Delta-\delta)}\log\Big(\frac{3TN|\mathcal{S}||\mathcal{A}|}{\eta}\Big) &\overset{(a)}{\le} 2\sqrt{2|\mathcal{S}|}\cdot\frac{\sqrt{2}L_\lambda|\mathcal{S}||\mathcal{A}|}{\Delta-\delta}\sqrt{\frac{1}{TN}\log\Big(\frac{3TN|\mathcal{S}||\mathcal{A}|}{\eta}\Big)} \\
&\overset{(b)}{\le} \frac{D\epsilon}{12} \quad (124)
\end{aligned}
$$

where (a) uses $\frac{4}{3} < \sqrt{2}$ and $y \leq \sqrt{y}$ for $y = \frac{1}{TN}\log\Big(\frac{3TN|\mathcal{S}||\mathcal{A}|}{\eta}\Big) \leq 1$ (Eq. (123)), and (b) uses $T \geq 1$ and Eq. (122). By substituting the hyperparameter choices (116)-(121) as well as Eqs. (122) and (124) into Eq. (115), we have

$$
\begin{aligned}
&\max_{\tilde{\pi}\in\Pi_\Delta}\big\langle \nabla_\pi V_{\lambda,\pi_{\widetilde{T}}}^{\pi_{\widetilde{T}}}, \tilde{\pi}-\pi_{\widetilde{T}}\big\rangle \\
\leq& \frac{1+\lambda\log|\mathcal{A}|}{T\beta(1-\gamma)} + \frac{\ell_\lambda|\mathcal{S}|\beta}{\Delta} + 2\sqrt{2|\mathcal{S}|}\Big[\frac{2|\mathcal{S}||\mathcal{A}|\epsilon_V}{\delta} \\
&+ \frac{4L_\lambda|\mathcal{S}||\mathcal{A}|}{3TN(\Delta-\delta)}\log\Big(\frac{3TN|\mathcal{S}||\mathcal{A}|}{\eta}\Big) + \frac{L_\lambda|\mathcal{S}||\mathcal{A}|}{\Delta-\delta}\sqrt{\frac{2}{N}\log\Big(\frac{3TN|\mathcal{S}||\mathcal{A}|}{\eta}\Big)} + \frac{\delta\ell_\lambda}{\Delta-\delta}\Big] \\
\leq& \frac{1+\lambda\log|\mathcal{A}|}{\beta(1-\gamma)}\frac{\epsilon\beta(1-\gamma)}{12D(1+\lambda\log|\mathcal{A}|)} + \frac{\ell_\lambda|\mathcal{S}|}{\Delta}\cdot\frac{\Delta\epsilon}{12D\ell_\lambda|\mathcal{S}|} \\
&+ \frac{4\sqrt{2|\mathcal{S}|}|\mathcal{S}||\mathcal{A}|}{\delta}\cdot\frac{\delta\epsilon}{48D|\mathcal{S}||\mathcal{A}|\sqrt{2|\mathcal{S}|}} + \frac{\epsilon}{12D} + \frac{\epsilon}{12D} + \frac{2\sqrt{2|\mathcal{S}|}\ell_\lambda}{\Delta/2}\cdot\frac{\Delta\epsilon}{48\sqrt{2|\mathcal{S}|}D\ell_\lambda} \\
=& \frac{D\epsilon}{2} \overset{(a)}{\leq} \frac{D\lambda}{5|\mathcal{A}|(1-\gamma)},
\end{aligned}
$$

where (a) uses $\epsilon \leq \frac{2\lambda D^2}{5|\mathcal{A}|(1-\gamma)}$. Then based on Proposition 2, the inequality above implies that

$$\max_{\tilde{\pi}\in\Pi}\big\langle \nabla_\pi V_{\lambda,\pi_{\widetilde{T}}}^{\pi_{\widetilde{T}}}, \tilde{\pi}-\pi_{\widetilde{T}}\big\rangle \leq D\epsilon,$$

which means $\pi_{\widetilde{T}}$ is a $D\epsilon$-stationary policy. Then if $\mu \geq 0$, Corollary 1 implies that $\pi_{\widetilde{T}}$ is also an ϵ-PO policy.

M Adjusting Our Results to the Existing Quadratic Regularizer

In Section 4, we have proposed a 0-FW algorithm and obtain its finite-time convergence result to the desired PO policy for our entropy-regularized value function (6). We will briefly show that 0-FW algorithm can also converge to PO for the existing performative reinforcement learning defined by the value function (1) with quadratic regularizer $\mathcal{H}_{\pi'}(\pi) = \frac{1}{2}\|d_{\pi,p_{\pi'}}\|^2$ (Mandal et al., 2023; Rank et al., 2024; Pollatos et al., 2025). The *performative value function* can be rewritten as the following λ-strongly concave function of d_{π,p_π}.

$$V_{\lambda,\pi}^{\pi} = \langle d_{\pi,p_\pi}, r_\pi\rangle - \lambda\|d_{\pi,p_\pi}\|^2. \tag{125}$$

We can prove the *performative value function* above also satisfies Theorem 1 (gradient dominance) with a different μ, following the same proof logic, since both regularizers $\mathcal{H}_\pi(\pi)$ are strongly convex functions of d_{π,p_π} which implies that $V_{\lambda,\pi_\alpha}^{\pi_\alpha}$ is a μ-strongly concave function of α as shown in the proof of Theorem 1 in Appendix F. By direct calculation, we can also show that $V_{\lambda,\pi}^{\pi}$ above is a Lipschitz continuous and Lipschitz smooth function of $\pi \in \Pi$. With these two properties, we can follow the proof logic of Theorem 4 to show that the 0-FW algorithm (with the same procedure as that of Algorithm 1 except the different values of $V_{\lambda,\pi_\alpha}^{\pi_\alpha}$ in the policy evaluation step) converges to a stationary policy of the *performative value function* (125), which by gradient dominance is a PO policy when the new value of μ satisfies $\mu \geq 0$.

N Use of Large Language Models (LLMs)

We used LLMs to generate some functions in the experimental code, and then checked and edited the code to ensure that it exactly implements the algorithms.