

# PERFORMATIVE POLICY GRADIENT: ASCENT TO OPTIMALITY IN PERFORMATIVE REINFORCEMENT LEARNING

Uddalak Mukherjee

ACMU, Indian Statistical Institute, Kolkata,  
West Bengal 700108, India

Debabrota Basu, Udvash Das, Brahim Driss

Univ. Lille, Inria, CNRS,  
Centrale Lille, UMR 9189 – CRISTAL, F-59000 Lille, France

## ABSTRACT

Post-deployment machine learning algorithms often influence the environments they act in, and thus *shift* the underlying dynamics that the standard reinforcement learning (RL) methods ignore. While designing optimal algorithms in this *performative* setting has recently been studied in supervised learning, the RL counterpart remains under-explored. In this paper, we prove the performative counterparts of the performance difference lemma and the policy gradient theorem in RL, and further introduce the **Performative Policy Gradient** algorithm (**PePG**). **PePG** is the first policy gradient algorithm designed to account for performativity in RL. Under softmax parametrisation, and also with and without entropy regularisation, we prove that **PePG** converges to *performatively optimal policies*, i.e. policies that remain optimal under the distribution shifts induced by themselves. Thus, **PePG** significantly extends the prior works in Performative RL that achieves *performative stability* but not optimality. Furthermore, our empirical analysis on standard performative RL environments validate that **PePG** outperforms standard policy gradient algorithms and the existing performative RL algorithms aiming for stability.

## 1 INTRODUCTION

Reinforcement Learning (RL) studies the dynamic decision making problems under incomplete information (Sutton & Barto, 1998). Since an RL algorithm tries and optimises an utility function over a sequence of interactions with an unknown environment, RL has emerged as a powerful tool for algorithmic decision making. Specially, in the last decade, RL has underpinned some of the celebrated successes of AI, such as championing Go with AlphaGo (Silver et al., 2014), controlling particle accelerators (St. John et al., 2021), aligning Large Language Models (LLMs) (Bai et al., 2022), reasoning (Havrilla et al.), to name a few. But the existing paradigm of RL assumes that the underlying environment with which the algorithm interacts stays static over time and the goal of the algorithm is to find the utility-maximising, aka optimal policy for choosing actions over time for this specific environment. But *this assumption does not hold universally*.

In this digital age, algorithms are not passive. Their decisions also shape the environment they interact with, inducing distribution shifts. This phenomenon that predictive AI models often trigger actions that influences their own outcomes is termed as *performativity*. In the supervised learning setting, the study of *performative prediction* is pioneered by Perdomo et al. (2020), and then followed by an extensive literature encompassing stochastic optimisation, control, multi-agent RL, games (Izzo et al., 2021; 2022; Miller et al., 2021; Li & Wai, 2022; Narang et al., 2023; Pilouras & Yu, 2023; Góis et al., 2024; Barakat et al., 2025) etc. There has been several attempts to achieve performative optimality or stability for real-life tasks— recommendation systems (Eilat & Rosenfeld, 2023), measuring the power of firms (Hardt et al., 2022; Mofakhami et al., 2023), healthcare (Zhang et al., 2022) etc. Performativity of algorithms is also omnipresent in practically deployed RL systems. For example, an RL algorithm deployed in a recommender system does not only aim to maximise the user satisfaction but also shifts the preferences of the users in the long-term (Chaney et al., 2018; Mansouri et al., 2020). To clarify the impact of performativity, let us consider an example.

**Example 1** (Performative RL in loan approval). *Let us consider a loan approval problem, where an applicant obtains a loan (or get rejected) according to their credit score  $x$ , and  $x$  depends on the capital of the applicant and that of the population. At each time  $t$ , a loan applicant arrives with a credit score  $x_t$  sampled from  $\mathcal{N}(\mu_t, \sigma^2)$ . The bank chooses whom to give a loan by applying a softmax binary classifier  $\pi_\theta : \mathbb{R} \rightarrow$*

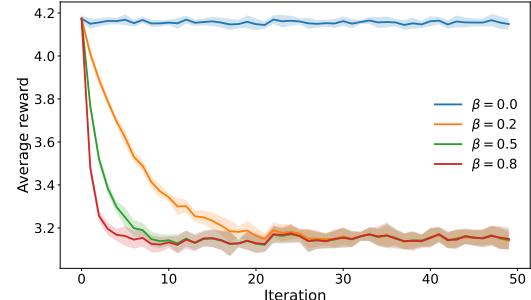


Figure 1: Average reward (over 10 runs) obtained by ERM and Performative Optimal policies across performative strength  $\beta$ .

054 {0, 1} on  $x$  with threshold parameter  $\theta$ . This decision has two effects. (a) The bank receives a positive payoff  $R$ , if the loan  
 055 applicant who was granted a loan repays, or else, loses by  $L$ . Thus, the bank's expected utility for policy  $\pi_\theta$  is  $U(\theta, \mu) =$   
 056  $\mathbb{E}_{x \sim \mathcal{N}(\mu, \sigma^2)} [\pi_\theta(x)(\mathbb{P}(\text{repayment}|x)R - (1 - \mathbb{P}(\text{repayment}|x))L)]$ . (b) Since the amount of capital both the applicant and  
 057 the population influence the credit score, we model that the change in the population mean  $\mu_{t+1}$  depends on the bank's policy,  
 058 via a grant rate  $\mathbb{E}_{x \sim \mathcal{N}(\mu_t, \sigma_t^2)} [\pi_\theta(x)]$ . Specifically,  $\mu_{t+1} = (1 - \beta)\mu_t + \beta f(\mathbb{E}_{x \sim \mathcal{N}(\mu_t, \sigma_t^2)} [\pi_\theta(x)])$ , where  $\beta \in [0, 1]$  is the  
 059 performative strength and  $f : \mathbb{R} \rightarrow [-M, M]$ . Now, if one ignores the performative nature of this decision making problem,  
 060 and try to find out the optimal with respect to a static credit distribution, it obtains  $\theta^{\text{ERM}} \triangleq \arg \max_\theta U(\theta, \mu_0)$ . In contrast,  
 061 if it considers performativity, it obtains  $\theta^{\text{Perf}} \triangleq \arg \max_\theta U(\theta, \mu^*(\theta))$ . In Figure 1, we show that the average reward  
 062 obtained by  $\theta^{\text{ERM}}$  and  $\theta^{\text{Perf}}$  are significantly different. This demonstrates why performativity is a common phenomenon  
 063 across algorithmic decision making problems, and how it changes the resulting optimal solution. Further details are in  
 064 Appendix B.  
 065

066 These problem scenarios have motivated the study of performative RL. Though Bell et al. (2021) were the first to propose a  
 067 setting where the transition and reward of an underlying MDP depend non-deterministically on the deployed policy, Mandal  
 068 et al. (2023) formally introduced *Performative RL*, and its solution concepts, i.e., performatively stable and optimal policies.  
 069 Performative stable policies do not get affected or changed due to distribution shifts after deployment. Performatively optimal  
 070 policies yield the highest expected return once deployed in the performative RL environment. Mandal et al. (2023) proposed  
 071 direct optimization and ascent based techniques that attains performative stability upon repeated retraining. Extending this  
 072 work, Rank et al. (2024) and Mandal & Radanovic (2024) manage to solve the same problem with delayed retraining for  
 073 gradually shifting and linear MDPs. However, there exists no algorithm yet in performative RL that provably converges to the  
 074 performative optimal policy.

075 As we know from the RL literature, the Policy Gradient (PG) type of algorithms that treats policy as a parametric function  
 076 and updates the parameters through gradient ascent algorithms are efficient and scalable (Williams, 1992; Sutton et al., 1999;  
 077 Kakade, 2001). Some examples of successful and popular policy gradient methods include TRPO (Schulman et al., 2015),  
 078 PPO (Schulman et al., 2017), NPG (Kakade, 2001), which are widely used in modern RL applications. Recent theoretical  
 079 advances also establish finite-sample convergence guarantees and complexity analyses (Agarwal et al., 2021; Yuan et al.,  
 080 2022) of PG algorithms. Motivated by the simplicity and universality of the PG algorithms, we ask these two questions in the  
 081 context of performative RL:

- 082    1. How to design PG-type algorithms for performative RL environments to achieve optimality?  
 083    2. What are the minimal conditions under which PG-type algorithms converge to the performatively optimal policy?

084 Our contributions address these questions affirmatively, and showcases the difference of optimality-seeking and stability-  
 085 seeking algorithms in performative RL.

086 **I. Algorithm Design:** We propose the first Performative Policy Gradient algorithm, PePG, for performative RL environments.  
 087 Specifically, we extend the classical vanilla PG and entropy-regularised PG algorithms to Performative RL settings. Though  
 088 the general algorithm design stays same, we derive a performative policy gradient theorem that shows, evaluation of the  
 089 gradient involves two novel gradient terms in performative RL – (a) the expected gradient of reward, and (b) the expected  
 090 gradient of log-transition probabilities times its impact on the expected cumulative return. We leverage this theorem to propose  
 091 an estimator of the performative policy gradient under any differentiable parametrisation.

092 **II. Convergence to Performative Optimality.** We further analyse PePG (with and without entropy regularisation) for  
 093 softmax policies, and softmax Performative Markov Decision Processes (PeMDPs), i.e. the MDPs with softmax transition  
 094 probabilities and linear rewards with respect to the parameters of the softmax policy. We provide a minimal recipe to prove  
 095 convergence of PePG using (a) smoothness of the performative value function, and (b) approximate gradient domination  
 096 lemma for performative policy gradients. This allows us to show that PePG converges to an  $\epsilon$ -ball around performative  
 097 optimal policy in  $\Omega\left(\frac{|\mathcal{S}||\mathcal{A}|^2}{\epsilon^2(1-\gamma)}\right)$  iterations, where  $|\mathcal{S}|$  and  $|\mathcal{A}|$  are the number of states and actions, respectively.

098 Specifically, Mandal et al. (2023) frames the question of using policy gradient to find stable policies as an open problem. The  
 099 authors further contemplate, as PG functions in the policy space, whether it is possible to converge towards a stable policy. In  
 100 this paper, we affirmatively solve an extension to this open problem for tabular softmax PeMDPs with softmax policies.

101 **III. Stability- vs. Optimality-seeking Algorithms in Performative RL.** We further theoretically and numerically con-  
 102 trast the performances of stability-seeking and optimality-seeking algorithms. Theoretically, we derive the performative  
 103 performance difference lemma that distinguished the effect of policy update in these two types of algorithms. Numerically,  
 104 we compare the performances of PePG with the state-of-the-art MDRR (Mixed Delayed Repeated Retraining (Rank et al.,  
 105 2024)) algorithm for finding performatively stable policies in the multi-agent environment proposed by (Mandal et al., 2023).

We show that PePG yields significantly higher values functions than MDRR, while MDRR achieves either similar or lower distance from stable state-action distribution than PePG .

## 2 PRELIMINARIES: FROM RL TO PERFORMATIVE RL

Now, we formalise the RL and performative RL problems, and provide the basics of policy gradient algorithms in RL.

### 2.1 RL: INFINITE-HORIZON DISCOUNTED MDPs

In RL, we mostly study Markov Decision Processes (MDPs) defined via the tuple  $(\mathcal{S}, \mathcal{A}, \mathbf{P}, r, \gamma)$ , where  $\mathcal{S} \subseteq \mathbb{R}^d$  is the state space and  $\mathcal{A} \subseteq \mathbb{R}^d$  is the action space. Both the spaces are assumed to be compact. At any time step  $t \in \mathbb{N}$ , an agent plays an action  $a_t \in \mathcal{A}$  at a state  $s_t \in \mathcal{S}$ . It transits the MDP environment to a state  $s_{t+1}$  according to a transition kernel  $\mathbf{P}(\cdot | s_t, a_t) \in \Delta(\mathcal{S})$ . The agent further receives a reward  $r(s_t, a_t) \in \mathbb{R}$  quantifying the goodness of taking action  $a_t$  at  $s_t$ . The strategy to take an action is represented by a stochastic map, called *policy*, i.e.  $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$ . Given an initial state distribution  $\rho \in \Delta(\mathcal{S})$ , the *goal is to find the optimal policy  $\pi^*$  that maximises the expected discounted sum of rewards*, i.e., the *value function*:  $V_\pi(\rho) \triangleq \mathbb{E}_{s_0 \sim \rho, s_{t+1} \sim \mathbb{P}(\cdot | s_t, \pi(s_t))} [\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t))]$ , where  $\gamma \in (0, 1)$  is called the *discount factor*.  $\gamma$  indicates how much a previous reward matters in the next step, and bounds the effective horizon of a policy to  $\frac{1}{1-\gamma}$ .

---

#### Algorithm 1 Vanilla Policy Gradient

---

```

1: Input: Learning rate  $\eta > 0$ .
2: Initialize: Policy parameter  $\theta_0(s, a) \forall s \in \mathcal{S}, a \in \mathcal{A}$ .
3: for  $t = 1$  to T do
4:   Estimate the gradient  $\nabla_\theta V^\pi(\rho) |_{\theta=\theta_t}$ 
5:   Gradient ascent step:  $\theta_{t+1} \leftarrow \theta_t + \eta \nabla_\theta V^\pi(\rho) |_{\theta=\theta_t}$ 
6: end for
```

---

$V^\pi(\rho)$ , we update  $\theta$  towards  $\nabla_\theta V^\pi(\rho)$ , which is the direction improving the value  $V^\pi(\rho)$  with a fixed learning rate  $\eta > 0$ . For vanilla PG, the policy gradient takes the convenient form leading to estimators computable only with policy rollouts.

**Theorem 1** (Policy Gradient Theorem (Sutton et al., 1999)). *Fix a differentiable parameterisation  $\theta \mapsto \pi_\theta(a | s)$  and an initial distribution  $\rho$ . Let us define the Q-value function  $Q^{\pi_\theta}(s, a) \triangleq \mathbb{E}_{s_{t+1} \sim \mathbb{P}_\pi(\cdot | s_t, \pi(s_t))} [\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t)) | s_0 = s, a_0 = a]$ , and advantage function  $A^{\pi_\theta}(s, a) \triangleq Q^{\pi_\theta}(s, a) - V^{\pi_\theta}(s)$ . Then,*

$$\nabla_\theta V^{\pi_\theta}(\rho) = \frac{1}{1-\gamma} \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_\theta}} \left[ \sum_{t=0}^{\infty} \gamma^t Q^{\pi_\theta}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t | s_t) \right] = \frac{1}{1-\gamma} \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_\theta}} \left[ \sum_{t=0}^{\infty} \gamma^t A^{\pi_\theta}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t | s_t) \right].$$

Since the value function is not concave in the policy parameters, achieving optimality with PG has been a challenge. But practical scalability and efficiency of these algorithms has motivated a long-line of work to understand the minimum conditions and parametric forms of policies leading to convergence to the optimal policy (Agarwal et al., 2021; Mei et al., 2020; Wang & Zou, 2022; Yuan et al., 2022). Our work extends these algorithmic techniques and theoretical insights to performative RL.

### 2.2 PERFORMATIVE RL: INFINITE-HORIZON DISCOUNTED PEMDPs

Given a policy set  $\Pi$ , we denote the Performative Markov Decision Process (PeMDP) is defined as the set of MDPs  $\{\mathcal{M}(\pi) | \pi \in \Pi\}$ , where each MDP is a tuple  $\mathcal{M}(\pi) \triangleq (\mathcal{S}, \mathcal{A}, \mathbf{P}_\pi, \mathbf{r}_\pi, \gamma)$ . Note, that the transition kernel and rewards distribution are no more invariant with respect to the policy. They shift with the deployed policy  $\pi \in \Delta(\mathcal{A})$  (Mandal et al., 2023; Mandal & Radanovic, 2024). In this setting, the probability of generating a trajectory  $\tau_\pi \triangleq (s_t, a_t)_{t=0}^\infty$  under policy  $\pi$  with underlying MDP  $\mathcal{M}(\pi')$  is given by<sup>1</sup>  $\mathbb{P}_{\pi'}^\pi(\tau | \rho) \triangleq \rho(s_0) \prod_{t=0}^{\infty} \pi(a_t | s_t) \mathbf{P}_{\pi'}(s_{t+1} | s_t, a_t)$ , where  $\rho \in \Delta(\mathcal{S})$  is the initial state distribution. Furthermore, the state-action occupancy measure for deployed policy  $\pi$  and environment-inducing policy  $\pi'$  is defined as  $d_{\pi', \rho}^\pi \triangleq (1 - \gamma) \mathbb{E}_{\tau \sim \mathbb{P}_{\pi'}} [\sum_{t=0}^{\infty} \gamma^t \mathbb{1}(s_t = s, a_t = a) | s_0 \sim \rho]$ . Now, we are ready to define the performative expected return, referred as the performative value function that we aim to maximise while solving PeMDP.

<sup>1</sup>Hereafter, for relevant quantities,  $\pi$  in superscript denotes the deployed policy, and  $\pi'$  in the subscript denoted the environment-inducing, i.e. the policy inducing the transition kernel and reward function that the algorithm interacts with.

**Definition 1** (Performative Value Function). *Given a policy  $\pi \in \Pi$  and an initial state distribution  $\rho \in \Delta(S)$ , the performative value function  $V_\pi^\pi(\rho)$  is*

$$V_\pi^\pi(\rho) \triangleq \mathbb{E}_{\tau \sim \mathbb{P}_\pi^\pi} \left[ \sum_{t=0}^{\infty} \gamma^t r_\pi(s_t, \pi(s_t)) \mid s_0 \sim \rho \right]. \quad (1)$$

Equation (2) gives the total expected return that captures the performativity aspect in PeMDPs as the underlying dynamics changes with a deployed policy  $\pi(\cdot \mid s)$ .

On a similar note, we define the performative Q-value function (or action-value function) of a policy  $\pi$  as follows.

**Definition 2** (Performative Q-value). *Given a policy  $\pi \in \Pi$  and a state-action pair  $(s, a) \in (\mathcal{S}, \mathcal{A})$ , the performative Q-value function  $Q_\pi^\pi(s, a)$  is*

$$Q_\pi^\pi(s, a) \triangleq \mathbb{E}_{\tau \sim \mathbb{P}_\pi^\pi} \left[ \sum_{t=0}^{\infty} \gamma^t r_\pi(s_t, a_t) \mid s_0 = s, a_0 = a \right] \quad (2)$$

The Q-value satisfies the following Bellman equation:

$$Q_\pi^\pi(s, a) = r_\pi(s, a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}_\pi(\cdot \mid s, a)} [V_\pi^\pi(s')] \quad (3)$$

Note that, we can maximise performative value function in two ways: (i) considering  $\pi$  as both the environment-inducing policy and the policy the RL agent deploys, or (ii) deploying  $\pi$  to fix it as the environment-inducing policy and agent plays another policy  $\pi'$ . At this vantage point, let us introduce the notion of optimality and stability of policies in PeMDPs (Mandal et al., 2023).

**Definition 3** (Performative Optimality). *A policy  $\pi_o^*$  is performatively optimal if it maximizes the performative value function.*

$$\pi_o^* \in \arg \max_{\pi \in \Delta(\mathcal{A})} V_\pi^\pi(\rho). \quad (4)$$

Thus, if we play the policy  $\pi$  in the environment induced by policy  $\pi$  to maximise the expected return, we land on the performatively optimal policy.

**Definition 4** (Performative Stability). *A policy  $\pi_s^*$  is performatively stable if there is no gain in performative value function due to deploying any other policy than  $\pi_s^*$  in the environment induced by  $\pi_s^*$ .*

$$\pi_s^* \in \arg \max_{\pi \in \Delta(\mathcal{A})} V_{\pi_s^*}^\pi(\rho). \quad (5)$$

As noted by Mandal et al. (2023), a performatively optimal policy may not be performatively stable, i.e.,  $\pi_o^*$  may not be optimal for a changed underlying environment  $\mathcal{M}(\pi_o^*)$ , when it is deployed. Also, in general, the performative value function of  $\pi_o^*$  might be equal to or higher than that of  $\pi_s^*$ . In this paper, we design PG algorithms computing the performative optimal policy for a given set of MDPs, and reinstate their differences with performatively stable policies.

The existing literature on PeMDPs (Mandal et al., 2023; Mandal & Radanovic, 2024; Rank et al., 2024; Pollatos et al., 2025; Chen et al., 2024) focused primarily on finding a performatively stable policy, i.e. a  $\pi_s^*$  according to Definition 4. In practice, while the notion of stable policies matters for very specific applications, a stable policy may not always suffice. But they might show large sub-optimality gaps, which are often not desired for real-life tasks. We fill up this gap in literature and propose the first provably converging and computationally efficient PG algorithm for PeMDPs. Later on, we also empirically show the deficiency of the existing stability finding algorithms if we aim for optimality (Section 5).

**Entropy Regularised PeMDPs.** Entropy regularisation has emerged as a simple but powerful technique in classical RL to design smooth and efficient algorithms with sufficient exploration. Thus, we study another variant of the performative value function that is regularised using discounted entropy (Mei et al., 2020; Neu et al., 2017; Liu et al., 2019; Zhao et al., 2019). In this setting, the original value function in Definition 1 is regularised using the discounted entropy  $H_\pi(\rho) \triangleq \mathbb{E}_{\tau \sim \mathbb{P}_\pi^\pi} [-\sum_{t=0}^{\infty} \gamma^t \log \pi(a_t \mid s_t)]$ . This is equivalent to maximising the expected reward with a shifted reward function  $\tilde{r}_\pi(\pi(s_t), s_t) = r_\pi(\pi(s_t), s_t) - \lambda \log(\pi(a_t \mid s_t))$  for some  $\lambda \geq 0$ .  $\tilde{r}_\pi$  is referred as the “soft-reward” in MDP literature (Wang & Uchibe, 2024; Herman et al., 2016; Shi et al., 2019). This allows us to define the soft performative value function.

**Definition 5** (Entropy Regularised (or Soft) Performative Value Function). *Given a policy  $\pi \in \Pi$ , a starting state distribution  $\rho \in \Delta(S)$ , and a regularisation parameter  $\lambda \geq 0$ , the soft performative value function  $\tilde{V}_\pi^\pi(\rho)$  is*

$$\tilde{V}_\pi^\pi(\rho) \triangleq \mathbb{E}_{\tau \sim \mathbb{P}_\pi^\pi} \left[ \sum_{t=0}^{\infty} \gamma^t (r_\pi(s_t, \pi(s_t)) - \lambda \log \pi(a_t \mid s_t)) \mid s_0 \sim \rho \right] = \mathbb{E}_{\tau \sim \mathbb{P}_\pi^\pi} \left[ \sum_{t=0}^{\infty} \gamma^t \tilde{r}_\pi(s_t, \pi(s_t)) \mid s_0 \sim \rho \right]. \quad (6)$$

Since policies belong to the probability simplex, the entropy regularisation naturally lends to smoother and stable PG algorithms. Later, we show that the discounted entropy is a smooth function of the policy parameters for PeMDPs extending the optimization-wise benefits of entropy regularisation to PeMDPs. Additionally, using the notion of soft rewards, we can further define soft performatively optimal and stable policies for entropy regularised PeMDPs. Leveraging it, we unifiedly design PG algorithms for both the unregularised and the entropy regularised PeMDPs.

### 3 POLICY GRADIENT ALGORITHMS IN PERFORMATIVE RL

In this section, we first study the impact of policy updates in PeMDPs. Then, we leverage it to derive the performative policy gradient theorem and design Performative Policy Gradient (**PePG**) algorithm for any differentiable parametric policy class.

#### 3.1 IMPACT OF POLICY UPDATES ON PEMDPs

Performance difference lemma has been central in RL to understand the impact of changing policies in terms of value functions (Kakade & Langford, 2002a). It has been also central to analysing and developing PG-type methods (Agarwal et al., 2021; Silver et al., 2014; Kallel et al., 2024). But the existing versions of performance difference cannot handle performativity. Here, we derive the performative version of the performance difference lemma that quantifies the shift in the performative value function due to change the deployed and environment-inducing policies.

**Lemma 1** (Performative Performance Difference Lemma). *The difference in performative value functions induced by  $\pi$  and  $\pi' \in \Pi$  while starting from the initial state distribution  $\rho$  is*

$$\begin{aligned} V_\pi^\pi(\rho) - V_{\pi'}^{\pi'}(\rho) &= \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi',\rho}^{\pi'}} [A_{\pi'}^{\pi'}(s, a)] \\ &\quad + \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi',\rho}^{\pi'}} [(r_\pi(s, a) - r_{\pi'}(s, a)) + \gamma(\mathbf{P}_\pi(\cdot|s, a) - \mathbf{P}_{\pi'}(\cdot|s, a))^\top V_\pi^\pi(\cdot)]. \end{aligned} \quad (7)$$

where  $A_{\pi'}^{\pi'}(s, a) \triangleq Q_{\pi'}^{\pi'}(s, a) - V_{\pi'}^{\pi'}(s)$  is the performative advantage function for any state  $s \in \mathcal{S}$  and action  $a \in \mathcal{A}$ .

The crux of the proof is decomposing the performative value through environment-inducing and deployed policies

$$V_\pi^\pi(s_0) - V_{\pi'}^{\pi'}(s_0) = \underbrace{V_\pi^\pi(s_0) - V_{\pi'}^{\pi'}(s_0)}_{\text{performative shift term}} + \underbrace{V_{\pi'}^{\pi'}(s_0) - V_{\pi'}^{\pi'}(s_0)}_{\text{performance difference term}}.$$

(1) *Connection to Classical RL.* In classical RL, the performance difference lemma yields  $V^\pi(\rho) - V^{\pi'}(\rho) = \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_\rho^{\pi'}} [A^{\pi'}(s, a)]$ . The first term in Lemma 1 is equivalent to the classical result in the environment induced by  $\pi'$ . But due to environment shift, two more terms appear in the performative performance difference incorporating the impacts of reward shifts and transition shifts. (2) *Connection to Performative Stability.* If we ignore the reward and transition shift terms, the performance difference term  $V_\pi^\pi(s_0) - V_{\pi'}^{\pi'}(s_0)$  quantifies the impact of changing the deployed policy from  $\pi'$  to  $\pi$  in an environment induced by  $\pi'$ . Thus, a stability seeking algorithm would like to minimise this term, while an optimality seeking algorithm has to incorporate all of the terms.

Now, we ask: *how much do the new environment shift terms change the performative performance difference?*

For simplicity, we focus on the commonly studied PeMDPs with bounded rewards and gradually shifting environments, i.e. the ones with Lipschitz transitions and rewards with respect to the deployed policies (Rank et al., 2024).

**Assumption 1** (Bounded reward). *We assume that the rewards are bounded in  $[-R_{\max}, R_{\max}]$ .*

This is the only assumption needed through the paper and is standard in MDP literature (Mei et al., 2020; Li & Yang, 2023).

**Lemma 2** (Bounding Performative Performance Difference for Gradually Shifting Environments). *Let us assume that both rewards and transitions are Lipschitz functions of policy, i.e.  $\|r_\pi - r_{\pi'}\|_\infty \leq L_r \|\pi - \pi'\|_1$  and  $\|\mathbf{P}_\pi - \mathbf{P}_{\pi'}\|_1 \leq L_{\mathbf{P}} \|\pi - \pi'\|_1$ , for some  $L_r, L_{\mathbf{P}} \geq 0$ . Then, under Assumption 1, the performative shift in the sub-optimality gap of a policy  $\pi_\theta$  satisfies*

$$\left| V_{\pi_\theta^*}^{\pi_\theta^*}(\rho) - V_{\pi_\theta}^{\pi_\theta}(\rho) - \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi_\theta^*,\rho}^{\pi_\theta}} [A_{\pi_\theta}^{\pi_\theta}(s, a)] \right| \leq \frac{2\sqrt{2}}{1-\gamma} (L_r + \frac{\gamma}{1-\gamma} L_{\mathbf{P}} R_{\max}) \mathbb{E}_{s_0 \sim \rho} D_H(\pi_\theta^*(\cdot|s_0) \| \pi_\theta(\cdot|s_0)). \quad (8)$$

where  $D_H(\mathbf{x} \| \mathbf{y})$  denotes the Hellinger distance between  $\mathbf{x}$  and  $\mathbf{y}$ .

---

**Algorithm 2** PePG: Performative Policy Gradient

---

1: **Input:** Transition Feature Map  $\psi(s) \forall s \in \mathcal{S}, \xi \in [-R_{\max}, R_{\max}]$  and discount factor  $\gamma$ .  
2: **Initialize:** Initial policy parameters  $\theta_0$ , initial value function parameters  $\phi_0$   
3: **for**  $k = 1, 2, \dots$  **do**  
4:   **Collect trajectories:**  $\mathcal{D}_k = \{\tau_i\}_{i=1}^I$ , where each  $\tau_i \triangleq \{(s_{i,t}, a_{i,t}, s_{i,t+1}, r_{i,t})\}_{t=0}^{T-1}$  by playing  $\pi_{\theta_k} = \pi(\theta_k)$   
5:   Compute returns  $R_k \triangleq \{R_{k,i}\}_{i=1}^I$ , where  $R_{k,i} = \{R_{k,i,t}\}_{t=0}^{T-1}$   
6:   Compute advantage estimates  $\hat{A}_k(\tau_i)$  using value function  $\hat{V}_{\phi_k}(\tau_i)$  for each  $\tau_i \in \mathcal{D}_k$  (estimate of  $V_{\pi_{\theta_k}}^{\pi_{\theta_k}}(\tau_i)$  obtained from fitted value network with parameters  $\phi_k$ )  
7:   **Gradient estimation:** Estimate policy gradient using (12)  
8:   **Gradient ascent step:** Update policy parameters using (9)  
9:   Fit value function  $V_{\phi_{k+1}}$ :

$$\phi_{k+1} \leftarrow \arg \min_{\phi} \frac{1}{I \cdot T} \sum_{i=1}^I \sum_{t=0}^{T-1} \left( \hat{V}_{\phi_k}(s_t \in \tau_i) - R_{k,i,t} \right)^2$$

10: **end for**

---

*Implication.* Lemma 2 shows novel characterisation of the *extra cost* we pay to adapt to performativity of the environment in terms of Hellinger distance between the true performatively optimal policy  $\pi_o^*$  and any other parametrised policy  $\pi_\theta$ . This implies that the order of difference between the optimal performative value function and that of any stability-seeking algorithm is  $\Theta(\frac{1}{1-\gamma})$ . This significantly improves the known order of sub-optimality achieved by existing algorithms. Specifically, [Mandal et al. \(2023\)](#) show that using repeated policy optimisation algorithms converges to a suboptimality gap  $\mathcal{O}\left(\max\left\{\frac{S^{5/3} A^{1/3} \epsilon^{2/3}}{(1-\gamma)^{14/3}}, \frac{\epsilon S}{(1-\gamma)^4}\right\}\right)$ . Thus, we see an opportunity to improve on the existing works and design algorithms that can achieve suboptimality gap of order  $\Theta(\frac{1}{1-\gamma})$ .

Additionally, we note that an optimality-seeking algorithm tries to minimise both the advantage function and the effect of the shifts in the environment quantified by the Hellinger distance, i.e.,  $D_H(\pi_o^*(\cdot|s_0) \| \pi_\theta(\cdot|s_0))$ . While it suffices for a stability-seeking algorithm to minimise the advantage function, and thus, we cannot minimise the RHS of Equation (8) lower than  $D_H(\pi_o^*(\cdot|s_0) \| \pi_\theta(\cdot|s_0))$ . Thus, optimality-seeking algorithms can achieve a lower performative performance difference than the stability-seeking algorithms if they also learn and incorporate the performative shifts in the environment.

### 3.2 ALGORITHM DESIGN: PERFORMATIVE POLICY GRADIENT (PePG)

To achieve performative optimality, the goal is to maximise value function at the end of learning process. Gradient ascent is a standard first-order optimisation method to find maxima of a function. Similar to Algorithm 1, the crux of performative policy gradient method lies in the ascent step:

$$\theta_{t+1} \leftarrow \begin{cases} \theta_t + \eta_t \nabla_{\theta} V_{\pi_{\theta}}^{\pi_{\theta}}(\tau) |_{\theta=\theta_t}, & \text{for unregularised objective} \\ \theta_t + \eta_t \nabla_{\theta} \tilde{V}_{\pi_{\theta}}^{\pi_{\theta}}(\tau) |_{\theta=\theta_t}, & \text{for Entropy-regularised objective.} \end{cases} \quad (9)$$

Given this ascent step, we have to evaluate the gradient at each time step from the rollouts of the present policy. In classical PG, the policy gradient theorem serves this purpose ([Williams, 1992](#); [Sutton et al., 1999](#); [Silver et al., 2014](#)). Thus, we derive the performative counterpart of the classic policy gradient theorem.

**Theorem 2** (Performative Policy Gradient Theorem). *The gradient of the performative value function w.r.t  $\theta$  is as follows:*

(a) *For the unregularised objective,*

$$\nabla_{\theta} V_{\pi_{\theta}}^{\pi_{\theta}}(\tau) = \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}} \left[ \sum_{t=0}^{\infty} \gamma^t (A_{\pi_{\theta}}^{\pi_{\theta}}(s_t, a_t) (\nabla_{\theta} \log \pi_{\theta}(a_t | s_t) + \nabla_{\theta} \log P_{\pi_{\theta}}(s_{t+1} | s_t, a_t)) + \nabla_{\theta} r_{\pi_{\theta}}(s_t, a_t)) \right], \quad (10)$$

(b) *For the entropy-regularised objective, we define the soft advantage, soft Q, and soft value functions with respect to the soft rewards  $\tilde{r}_{\pi_{\theta}}$  satisfying  $\tilde{A}_{\pi_{\theta}}^{\pi_{\theta}}(s, a) = \tilde{Q}_{\pi_{\theta}}^{\pi_{\theta}}(s, a) - \tilde{V}_{\pi_{\theta}}^{\pi_{\theta}}(s)$  that further yields*

$$\nabla_{\theta} \tilde{V}_{\pi_{\theta}}^{\pi_{\theta}}(\tau) = \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}} \left[ \sum_{t=0}^{\infty} \gamma^t (\tilde{A}_{\pi_{\theta}}^{\pi_{\theta}}(s_t, a_t) (\nabla_{\theta} \log \pi_{\theta}(a_t | s_t) + \nabla_{\theta} \log P_{\pi_{\theta}}(s_{t+1} | s_t, a_t)) + \nabla_{\theta} \tilde{r}_{\pi_{\theta}}(s_t, a_t | \theta)) \right]. \quad (11)$$

**PePG:** To elaborate on the design of PePG (Algorithm 2), we focus only on the REINFORCE update and softmax policy parametrisation. With the appropriate parameter choices, and initialisation of the policy parameter  $\theta$  and value function parameter  $\phi$ , for each episode  $k = 1, 2, \dots$ , PePG collects  $I$  trajectories to calculate return  $R^i$  and estimates advantage function  $\hat{A}_k$  (Line 4-6). For a particular trajectory  $\tau_i$ , the estimated advantage for a given state-action is  $\widehat{A_{\pi_{\theta_k}}}(s_t^i, a_t^i) = R_{t,k}^i - V_{\phi_k}(s_t^i)$ , where  $R^i = \sum_{t=0}^{T-1} \gamma^t r_{\pi_{\theta_k}}(s_t^i, a_t^i)$ .

**Gradient Estimation (Line 7).** With the necessary estimates in hand for all the collected  $I$  trajectories, PePG computes average gradient estimate over all the trajectories using

$$\widehat{\nabla_{\theta_k} V_{\pi_{\theta_k}}} = \frac{1}{I} \sum_{i=1}^I \sum_{t=0}^T \gamma^t (\widehat{A_{\pi_{\theta_k}}}(s_t^i, a_t^i)) \left( \nabla_{\theta_k} \log \pi_{\theta_k}(a_t^i | s_t^i) + \nabla_{\theta_k} \log P_{\pi_{\theta_k}}(s_{t+1}^i | s_t^i, a_t^i) \right) + \nabla_{\theta_k} r_{\pi_{\theta_k}}(s_t^i, a_t^i | \theta_k) \quad (12)$$

where all the individual gradients  $\nabla_{\theta_k} \log P_{\pi_{\theta_k}}$ ,  $\nabla_{\theta_k} r_{\pi_{\theta_k}}$  and  $\nabla_{\theta_k} \log \pi_{\theta_k}$  have the closed form expressions for softmax parametrisation according to Equation (35). Further, in Line 8, PePG updates the policy parameter for the next episode using a gradient ascent step leveraging the estimated average gradient over all  $I$  trajectories. Specifically, we plug in  $\widehat{\nabla_{\theta_k} V_{\pi_{\theta_k}}}$  to both the unregularised and entropy-regularised update rules are given in Equation (9). For the next episode, we again run a regression to update the value network plugging in the current estimates and resume the learning process further.

## 4 CONVERGENCE ANALYSIS OF PePG: SOFTMAX POLICIES AND SOFTMAX PeMDPs

For rigorous theoretical analysis of PePG, we restrict ourselves to *softmax policy class*, and *softmax PeMDPs*. We define the softmax PeMDPs as the ones having softmax transition kernels with feature map  $\psi(\cdot) : \mathcal{S} \rightarrow \mathbb{R}$ , and linear reward functions with respect to the policy parameters, for all state  $s \in \mathcal{S}$  and action  $a \in \mathcal{A}$ . Specifically, the class of softmax PeMDPs is  $\{\mathcal{M}(\theta) = \mathcal{M}(\pi_\theta) \mid \theta \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}\}$  such that

$$\pi_\theta(a|s) = \frac{e^{\theta_{s,a}}}{\sum_{a'} e^{\theta_{s,a'}}}, \quad \mathbf{P}_{\pi_\theta}(s'|s, a) = \frac{e^{\theta_{s,a}\psi(s')}}{\sum_{s''} e^{\theta_{s,a}\psi(s'')}}, \quad r_{\pi_\theta}(s, a) = \mathcal{P}_{[-R_{\max}, R_{\max}]}[\xi \theta_{s,a}], \quad (13)$$

where  $\psi$  is non-negative and upper bounded by  $\psi_{\max} > 0$ , and  $\xi \in [0, R_{\max}]$  to align with Assumption 1.

Thus, we derive the derivatives of policy, transitions, and rewards as

$$\begin{aligned} \frac{\partial}{\partial \theta_{s',a'}} \log \pi_\theta(a|s) &= \mathbb{1}[s = s', a = a'] - \pi_\theta(a'|s) \mathbb{1}[s = s'], \\ \frac{\partial}{\partial \theta_{s',a'}} \log \mathbf{P}_{\pi_\theta}(s''|s, a) &= \psi(s'') \mathbb{1}[s = s', a = a'] (1 - \mathbf{P}_{\pi_\theta}(s''|s, a)), \quad \frac{\partial}{\partial \theta_{s',a'}} r_{\pi_\theta}(s, a) = \xi \mathbb{1}[s = s', a = a']. \end{aligned} \quad (14)$$

Given the derivatives, we can now readily estimate the policy gradient and deploy PePG for softmax PeMDPs.

**Convergence Analysis: Challenges and Three Step Analysis.** The main challenge to prove convergence of PePG is that the performative value function is not concave in the parameterisation  $\theta$ , in general, and also in softmax PeMDPs. The similar issue occurs while proving convergence of PG-type algorithms in classical RL, which has been overcome by leveraging smoothness properties of the value functions and by deriving the local Polyak-Lojasiewicz (PL)-type conditions, known as *gradient domination*, with respect to the policy parameterisation. Leveraging these insights, we devise a three step convergence analysis for PePG.

**Step 1: Smoothness of Performative Value Functions.** First, we prove that the unregularised performative value function is  $\mathcal{O}(\frac{|\mathcal{A}|}{(1-\gamma)^2})$  smooth. As we show that the entropy is also a smooth function for softmax PeMDPs, then under proper choice of the regularisation parameter, i.e.,  $\lambda = \frac{1-\gamma}{1+2\log|\mathcal{A}|}$ , entropy regularised performative value function is also  $\mathcal{O}(\frac{|\mathcal{A}|}{(1-\gamma)^2})$  smooth. Since gradient ascent/descent methods can work well in smooth functions, we proceed thoroughly.

**Step 2: Gradient Domination for Softmax PeMDPs.** Now, the next step is to relate the performative performance difference with the performative policy gradient. This allows us to connect the per iteration improvement in the performative value function with the performative gradient descent at that step. These are known as PL-type inequalities. For non-concave objectives, PL inequalities guarantee convergence to global maxima by showing that the gradient of the objective at any parameter dominates the sub-optimality w.r.t. that parameter.

**Lemma 3** (Performative Gradient Domination for Softmax PeMDPs). *Let us consider PeMDPs defined in (13).*

	Algorithms	Regulariser $\lambda$	Min. #samples	Environment
378	RPO FS (Mandal et al., 2023)	$\mathcal{O}\left(\frac{ \mathcal{S}  + \gamma  \mathcal{S} ^{5/2}}{(1-\omega)(1-\gamma)^4}\right)$	$\frac{ \mathcal{A} ^2  \mathcal{S} ^3}{\epsilon^4 (1-\gamma)^6 \lambda^2} \ln(\#\text{iter})$	Direct PeMDPs + quadratic-regul. on occupancy $\omega$ -dependence between two envs.
379	MDRR (Rank et al., 2024)	$\mathcal{O}\left(\frac{ \mathcal{S}  + \gamma  \mathcal{S} ^{5/2}}{(1-\omega)(1-\gamma)^4}\right)$	$\frac{ \mathcal{A} ^2  \mathcal{S} ^3}{\epsilon^4 (1-\gamma)^6 \lambda^2} \ln(\#\text{iter})$	Direct PeMDPs + quadratic-regul. on occupancy $\omega$ -dependence between two envs.
380	PePG (This paper)	$\frac{R_{\max}(1-\gamma)}{1+\log( \mathcal{A} )}$	$\frac{ \mathcal{S}   \mathcal{A} ^2}{\epsilon^2 (1-\gamma)^3}$	softmax PeMDPs + entropy regul. on policy
381	PePG (This paper)	0	$\frac{ \mathcal{S}   \mathcal{A} }{\epsilon^2} \max\left\{\frac{\gamma R_{\max}  \mathcal{A} }{(1-\gamma)^3}, \frac{\gamma^2}{(1-\gamma)^4}\right\}$	unregularised softmax PeMDPs

Table 1: Comparison of theoretical performance of SOTA stability-seeking algorithms against PePG.

(a) For unregularised value function,

$$V_{\pi_o^*}(\rho) - V_{\pi_\theta}(\rho) \leq \sqrt{|\mathcal{S}||\mathcal{A}|} \left\| \frac{d_{\pi_\theta, \rho}^{\pi_o^*}}{d_{\pi_\theta, \nu}^{\pi_o^*}} \right\| \|\nabla_\theta V_{\pi_\theta}(\nu)\|_2 + \frac{R_{\max}}{1-\gamma} \left( 1 + \frac{2\gamma}{1-\gamma} \psi_{\max} \right). \quad (15)$$

(b) For entropy-regularised value function,  $\tilde{V}_{\pi_o^*}(\rho) - \tilde{V}_{\pi_\theta}(\rho) \leq$ 

$$\sqrt{|\mathcal{S}||\mathcal{A}|} \left\| \frac{d_{\pi_\theta, \rho}^{\pi_o^*}}{d_{\pi_\theta, \nu}^{\pi_o^*}} \right\| \|\nabla_\theta V_{\pi_\theta}(\nu)\|_2 + \frac{R_{\max}}{1-\gamma} \left( 1 + \frac{2\gamma}{1-\gamma} \psi_{\max} \left( 1 + \frac{\lambda}{R_{\max}} \log |\mathcal{A}| \right) \right) + \frac{\lambda}{1-\gamma} (1 + 2 \log |\mathcal{A}|). \quad (16)$$

**Step 3: Iterative Application of Gradient Domination for Smooth Functions.** Now, we can apply gradient domination along with the classic iterative convergence proof of gradient ascent for smooth functions. The intuition is that since the per-step sub-optimality is dominated by the gradient and the smooth functions are bounded by quadratic envelopes of parameters, applying gradient ascent iteratively would bring the sub-optimality down to small error level after enough iterations. We formalise this in Theorem 3.

**Theorem 3** (Convergence of PePG in softmax PeMDPs). *Let  $\text{Cov} \triangleq \max_{\theta, \nu} \left\| \frac{d_{\pi_\theta, \rho}^{\pi_o^*}}{d_{\pi_\theta, \nu}^{\pi_o^*}} \right\|_\infty$ . The gradient ascent algorithm on  $V_{\pi_\theta}(\rho)$  (Equation (9)) satisfies, for all distributions  $\rho \in \Delta(\mathcal{S})$ .*

(a) in the unregularised case with  $\eta = \Omega(\min\{\frac{(1-\gamma)^2}{\gamma|\mathcal{A}|}, \frac{(1-\gamma)^3}{\gamma^2}\})$ ,  $\min_{t < T} \left\{ V_{\pi_o^*}(\rho) - V_{\pi_{\theta_t}}(\rho) \right\} \leq \epsilon + \mathcal{O}\left(\frac{1}{1-\gamma}\right)$  when  $T = \Omega\left(\frac{|\mathcal{S}||\mathcal{A}|\text{Cov}^2}{\epsilon^2} \max\left\{\frac{\gamma R_{\max} |\mathcal{A}|}{(1-\gamma)^3}, \frac{\gamma^2}{(1-\gamma)^4}\right\}\right)$ .

(b) in the entropy regularisation scenario with  $\lambda = \frac{(1-\gamma)R_{\max}}{1+2\log|\mathcal{A}|}$  and  $\eta = \Omega\left(\frac{(1-\gamma)^2}{\gamma|\mathcal{A}|}\right)$ ,  $\min_{t < T} \left\{ \tilde{V}_{\pi_o^*}(\rho) - \tilde{V}_{\pi_{\theta_t}^{(t)}}(\rho) \right\} \leq \epsilon + \mathcal{O}\left(\frac{1}{1-\gamma}\right)$  when  $T = \Omega\left(\frac{|\mathcal{S}||\mathcal{A}|^2 \text{Cov}^2}{\epsilon^2 (1-\gamma)^3}\right)$ .

**Implications.** (1) We observe that PePG converges to an  $\epsilon$ -optimal policy in  $\frac{|\mathcal{S}||\mathcal{A}|^2}{\epsilon^2 (1-\gamma)^3}$  iterations. This reduces the sample complexity required for the existing stability-seeking algorithms by at least an order  $\frac{|\mathcal{S}|^2}{\epsilon^2 (1-\gamma)^3}$ , and shows efficiency of using PePG than the algorithms directly optimising the occupancy measures. (2) Additionally, the regularisation parameters needed for the existing algorithms are pretty big and bigger than  $\frac{|\mathcal{S}|}{(1-\gamma)^4}$ . This is counter-intuitive and does not match the experimental observations. Here, we prove that setting the regularisation parameter to  $\frac{(1-\gamma)R_{\max}}{1+2\log|\mathcal{A}|}$  suffices for proving convergence to optimality. (3) The minimum number of samples required to achieve convergence is proportional to the square of coverage for the softmax PeMDP. This is a ubiquitous quantity dictating convergence of PG-methods in classical RL (Agarwal et al., 2021; Mei et al., 2020), and retraining methods in performative RL (Mandal et al., 2023; Rank et al., 2024). (4) The  $\mathcal{O}\left(\frac{1}{1-\gamma}\right)$  suboptimality gap appearing in Theorem 3 is analogous to the effect of using relaxed weak gradient domination result (Yuan et al., 2022, Corollary 3.7). It argues that if the policy gradient in classical MDPs satisfies the relaxed weak gradient domination, i.e.,  $\epsilon' + \|\nabla_\theta V(\theta)\| \geq 2\sqrt{\mu}(V^* - V(\theta))$  for some  $\mu > 0$  and  $\epsilon' > 0$ , then the corresponding policy gradient method guarantees  $\min_{t \in \{0, \dots, T\}} (V^* - V(\theta_t)) \leq \mathcal{O}(\epsilon) + \mathcal{O}(\epsilon')$  for big enough  $T$ . Lemma 3 constructs the performative counterpart

of this relaxed weak gradient domination property with  $\epsilon' = \mathcal{O}\left(\frac{1}{1-\gamma}\right)$ . Similarly, (Sahitaj et al., 2025) also supports existence of such a gap empirically for Markov potential games. Thus, this indicates an inherent property of performative policy gradient which has to incorporate gradients of transitions and rewards along with gradients of policies at every step.

## 5 EXPERIMENTAL ANALYSIS

In this section, we empirically compare the performance of PePG in the performative reinforcement learning setting and analyse its behaviour against the state-of-the-art stability-finding methods.<sup>2</sup>

**Performative RL Environment.** We evaluate PePG in the Gridworld test-bed (Mandal et al., 2023), which has become a standard benchmark in performative RL. This environment consists of a grid where two agents  $A_1$  (the principal) and  $A_2$  (the follower), jointly control an actor navigating from start positions (S) to the goal (G) while avoiding hazards. The environment dynamics are as follows: Agent  $A_1$  proposes a control policy for the actor by selecting one of four directional actions. Agent  $A_2$  can either accept this action (not intervene) or override it with its own directional choice. *This creates a performative environment for  $A_1$ , as its effective policy outcomes depend on  $A_2$ 's responses to its deployed strategy.*

The cost structure follows: visiting blank cells (S) incurs penalty of  $-0.01$ , goal cells (F) cost  $-0.02$ , hazard cells (H) impose a severe penalty of  $-0.5$ , and any intervention by  $A_2$  results in an additional cost of  $-0.05$  for the intervening agent. The response model also follows that of Mandal et al. (2023), i.e., the agent  $A_2$  responds to  $A_1$ 's policy using a Boltzmann softmax operator. Given  $A_1$ 's current policy  $\pi_1$ , we compute the optimal Q-function  $Q^*|\pi_1$  for each follower agent  $A_j$  relative to a perturbed version of the grid world, where each cell types matches  $A_1$ 's environment with probability 0.7. We then define an average Q-function over the follower agents and determine the collective response policy via Boltzmann softmax  $Q^*|\pi_1(s, a) = \frac{1}{n} \sum_{j=2}^{n+1} Q_j^*|\pi_1(s, a), \pi_2(a|s) = \frac{\exp(\beta \cdot Q^*|\pi_1(s, a))}{\sum_{a'} \exp(\beta \cdot Q^*|\pi_1(s, a'))}$ .

It is important to note that our experimental setup deliberately uses the immediate response model from the original performative RL framework, rather than the gradually shifting environment introduced by Rank et al. (2024) that assumes slow shifts in the environment. Our choice to use the immediate response model presents a more challenging performative setting where the environment responds instantaneously to policy changes. This allows us to demonstrate that unlike MDRR (Rank et al., 2024), PePG can handle the fundamental performative challenge without requiring environmental assumptions that artificially slows down the feedback loop, thereby highlighting the robustness of the proposed PePG approach.

**Experimental Setup.** We evaluate PePG (with and without entropy regularisation) alongside Mixed Delayed Repeating Retraining (MDRR), which represents the current state-of-the-art in performative reinforcement learning under gradually shifting environments (Rank et al., 2024), and Repeated Policy Optimization with Finite Samples (RPO FS). MDRR has demonstrated significant improvements over traditional repeated retraining methods, by leveraging historical data from multiple deployments, while RPO FS is included as the baseline method from (Mandal et al., 2023) for direct comparison with the original performative RL approach.

All experiments use a  $8 \times 8$  grid with  $\gamma = 0.9$ , exploration parameter  $\epsilon = 0.5$  for initial policy construction, one follower agent  $A_2$ , and 100 trajectory samples per iteration. The algorithms share common parameters of  $T = 100$  iterations. For regularization, RPO FS and MDRR use  $\lambda = 0.1$  from their original experiments, while entropy-regularized PePG uses  $\lambda = 2.0$  (ablation studies for this choice are provided in the appendix). PePG uses learning rate  $\eta = 0.1$ , MDRR employs memory weight  $v = 1.1$  for historical data utilization, delayed round parameter  $k = 3$ , and FTRL parameters  $N = B = 10$ , while RPO FS follows the finite-sample optimization from Mandal et al.

**Results and Observations.** Our experimental evaluation across 200 iterations reveals fundamental differences between PePG and MDRR and RPO in the immediate response performative setting. We used shorter training compared to (Rank et al., 2024), as this time-frame sufficiently demonstrates RPO and MDRR's stability convergence and PePG's progression toward optimality.

**I. Results: Optimality:** The left panel reveals a clear performance hierarchy among the four methods. PePG achieves the highest value function performance, with standard PePG approaching zero (potentially crossing it in the near future) and regularized PePG (Reg PePG) reaching a value close to 0.3, both showing consistent improvement from initial values around  $-0.15$  and still progressing upward at the end of the 100 iteration window. This steady upward progression highlights PePG's effectiveness in discovering better performative equilibria rather than settling for the first stable solution encountered. MDRR remains relatively stable around  $-0.2$  throughout training, while RPO FS stabilizes at the lowest performance level of approximately  $-0.3$  and remains flat throughout training.

<sup>2</sup>Anonymous code repository of PePG implementation is [Link](#). Further ablation studies w.r.t. hyperparameters are in Appendix H.

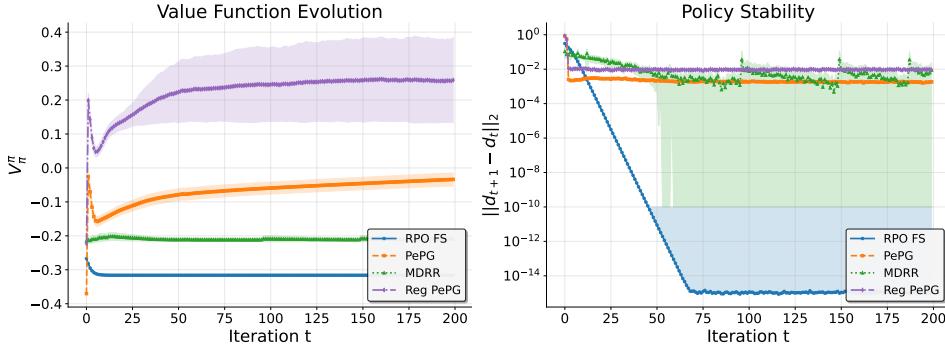


Figure 2: Comparison of evolution in expected average return (both regularised and unregularised) and stability of **PePG** with SOTA stability-achieving methods. Each algorithm is run for 20 random seeds and 200 iterations.

**II. Results: Comparison of Optimality- and Stability-seeking Algorithms.** The results expose a critical limitation of algorithms designed primarily for stability rather than optimality. MDRR successfully achieves its design goal, with the right panel showing decreasing toward zero in the stability metric  $\|d_{t+1} - d_t\|_2$  (the  $L_2$  distance between occupancy measures of consecutive policy iterations), indicating policy stabilization. However, this stability comes at the cost of solution quality, as MDRR becomes trapped in a suboptimal point. The method prioritised finding any stable point over finding an optimal solution. In contrast, both PePG variants exhibit higher policy variability as they actively explore for better solutions. RPO FS maintains high stability with successive occupancy difference below  $10^{-14}$  but with limited performance improvement.

## 6 DISCUSSIONS, LIMITATIONS, AND FUTURE WORKS

We study the problem of Performative Reinforcement learning in tabular MDPs (PeMDPs) using softmax parametrised policies with entropy-regularised objective function, where any action taken by the agent cause potential shift in the MDP's underlying reward and transition dynamics. We are the first to develop PG-type algorithm, **PePG**, that attains performatively optimality against the existing performative stability-seeking algorithms, affirmatively solving an extended open problem in (Mandal et al., 2023). We also derive the novel performative counterpart of classic Performance Difference Lemma and Policy Gradient Theorem that affirmatively captures this performative nature of the environment we act. We provide a sufficient conditions to prove that **PePG** converges to an  $(\epsilon + \frac{1}{1-\gamma})$ -ball around performative optimal policy in  $\Omega(\frac{|\mathcal{S}||\mathcal{A}|^2}{\epsilon^2(1-\gamma)^3})$  iterations.

As we develop a PG-type algorithm, it will be interesting to see how much can we reduce the variance (Wu et al., 2018; Papini et al., 2018) in estimation while achieving optimality. We are still in the tabular setting with finite set of state-actions. A potential future direction would be to scale **PePG** to continuous state-space with large number of state-actions.

## REFERENCES

- Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021. [2](#), [3](#), [5](#), [8](#), [17](#)
- 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 preprint arXiv:2204.05862*, 2022. [1](#)
- Anas Barakat, John Lazarsfeld, Georgios Piliouras, and Antonios Varvitsiotis. Multi-agent online control with adversarial disturbances. *arXiv preprint arXiv:2506.18814*, 2025. [1](#), [17](#)
- James Bell, Linda Linsefors, Caspar Oesterheld, and Joar Skalse. Reinforcement learning in newcomblike environments. *Advances in Neural Information Processing Systems*, 34:22146–22157, 2021. [2](#), [17](#)
- Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *International conference on artificial intelligence and statistics*, pp. 6045–6061. PMLR, 2022. [17](#)
- Songfu Cai, Fei Han, and Xuanyu Cao. Performative control for linear dynamical systems. *Advances in Neural Information Processing Systems*, 37:70617–70658, 2024. [17](#)

- 540 Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. How algorithmic confounding in recommendation systems  
 541 increases homogeneity and decreases utility. In *Proceedings of the 12th ACM conference on recommender systems*, pp.  
 542 224–232, 2018. 1
- 543 Qianyi Chen, Ying Chen, and Bo Li. Practical performative policy learning with strategic agents. *arXiv preprint*  
 544 *arXiv:2412.01344*, 2024. 4
- 546 Kamil Ciosek and Shimon Whiteson. Expected policy gradients for reinforcement learning. *Journal of Machine Learning*  
 547 *Research*, 21(52):1–51, 2020. 3
- 549 Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience [John Wiley & Sons], Hoboken,  
 550 NJ, 2nd edition, 2006. ISBN 978-0-471-24195-9. 35
- 551 Itay Eilat and Nir Rosenfeld. Performative recommendation: diversifying content via strategic incentives. In *International*  
 552 *Conference on Machine Learning*, pp. 9082–9103. PMLR, 2023. 1, 17
- 554 Mohammad Ghavamzadeh and Yaakov Engel. Bayesian policy gradient algorithms. *Advances in neural information process-*  
 555 *ing systems*, 19, 2006. 3
- 556 António Góis, Mehrnaz Mofakhami, Fernando P Santos, Gauthier Gidel, and Simon Lacoste-Julien. Performative prediction  
 557 on games and mechanism design. *arXiv preprint arXiv:2408.05146*, 2024. 1, 17
- 559 Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner. Performative power. *Advances in Neural Information*  
 560 *Processing Systems*, 35:22969–22981, 2022. 1, 17
- 561 Alexander Havrilla, Yuqing Du, Sharath Chandra Raparth, Christoforos Nalmpantis, Jane Dwivedi-Yu, Eric Hambro, Sain-  
 562 bayar Sukhbaatar, and Roberta Raileanu. Teaching large language models to reason with reinforcement learning. In *AI for*  
 563 *Math Workshop@ ICML 2024*. 1
- 565 Michael Herman, Tobias Gindele, Jörg Wagner, Felix Schmitt, and Wolfram Burgard. Inverse reinforcement learning with  
 566 simultaneous estimation of rewards and dynamics. In *Artificial intelligence and statistics*, pp. 102–110. PMLR, 2016. 4
- 567 Zachary Izzo, Lexing Ying, and James Zou. How to learn when data reacts to your model: performative gradient descent. In  
 568 *International Conference on Machine Learning*, pp. 4641–4650. PMLR, 2021. 1, 17
- 570 Zachary Izzo, James Zou, and Lexing Ying. How to learn when data gradually reacts to your model. In *International*  
 571 *Conference on Artificial Intelligence and Statistics*, pp. 3998–4035. PMLR, 2022. 1, 17
- 572 Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nine-*  
 573 *teenth international conference on machine learning*, pp. 267–274, 2002a. 5
- 575 Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001. 2
- 576 Sham M. Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems 14 (NIPS 2001)*, pp.  
 577 1531–1538, 2002. 17
- 579 Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Confer-*  
 580 *ence on Machine Learning*, 2002b. URL <https://api.semanticscholar.org/CorpusID:31442909>. 18
- 582 Mahdi Kallel, Debabrota Basu, Riad Akroud, and Carlo D’Eramo. Augmented bayesian policy search. In *The Twelfth*  
 583 *International Conference on Learning Representations (ICLR)*, 2024. URL <https://openreview.net/forum?id=OvlcyABNQT>. 5
- 585 Vijay R. Konda and John N. Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems 13*  
 586 (*NIPS 2000*), pp. 1008–1014, 2000. 17
- 588 Qiang Li and Hoi-To Wai. State dependent performative prediction with stochastic approximation. In *International Confer-*  
 589 *ence on Artificial Intelligence and Statistics*, pp. 3164–3186. PMLR, 2022. 1, 17
- 590 Qiang Li, Chung-Yiu Yau, and Hoi-To Wai. Multi-agent performative prediction with greedy deployment and consensus  
 591 seeking agents. *Advances in Neural Information Processing Systems*, 35:38449–38460, 2022. 17
- 593 Shengshi Li and Lin Yang. Horizon-free learning for markov decision processes and games: stochastically bounded rewards  
 and improved bounds. In *International Conference on Machine Learning*, pp. 20221–20252. PMLR, 2023. 5

- 594 Jingbin Liu, Xinyang Gu, and Shuai Liu. Policy optimization reinforcement learning with entropy regularization. *arXiv*  
 595 *preprint arXiv:1912.01557*, 2019. 4
- 596
- 597 Debmalya Mandal and Goran Radanovic. Performative reinforcement learning with linear markov decision process. *arXiv*  
 598 *preprint arXiv:2411.05234*, 2024. 2, 3, 4, 17
- 599 Debmalya Mandal, Stelios Triantafyllou, and Goran Radanovic. Performative reinforcement learning, 2023. URL <https://arxiv.org/abs/2207.00046>. 2, 3, 4, 6, 8, 9, 10, 17
- 600
- 601 Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. Feedback loop and  
 602 bias amplification in recommender systems. In *Proceedings of the 29th ACM international conference on information &*  
 603 *knowledge management*, pp. 2145–2148, 2020. 1
- 604
- 605 Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy  
 606 gradient methods. In *International conference on machine learning*, pp. 6820–6829. PMLR, 2020. 3, 4, 5, 8, 24, 34
- 607
- 608 Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33:4929–4939, 2020. 17
- 609
- 610 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. PMLR, 2021. 1, 17
- 611
- 612 Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and  
 613 Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 1928–1937, 2016. 17
- 614
- 615 Mehrnaz Mofakhami, Ioannis Mitliagkas, and Gauthier Gidel. Performative prediction with neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 11079–11093. PMLR, 2023. 1, 17
- 616
- 617 Adhyyan Narang, Evan Faulkner, Dmitriy Drusvyatskiy, Maryam Fazel, and Lillian J Ratliff. Multiplayer performative prediction: Learning in decision-dependent games. *Journal of Machine Learning Research*, 24(202):1–56, 2023. 1, 17
- 618
- 619 Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017. 4
- 620
- 621 Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *International conference on machine learning*, pp. 4026–4035. PMLR, 2018. 10
- 622
- 623 Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020. 1, 17
- 624
- 625 Georgios Piliouras and Fang-Yi Yu. Multi-agent performative prediction: From global stability and optimality to chaos. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pp. 1047–1074, 2023. 1, 17
- 626
- 627 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. 4
- 628
- 629 Ben Rank, Stelios Triantafyllou, Debmalya Mandal, and Goran Radanovic. Performative reinforcement learning in gradually shifting environments. *arXiv preprint arXiv:2402.09838*, 2024. 2, 4, 5, 8, 9, 17
- 630
- 631 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. 17
- 632
- 633 Rilind Sahitaj, Paulius Sasnauskas, Yiğit Yalın, Debmalya Mandal, and Goran Radanović. Independent learning in performative markov potential games. *arXiv preprint arXiv:2504.20593*, 2025. 9
- 634
- 635 John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 1889–1897, 2015. 2, 17
- 636
- 637 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. In *arXiv preprint arXiv:1707.06347*, 2017. 2, 17
- 638
- 639 Wenjie Shi, Shiji Song, and Cheng Wu. Soft policy gradient method for maximum entropy deep reinforcement learning. *arXiv preprint arXiv:1909.03198*, 2019. 4

- 648 David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient  
 649 algorithms. In *International conference on machine learning*, pp. 387–395. Pmlr, 2014. 1, 5, 6  
 650
- 651 Jason St. John, Christian Herwig, Diana Kafkes, Jovan Mitrevski, William A Pellico, Gabriel N Perdue, Andres Quintero-  
 652 Parra, Brian A Schupbach, Kiyomi Seiya, Nhan Tran, et al. Real-time artificial intelligence for accelerator control: A study  
 653 at the fermilab booster. *Physical Review Accelerators and Beams*, 24(10):104601, 2021. 1
- 654 Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press, 1998. 1  
 655
- 656 Richard S. Sutton, David A. McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement  
 657 learning with function approximation. In *Advances in Neural Information Processing Systems 12 (NIPS 1999)*, pp. 1057–  
 658 1063. The MIT Press, 1999. 2, 3, 6, 17
- 659 Jixin Wang and Eiji Uchibe. Reward-punishment reinforcement learning with maximum entropy. In *2024 International  
 660 Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2024. 4  
 661
- 662 Xiaolu Wang, Chung-Yiu Yau, and Hoi To Wai. Network effects in performative prediction games. In *International Confer-  
 663 ence on Machine Learning*, pp. 36514–36540. PMLR, 2023. 17
- 664 Yue Wang and Shaofeng Zou. Policy gradient method for robust reinforcement learning. In *International conference on  
 665 machine learning*, pp. 23484–23526. PMLR, 2022. 3  
 666
- 667 Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine  
 668 Learning*, 8:229–256, 1992. 2, 3, 6, 17
- 669 Cathy Wu, Aravind Rajeswaran, Yan Duan, Vikash Kumar, Alexandre M Bayen, Sham Kakade, Igor Mordatch, and  
 670 Pieter Abbeel. Variance reduction for policy gradient with action-dependent factorized baselines. *arXiv preprint  
 671 arXiv:1803.07246*, 2018. 10  
 672
- 673 Tengyu Xu, Haoyang Wang, and Yingbin Liang. Improved sample complexity for policy gradient via Poincaré inequality. In  
 674 *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 15728–15739, 2020. 17  
 675
- 676 Rui Yuan, Robert M Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient. In  
 677 *International Conference on Artificial Intelligence and Statistics*, pp. 3332–3380. PMLR, 2022. 2, 3, 8, 17  
 678
- 679 Angela Zhang, Lei Xing, James Zou, and Joseph C Wu. Shifting machine learning for healthcare from development to  
 deployment and from models to data. *Nature biomedical engineering*, 6(12):1330–1345, 2022. 1, 17  
 680
- 681 Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. On the global convergence of policy gradient methods with variance  
 682 reduction. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 11106–11115. PMLR,  
 2020. 17  
 683
- 684 Rui Zhao, Xudong Sun, and Volker Tresp. Maximum entropy-regularized multi-goal reinforcement learning. In *International  
 685 Conference on Machine Learning*, pp. 7553–7562. PMLR, 2019. 4  
 686
- 687
- 688
- 689
- 690
- 691
- 692
- 693
- 694
- 695
- 696
- 697
- 698
- 699
- 700
- 701

702

703

704

705

706

707

708

# Appendix

## Table of Contents

---

709	<b>A Notations</b>	<b>15</b>
710		
711	<b>B Details of the Toy Example: Loan Approvalment Problem</b>	<b>16</b>
712		
713	<b>C Extended Related Works</b>	<b>17</b>
714		
715	<b>D Impact of Policy Updates on PeMDPs (Section 3.1)</b>	<b>18</b>
716		
717	<b>E Smoothness of Performative Value Function and Entropy Regulariser</b>	<b>21</b>
718		
719	<b>F Derivation of Performative Policy Gradients</b>	<b>27</b>
720		
721	<b>G Convergence of PePG : Proofs of Section 4</b>	<b>29</b>
722	G.1 Proofs for Unregularised Value Function . . . . .	29
	G.2 Proofs for Entropy-regularised or Soft Value Function . . . . .	33
723		
724	<b>H Ablation study on entropy regularisation</b>	<b>38</b>
725		
726	<b>I Technical Lemmas</b>	<b>39</b>

---

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756

## A NOTATIONS

757

758

759

760

761

762

763

764

765

766

767

768

769

770

771

772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

Notation	Description
$\mathcal{S}$	state space
$\mathcal{A}$	action space
$\gamma$	discount factor
$\pi_\theta$	policy parametrized by $\theta$
$\Pi(\Theta)$	policy space
$\mathbf{P}_\pi$	transition under the environment induced by policy $\pi$
$r_\pi$	reward under the environment induced by policy $\pi$
$\pi_s^*$	performatively stable policy
$\pi_o^*$	performatively optimal policy
$\mathbf{P}_{\pi_o^*}$	reward under the environment induced by performatively optimal policy
$r_{\pi_o^*}$	reward under the environment induced by performatively optimal policy
$d_{\pi_o^*}^{\pi_o^*}$	state-action occupancy of optimal policy
$V_{\pi_o^*}^{\pi_o^*}$	value function of optimal policy
$d_{\pi_2}^{\pi_1}$	state-action occupancy of playing policy $\pi_2$ in the environment induced by policy $\pi_1$
$V_{\pi_1}^{\pi_2}$	value function for playing policy $\pi_2$ in the environment induced by policy $\pi_1$
$Q_{\pi_1}^{\pi_2}$	Q-value function for playing policy $\pi_2$ in the environment induced by policy $\pi_1$
$A_{\pi_1}^{\pi_2}$	advantage function for playing policy $\pi_2$ in the environment induced by policy $\pi_1$
$\Delta_K$	$K$ -dimensional simplex
$\rho$	Initial state distribution $\in \Delta_{\mathcal{S}}$

## B DETAILS OF THE TOY EXAMPLE: LOAN APPROVEMENT PROBLEM

*Environment.* We consider a population of loan applicants represented by a scalar feature  $x \in \mathbb{R}$ , distributed as  $x \sim \mathcal{N}(\mu, \sigma^2)$ , where  $\mu$  is the population mean and  $\sigma > 0$  is fixed.

*Bank's Policy.* The bank chooses a *threshold policy* parameterized by  $\theta \in \mathbb{R}$ . A loan is granted to an applicant  $x$  if  $x \geq \theta$ . To smooth analysis, we use a differentiable approximation:  $\pi_\theta(x) = \sigma(k(x - \theta))$ , where  $\sigma(z) = \frac{1}{1+e^{-z}}$  is the logistic sigmoid and  $k > 0$  controls smoothness.

*Rewards.* If a loan is granted to applicant  $x$ , the bank receives a random payoff:

$$r(x) = \begin{cases} +R & \text{if applicant repays,} \\ -L & \text{if applicant defaults,} \end{cases}$$

with repayment probability  $\mathbb{P}(\text{repay} \mid x) = \sigma(\gamma x - c)$ , where  $\gamma > 0$  controls sensitivity and  $c$  is a calibration constant. The expected reward from granting to  $x$  is

$$u(x) = \sigma(\gamma x - c) \cdot R - (1 - \sigma(\gamma x - c)) \cdot L.$$

*Expected Utility.* Given distribution  $x \sim \mathcal{N}(\mu, \sigma^2)$ , the bank's expected utility for policy  $\theta$  is

$$U(\theta, \mu) = \mathbb{E}_{x \sim \mathcal{N}(\mu, \sigma^2)} [\pi_\theta(x) \cdot u(x)].$$

*Performative Feedback.* The population mean  $\mu$  depends on the bank's policy, via the grant rate:  $g(\theta, \mu) = \mathbb{E}_{x \sim \mathcal{N}(\mu, \sigma^2)} [\pi_\theta(x)]$ .

We assume a bounded performative update rule:  $\mu_{t+1} = (1 - \beta)\mu_t + \beta \cdot f(g(\theta, \mu_t))$ , where  $\beta \in [0, 1]$  is the performative strength and  $f(g) \in [-M, M]$  maps the grant rate to a feasible population mean.

At equilibrium, the induced feature distribution satisfies the fixed point condition:

$$\mu^*(\theta) = (1 - \beta)\mu^*(\theta) + \beta f(g(\theta, \mu^*(\theta))).$$

*Optimization Problems.* **ERM Optimum.** Ignoring performative effects (i.e. assuming  $\mu = \mu_0$  is fixed), the ERM-optimal policy solves

$$\theta^{\text{ERM}} = \arg \max_{\theta} U(\theta, \mu_0).$$

**Performative Optimum.** Accounting for performative feedback, the performative-optimal policy solves

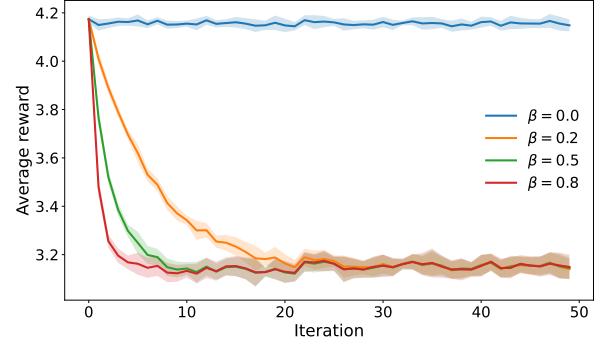
$$\theta^{\text{Perf}} = \arg \max_{\theta} U(\theta, \mu^*(\theta)).$$

### Learning via Reinforcement Learning

An RL agent plays policies  $\theta_t$  sequentially. At each round  $t$ :

1. Sample  $x \sim \mathcal{N}(\mu_t, \sigma^2)$ .
2. Grant loan with probability  $\pi_{\theta_t}(x)$ .
3. Observe reward  $r_t$ .
4. Update  $\theta_{t+1}$  using policy gradient (REINFORCE).
5. Update population mean via performative dynamics:

$$\mu_{t+1} = (1 - \beta)\mu_t + \beta f(g(\theta_t, \mu_t)).$$



864

## C EXTENDED RELATED WORKS

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

**Performative Prediction.** The study of performative prediction started with the pioneering work of (Perdomo et al., 2020), where they leveraged repeated retraining with the aim to converge towards a performatively stable point. We see extension of this work trying to achieve performative optimality (Izzo et al., 2021; 2022; Miller et al., 2021). This further opened a plethora of works in various other domains such as Multi-agent systems (Narang et al., 2023; Li et al., 2022; Piliouras & Yu, 2023), control systems (Cai et al., 2024; Barakat et al., 2025), stochastic optimisation (Li & Wai, 2022; Mendler-Dünner et al., 2020), games (Wang et al., 2023; Góis et al., 2024) etc. There has been several attempt of achieve performative optimality or stability for real-life tasks like recommendation (Eilat & Rosenfeld, 2023), to measure the power of firms (Hardt et al., 2022; Mofakhami et al., 2023), in healthcare (Zhang et al., 2022) etc. Another interesting setting is the *stateful* performative prediction i.e. prediction under gradual shifts in the distribution (Brown et al., 2022; Izzo et al., 2022; Ray et al., 2022), that paved the way for incorporating performative prediction in Reinforcement Learning.

**Performative Reinforcement Learning.** Bell et al. (2021) were the first to propose a setting where the transition and reward of an underlying MDP depend non-deterministically on the deployed policy, thus capturing the essence of performativity to some extent. However, Mandal et al. (2023) can be considered the pioneer in introducing the notion of “*Performative Reinforcement Learning*” and its solution concepts, performatively stable and optimal policy. They propose direct optimization and ascent based techniques which manage to attain performative stability upon repeated retraining. Extensions to this work, Rank et al. (2024) and Mandal & Radanovic (2024) manage to solve the same problem with delayed retraining for linear MDPs. However, there exists no literature that proposes a performative RL algorithm that converges to the performative optimal policy.

Specifically, Mandal et al. (2023) frames the question of using policy gradient to find stable policies as an open problem. The authors further contemplate, as PG functions in the policy space, whether it is possible to converge towards a stable policy. Thus, in this paper, we affirmatively solve an extension (rather a harder problem) of this open problem for tabular MDPs with softmax policies.

**Policy Gradient Algorithms.** Policy gradient algorithms build a central paradigm in reinforcement learning, directly optimizing parametrised policies by estimating the gradient of expected return. The foundational policy gradient theorem (Sutton et al., 1999) established an expression for this gradient in terms of the score and action-value function, while Williams (1992) introduced the REINFORCE algorithm, providing an unbiased likelihood-ratio estimator. Convergence properties of stochastic gradient ascent in policy space were analysed in these early works. Subsequently, Konda & Tsitsiklis (2000) formalized actor–critic methods via two-timescale stochastic approximation, and Kakade (2002) proposed the natural policy gradient, leveraging the Fisher information geometry to accelerate learning. Extensions to trust region methods (Schulman et al., 2015), proximal policy optimization (Schulman et al., 2017), and entropy-regularized objectives (Mnih et al., 2016) have made policy gradient methods widely practical in high-dimensional settings. Recent theoretical advances provide finite-sample convergence guarantees and complexity analyses (Agarwal et al., 2021; Yuan et al., 2022), as well as robustness to distributional shift and adversarial perturbations (Zhang et al., 2020; Xu et al., 2020). Collectively, this body of work establishes policy gradient methods as both practically effective and theoretically grounded method for solving MDP.

## D IMPACT OF POLICY UPDATES ON PEMDPs (SECTION 3.1)

**Lemma 1** (Performative Performance Difference Lemma). *The difference in performative value functions induced by  $\pi$  and  $\pi' \in \Pi$  while starting from the initial state distribution  $\rho$  is*

$$(1) \quad V_{\pi}^{\pi}(\rho) - V_{\pi'}^{\pi'}(\rho) = \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi',\rho}} [A_{\pi'}^{\pi'}(s, a)] + \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi',\rho}} [(r_{\pi}(s, a) - r_{\pi'}(s, a)) + \gamma (\mathbf{P}_{\pi}(\cdot|s, a) - \mathbf{P}_{\pi'}(\cdot|s, a))^{\top} V_{\pi}^{\pi}(\cdot)]. \quad (17)$$

where  $A_{\pi'}^{\pi'}(s, a) \triangleq Q_{\pi'}^{\pi'}(s, a) - V_{\pi'}^{\pi'}(s)$  is the performative advantage function for any state  $s \in \mathcal{S}$  and action  $a \in \mathcal{A}$ .

$$(2) \quad V_{\pi}^{\pi}(\rho) - V_{\pi'}^{\pi'}(\rho) = \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi',\rho}} [A_{\pi'}^{\pi'}(s, a)] + \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi',\rho}} [(r_{\pi}(s, a) - r_{\pi'}(s, a)) + \gamma (\mathbf{P}_{\pi}(\cdot|s, a) - \mathbf{P}_{\pi'}(\cdot|s, a))^{\top} V_{\pi'}^{\pi'}(\cdot)]. \quad (18)$$

where  $A_{\pi'}^{\pi'}(s, a) \triangleq Q_{\pi'}^{\pi'}(s, a) - V_{\pi'}^{\pi'}(s)$  is the performative advantage function for any state  $s \in \mathcal{S}$  and action  $a \in \mathcal{A}$ .

$$(3) \quad V_{\pi}^{\pi}(\rho) - V_{\pi'}^{\pi'}(\rho) = \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi',\rho}} [A_{\pi'}^{\pi'}(s, a)] \quad (19)$$

$$+ \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi',\rho}} [(r_{\pi}(s, a) - r_{\pi'}(s, a)) + \gamma (\mathbf{P}_{\pi}(\cdot|s, a) - \mathbf{P}_{\pi'}(\cdot|s, a))^{\top} V_{\pi'}^{\pi'}(\cdot)]. \quad (20)$$

where  $A_{\pi'}^{\pi'}(s, a) \triangleq Q_{\pi'}^{\pi'}(s, a) - V_{\pi'}^{\pi'}(s)$  is the performative advantage function for any state  $s \in \mathcal{S}$  and action  $a \in \mathcal{A}$ .

We only use the first version of this lemma in the main draft, and also hereafter, for the proofs.

*Proof of Lemma 1.* We do this proof in two steps. First step involves a decomposition of the difference in value function into two terms : (i) difference in value function after deploying the same policy while agent plays two different policies i.e. the difference that explains stability of the deployed policy, and (ii) difference in value function for deploying two different policies i.e. performance difference for changing the deployed policy. While the second term can be bounded using classic performance difference lemma, in the next and final step, we control the stability inducing term (i).

**Part(1) – Step 1: Decomposition.** We start by decomposing the performative performance difference to get a stability and a performance difference terms separately.

$$\begin{aligned} V_{\pi}^{\pi}(s_0) - V_{\pi'}^{\pi'}(s_0) &= \underbrace{V_{\pi}^{\pi}(s_0) - V_{\pi'}^{\pi'}(s_0)}_{\text{performative shift term}} + \underbrace{V_{\pi'}^{\pi'}(s_0) - V_{\pi'}^{\pi'}(s_0)}_{\text{performance difference term}} \\ &= V_{\pi}^{\pi}(s_0) - V_{\pi'}^{\pi'}(s_0) + \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi',\cdot|s_0}} [A_{\pi'}^{\pi'}(s, a)] \end{aligned} \quad (21)$$

The last equality is a consequence of the classical performance difference lemma (Kakade & Langford, 2002b).

**Step 2: Controlling the performative shift term.** First, let us define  $\mathbf{P}_{\pi}^{\pi}(s', s) \triangleq \sum_{a \in \mathcal{A}} \mathbf{P}_{\pi}(s'|s, a) \pi(a|s)$ , and  $\langle \mathbf{P}_{\pi}^{\pi}(\cdot, s_0), V_{\pi}^{\pi}(\cdot) \rangle \triangleq \sum_{s \in \mathcal{S}} V_{\pi}^{\pi}(s) \mathbf{P}_{\pi}^{\pi}(s, s_0)$ .

We first observe that

$$\begin{aligned} V_{\pi}^{\pi}(s_0) - V_{\pi'}^{\pi'}(s_0) &= \mathbb{E}_{a \sim \pi(\cdot|s_0)} [r_{\pi}(s_0, a) - r_{\pi'}(s_0, a)] + \gamma \mathbb{E}_{s \sim \mathbf{P}_{\pi}^{\pi}(\cdot, s_0)} [V_{\pi}^{\pi}(s)] - \gamma \mathbb{E}_{s \sim \mathbf{P}_{\pi'}^{\pi'}(\cdot, s_0)} [V_{\pi'}^{\pi'}(s)] \\ &= \mathbb{E}_{a \sim \pi(\cdot|s_0)} [r_{\pi}(s_0, a) - r_{\pi'}(s_0, a)] \\ &\quad + \gamma \sum_s (\mathbf{P}_{\pi}^{\pi}(s, s_0) - \mathbf{P}_{\pi'}^{\pi'}(s, s_0)) V_{\pi}^{\pi}(s) + \gamma \sum_s \mathbf{P}_{\pi'}^{\pi'}(s, s_0) (V_{\pi}^{\pi}(s) - V_{\pi'}^{\pi'}(s)) \\ &= \mathbb{E}_{(s,a) \sim d_{\pi',\cdot|s_0}} [r_{\pi}(s, a) - r_{\pi'}(s, a) + \gamma (\mathbf{P}_{\pi}(\cdot|s, a) - \mathbf{P}_{\pi'}(\cdot|s, a))^{\top} V_{\pi}^{\pi}(\cdot)] \end{aligned}$$

972 The last equality is obtained by recurring the preceding step iteratively.  
 973  
 974

**Combining** steps 1 and 2 and taking expectation over  $s_0 \sim \rho$ , we get

$$975 \quad V_{\pi}(\rho) - V_{\pi'}(\rho) = \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi',\rho}} \left[ A_{\pi'}(s,a) + (r_{\pi}(s,a) - r_{\pi'}(s,a)) + \gamma (\mathbf{P}_{\pi}(\cdot|s,a) - \mathbf{P}_{\pi'}(\cdot|s,a))^{\top} V_{\pi}(\cdot) \right].$$

977 *Part(2)* – The second equality is obtained by changing the Step 2 as follows:  
 978

$$979 \quad V_{\pi}(\rho) - V_{\pi'}(\rho) = \mathbb{E}_{a \sim \pi(\cdot|s_0)} \left[ r_{\pi}(s_0, a) - r_{\pi'}(s_0, a) \right] + \gamma \mathbb{E}_{s \sim \mathbf{P}_{\pi}(\cdot,s_0)} [V_{\pi}(\rho)] - \gamma \mathbb{E}_{s \sim \mathbf{P}_{\pi'}(\cdot,s_0)} [V_{\pi'}(\rho)]$$

$$980 \quad = \mathbb{E}_{a \sim \pi(\cdot|s_0)} \left[ r_{\pi}(s_0, a) - r_{\pi'}(s_0, a) \right]$$

$$981 \quad + \gamma \sum_s (\mathbf{P}_{\pi}(s, s_0) - \mathbf{P}_{\pi'}(s, s_0)) V_{\pi'}(s) + \gamma \sum_s \mathbf{P}_{\pi}(s, s_0) (V_{\pi}(\rho) - V_{\pi'}(\rho))$$

$$982 \quad = \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi',\rho}} \left[ A_{\pi'}(s,a) \right]$$

$$983 \quad + \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi',\rho}} \left[ (r_{\pi}(s,a) - r_{\pi'}(s,a)) + \gamma (\mathbf{P}_{\pi}(\cdot|s,a) - \mathbf{P}_{\pi'}(\cdot|s,a))^{\top} V_{\pi'}(\cdot) \right].$$

984 The last equality is obtained by recurring the preceding step iteratively.  
 985  
 986

*Part(3)* – The third equality is obtained through the following steps.

$$987 \quad V_{\pi}(\rho) - V_{\pi'}(\rho) = V_{\pi}(\rho) - V_{\pi'}(\rho) + V_{\pi}(\rho) - V_{\pi'}(\rho)$$

$$988 \quad = \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi}(\cdot|s_0)} [A_{\pi'}(s,a)] + V_{\pi}(\rho) - V_{\pi'}(\rho)$$

$$989 \quad = \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi}(\cdot|s_0)} [A_{\pi'}(s,a)] + \mathbb{E}_{a \sim \pi'(\cdot|s_0)} [r_{\pi}(s_0, a) - r_{\pi'}(s_0, a)]$$

$$990 \quad + \gamma \sum_s (\mathbf{P}_{\pi'}(s, s_0) - \mathbf{P}_{\pi'}(s, s_0)) V_{\pi'}(s) + \gamma \sum_s \mathbf{P}_{\pi'}(s, s_0) (V_{\pi'}(\rho) - V_{\pi'}(s))$$

$$991 \quad = \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi,\rho}} \left[ A_{\pi'}(s,a) \right]$$

$$992 \quad + \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi',\rho}} \left[ (r_{\pi}(s,a) - r_{\pi'}(s,a)) + \gamma (\mathbf{P}_{\pi}(\cdot|s,a) - \mathbf{P}_{\pi'}(\cdot|s,a))^{\top} V_{\pi'}(\cdot) \right].$$

1004  $\square$

1005 **Lemma 2** (Bounding Performative Performance Difference for Gradually Shifting Environments). *Let us assume that both re-  
 1006 wards and transitions are Lipschitz functions of policy, i.e.  $\|r_{\pi} - r_{\pi'}\| \leq L_r \|\pi - \pi'\|$  and  $\|\mathbf{P}_{\pi} - \mathbf{P}_{\pi'}\| \leq L_{\mathbf{P}} \|\pi - \pi'\|$ ,  
 1007 for some  $L_r, L_{\mathbf{P}} \geq 0$ . Then, under Assumption 1, the performative shift in the sub-optimality gap of a policy  $\pi_{\theta}$  satisfies*

$$1008 \quad \left| V_{\pi_o^*}(\rho) - V_{\pi_{\theta}}(\rho) - \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi_{\theta},\rho}} [A_{\pi_{\theta}}(s,a)] \right| \leq \frac{2\sqrt{2}}{1-\gamma} (L_r + \frac{\gamma}{1-\gamma} L_{\mathbf{P}} R_{\max}) \mathbb{E}_{s_0 \sim \rho} D_{\text{H}}(\pi_o^*(\cdot|s_0) \| \pi_{\theta}(\cdot|s_0)). \quad (22)$$

1012 where  $D_{\text{H}}(\mathbf{x} \| \mathbf{y})$  denotes the Hellinger distance between  $\mathbf{x}$  and  $\mathbf{y}$ .

1013  
 1014 *Proof of Lemma 2.* We do this proof in three steps. We start from the final expression in Lemma 1, then in step 2 we impose  
 1015 bounds on reward and transition differences leveraging the Lipschitz assumption. Lastly, we bound the policy difference in  
 1016 first order norm using relation between Total Variation (TV) and Hellinger distance.

1017  
 1018 **Step 1:** From Lemma 1, we get

$$1019 \quad V_{\pi_o^*}(\rho) - V_{\pi_{\theta}}(\rho) = \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi_{\theta},\rho}} \left[ A_{\pi_{\theta}}(s,a) + (r_{\pi_o^*}(s,a) - r_{\pi_{\theta}}(s,a)) + \gamma (\mathbf{P}_{\pi_o^*}(\cdot|s,a) - \mathbf{P}_{\pi_{\theta}}(\cdot|s,a))^{\top} V_{\pi_o^*}(\cdot) \right].$$

1020  
 1021 Thus,

$$1022 \quad \left| V_{\pi_o^*}(\rho) - V_{\pi_{\theta}}(\rho) - \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi_{\theta},\rho}} [A_{\pi_{\theta}}(s,a)] \right|$$

$$= \frac{1}{1-\gamma} \left| \mathbb{E}_{(s,a) \sim d_{\pi_\theta, \rho}^{\pi_o^*}} (r_{\pi_o^*}(s, a) - r_{\pi_\theta}(s, a)) + \gamma (\mathbf{P}_{\pi_o^*}(\cdot | s, a) - \mathbf{P}_{\pi_\theta}(\cdot | s, a))^{\top} V_{\pi_o^*}^{\pi_o^*}(\cdot) \right| \quad (23)$$

**Step 2:** Using Jensen's inequality together with the fact that  $d_{\pi_\theta, \rho}^{\pi_o^*}(s, a | s_0) \leq 1$ , for rewards, we get

$$\left| \mathbb{E}_{(s,a) \sim d_{\pi_\theta, \rho}^{\pi_o^*}(\cdot, \cdot | s_0)} [r_{\pi_o^*}(s, a) - r_{\pi_\theta}(s, a)] \right| \leq \mathbb{E}_{(s,a) \sim d_{\pi_\theta, \rho}^{\pi_o^*}(\cdot, \cdot | s_0)} |r_{\pi_o^*}(s, a) - r_{\pi_\theta}(s, a)| \leq \|r_{\pi_o^*} - r_{\pi_\theta}\|_1$$

Similarly for transitions, we get

$$\begin{aligned} \left| \mathbb{E}_{(s,a) \sim d_{\pi_\theta, \rho}^{\pi_o^*}(\cdot, \cdot | s_0)} [(\mathbf{P}_{\pi_o^*} - \mathbf{P}_{\pi_\theta})^{\top} V_{\pi}^{\pi}] \right| &\leq \mathbb{E}_{(s,a) \sim d_{\pi_\theta, \rho}^{\pi_o^*}(\cdot, \cdot | s_0)} |(\mathbf{P}_{\pi_o^*} - \mathbf{P}_{\pi_\theta})^{\top} V_{\pi}^{\pi}| \\ &\stackrel{(a)}{\leq} \mathbb{E}_{(s,a) \sim d_{\pi_\theta, \rho}^{\pi_o^*}(\cdot, \cdot | s_0)} [\|\mathbf{P}_{\pi_o^*} - \mathbf{P}_{\pi_\theta}\|_1 \cdot \|V_{\pi_o^*}^{\pi_o^*}\|_{\infty}] \\ &= \|\mathbf{P}_{\pi_o^*} - \mathbf{P}_{\pi_\theta}\|_1 \cdot \|V_{\pi_o^*}^{\pi_o^*}\|_{\infty}, \end{aligned}$$

(a) holds due to Hölder's inequality.

Now, leveraging the triangle inequality and Lipschitzness assumption on reward and transitions, we further get

$$\left| \mathbb{E}_{(s,a) \sim d_{\pi_\theta, \rho}^{\pi_o^*}(\cdot, \cdot | s_0)} [r_{\pi_o^*}(s, a) - r_{\pi_\theta}(s, a) + \gamma (\mathbf{P}_{\pi_o^*} - \mathbf{P}_{\pi_\theta})^{\top} V_{\pi}^{\pi}] \right| \leq L_r \|\pi_o^* - \pi_\theta\|_1 + \gamma L_P \left\| \mathbf{V}_{\pi_o^*}^{\pi_o^*} \right\|_{\infty} \|\pi_o^* - \pi_\theta\|_1$$

Finally, due to Assumption 1, we get  $\left\| \mathbf{V}_{\pi_o^*}^{\pi_o^*} \right\|_{\infty} \leq \frac{R_{\max}}{1-\gamma}$ , and thus,

$$\left| \mathbb{E}_{(s,a) \sim d_{\pi_\theta, \rho}^{\pi_o^*}(\cdot, \cdot | s_0)} [r_{\pi_o^*}(s, a) - r_{\pi_\theta}(s, a) + \gamma (\mathbf{P}_{\pi_o^*} - \mathbf{P}_{\pi_\theta})^{\top} V_{\pi_o^*}^{\pi_o^*}] \right| \leq L_r \|\pi_o^* - \pi_\theta\|_1 + \frac{\gamma}{1-\gamma} L_P R_{\max} \|\pi_o^* - \pi_\theta\|_1$$

**Step 3:** We know  $\|\pi_o^* - \pi_\theta\|_1 = 2\text{TV}(\pi_o^* \parallel \pi_\theta) \leq 2\sqrt{2}D_H(\pi_o^* \parallel \pi_\theta)$ . Thus,

$$\begin{aligned} &\left| \mathbb{E}_{(s,a) \sim d_{\pi_\theta, \rho}^{\pi_o^*}(\cdot, \cdot | s_0)} [r_{\pi_o^*}(s, a) - r_{\pi_\theta}(s, a) + \gamma (\mathbf{P}_{\pi_o^*} - \mathbf{P}_{\pi_\theta})^{\top} V_{\pi_o^*}^{\pi_o^*}] \right| \\ &\leq 2\sqrt{2} \left( L_r + \frac{\gamma}{1-\gamma} L_P R_{\max} \right) D_H(\pi_o^* \parallel \pi_\theta) \end{aligned} \quad (24)$$

We conclude this proof by putting the upper bound in Equation (24) in Equation (23) and taking expectation over  $s_0 \sim \rho$  to get the desired expression.  $\square$

1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079

## 1080 E SMOOTHNESS OF PERFORMATIVE VALUE FUNCTION AND ENTROPY REGULARISER

1081  
**Lemma 4** (Performative Smoothness Lemma). Let  $\pi_\alpha \triangleq \pi_{\theta+\alpha u}$ , and let  $V_\alpha^\alpha(s_0)$  be the corresponding value at a fixed state  
1082  $s_0$ , i.e.,  $V_\alpha^\alpha(s_0) \triangleq V_{\pi_\alpha}^\alpha(s_0)$ . If the following conditions hold true,  
1083

$$\begin{aligned} \sum_{a \in \mathcal{A}} \left| \frac{d\pi_\alpha(a | s_0)}{d\alpha} \right|_{\alpha=0} &\leq C_1, \quad \sum_{a \in \mathcal{A}} \left| \frac{d^2\pi_\alpha(a | s_0)}{d\alpha^2} \right|_{\alpha=0} \leq C_2, \sum_{s \in \mathcal{S}} \left| \frac{d\mathbf{P}_\alpha(s | s_0, a_0)}{d\alpha} \right|_{\alpha=0} \leq T_1, \\ \sum_{s \in \mathcal{S}} \left| \frac{d^2\mathbf{P}_\alpha(s | s_0, a_0)}{d\alpha^2} \right|_{\alpha=0} &\leq T_2, \sum_{a \in \mathcal{A}} \left| \frac{dr_\alpha(s_0, a)}{d\alpha} \right|_{\alpha=0} \leq R_1, \quad \sum_{a \in \mathcal{A}} \left| \frac{d^2r_\alpha(s_0, a)}{d\alpha^2} \right|_{\alpha=0} \leq R_2, \end{aligned}$$

1090 we get

$$\max_{\|u\|_2=1} \left\| \frac{d^2V_\alpha^\alpha(s_0)}{d\alpha^2} \right\|_{\alpha=0} \leq \frac{C_2}{1-\gamma} + 2C_1\beta_1 + C_2\beta_2 \triangleq L,$$

1091 where  $\beta_1 = \frac{\gamma}{(1-\gamma)^2}(C_1 + T_1) + \frac{R_1}{1-\gamma}$  and  $\beta_2 = \frac{2\gamma^2}{(1-\gamma)^3}(C_1 + T_1)^2 + \frac{\gamma}{(1-\gamma)^2}(C_2 + 2C_1T_1 + T_2) + \frac{2\gamma R_1}{(1-\gamma)^2}(C_2 + 2C_1T_1 + T_2) + \frac{R_2}{1-\gamma} + \frac{\gamma C_1 R_1}{(1-\gamma)^2}$ .

1092 **Proof. Step 1:** To prove the second order smoothness of the value function we start by taking its second derivative. Consider  
1093 the expected return under policy  $\pi_\alpha$ :

$$V_\alpha^\alpha(s_0) = \sum_a \pi_\alpha(a | s_0) Q_\alpha^\alpha(s_0, a)$$

1102 Differentiating twice with respect to  $\alpha$ , we obtain:

$$\frac{d^2V_\alpha^\alpha(s_0)}{d\alpha^2} = \sum_a \frac{d^2\pi_\alpha(a | s_0)}{d\alpha^2} Q_\alpha^\alpha(s_0, a) + 2 \sum_a \frac{d\pi_\alpha(a | s_0)}{d\alpha} \frac{dQ_\alpha^\alpha(s_0, a)}{d\alpha} + \sum_a \pi_\alpha(a | s_0) \frac{d^2Q_\alpha^\alpha(s_0, a)}{d\alpha^2}$$

1107  $Q_\alpha^\alpha(s_0, a_0)$  is the Q-function corresponding to the policy  $\pi_\alpha$  at state  $s_0$  and action  $a_0$ . Observe that  $Q_\alpha^\alpha(s_0, a_0)$  can further  
1108 be written as:

$$Q_\alpha^\alpha(s_0, a_0) = e_{(s_0, a_0)}^\top (I - \gamma \tilde{\mathbf{P}}(\alpha))^{-1} r_\alpha = e_{(s_0, a_0)}^\top M(\alpha) r_\alpha$$

1111 where  $M(\alpha) \triangleq (I - \gamma \mathbf{P}(\alpha))^{-1}$  and  $\tilde{\mathbf{P}}(\alpha)$  is the state-action transition matrix under policy  $\pi_\alpha$ , defined as:

$$[\tilde{\mathbf{P}}(\alpha)](s', a' | s, a) \triangleq \pi_\alpha(a' | s') \mathbf{P}_\alpha(s' | s, a)$$

1114 Differentiating  $Q_\alpha^\alpha(s, a)$  with respect to  $\alpha$  gives:

$$\frac{dQ_\alpha^\alpha(s_0, a_0)}{d\alpha} = \gamma e_{(s_0, a_0)}^\top M(\alpha) \frac{d\tilde{\mathbf{P}}(\alpha)}{d\alpha} M(\alpha) r_\alpha + e_{(s_0, a_0)}^\top M(\alpha) \frac{dr_\alpha}{d\alpha}$$

1119 And correspondingly,

$$\begin{aligned} \frac{d^2Q_\alpha^\alpha(s_0, a_0)}{d\alpha^2} &= 2\gamma^2 e_{(s_0, a_0)}^\top M(\alpha) \frac{d\tilde{\mathbf{P}}(\alpha)}{d\alpha} M(\alpha) \frac{d\tilde{\mathbf{P}}(\alpha)}{d\alpha} M(\alpha) r_\alpha + \gamma e_{(s_0, a_0)}^\top M(\alpha) \frac{d^2\tilde{\mathbf{P}}(\alpha)}{d\alpha^2} M(\alpha) r_\alpha \\ &\quad + \gamma e_{(s_0, a_0)}^\top M(\alpha) \frac{d\tilde{\mathbf{P}}(\alpha)}{d\alpha} M(\alpha) \frac{dr_\alpha}{d\alpha} + e_{(s_0, a_0)}^\top M(\alpha) \frac{d^2r_\alpha}{d\alpha^2} \\ &\quad + \gamma e_{(s_0, a_0)}^\top M(\alpha) \frac{d\tilde{\mathbf{P}}(\alpha)}{d\alpha} M(\alpha) \frac{dr_\alpha}{d\alpha} \end{aligned} \tag{25}$$

1129 **Step 2:** Now we need to find the derivative of  $\tilde{\mathbf{P}}(\alpha)$  w.r.t  $\alpha$  in order to substitute in (25). Hence, we can differentiate  $\tilde{\mathbf{P}}(\alpha)$   
1130 with respect to  $\alpha$  to obtain:

$$\frac{d\tilde{\mathbf{P}}(\alpha)}{d\alpha} \Big|_{\alpha=0} (s', a' | s, a) = \frac{d\pi_\alpha(a' | s')}{d\alpha} \Big|_{\alpha=0} \mathbf{P}_\alpha(s' | s, a) + \frac{d\mathbf{P}_\alpha(s' | s, a)}{d\alpha} \Big|_{\alpha=0} \pi_\alpha(a' | s')$$

1134 Now, for an arbitrary vector  $\mathbf{x}$ , we have:

$$\begin{aligned} 1136 \quad & \left[ \frac{d\tilde{\mathbf{P}}(\alpha)}{d\alpha} \Big|_{\alpha=0} \mathbf{x} \right]_{(s,a)} = \sum_{s',a'} \frac{d\pi_\alpha(a' | s')}{d\alpha} \Big|_{\alpha=0} \mathbf{P}_\alpha(s' | s, a) \mathbf{x}_{s',a'} + \sum_{s',a'} \frac{d\mathbf{P}_\alpha(s' | s, a)}{d\alpha} \Big|_{\alpha=0} \pi_\alpha(a' | s') \mathbf{x}_{s',a'} \\ 1137 \\ 1138 \end{aligned}$$

1139 Taking the maximum over unit vectors  $\mathbf{u}$  in  $\ell_2$ -norm:

$$\begin{aligned} 1141 \quad & \max_{\|\mathbf{u}\|_2=1} \left\| \frac{d\tilde{\mathbf{P}}(\alpha)}{d\alpha} \Big|_{\alpha=0} \mathbf{x} \right\|_\infty \leq \max_{\|\mathbf{u}\|_2=1} \left| \sum_{s',a'} \frac{d\pi_\alpha(a' | s')}{d\alpha} \Big|_{\alpha=0} \mathbf{P}_\alpha(s' | s, a) \mathbf{x}_{s',a'} \right| \\ 1142 \\ 1143 \quad & + \max_{\|\mathbf{u}\|_2=1} \left| \sum_{s',a'} \frac{d\mathbf{P}_\alpha(s' | s, a)}{d\alpha} \Big|_{\alpha=0} \pi_\alpha(a' | s') \mathbf{x}_{s',a'} \right| \\ 1144 \\ 1145 \quad & \leq \max_{s,a} \sum_{s'} \mathbf{P}_\alpha(s' | s, a) \sum_{a'} \left| \frac{d\pi_\alpha(a' | s')}{d\alpha} \Big|_{\alpha=0} \right| \cdot \|\mathbf{x}\|_\infty \\ 1146 \\ 1147 \quad & + \max_{s,a} \sum_{a'} \pi_\alpha(a' | s') \sum_{s'} \left| \frac{d\mathbf{P}_\alpha(s' | s, a)}{d\alpha} \Big|_{\alpha=0} \right| \cdot \|\mathbf{x}\|_\infty \\ 1148 \\ 1149 \quad & \leq \max_{s,a} \sum_{s'} \mathbf{P}_\alpha(s' | s, a) \|\mathbf{x}\|_\infty C_1 + \max_{s,a} \sum_{a'} \pi_\alpha(a' | s') \|\mathbf{x}\|_\infty T_1 \\ 1150 \\ 1151 \quad & \leq C_1 \|\mathbf{x}\|_\infty + T_1 \|\mathbf{x}\|_\infty = (C_1 + T_1) \|\mathbf{x}\|_\infty \\ 1152 \\ 1153 \\ 1154 \\ 1155 \\ 1156 \\ 1157 \end{aligned}$$

By the definition of the  $\ell_\infty$ -norm, we conclude:

$$\max_{\|\mathbf{u}\|_2=1} \left\| \frac{d\mathbf{P}_\alpha}{d\alpha} \Big|_{\alpha=0} \mathbf{x} \right\|_\infty \leq (C_1 + T_1) \|\mathbf{x}\|_\infty \quad (26)$$

1161 Similarly, differentiating  $\tilde{\mathbf{P}}(\alpha)$  twice w.r.t.  $\alpha$ , we get

$$\begin{aligned} 1164 \quad & \left[ \frac{d^2\tilde{\mathbf{P}}(\alpha)}{d\alpha^2} \Big|_{\alpha=0} \right]_{(s,a) \rightarrow (s',a')} = \frac{d^2\pi_\alpha(a' | s')}{(d\alpha)^2} \Big|_{\alpha=0} \mathbf{P}_\alpha(s' | s, a) + \frac{d^2\mathbf{P}_\alpha(s' | s, a)}{d\alpha^2} \Big|_{\alpha=0} \pi_\alpha(a' | s') \\ 1165 \\ 1166 \quad & + 2 \frac{d\pi_\alpha(a' | s')}{d\alpha} \Big|_{\alpha=0} \frac{d\mathbf{P}_\alpha(s' | s, a)}{d\alpha} \Big|_{\alpha=0} \\ 1167 \\ 1168 \\ 1169 \end{aligned}$$

1170 Hence, we can consider the following norm bound:

$$\max_{\|\mathbf{u}\|_2=1} \left\| \frac{d^2\tilde{\mathbf{P}}(\alpha)}{d\alpha^2} \Big|_{\alpha=0} \mathbf{x} \right\|_\infty \leq C_2 \|\mathbf{x}\|_\infty + 2C_1 T_1 \|\mathbf{x}\|_\infty + T_2 \|\mathbf{x}\|_\infty = (C_2 + 2C_1 T_1 + T_2) \|\mathbf{x}\|_\infty \quad (27)$$

1175 **Step 3:** Now we need to put the pieces back together in order to calculate the second derivative of  $V_\alpha^\alpha$  w.r.t  $\alpha$ . Let us recall  
1176  $M(\alpha)$ . Using the power series expansion of the matrix inverse, we can write  $M(\alpha)$  as:  
1177

$$M(\alpha) = (I - \gamma \tilde{\mathbf{P}}(\alpha))^{-1} = \sum_{n=0}^{\infty} \gamma^n \tilde{\mathbf{P}}(\alpha)^n$$

1180 which implies that  $M(\alpha) \geq 0$  (component-wise), and

$$M(\alpha) \mathbf{1} = \frac{1}{1-\gamma} \mathbf{1},$$

1184 i.e., each row of  $M(\alpha)$  is positive and sums to  $\frac{1}{1-\gamma}$ .

1186 This implies:

$$\max_{\|\mathbf{u}\|_2=1} \|M(\alpha) \mathbf{x}\|_\infty \leq \frac{1}{1-\gamma} \|\mathbf{x}\|_\infty.$$

1188 This gives, using the expressions for  $\frac{d^2Q_\alpha^\alpha(s_0, a_0)}{d\alpha^2}$  and  $\frac{dQ_\alpha^\alpha(s_0, a_0)}{d\alpha}$ , an upper bound on their magnitudes based on  $\|\mathbf{x}\|_\infty$  and  
 1189 constants arising from bounds on the derivatives of  $\tilde{\mathbf{P}}(\alpha)$  and  $r_\alpha$ .  
 1190

$$\begin{aligned} & \max_{\|\mathbf{u}\|_2=1} \left\| \frac{d^2Q_\alpha^\alpha(s_0, a_0)}{d\alpha^2} \right\|_\infty \\ & \leq 2\gamma^2 \left\| M(\alpha) \frac{d\tilde{\mathbf{P}}(\alpha)}{d\alpha} M(\alpha) \frac{d\tilde{\mathbf{P}}(\alpha)}{d\alpha} M(\alpha) r_\alpha \right\|_\infty + \gamma \left\| M(\alpha) \frac{d^2\tilde{\mathbf{P}}(\alpha)}{d\alpha^2} M(\alpha) r_\alpha \right\|_\infty \\ & \quad + \gamma \left\| M(\alpha) \frac{d^2\tilde{\mathbf{P}}(\alpha)}{d\alpha^2} M(\alpha) \frac{dr_\alpha}{d\alpha} \right\|_\infty + \left\| M(\alpha) \frac{d^2r_\alpha}{d\alpha^2} \right\|_\infty + 2\gamma \left\| M(\alpha) \frac{d\tilde{\mathbf{P}}(\alpha)}{d\alpha} M(\alpha) \frac{dr_\alpha}{d\alpha} \right\|_\infty \end{aligned}$$

1201 Bounding using known bounds on transitions and rewards:  
 1202

$$\begin{aligned} & \max_{\|\mathbf{u}\|_2=1} \left\| \frac{d^2Q_\alpha^\alpha(s_0, a_0)}{d\alpha^2} \right\|_\infty \leq \frac{2\gamma^2}{(1-\gamma)^3} (C_1 + T_1)^2 + \frac{\gamma}{(1-\gamma)^2} (C_2 + 2C_1T_1 + T_2) \\ & \quad + \frac{2\gamma R_1}{(1-\gamma)^2} (C_2 + 2C_1T_1 + T_2) + \frac{R_2}{1-\gamma} + \frac{\gamma C_1 R_1}{(1-\gamma)^2} = \beta_2 \end{aligned}$$

1209 Corresponding bound on the first derivative is:  
 1210

$$\begin{aligned} & \max_{\|\mathbf{u}\|_2=1} \left\| \frac{dQ_\alpha^\alpha(s_0, a_0)}{d\alpha} \right\|_\infty \leq \gamma \left\| M(\alpha) \frac{d\tilde{\mathbf{P}}(\alpha)}{d\alpha} M(\alpha) \frac{dr_\alpha}{d\alpha} \right\|_\infty + \left\| M(\alpha) \frac{dr_\alpha}{d\alpha} \right\|_\infty \\ & \leq \frac{\gamma}{(1-\gamma)^2} (C_1 + T_1) + \frac{R_1}{1-\gamma} = \beta_1 \end{aligned}$$

1217 **Step 4:** Finally, putting all the bounds together to evaluate the upper bound of the desired quantity, we get,  
 1218

$$\max_{\|\mathbf{u}\|_2=1} \left\| \frac{d^2V_\alpha^\alpha(s_0)}{d\alpha^2} \right\|_\infty \leq \frac{C_2}{1-\gamma} + 2C_1\beta_1 + \beta_2 \quad (28)$$

1223  $\square$

1224 **Corollary 1.** For softmax PeMDPs, we characterise  
 1225

$$C_1 = 2, \quad C_2 = 6, \quad T_1 = L_{\mathbf{P}} = \max_s |\psi(s)| \triangleq \psi_{\max}, \quad T_2 = \max_s |\psi(s)|^2, \quad R_1 = L_r |\mathcal{A}| = \xi |\mathcal{A}|, \quad R_2 = 0$$

1228 *Thus,*

$$\max_{\|\mathbf{u}\|_2=1} \left\| \frac{d^2V_\alpha^\alpha(s_0)}{d\alpha^2} \Big|_{\alpha=0} \right\| \leq \mathcal{O} \left( \max \left\{ \frac{\gamma R_{\max} |\mathcal{A}|}{(1-\gamma)^2}, \frac{\gamma^2}{(1-\gamma)^3} \right\} \right) \triangleq \mathcal{O}(L). \quad (29)$$

1233 *Proof.* We use the expressions already found in (35) to state the following:  
 1234

$$\sum_{a \in \mathcal{A}} \left| \frac{d}{d\alpha} \pi_{\theta+\alpha\mathbf{u}}(a | s) \Big|_{\alpha=0} \right| \leq \sum_{a \in \mathcal{A}} \pi_\theta(a | s) |\mathbf{u}_s^\top (\mathbf{e}_a - \pi(\cdot | s))| \leq \max_{a \in \mathcal{A}} (\mathbf{u}_s^\top \mathbf{e}_a + \mathbf{u}_s^\top \pi(\cdot | s)) \leq 2.$$

1239 Similarly, differentiating once again w.r.t.  $\alpha$ , we get  
 1240

$$\sum_{a \in \mathcal{A}} \left| \frac{d^2}{d\alpha^2} \pi_{\theta+\alpha\mathbf{u}}(a | s) \Big|_{\alpha=0} \right| \leq \max_{a \in \mathcal{A}} (\mathbf{u}_s^\top \mathbf{e}_a \mathbf{e}_a^\top \mathbf{u}_s + \mathbf{u}_s^\top \mathbf{e}_a \pi(\cdot | s)^\top \mathbf{u}_s + \mathbf{u}_s^\top \pi(\cdot | s) \mathbf{e}_a^\top \mathbf{u}_s)$$

$$+ 2 \mathbf{u}_s^\top \boldsymbol{\pi}(\cdot | s) \boldsymbol{\pi}(\cdot | s)^\top \mathbf{u}_s + \mathbf{u}_s^\top \text{diag}(\boldsymbol{\pi}(\cdot | s)) \mathbf{u}_s \Big) \leq 6.$$

And hence for transition we get,

$$\sum_{s' \in \mathcal{S}} \left| \frac{d}{d\alpha} \mathbf{P}_{\boldsymbol{\pi}_{\theta+\alpha u}}(a | s) \right|_{\alpha=0} \leq \sum_{s' \in \mathcal{S}} |\psi(s')| \mathbf{P}_{\boldsymbol{\pi}_{\theta}}(s' | s, a) |\mathbf{u}_{s,a}(1 - \mathbf{P}_{\boldsymbol{\pi}_{\theta}}(\cdot | s, a))| \leq |\mathbf{u}_{s,a}| \max_s |\psi(s)| \leq \max_s |\psi(s)|$$

And similarly, it can be shown that:

$$\sum_{a \in \mathcal{A}} \left| \frac{d^2}{d\alpha^2} \mathbf{P}_{\boldsymbol{\pi}_{\theta+\alpha u}}(a | s) \right|_{\alpha=0} \leq |\mathbf{u}_{s,a}|^2 \max_s |\psi(s)|^2 \leq \max_s |\psi(s)|^2$$

Similarly for rewards we get:

$$\sum_{a \in \mathcal{A}} \left| \frac{d}{d\alpha} r_{\boldsymbol{\pi}_{\theta+\alpha u}}(a | s) \right|_{\alpha=0} \leq \xi |\mathcal{A}| , \quad \sum_{a \in \mathcal{A}} \left| \frac{d^2}{d\alpha^2} r_{\boldsymbol{\pi}_{\theta+\alpha u}}(a | s) \right|_{\alpha=0} = 0$$

Hence, we can use the following choice of constants for softmax parametrization,

$$\begin{aligned} C_1 &= 2 , \quad C_2 = 6 \\ T_1 &= L_{\mathbf{P}} = \max_s |\psi(s)| , \quad T_2 = \max_s |\psi(s)|^2 \\ R_1 &= L_r |\mathcal{A}| = \xi |\mathcal{A}| , \quad R_2 = 0 \end{aligned}$$

to get the desired order of  $\max_{\|u\|_2=1} \left\| \frac{d^2 V_{\alpha}^{\theta}(s_0)}{d\alpha^2} \right\|_{\alpha=0}$ .

□

**Lemma 5** (Smoothness of Entropy Regularizer). Define the discounted entropy regularizer as:

$$\mathcal{H}_{\boldsymbol{\pi}_{\theta_{\alpha}}}^{\theta_{\alpha}}(s) = \mathbb{E}_{\tau \sim \mathbf{P}_{\boldsymbol{\pi}}} \left[ \sum_{t=0}^{\infty} -\gamma^t \log \boldsymbol{\pi}_{\theta_{\alpha}}(a_t | s_t) \right]$$

Under the same assumptions as 4, the following holds:

$$\max_{\|u\|_2=1} \left\| \frac{\partial^2 \mathcal{H}_{\boldsymbol{\pi}_{\theta_{\alpha}}}^{\theta_{\alpha}}(s)}{\partial \alpha^2} \right\|_{\alpha=0} \leq \beta_{\lambda}$$

where

$$\beta_{\lambda} = 2\gamma^2 \frac{3(1 + \log |\mathcal{A}|)}{1 - \gamma} + \gamma \frac{2 \log |\mathcal{A}|}{(1 - \gamma)^2} (C_1 + T_1) + 2\gamma \frac{\log |\mathcal{A}|}{(1 - \gamma)^2} (C_2 + 2C_1 T_1 + T_2) + \frac{\log |\mathcal{A}|}{(1 - \gamma)^3} (C_1 + T_1)^2 .$$

*Proof.* **Step 1:** Define the state-wise entropy term:

$$h_{\boldsymbol{\theta}_{\alpha}}(s) = - \sum_a \boldsymbol{\pi}_{\boldsymbol{\theta}_{\alpha}}(a | s) \log \boldsymbol{\pi}_{\boldsymbol{\theta}_{\alpha}}(a | s).$$

From [Mei et al. \(2020\)](#) (Lemma 7) we report that,

$$\left\| \frac{\partial h_{\boldsymbol{\theta}_{\alpha}}}{\partial \alpha} \right\|_{\infty} \leq 2 \cdot \log |\mathcal{A}| \cdot \|u\|_2, \quad \left\| \frac{\partial^2 h_{\boldsymbol{\theta}_{\alpha}}}{\partial \alpha^2} \right\|_{\infty} \leq 3 \cdot (1 + \log |\mathcal{A}|) \cdot \|u\|_2^2. \quad (30)$$

Additionally, [Mei et al. \(2020\)](#) also presents a second result expressing the second derivative of the entropy w.r.t  $\alpha$ ,

$$\frac{\partial^2 \mathcal{H}_{\boldsymbol{\pi}_{\theta_{\alpha}}}^{\theta_{\alpha}}(s)}{\partial \alpha^2} = 2\gamma^2 \mathbf{e}_s^\top M(\alpha) \frac{\partial \mathbf{P}(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial \mathbf{P}(\alpha)}{\partial \alpha} M(\alpha) h_{\boldsymbol{\theta}_{\alpha}}$$

$$+ \gamma \mathbf{e}_s^\top M(\alpha) \frac{\partial^2 \mathbf{P}(\alpha)}{\partial \alpha^2} M(\alpha) h_{\theta_\alpha} + 2\gamma \mathbf{e}_s^\top M(\alpha) \frac{\partial \mathbf{P}(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial h_{\theta_\alpha}}{\partial \alpha} + \mathbf{e}_s^\top M(\alpha) \frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2}.$$

**Step 2:** Now we proceed with bounding the absolute value of each term which will contribute towards bounding the overall second derivative of the regulariser.

For the last term,

$$\begin{aligned} \left| \mathbf{e}_s^\top M(\alpha) \frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2} \Big|_{\alpha=0} \right| &\leq \|\mathbf{e}_s^\top\|_1 \cdot \left\| M(\alpha) \frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2} \Big|_{\alpha=0} \right\|_\infty \\ &\leq \frac{1}{1-\gamma} \cdot \left\| \frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2} \Big|_{\alpha=0} \right\|_\infty \\ &\leq \frac{3 \cdot (1 + \log |\mathcal{A}|)}{1-\gamma} \cdot \|\mathbf{u}\|_2^2. \end{aligned}$$

For the second last term,

$$\begin{aligned} \left| \mathbf{e}_s^\top M(\alpha) \frac{\partial \mathbf{P}(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial h_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right| &\leq \left\| M(\alpha) \frac{\partial \mathbf{P}(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial h_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right\|_\infty \\ &\leq \frac{1}{1-\gamma} \cdot \left\| \frac{\partial \mathbf{P}(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial h_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right\|_\infty \\ &\leq \frac{(C_1 + T_1) \cdot \|\mathbf{u}\|_2}{1-\gamma} \cdot \left\| M(\alpha) \frac{\partial h_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right\|_\infty \\ &\leq \frac{(C_1 + T_1) \cdot \|\mathbf{u}\|_2}{(1-\gamma)^2} \cdot \left\| \frac{\partial h_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right\|_\infty \\ &\leq \frac{2 \cdot \log |\mathcal{A}|}{(1-\gamma)^2} (C_1 + T_1) \cdot \|\mathbf{u}\|_2^2. \end{aligned}$$

For the second term,

$$\begin{aligned} \left| \mathbf{e}_s^\top M(\alpha) \frac{\partial^2 \mathbf{P}(\alpha)}{\partial \alpha^2} M(\alpha) h_{\theta_\alpha} \Big|_{\alpha=0} \right| &\leq \left\| M(\alpha) \frac{\partial^2 \mathbf{P}(\alpha)}{\partial \alpha^2} M(\alpha) h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \\ &\leq \frac{1}{1-\gamma} \cdot \left\| \frac{\partial^2 \mathbf{P}(\alpha)}{\partial \alpha^2} M(\alpha) h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \\ &\leq \frac{\|\mathbf{u}\|_2^2}{1-\gamma} \cdot \left\| M(\alpha) h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty (C_2 + 2C_1T_1 + T_2) \\ &\leq \frac{\|\mathbf{u}\|_2^2}{(1-\gamma)^2} \cdot \left\| h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty (C_2 + 2C_1T_1 + T_2) \\ &\leq \frac{\log |\mathcal{A}|}{(1-\gamma)^2} (C_2 + 2C_1T_1 + T_2) \cdot \|\mathbf{u}\|_2^2. \end{aligned}$$

For the first term,

$$\begin{aligned} \left| \mathbf{e}_s^\top M(\alpha) \frac{\partial \mathbf{P}(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial \mathbf{P}(\alpha)}{\partial \alpha} M(\alpha) h_{\theta_\alpha} \Big|_{\alpha=0} \right| &\leq \left\| M(\alpha) \frac{\partial \mathbf{P}(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial \mathbf{P}(\alpha)}{\partial \alpha} M(\alpha) h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \\ &\leq \frac{1}{1-\gamma} \cdot \|\mathbf{u}\|_2 \cdot \frac{1}{1-\gamma} \cdot \|\mathbf{u}\|_2 \cdot \frac{1}{1-\gamma} \cdot \log |\mathcal{A}| \cdot (C_1 + T_1)^2 \\ &= \frac{\log |\mathcal{A}|}{(1-\gamma)^3} (C_1 + T_1)^2 \cdot \|\mathbf{u}\|_2^2. \end{aligned}$$

1350    **Step 3:** Now combining all the above equations, we get the final expression,  
 1351

$$1353 \quad \max_{\|\mathbf{u}\|_2=1} \left\| \frac{\partial^2 \mathcal{H}_{\pi_{\theta_\alpha}}^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2} \Big|_{\alpha=0} \right\|_\infty \leq \beta_\lambda$$

1356    where  
 1357

$$1358 \quad \beta_\lambda = 2\gamma^2 \cdot \frac{3 \cdot (1 + \log |\mathcal{A}|)}{1 - \gamma} + \gamma \cdot \frac{2 \cdot \log |\mathcal{A}|}{(1 - \gamma)^2} (C_1 + T_1) \\ 1359 \quad + 2\gamma \cdot \frac{\log |\mathcal{A}|}{(1 - \gamma)^2} (C_2 + 2C_1T_1 + T_2) + \frac{\log |\mathcal{A}|}{(1 - \gamma)^3} (C_1 + T_1)^2$$

□

1363    By definition of smoothness, the “soft performative value function”  $\tilde{V}_\pi^\pi$  is Lipschitz smooth with Lipschitz constant  $L_\lambda$   
 1364    where  $L_\lambda \triangleq L + \beta_\lambda$ . Once again, we can choose  $C_1, C_2, T_1, T_2$  according to Corollary 1 for simplification to get the order  
 1365     $\beta_\lambda = \mathcal{O}\left(\frac{\log |\mathcal{A}|}{(1-\gamma)^3} \psi_{\max}^2\right)$ . Thus, the final bound for  $L_\lambda$  as  
 1366

$$1367 \quad L_\lambda = \mathcal{O}(\max\{L, \lambda\beta_\lambda\}) = \mathcal{O}\left(\max\left\{\frac{\gamma R_{\max} |\mathcal{A}|}{(1-\gamma)^2}, \frac{\lambda \log |\mathcal{A}| \psi_{\max}^2}{(1-\gamma)^3}\right\}\right). \quad (31)$$

1368  
 1369  
 1370  
 1371  
 1372  
 1373  
 1374  
 1375  
 1376  
 1377  
 1378  
 1379  
 1380  
 1381  
 1382  
 1383  
 1384  
 1385  
 1386  
 1387  
 1388  
 1389  
 1390  
 1391  
 1392  
 1393  
 1394  
 1395  
 1396  
 1397  
 1398  
 1399  
 1400  
 1401  
 1402  
 1403

## F DERIVATION OF PERFORMATIVE POLICY GRADIENTS

**Theorem 2** (Performative Policy Gradient Theorem). *The gradient of the performative value function w.r.t  $\theta$  is as follows:*

(a) *For the unregularised objective,*

$$\nabla_{\theta} V_{\pi_{\theta}}^{\pi_{\theta}}(\tau) = \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}} \left[ \sum_{t=0}^{\infty} \gamma^t (A_{\pi_{\theta}}^{\pi_{\theta}}(s_t, a_t) (\nabla_{\theta} \log \pi_{\theta}(a_t | s_t) + \nabla_{\theta} \log P_{\pi_{\theta}}(s_{t+1} | s_t, a_t)) + \nabla_{\theta} r_{\pi_{\theta}}(s_t, a_t)) \right]. \quad (32)$$

(b) *For the entropy-regularised objective, we define the soft advantage, soft Q, and soft value functions with respect to the soft rewards  $\tilde{r}_{\pi_{\theta}}$  satisfying  $\tilde{A}_{\pi_{\theta}}^{\pi_{\theta}}(s, a) = \tilde{Q}_{\pi_{\theta}}^{\pi_{\theta}}(s, a) - \tilde{V}_{\pi_{\theta}}^{\pi_{\theta}}(s)$  that further yields*

$$\nabla_{\theta} \tilde{V}_{\pi_{\theta}}^{\pi_{\theta}}(\tau) = \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}} \left[ \sum_{t=0}^{\infty} \gamma^t (\tilde{A}_{\pi_{\theta}}^{\pi_{\theta}}(s_t, a_t) (\nabla_{\theta} \log \pi_{\theta}(a_t | s_t) + \nabla_{\theta} \log P_{\pi_{\theta}}(s_{t+1} | s_t, a_t)) + \nabla_{\theta} \tilde{r}_{\pi_{\theta}}(s_t, a_t | \theta)) \right]. \quad (33)$$

*Proof of Theorem 2.* We prove each part of this theorem separately.

*Proof of part (a).* First, we derive explicit closed form gradient for unregularised performative value function.

**Step 1.** Given a trajectory  $\tau = \{s_0, a_0, \dots, s_t, a_t, \dots\}$ , let us denote the unregularised objective function as

$$f_{\theta}(\tau) = \sum_{t=0}^{\infty} \gamma^t r_{\pi_{\theta}}(s_t, a_t)$$

Thus,

$$\begin{aligned} \nabla_{\theta} V_{\pi_{\theta}}^{\pi_{\theta}}(\tau) &= \nabla_{\theta} \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}} [f_{\theta}(\tau)] = \nabla_{\theta} \sum_{\tau} \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}(\tau) f_{\theta}(\tau) \\ &= \sum_{\tau} \nabla_{\theta} (\mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}(\tau) f_{\theta}(\tau)) \\ &= \sum_{\tau} (\nabla_{\theta} \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}(\tau)) f_{\theta}(\tau) + \sum_{\tau} \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}(\tau) (\nabla_{\theta} f_{\theta}(\tau)) \\ &\stackrel{(a)}{=} \sum_{\tau} \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}(\tau) (\nabla_{\theta} \log \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}(\tau)) f_{\theta}(\tau) + \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}} [\nabla_{\theta} f_{\theta}(\tau)] \\ &= \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}} [(\nabla_{\theta} \log \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}(\tau)) f_{\theta}(\tau)] + \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}} [\nabla_{\theta} f_{\theta}(\tau)]. \end{aligned}$$

(a) holds since  $\nabla_{\theta} \log \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}(\tau) = \frac{\nabla_{\theta} \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}(\tau)}{\mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}(\tau)}$ .

**Step 2.** Given the initial state distribution  $\rho$ , we further have

$$\log \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}(\tau) = \log \rho(s_0) + \sum_{t=0}^{\infty} \log \pi_{\theta}(a_t | s_t) + \sum_{t=0}^{\infty} \log P_{\pi_{\theta}}(s_{t+1} | s_t, a_t)$$

Taking the gradient with respect to  $\theta$ , we obtain

$$\nabla_{\theta} \log \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}(\tau) = \sum_{t=0}^{\infty} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) + \sum_{t=0}^{\infty} \nabla_{\theta} \log P_{\pi_{\theta}}(s_{t+1} | s_t, a_t)$$

**Step 3.** Now, by substituting the value of  $\nabla_{\theta} \log(P_{\pi_{\theta}}^{\pi_{\theta}})$  in  $\nabla_{\theta} V_{\pi_{\theta}}^{\pi_{\theta}}(\tau)$ , we get,

$$\begin{aligned} \nabla_{\theta} V_{\pi_{\theta}}^{\pi_{\theta}}(\tau) &= \nabla_{\theta} \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}} [f_{\theta}(\tau)] = \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}} \left[ \left( \sum_{t=0}^{\infty} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right) \cdot \left( \sum_{t=0}^{\infty} \gamma^t r_{\pi_{\theta}}(s_t, a_t) \right) \right] \\ &\quad + \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}} \left[ \left( \sum_{t=1}^{\infty} \nabla_{\theta} \log P_{\pi_{\theta}}(s_t | s_{t-1}, a_{t-1}) \right) \cdot \left( \sum_{t=0}^{\infty} \gamma^t r_{\pi_{\theta}}(s_t, a_t) \right) \right] \end{aligned}$$

$$\begin{aligned}
& + \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}} \left[ \sum_{t=0}^{\infty} \gamma^t \nabla_{\theta} r_{\pi_{\theta}}(s_t, a_t) \right] \\
& = \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}} \left[ \sum_{t=0}^{\infty} \gamma^t A_{\pi_{\theta}}^{\pi_{\theta}}(s_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right] \\
& + \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}} \left[ \sum_{t=1}^{\infty} \gamma^t A_{\pi_{\theta}}^{\pi_{\theta}}(s_t, a_t) \nabla_{\theta} \log \mathbf{P}_{\pi_{\theta}}(s_t | s_{t-1}, a_{t-1}) \right] \\
& + \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}} \left[ \sum_{t=0}^{\infty} \gamma^t \nabla_{\theta} r_{\pi_{\theta}}(s_t, a_t) \right].
\end{aligned}$$

The last equality is due to the definition of advantage function

$$A_{\pi_{\theta}}^{\pi_{\theta}}(s_t, a_t) \triangleq \sum_{i=t}^{\infty} \gamma^{t-i} r_{\pi_{\theta}}(s_i, \pi_{\theta}(s_i)) - \mathbb{E}_{s_{t'+1} \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}(\cdot | s_{t'}, a_{t'})} \left[ \sum_{i=t}^{\infty} \gamma^{t-i} r_{\pi_{\theta}}(s_i, \pi_{\theta}(s_i)) \right] \triangleq Q_{\pi_{\theta}}^{\pi_{\theta}}(s_t) - V_{\pi_{\theta}}^{\pi_{\theta}}(s_t)$$

as in classical policy gradient theorem. Hence, we conclude the proof for part (a) of the theorem.

*Proof of part (b).* Now, we derive explicit gradient form for entropy-regularised value function.

Let us define the soft reward as  $\tilde{r}_{\pi_{\theta}}(s_t, a_t) \triangleq r_{\pi_{\theta}}(s_t, a_t) - \lambda \log \pi_{\theta}(a_t | s_t)$ . Again, we start by defining regularised objective function

$$\tilde{f}_{\theta}(\tau) = \sum_{t=0}^{\infty} \gamma^t \tilde{r}_{\pi_{\theta}}(s_t, a_t)$$

Following the same steps as that of *Part (a)*, we get

$$\begin{aligned}
\nabla_{\theta} \tilde{V}_{\pi_{\theta}}^{\pi_{\theta}}(\tau) &= \nabla_{\theta} \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}} [\tilde{f}_{\theta}(\tau)] = \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}} \left[ \sum_{t=0}^{\infty} \gamma^t \tilde{A}_{\pi_{\theta}}^{\pi_{\theta}}(s_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right] \\
&+ \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}} \left[ \sum_{t=1}^{\infty} \gamma^t \tilde{A}_{\pi_{\theta}}^{\pi_{\theta}}(s_t, a_t) \nabla_{\theta} \log \mathbf{P}_{\pi_{\theta}}(s_t | s_{t-1}, a_{t-1}) \right] \\
&+ \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}} \left[ \sum_{t=0}^{\infty} \gamma^t \nabla_{\theta} \tilde{r}_{\pi_{\theta}}(s_t, a_t) \right]. \\
&= \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}} \left[ \sum_{t=0}^{\infty} \gamma^t \tilde{A}_{\pi_{\theta}}^{\pi_{\theta}}(s_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right] \\
&+ \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}} \left[ \sum_{t=1}^{\infty} \gamma^t \tilde{A}_{\pi_{\theta}}^{\pi_{\theta}}(s_t, a_t) \nabla_{\theta} \log \mathbf{P}_{\pi_{\theta}}(s_t | s_{t-1}, a_{t-1}) \right] \\
&+ \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}} \left[ \sum_{t=0}^{\infty} \gamma^t \nabla_{\theta} r_{\pi_{\theta}}(s_t, a_t) \right] - \lambda \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}} \left[ \sum_{t=0}^{\infty} \gamma^t \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right]
\end{aligned}$$

Here,

$$\tilde{A}_{\pi_{\theta}}^{\pi_{\theta}}(s_t, a_t) \triangleq \sum_{i=t}^{\infty} \gamma^{t-i} \tilde{r}_{\pi_{\theta}}(s_i, \pi_{\theta}(s_i)) - \mathbb{E}_{s_{t'+1} \sim \mathbb{P}_{\pi_{\theta}}^{\pi_{\theta}}(\cdot | s_{t'}, a_{t'})} \left[ \sum_{i=t}^{\infty} \gamma^{t-i} \tilde{r}_{\pi_{\theta}}(s_i, \pi_{\theta}(s_i)) \right] \triangleq \tilde{Q}_{\pi_{\theta}}^{\pi_{\theta}}(s_t, a_t) - \tilde{V}_{\pi_{\theta}}^{\pi_{\theta}}(s_t)$$

denotes the advantage function with soft rewards, or in brief, the soft advantage function. Hence, we conclude proof of part (b).  $\square$

## 1512 G CONVERGENCE OF PePG : PROOFS OF SECTION 4

### 1513 G.1 PROOFS FOR UNREGULARISED VALUE FUNCTION

1514 **Lemma 6** (Performative Policy Gradient for Softmax PeMDPs). *Given softmax PeMDPs defined by (13), for all  $(s, a, s') \in$*   
 1515  *$(\mathcal{S}, \mathcal{A}, \mathcal{S})$ , derivative of the performative value function w.r.t  $\theta_{s,a}$  satisfies:*

$$1519 \frac{\partial V_{\pi_\theta}(\rho)}{\partial \theta_{s,a}} \geq \frac{1}{1-\gamma} d_{\pi_\theta}(s, a | \rho) (A_{\pi_\theta}(s, a) + \xi) . \quad (34)$$

1522 *Proof.* First, we note that

$$\begin{aligned} 1523 \frac{\partial}{\partial \theta_{s',a'}} \log \pi_\theta(a|s) &= \mathbb{1}[s = s', a = a'] - \pi_\theta(a'|s) \mathbb{1}[s = s'] \\ 1524 \frac{\partial}{\partial \theta_{s',a'}} \log \mathbf{P}_{\pi_\theta}(s''|s, a) &= \psi(s'') \mathbb{1}[s = s', a = a'] (1 - \mathbf{P}_{\pi_\theta}(s''|s, a)) \\ 1525 \frac{\partial}{\partial \theta_{s',a'}} r_{\pi_\theta}(s, a) &= \xi \mathbb{1}[s = s', a = a'] . \end{aligned} \quad (35)$$

1531 In this proof, we further substitute the expressions of individual gradients in Equation (35) into Equation (10).

1532 Therefore, for a given initial state distribution  $\rho$ , we get

$$\begin{aligned} 1533 \frac{\partial}{\partial \theta_{s,a}} V_{\pi_\theta}(\rho) &= \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_\theta}^{\pi_\theta}} \left[ \sum_{t=0}^{\infty} \gamma^t \left( A_{\pi_\theta}(s_t, a_t) \frac{\partial}{\partial \theta_{s,a}} \log \pi_\theta(a_t | s_t) \right. \right. \\ 1534 &\quad + A_{\pi_\theta}(s_t, a_t) \frac{\partial}{\partial \theta_{s,a}} \log P_{\pi_\theta}(s_{t+1}|s_t, a_t) \\ 1535 &\quad \left. \left. + \frac{\partial}{\partial \theta_{s,a}} r_{\pi_\theta}(s_t, a_t) \right) \right] \\ 1536 &= \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_\theta}^{\pi_\theta}} \left[ \sum_{t=0}^{\infty} \gamma^t \left( A_{\pi_\theta}(s_t, a_t) (\mathbb{1}[s_t = s, a_t = a] - \pi_\theta(a|s) \mathbb{1}[s_t = s]) \right. \right. \\ 1537 &\quad + A_{\pi_\theta}(s_t, a_t) \psi(s_{t+1}) \mathbb{1}[s_t = s, a_t = a] (1 - \mathbf{P}_{\pi_\theta}(s_{t+1}|s, a)) \\ 1538 &\quad \left. \left. + \xi \mathbb{1}[s_t = s, a_t = a] \right) \right] \\ 1539 &\stackrel{(a)}{\geq} \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_\theta}^{\pi_\theta}} \left[ \sum_{t=0}^{\infty} \gamma^t A_{\pi_\theta}(s_t, a_t) \mathbb{1}[s_t = s, a_t = a] \right] - \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_\theta}^{\pi_\theta}} \left[ \sum_{t=0}^{\infty} \gamma^t \pi_\theta(a|s) \mathbb{1}[s_t = s] A_{\pi_\theta}(s_t, a_t) \right] \\ 1540 &\quad + \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_\theta}^{\pi_\theta}} \left[ \sum_{t=0}^{\infty} \gamma^t \xi \mathbb{1}[s_t = s, a_t = a] \right] \\ 1541 &\stackrel{(b)}{=} \frac{1}{1-\gamma} d_{\pi_\theta, \rho}(s, a) A_{\pi_\theta}(s, a) + \frac{1}{1-\gamma} \xi d_{\pi_\theta, \rho}(s, a) \end{aligned}$$

1542 (a) is due to the fact that  $1 - \mathbf{P}_{\pi_\theta}(s, a) \geq 0$  for all  $s, a$ . (b) is due to  $\mathbb{E}_{\tau \sim \mathbb{P}_{\pi_\theta}^{\pi_\theta}} \left[ \sum_{t=0}^{\infty} \gamma^t \pi_\theta(a|s) \mathbb{1}[s_t = s] A_{\pi_\theta}(s_t, a_t) \right] = 0$ .

1543  $\square$

1544 **Lemma 3.** *Performative Gradient Domination for Softmax PeMDPs* Let us consider PeMDPs defined in (13).

1545 (a) For unregularised value function,

$$1546 V_{\pi_\theta^*}(\rho) - V_{\pi_\theta}(\rho) \leq \sqrt{|\mathcal{S}| |\mathcal{A}|} \left\| \frac{d_{\pi_\theta^*, \rho}^*}{d_{\pi_\theta, \nu}} \right\|_\infty \|\nabla_\theta V_{\pi_\theta}(\nu)\|_2 + \frac{R_{\max}}{1-\gamma} \left( 1 + \frac{2\gamma}{1-\gamma} \psi_{\max} \right) . \quad (36)$$

1547 *Proof of Lemma 3–Part (a).* This proof is divided into two parts. In the first part we bound the expected advantage term  
 1548 from Lemma 2 with the norm of the gradient of value function. During this step, we need to express the expected advantage

1566 as a linear combination of the advantage itself and the occupancy measure over all states and actions like in equation (34).  
1567 The expectation however is taken w.r.t the occupancy measure  $d_{\pi_\theta}^{\pi_o^*}$ , thus we need to perform a change of measure which  
1568 introduces a coverage term as shown below. In the second step we directly use the bound of rewards and transitions obtained  
1569 from their Lipchitzness in lemma 2. We know by Lemma 1 that  
1570

$$\begin{aligned} V_{\pi_o^*}^*(\rho) - V_{\pi_\theta}^*(\rho) &= \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi_\theta, \rho}^{\pi_o^*}(\cdot | \rho)} [A_{\pi_\theta}^*(s, a)] \\ &\quad + \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi_\theta, \rho}^{\pi_o^*}} [(r_{\pi_o^*}(s, a) - r_{\pi_\theta}(s, a)) + \gamma (\mathbf{P}_{\pi_o^*}(\cdot | s, a) - \mathbf{P}_{\pi_\theta}(\cdot | s, a))^T V_{\pi_o^*}^*(\cdot)]. \end{aligned}$$

### Step 1: Upper bounding Term 1.

$$\begin{aligned} \mathbb{E}_{(s,a) \sim d_{\pi_\theta, \rho}^{\pi_o^*}} [A_{\pi_\theta}^*(s, a)] &= \sum_{s,a} d_{\pi_\theta}^{\pi_o^*}(s, a | \rho) A_{\pi_\theta}^*(s, a) = \sum_{s,a} \frac{d_{\pi_\theta}^{\pi_o^*}(s, a | \rho)}{d_{\pi_\theta}^{\pi_o^*}(s, a | \nu)} d_{\pi_\theta}^{\pi_\theta}(s, a | \nu) A_{\pi_\theta}^*(s, a) \\ &\leq \left\| \frac{d_{\pi_\theta, \rho}^{\pi_o^*}}{d_{\pi_\theta, \nu}^{\pi_o^*}} \right\|_\infty \sum_{s,a} d_{\pi_\theta}^{\pi_\theta}(s, a | \nu) A_{\pi_\theta}^*(s, a) \end{aligned} \quad (37)$$

Now, we leverage the gradient of softmax performatve MDPs to obtain

$$\begin{aligned} \sum_{s,a} d_{\pi_\theta}^{\pi_\theta}(s, a | \nu) A_{\pi_\theta}^*(s, a) &\leq (1-\gamma) \sum_{s,a} \frac{\partial V_{\pi_\theta}^*(\nu)}{\partial \theta_{s,a}} - \xi \\ &= (1-\gamma) \mathbf{1}^\top \nabla_\theta V_{\pi_\theta}^*(\nu) - \xi \\ &\leq (1-\gamma) \sqrt{|\mathcal{S}| |\mathcal{A}|} \|\nabla_\theta V_{\pi_\theta}^*(\nu)\|_2 - \xi \end{aligned}$$

The last inequality is obtained by applying Cauchy-Schwarz inequality.

Now, substituting the above result back in Equation (37), we get

$$\frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi_\theta, \rho}^{\pi_o^*}(\cdot | s_0)} [A_{\pi_\theta}^*(s, a)] \leq \sqrt{|\mathcal{S}| |\mathcal{A}|} \left\| \frac{d_{\pi_\theta, \rho}^{\pi_o^*}}{d_{\pi_\theta, \nu}^{\pi_o^*}} \right\|_\infty \|\nabla_\theta V_{\pi_\theta}^*(\nu)\|_2 - \left\| \frac{d_{\pi_\theta, \rho}^{\pi_o^*}}{d_{\pi_\theta, \nu}^{\pi_o^*}} \right\|_\infty \frac{\xi}{1-\gamma} \quad (38)$$

**Step 2: Upper bounding Term 2.** For softmax rewards and transitions, we further obtain from Lemma 2,

$$\begin{aligned} \text{Term 2} &= \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi_\theta, \rho}^{\pi_o^*}} [(r_{\pi_o^*}(s, a) - r_{\pi_\theta}(s, a)) + \gamma (\mathbf{P}_{\pi_o^*}(\cdot | s, a) - \mathbf{P}_{\pi_\theta}(\cdot | s, a))^T V_{\pi_o^*}^*(\cdot)] \\ &\leq \frac{1}{1-\gamma} (\xi + \frac{\gamma}{1-\gamma} R_{\max} \psi_{\max}) \|\pi_o^*(\cdot | s_0) - \pi_\theta(\cdot | s_0)\|_1 \end{aligned} \quad (39)$$

$$\leq \frac{2}{1-\gamma} (\xi + \frac{\gamma}{1-\gamma} R_{\max} \psi_{\max}). \quad (40)$$

**Step 3:** Now, if we use Equation (38) and (40) together, we get

$$\begin{aligned} V_{\pi_o^*}^*(\rho) - V_{\pi_\theta}^*(\rho) &\leq \sqrt{|\mathcal{S}| |\mathcal{A}|} \left\| \frac{d_{\pi_\theta, \rho}^{\pi_o^*}}{d_{\pi_\theta, \nu}^{\pi_o^*}} \right\|_\infty \|\nabla_\theta V_{\pi_\theta}^*(\nu)\|_2 + \left( 2 - \left\| \frac{d_{\pi_\theta, \rho}^{\pi_o^*}}{d_{\pi_\theta, \nu}^{\pi_o^*}} \right\|_\infty \right) \frac{\xi}{1-\gamma} + \frac{2\gamma}{(1-\gamma)^2} R_{\max} \psi_{\max} \\ &\leq \sqrt{|\mathcal{S}| |\mathcal{A}|} \left\| \frac{d_{\pi_\theta, \rho}^{\pi_o^*}}{d_{\pi_\theta, \nu}^{\pi_o^*}} \right\|_\infty \|\nabla_\theta V_{\pi_\theta}^*(\nu)\|_2 + \frac{R_{\max}}{1-\gamma} + \frac{2\gamma}{(1-\gamma)^2} R_{\max} \psi_{\max} \\ &= \sqrt{|\mathcal{S}| |\mathcal{A}|} \left\| \frac{d_{\pi_\theta, \rho}^{\pi_o^*}}{d_{\pi_\theta, \nu}^{\pi_o^*}} \right\|_\infty \|\nabla_\theta V_{\pi_\theta}^*(\nu)\|_2 + \frac{R_{\max}}{1-\gamma} \left( 1 + \frac{2\gamma}{1-\gamma} \psi_{\max} \right) \end{aligned}$$

The last inequality is true since  $\left\| \frac{d_{\pi_\theta, \rho}^{\pi_o^*}}{d_{\pi_\theta, \nu}^{\pi_o^*}} \right\|_\infty \geq 1$  (Lemma 9) and  $\xi \leq R_{\max}$ .

1620  
1621  
1622  
1623  
1624  
1625

**Theorem 3** (Convergence of PePG in softmax PeMDPs – Part (a)). Let  $\text{Cov} \triangleq \max_{\theta, \nu} \left\| \frac{d_{\pi_\theta^*, \rho}}{d_{\pi_\theta, \nu}} \right\|_\infty$ . The gradient ascent algorithm on  $V_{\pi_\theta^*}(\rho)$  (Equation (9)) with step size  $\eta = \Omega(\min\{\frac{(1-\gamma)^2}{\gamma|\mathcal{A}|}, \frac{(1-\gamma)^3}{\gamma^2}\})$  satisfies, for all distributions  $\rho \in \Delta(\mathcal{S})$ .

(a) For unregularised case,

$$\min_{t < T} \left\{ V_{\pi_\theta^*}(\rho) - V_{\pi_{\theta_t}}(\rho) \right\} \leq \epsilon \text{ when } T = \Omega \left( \frac{|\mathcal{S}||\mathcal{A}|\text{Cov}^2}{\epsilon^2} \max \left\{ \frac{\gamma R_{\max} |\mathcal{A}|}{(1-\gamma)^3}, \frac{\gamma^2}{(1-\gamma)^4} \right\} \right), \text{ and } \epsilon = \Omega \left( \frac{1}{1-\gamma} \right).$$

1629

*Proof of Theorem 3– Part (a).* We proceed with this proof by dividing it in four steps. In the first step, we use the smoothness of the value function to prove an upper bound for the minimum squared gradient norm of the value over time which is a constant times  $1/T$ . In the second step, we derive a lower bound on the norm of gradient of value function using Lemma 3. In the final two steps, we combine the bounds obtained from the first two steps to derive lower bounds for  $T$  and  $\epsilon$ , i.e. the error threshold.

1634

**Step 1:** As  $V_{\pi_\theta}$  is  $L$ -smooth (Lemma 4), it satisfies

$$\left| V_{\pi_\theta}(\rho) - V_{\pi'_\theta}(\rho) - \langle \nabla_{\theta} V_{\pi_\theta}(\rho), \theta - \theta' \rangle \right| \leq \frac{L}{2} \|\theta - \theta'\|^2$$

Thus, taking  $\theta$  as  $\theta_{t+1}$  and  $\theta'$  as  $\theta_t$  and using the gradient ascent expression (Equation (9)) yields

$$\begin{aligned} \left| V_{\pi_{\theta^{(t+1)}}}(\rho) - V_{\pi_{\theta^{(t)}}}(\rho) - \eta \|\nabla_{\theta} V_{\pi_{\theta^{(t)}}}(\rho)\|^2 \right| &\leq \frac{L}{2} \|\theta_{t+1} - \theta_t\|^2 \\ \implies V_{\pi_{\theta^{(t+1)}}}(\rho) - V_{\pi_{\theta^{(t)}}}(\rho) &\geq \eta \|\nabla V_{\pi_{\theta^{(t)}}}(\rho)\|^2 - \frac{L}{2} \|\theta_{t+1} - \theta_t\|^2 \end{aligned}$$

This further implies that

$$\begin{aligned} V_{\pi_{\theta^{(t+1)}}}(\rho) - V_{\pi_{\theta^*}^*}(\rho) &\geq V_{\pi_{\theta^{(t)}}}(\rho) - V_{\pi_{\theta^*}^*}(\rho) + \eta \|\nabla_{\theta} V_{\pi_{\theta^{(t)}}}(\rho)\|^2 - \frac{L}{2} \|\theta_{t+1} - \theta_t\|^2 \\ &= V_{\pi_{\theta^{(t)}}}(\rho) - V_{\pi_{\theta^*}^*}(\rho) + \eta \left(1 - \frac{L\eta}{2}\right) \|\nabla V_{\pi_{\theta^{(t)}}}(\rho)\|^2 \end{aligned} \quad (41)$$

The last equality is due to Equation (9).

Now, telescoping Equation (41) leads to

$$\eta \left(1 - \frac{L\eta}{2}\right) \sum_{t=0}^{T-1} \|\nabla V_{\pi_{\theta^{(t)}}}(\rho)\|^2 \leq \left( V_{\pi_{\theta^*}^*}(\rho) - V_{\pi_{\theta^*}^0}(\rho) \right) - \left( V_{\pi_{\theta^*}^*}(\rho) - V_{\pi_{\theta^*}^T}(\rho) \right) \quad (42)$$

$$\leq \left( V_{\pi_{\theta^*}^*}(\rho) - V_{\pi_{\theta^*}^0}(\rho) \right) \quad (43)$$

Since  $\sum_{t=0}^{T-1} \|\nabla V_{\pi_{\theta^{(t)}}}(\rho)\|^2 \geq T \min_{t \in [T-1]} \|\nabla V_{\pi_{\theta^{(t)}}}(\rho)\|^2$ , we obtain

$$\min_{t \in [T-1]} \|\nabla V_{\pi_{\theta^{(t)}}}(\rho)\|^2 \leq \frac{1}{T\eta \left(1 - \frac{L\eta}{2}\right)} \left( V_{\pi_{\theta^*}^*}(\rho) - V_{\pi_{\theta^*}^0}(\rho) \right) \leq \frac{R_{\max}}{T\eta \left(1 - \frac{L\eta}{2}\right) (1-\gamma)}.$$

The last inequality comes from  $V_{\pi_{\theta^*}^*}(\rho) \leq \frac{R_{\max}}{1-\gamma}$  (Assumption 1).

**Step 2:** We derive from Equation (15) that

$$\begin{aligned} (V_{\pi_{\theta^*}^*}(\rho) - V_{\pi_{\theta^*}^0}(\rho))^2 &\leq \left( \sqrt{|\mathcal{S}||\mathcal{A}|} \left\| \frac{d_{\pi_{\theta^*}^*, \rho}}{d_{\pi_{\theta^*}^0, \rho}} \right\|_\infty \|\nabla_{\theta} V_{\pi_{\theta^*}^0}(\rho)\|_2 + \frac{2R_{\max}}{1-\gamma} \left( \frac{1}{2} + \frac{\gamma}{1-\gamma} \psi_{\max} \right) \right)^2 \\ &\leq 2|\mathcal{S}||\mathcal{A}| \left\| \frac{d_{\pi_{\theta^*}^*, \rho}}{d_{\pi_{\theta^*}^0, \rho}} \right\|_\infty^2 \|\nabla_{\theta} V_{\pi_{\theta^*}^0}(\rho)\|_2^2 + \frac{8R_{\max}^2}{(1-\gamma)^2} \left( \frac{1}{2} + \frac{\gamma}{1-\gamma} \psi_{\max} \right)^2. \end{aligned}$$

1674  
1675  
1676  
1677  
1678  
1679  
1680  
1681  
1682  
1683  
1684

Thus, we further get

$$\begin{aligned} \min_{t \in [T]} (V_{\pi_o^*}(\rho) - V_{\pi_\theta^{(t)}}(\rho))^2 &\leq 2|\mathcal{S}||\mathcal{A}| \min_{t \in [T]} \left\| \frac{d\pi_o^*}{d\pi_\theta^{(t)}, \nu} \right\|_\infty^2 \|\nabla_\theta V_{\pi_\theta^{(t)}}(\nu)\|_2^2 + \frac{8R_{\max}^2}{(1-\gamma)^2} \left( \frac{1}{2} + \frac{\gamma}{1-\gamma} \psi_{\max} \right)^2 \\ &\leq 2|\mathcal{S}||\mathcal{A}| \text{Cov}^2 \min_{t \in [T]} \|\nabla_\theta V_{\pi_\theta^{(t)}}(\nu)\|_2^2 + \frac{8R_{\max}^2}{(1-\gamma)^2} \left( \frac{1}{2} + \frac{\gamma}{1-\gamma} \psi_{\max} \right)^2 \\ &\leq 2|\mathcal{S}||\mathcal{A}| \text{Cov}^2 \frac{R_{\max}}{T\eta \left(1 - \frac{L\eta}{2}\right) (1-\gamma)} + \frac{8R_{\max}^2}{(1-\gamma)^2} \left( \frac{1}{2} + \frac{\gamma}{1-\gamma} \psi_{\max} \right)^2. \end{aligned}$$

1685 **Step 3:** Now, we set

$$\begin{aligned} \min_{t \in [T]} (V_{\pi_o^*}(\rho) - V_{\pi_\theta^{(t)}}(\rho))^2 &\leq 2|\mathcal{S}||\mathcal{A}| \text{Cov}^2 \frac{R_{\max}}{T\eta \left(1 - \frac{L\eta}{2}\right) (1-\gamma)} + \frac{8R_{\max}^2}{(1-\gamma)^2} \left( \frac{1}{2} + \frac{\gamma}{1-\gamma} \psi_{\max} \right)^2 \\ &\leq \left( \sqrt{2|\mathcal{S}||\mathcal{A}| \frac{R_{\max}}{T\eta \left(1 - \frac{L\eta}{2}\right) (1-\gamma)} \text{Cov}} + \frac{2\sqrt{2}R_{\max}}{(1-\gamma)} \left( \frac{1}{2} + \frac{\gamma}{1-\gamma} \psi_{\max} \right) \right)^2 \\ &\leq \left( \epsilon + \frac{\sqrt{2}R_{\max}}{(1-\gamma)} \left( 1 + \frac{2\gamma}{1-\gamma} \psi_{\max} \right) \right)^2, \end{aligned}$$

1697 and solve for  $T$  to get

$$T \geq \frac{2|\mathcal{S}||\mathcal{A}| \text{Cov}^2 R_{\max}}{\eta(1 - \frac{L\eta}{2})(1-\gamma)\epsilon^2} \quad (44)$$

1702 Choosing  $\eta = \frac{1}{L}$ , we get the final expression

$$T \geq \frac{4L|\mathcal{S}||\mathcal{A}| \text{Cov}^2 R_{\max}}{\epsilon^2(1-\gamma)}. \quad (45)$$

1703  
1704  
1705  
1706  
1707  
1708  
1709  
1710  
1711  
1712  
1713  
1714  
1715  
1716  
1717  
1718  
1719  
1720  
1721  
1722  
1723  
1724  
1725  
1726  
1727

for any  $\epsilon > 0$  and the smoothness constant  $L = \mathcal{O} \left( \max \left\{ \frac{\gamma R_{\max} |\mathcal{A}|}{(1-\gamma)^2}, \frac{\gamma^2}{(1-\gamma)^3} \right\} \right)$ .

Hence, we conclude that for  $T = \Omega \left( \frac{|\mathcal{S}||\mathcal{A}|}{\epsilon^2} \max \left\{ \frac{\gamma R_{\max} |\mathcal{A}|}{(1-\gamma)^3}, \frac{\gamma^2}{(1-\gamma)^4} \right\} \right)$  and  $\psi_{\max} = \mathcal{O}(\frac{1-\gamma}{\gamma})$ ,

$$\min_{t \in [T]} (V_{\pi_o^*}(\rho) - V_{\pi_\theta^{(t)}}(\rho)) \leq \epsilon + \mathcal{O} \left( \frac{1}{1-\gamma} \right).$$

□

1728            G.2 PROOFS FOR ENTROPY-REGULARISED OR SOFT VALUE FUNCTION  
1729  
1730            **Definition 6.** *The discounted state occupancy measure  $d_{\pi'}^{\pi}(s|s_0)$  induced by a policy  $\pi$  and an MDP environment defined by*  
1731             *$\pi'$  is defined as*

$$1733 \quad d_{\pi'}^{\pi}(s|s_0) \triangleq \sum_{a \in \mathcal{A}} d_{\pi'}^{\pi}(s, a|s_0) = (1 - \gamma) \sum_{a \in \mathcal{A}} \mathbb{E}_{\tau \sim \mathbb{P}_{\pi'}} \left[ \sum_{t=0}^{\infty} \gamma^t \mathbf{1}\{s_t = s, a_t = a\} \right].$$

1735            **Lemma 7** (Regularized Performative Policy Difference: Generic Upper Bound). *Under Assumption 1, the sub-optimality*  
1736            *gap of a policy  $\pi_{\theta}$  is*

$$\begin{aligned} 1738 \quad \tilde{V}_{\pi_o^*}^{\pi^*}(s_0) - \tilde{V}_{\pi_{\theta}}^{\pi}(s_0) &\leq \frac{1}{1 - \gamma} \mathbb{E}_{(s,a) \sim d_{\pi_{\theta}}^{\pi^*}(\cdot|s_0)} [\tilde{A}_{\pi_{\theta}}^{\pi}(s, a)] \\ 1739 \quad &+ \frac{2}{1 - \gamma} \left( \xi + \frac{\gamma}{1 - \gamma} \psi_{\max}(R_{\max} + \lambda \log |\mathcal{A}|) \right) \\ 1740 \quad &- \frac{\lambda}{1 - \gamma} \sum_s d_{\pi_{\theta}}^{\pi^*}(s|s_0) D_{\text{KL}}(\pi_o^*(\cdot|s) \parallel \pi_{\theta}(\cdot|s)) \end{aligned} \quad (46)$$

1745            *Proof.* This lemma follows the same sketch as Lemma 2 with an exception in the way the soft rewards are handled. The  
1746            difference in the soft rewards equals the difference of the original rewards with a lagrange dependent term. This term is the  
1747            expected KL divergence over the state visitation distribution. Lemma 1 for regularized rewards reduces to,

$$\begin{aligned} 1749 \quad \tilde{V}_{\pi}^{\pi}(s_0) - \tilde{V}_{\pi'}^{\pi'}(s_0) &= \frac{1}{1 - \gamma} \mathbb{E}_{(s,a) \sim d_{\pi'}^{\pi'}(\cdot|s_0)} [\tilde{A}_{\pi'}^{\pi'}(s, a)] \\ 1750 \quad &+ \frac{1}{1 - \gamma} \mathbb{E}_{(s,a) \sim d_{\pi'}^{\pi'}(\cdot|s_0)} \left( [\tilde{r}_{\pi}(s, a) - \tilde{r}_{\pi'}(s, a)] + \gamma (\mathbf{P}_{\pi} - \mathbf{P}_{\pi'})^{\top} \tilde{V}_{\pi}^{\pi}(s_0) \right). \end{aligned} \quad (47)$$

1753            Therefore,

$$1755 \quad \tilde{r}_{\pi_o^*}(s, a) - \tilde{r}_{\pi_{\theta}}(s, a) = r_{\pi_o^*}(s, a) - r_{\pi_{\theta}}(s, a) + \lambda \left( \log \pi_{\theta}(a|s) - \log \pi_o^*(a|s) \right)$$

1758            Therefore, we can write (47) in the following way,

$$\begin{aligned} 1760 \quad \tilde{V}_{\pi_o^*}^{\pi^*}(s_0) - \tilde{V}_{\pi_{\theta}}^{\pi}(s_0) &= \frac{1}{1 - \gamma} \mathbb{E}_{(s,a) \sim d_{\pi_{\theta}}^{\pi^*}(\cdot|s_0)} [\tilde{A}_{\pi_{\theta}}^{\pi}(s, a)] \\ 1761 \quad &+ \frac{1}{1 - \gamma} \mathbb{E}_{(s,a) \sim d_{\pi_{\theta}}^{\pi^*}(\cdot|s_0)} \left( [\tilde{r}_{\pi_o^*}(s, a) - \tilde{r}_{\pi_{\theta}}(s, a)] + \gamma (\mathbf{P}_{\pi_o^*} - \mathbf{P}_{\pi_{\theta}})^{\top} \tilde{V}_{\pi_o^*}^{\pi^*}(s_0) \right). \\ 1762 \quad &+ \frac{\lambda}{1 - \gamma} \sum_{s,a} [\log \pi_{\theta}(a|s) - \log \pi_o^*(a|s)] d_{\pi_{\theta}}^{\pi^*}(s, a|s_0) \\ 1763 \quad &= \frac{1}{1 - \gamma} \mathbb{E}_{(s,a) \sim d_{\pi_{\theta}}^{\pi^*}(\cdot|s_0)} [\tilde{A}_{\pi_{\theta}}^{\pi}(s, a)] \\ 1764 \quad &+ \frac{1}{1 - \gamma} \mathbb{E}_{(s,a) \sim d_{\pi_{\theta}}^{\pi^*}(\cdot|s_0)} \left( [\tilde{r}_{\pi_o^*}(s, a) - \tilde{r}_{\pi_{\theta}}(s, a)] + \gamma (\mathbf{P}_{\pi_o^*} - \mathbf{P}_{\pi_{\theta}})^{\top} \tilde{V}_{\pi_o^*}^{\pi^*}(s_0) \right) \\ 1765 \quad &+ \frac{\lambda}{1 - \gamma} \sum_{s,a} d_{\pi_{\theta}}^{\pi^*}(s|s_0) \pi_o^*(a|s) [\log \pi_{\theta}(a|s) - \log \pi_o^*(a|s)] \\ 1766 \quad &= \frac{1}{1 - \gamma} \mathbb{E}_{(s,a) \sim d_{\pi_{\theta}}^{\pi^*}(\cdot|s_0)} [\tilde{A}_{\pi_{\theta}}^{\pi}(s, a)] \\ 1767 \quad &+ \frac{1}{1 - \gamma} \mathbb{E}_{(s,a) \sim d_{\pi_{\theta}}^{\pi^*}(\cdot|s_0)} \left( [\tilde{r}_{\pi_o^*}(s, a) - \tilde{r}_{\pi_{\theta}}(s, a)] + \gamma (\mathbf{P}_{\pi_o^*} - \mathbf{P}_{\pi_{\theta}})^{\top} \tilde{V}_{\pi_o^*}^{\pi^*}(s_0) \right) \\ 1768 \quad &- \frac{\lambda}{1 - \gamma} \sum_s d_{\pi_{\theta}}^{\pi^*}(s|s_0) D_{\text{KL}}(\pi_o^*(\cdot|s) \parallel \pi_{\theta}(\cdot|s)) \\ 1769 \quad &\stackrel{\text{Holder's ineq.}}{\leq} \frac{1}{1 - \gamma} \mathbb{E}_{(s,a) \sim d_{\pi_{\theta}}^{\pi^*}(\cdot|s_0)} [\tilde{A}_{\pi_{\theta}}^{\pi}(s, a)] \end{aligned}$$

$$\begin{aligned}
& + \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi_\theta}^{\pi_o^*}(\cdot|s_0)} \left( [\tilde{r}_{\pi_o^*}(s,a) - \tilde{r}_{\pi_\theta}(s,a)] + \gamma \|\mathbf{P}_{\pi_o^*} - \mathbf{P}_{\pi_\theta}\|_1 \|\tilde{V}_{\pi_o^*}^{\pi_\theta}(s_0)\|_\infty \right) \\
& - \frac{\lambda}{1-\gamma} \sum_s d_{\pi_\theta}^{\pi_o^*}(s|s_0) D_{\text{KL}}(\pi_o^*(\cdot|s) \parallel \pi_\theta(\cdot|s)) \\
& \stackrel{(b)}{\leq} \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi_\theta}^{\pi_o^*}(\cdot|s_0)} [\tilde{A}_{\pi_\theta}(s,a)] \\
& + \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi_\theta}^{\pi_o^*}(\cdot|s_0)} \left( [\tilde{r}_{\pi_o^*}(s,a) - \tilde{r}_{\pi_\theta}(s,a)] + \gamma \|\mathbf{P}_{\pi_o^*} - \mathbf{P}_{\pi_\theta}\|_1 \frac{R_{\max} + \lambda \log |\mathcal{A}|}{1-\gamma} \right) \\
& - \frac{\lambda}{1-\gamma} \sum_s d_{\pi_\theta}^{\pi_o^*}(s|s_0) D_{\text{KL}}(\pi_o^*(\cdot|s) \parallel \pi_\theta(\cdot|s)) \\
& \stackrel{\text{Lipschitz } r \text{ & } \mathbf{P}}{\leq} \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi_\theta}^{\pi_o^*}(\cdot|s_0)} [\tilde{A}_{\pi_\theta}(s,a)] \\
& + \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi_\theta}^{\pi_o^*}(\cdot|s_0)} \left( L_r + L_{\mathbf{P}} \frac{\gamma(R_{\max} + \lambda \log |\mathcal{A}|)}{1-\gamma} \right) \|\pi_o^* - \pi_\theta\|_1 \\
& - \frac{\lambda}{1-\gamma} \sum_s d_{\pi_\theta}^{\pi_o^*}(s|s_0) D_{\text{KL}}(\pi_o^*(\cdot|s) \parallel \pi_\theta(\cdot|s)) \\
& \stackrel{(c)}{\leq} \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi_\theta}^{\pi_o^*}(\cdot|s_0)} [\tilde{A}_{\pi_\theta}(s,a)] \\
& + \frac{2}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\pi_\theta}^{\pi_o^*}(\cdot|s_0)} \left( L_r + L_{\mathbf{P}} \frac{\gamma(R_{\max} + \lambda \log |\mathcal{A}|)}{1-\gamma} \right) \\
& - \frac{\lambda}{1-\gamma} \sum_s d_{\pi_\theta}^{\pi_o^*}(s|s_0) D_{\text{KL}}(\pi_o^*(\cdot|s) \parallel \pi_\theta(\cdot|s))
\end{aligned}$$

The equality (a) holds since,

$$\mathbb{E}_{a \sim \pi_o^*(\cdot|s)} [\log \pi_\theta(a|s) - \log \pi_o^*(a|s)] = -D_{\text{KL}}(\pi_o^*(\cdot|s) \parallel \pi_\theta(\cdot|s))$$

The inequality (b) holds due to the result of [Mei et al. \(2020\)](#), i.e.

$$\|\tilde{V}_{\pi_o^*}^{\pi_\theta}\|_\infty \leq \frac{R_{\max} + \lambda \log |\mathcal{A}|}{1-\gamma} \quad (48)$$

Finally, (c) is due to the fact that  $\|\pi_o^* - \pi_\theta\|_1 \leq 2$ .

□

**Lemma 8** (Regularized Performative Policy gradient for softmax policies and softmax MDPs). *For a class of PeMDPs  $\mathcal{M} \triangleq (\mathcal{S}, \mathcal{A}, \pi, \mathbf{P}_\pi, r_\pi, \boldsymbol{\theta}, \rho)$  consider softmax parametrization for policy  $\pi_\theta \in \Delta(\boldsymbol{\theta} \in \Theta)$  and transition dynamics  $\mathbf{P}_{\pi_\theta}$  and linear parametrization for reward  $r_{\pi_\theta}$ . For all  $(s, a, s') \in (\mathcal{S}, \mathcal{A}, \mathcal{S})$ , derivative of the expected return w.r.t  $\boldsymbol{\theta}_{s,a}$  satisfies:*

$$\frac{\partial \tilde{V}_{\pi_\theta}^{\pi_\theta}(\rho)}{\partial \boldsymbol{\theta}_{s,a}} \geq \frac{1}{1-\gamma} d_{\pi_\theta}^{\pi_\theta}(s, a | \rho) \left( \tilde{A}_{\pi_\theta}^{\pi_\theta}(s, a) + \xi \right) - \frac{\lambda}{1-\gamma} (1 + \log |\mathcal{A}|). \quad (49)$$

*Proof.* This proof follows the same sketch as the proof of Theorem 3. However, we get two additional  $\lambda$ -dependent terms—(a) one from the log policy term in the soft advantage, and (b) the other from the log policy term in the soft rewards. We then simplify these terms to obtain the final expression.

First, let us note that

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}_{s',a'}} \log \pi_\theta(a|s) &= \mathbb{1}[s = s', a = a'] - \pi_\theta(a'|s) \mathbb{1}[s = s'] \\
\frac{\partial}{\partial \boldsymbol{\theta}_{s',a'}} \log \mathbf{P}_{\pi_\theta}(s''|s, a) &= \psi(s'') \mathbb{1}[s = s', a = a'] (1 - \mathbf{P}_{\pi_\theta}(s''|s, a)) \\
\frac{\partial}{\partial \boldsymbol{\theta}_{s',a'}} r_{\pi_\theta}(s, a) &= \xi \mathbb{1}[s = s', a = a'].
\end{aligned} \quad (50)$$

1836 Now, we get from Theorem 2,

$$\begin{aligned}
1838 \quad & \frac{\partial}{\partial \boldsymbol{\theta}_{s,a}} \tilde{V}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}(\rho) = \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}} \left[ \sum_{t=0}^{\infty} \gamma^t \left( \tilde{A}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}(s_t, a_t) \frac{\partial}{\partial \boldsymbol{\theta}_{s,a}} \log \pi_{\boldsymbol{\theta}}(a_t | s_t) \right. \right. \\
1839 \quad & \quad + \tilde{A}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}(s_t, a_t) \frac{\partial}{\partial \boldsymbol{\theta}_{s,a}} \log P_{\pi_{\boldsymbol{\theta}}}(s_{t+1} | s_t, a_t) \\
1840 \quad & \quad \left. \left. + \frac{\partial}{\partial \boldsymbol{\theta}_{s,a}} r_{\pi_{\boldsymbol{\theta}}}(s_t, a_t) - \lambda \frac{\partial}{\partial \boldsymbol{\theta}_{s,a}} \log \pi_{\boldsymbol{\theta}}(a_t | s_t) \right) \right] \\
1841 \quad & = \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}} \left[ \sum_{t=0}^{\infty} \gamma^t \left( \tilde{A}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}(s_t, a_t) (\mathbb{1}[s_t = s, a_t = a] - \pi_{\boldsymbol{\theta}}(a|s) \mathbb{1}[s_t = s]) \right. \right. \\
1842 \quad & \quad + \tilde{A}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}(s_t, a_t) \psi(s_{t+1}) \mathbb{1}[s_t = s, a_t = a] (1 - \mathbf{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1} | s, a)) \\
1843 \quad & \quad \left. \left. + \xi \mathbb{1}[s_t = s, a_t = a] - \lambda \mathbb{1}[s_t = s, a_t = a] + \lambda \pi_{\boldsymbol{\theta}}(a|s) \mathbb{1}[s_t = s] \right) \right] \\
1845 \quad & \geq \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}} \left[ \sum_{t=0}^{\infty} \gamma^t \tilde{A}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}(s_t, a_t) \mathbb{1}[s_t = s, a_t = a] \right] - \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}} \left[ \sum_{t=0}^{\infty} \gamma^t \pi_{\boldsymbol{\theta}}(a|s) \mathbb{1}[s_t = s] \tilde{A}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}(s_t, a_t) \right] \\
1846 \quad & \stackrel{(a)}{=} \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}} \left[ \sum_{t=0}^{\infty} \gamma^t \xi \mathbb{1}[s_t = s, a_t = a] \right] - \lambda \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}} \left[ \sum_{t=0}^{\infty} \gamma^t \mathbb{1}[s_t = s, a_t = a] \right] + \lambda \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}} \left[ \sum_{t=0}^{\infty} \gamma^t \pi_{\boldsymbol{\theta}}(a_t | s_t) \mathbb{1}[s_t = s] \right] \\
1847 \quad & = \frac{1}{1-\gamma} d_{\pi_{\boldsymbol{\theta}}, \rho}^{\pi_{\boldsymbol{\theta}}}(s, a) \tilde{A}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}(s, a) + \lambda \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}} \left[ \sum_{t=0}^{\infty} \gamma^t \pi_{\boldsymbol{\theta}}(a_t | s_t) \log \pi_{\boldsymbol{\theta}}(a_t | s_t) \mathbb{1}[s_t = s] \right] \\
1848 \quad & \quad + \frac{1}{1-\gamma} \xi d_{\pi_{\boldsymbol{\theta}}, \rho}^{\pi_{\boldsymbol{\theta}}}(s, a) - \frac{\lambda}{1-\gamma} d_{\pi_{\boldsymbol{\theta}}, \rho}^{\pi_{\boldsymbol{\theta}}}(s, a | s_0) + \lambda \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}} \left[ \sum_{t=0}^{\infty} \gamma^t \pi_{\boldsymbol{\theta}}(a_t | s_t) \sum_a \mathbb{1}[s_t = s, a_t = a] \right] \\
1849 \quad & = \frac{1}{1-\gamma} d_{\pi_{\boldsymbol{\theta}}, \rho}^{\pi_{\boldsymbol{\theta}}}(s, a) \tilde{A}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}(s, a) + \lambda \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}} \left[ \sum_{t=0}^{\infty} \gamma^t \pi_{\boldsymbol{\theta}}(a_t | s_t) \log \pi_{\boldsymbol{\theta}}(a_t | s_t) \sum_a \mathbb{1}[s_t = s, a_t = a] \right] \\
1850 \quad & \quad + \frac{1}{1-\gamma} \xi d_{\pi_{\boldsymbol{\theta}}, \rho}^{\pi_{\boldsymbol{\theta}}}(s, a) - \frac{\lambda}{1-\gamma} d_{\pi_{\boldsymbol{\theta}}, \rho}^{\pi_{\boldsymbol{\theta}}}(s, a) + \lambda \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}} \left[ \sum_a \pi_{\boldsymbol{\theta}}(a|s) \sum_{t=0}^{\infty} \gamma^t \mathbb{1}[s_t = s, a_t = a] \right] \\
1851 \quad & = \frac{1}{1-\gamma} d_{\pi_{\boldsymbol{\theta}}, \rho}^{\pi_{\boldsymbol{\theta}}}(s, a) \tilde{A}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}(s, a) + \lambda \mathbb{E}_{\tau \sim \mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}} \left[ \sum_a \pi_{\boldsymbol{\theta}}(a|s) \log \pi_{\boldsymbol{\theta}}(a|s) \sum_{t=0}^{\infty} \gamma^t \mathbb{1}[s_t = s, a_t = a] \right] \\
1852 \quad & \quad + \frac{1}{1-\gamma} \xi d_{\pi_{\boldsymbol{\theta}}, \rho}^{\pi_{\boldsymbol{\theta}}}(s, a) - \frac{\lambda}{1-\gamma} d_{\pi_{\boldsymbol{\theta}}, \rho}^{\pi_{\boldsymbol{\theta}}}(s, a) + \frac{\lambda}{1-\gamma} \sum_a d_{\pi_{\boldsymbol{\theta}}, \rho}^{\pi_{\boldsymbol{\theta}}}(s, a) \pi_{\boldsymbol{\theta}}(a|s) \\
1853 \quad & \geq \frac{1}{1-\gamma} d_{\pi_{\boldsymbol{\theta}}, \rho}^{\pi_{\boldsymbol{\theta}}}(s, a) \left( \tilde{A}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}(s, a) + \xi \right) - \frac{\lambda}{1-\gamma} d_{\pi_{\boldsymbol{\theta}}, \rho}^{\pi_{\boldsymbol{\theta}}}(s, a) - \frac{\lambda}{1-\gamma} \sum_a d_{\pi_{\boldsymbol{\theta}}, \rho}^{\pi_{\boldsymbol{\theta}}}(s, a) \pi_{\boldsymbol{\theta}}(a|s) \log \frac{1}{\pi_{\boldsymbol{\theta}}(a|s)} \\
1854 \quad & \geq \frac{1}{1-\gamma} d_{\pi_{\boldsymbol{\theta}}, \rho}^{\pi_{\boldsymbol{\theta}}}(s, a) \left( \tilde{A}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}(s, a) + \xi \right) - \frac{\lambda}{1-\gamma} d_{\pi_{\boldsymbol{\theta}}, \rho}^{\pi_{\boldsymbol{\theta}}}(s, a) (1 + \log |\mathcal{A}|). \\
1855 \quad & \stackrel{(b)}{\geq} \frac{1}{1-\gamma} d_{\pi_{\boldsymbol{\theta}}, \rho}^{\pi_{\boldsymbol{\theta}}}(s, a) \left( \tilde{A}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}(s, a) + \xi \right) - \frac{\lambda}{1-\gamma} d_{\pi_{\boldsymbol{\theta}}, \rho}^{\pi_{\boldsymbol{\theta}}}(s, a) (1 + \log |\mathcal{A}|).
\end{aligned}$$

1878 (b) holds from the following:

$$\begin{aligned}
1880 \quad & - \sum_a d_{\pi_{\boldsymbol{\theta}}}(s, a | s_0) \log \pi_{\boldsymbol{\theta}}(a | s) = d_{\pi_{\boldsymbol{\theta}}}(s | s_0) \left( - \sum_a \pi_{\boldsymbol{\theta}}(a | s) \log \pi_{\boldsymbol{\theta}}(a | s) \right) \\
1881 \quad & \quad \stackrel{(c)}{\leq} d_{\pi_{\boldsymbol{\theta}}}(s | s_0) \log |\mathcal{A}| \leq \log |\mathcal{A}|
\end{aligned}$$

1884 and (c) holds as entropy is upper bounded by  $\log |\mathcal{A}|$  (Cover & Thomas, 2006, Theorem 2.6.4).

□

1887 **Lemma 3** (Regularized Performative Gradient Domination: Part(b) of Lemma 3). *For regularized PeMDPs the following inequality holds:*

$$1889 \quad \tilde{V}_{\pi_{\boldsymbol{o}}^*}^{\pi_{\boldsymbol{o}}^*}(\rho) - \tilde{V}_{\pi_{\boldsymbol{\theta}}}^{\pi_{\boldsymbol{\theta}}}(\rho)$$

$$\leq \sqrt{|\mathcal{S}||\mathcal{A}|} \left\| \frac{\mathbf{d}_{\pi_\theta^\star, \rho}^{\pi_o^\star}}{\mathbf{d}_{\pi_\theta, \nu}^{\pi_o^\star}} \right\|_\infty \|\nabla_{\theta} \tilde{V}_{\pi_\theta}^{\pi_\theta}(\nu)\|_2 + \frac{R_{\max}}{1-\gamma} \left( 1 + \frac{2\gamma}{1-\gamma} \psi_{\max} \left( 1 + \frac{\lambda}{R_{\max}} \log |\mathcal{A}| \right) \right) + \frac{\lambda}{1-\gamma} (1 + \log |\mathcal{A}|). \quad (51)$$

**Proof. Step 1.** First, we observe that

$$-D_{\text{KL}}(\pi_o^\star(\cdot|s) \parallel \pi_\theta(\cdot|s)) \leq -\sum_{a \in \mathcal{A}} \pi_o^\star(a|s) \log \pi_o^\star(a|s) \leq \log |\mathcal{A}|$$

Hence, we get

$$-\sum_s \mathbf{d}_{\pi_\theta}^{\pi^\star}(s|s_0) D_{\text{KL}}(\pi_o^\star(\cdot|s) \parallel \pi_\theta(\cdot|s)) \leq \log |\mathcal{A}| \quad (52)$$

**Step 2.** Using Lemma 8 and applying Cauchy-Schwarz inequality, we get

$$\sum_{s,a} \mathbf{d}_{\pi_\theta}^{\pi^\star}(s,a) \tilde{A}_{\pi_\theta}^{\pi_\theta}(s,a) \leq \sqrt{|\mathcal{S}||\mathcal{A}|} (1-\gamma) \|\nabla_{\theta} \tilde{V}_{\pi_\theta}^{\pi_\theta}(\nu)\|_2 - \xi + \lambda(\log |\mathcal{A}| + 1) \quad (53)$$

**Step 3.** Now, substituting Equation (52) and (53) in Equation (46), we finally get

$$\begin{aligned} \tilde{V}_{\pi_o^\star}^{\pi^\star}(\rho) - \tilde{V}_{\pi_\theta}^{\pi_\theta}(\rho) &\leq \sqrt{|\mathcal{S}||\mathcal{A}|} \left\| \frac{\mathbf{d}_{\pi_\theta^\star, \rho}^{\pi_o^\star}}{\mathbf{d}_{\pi_\theta, \nu}^{\pi_o^\star}} \right\|_\infty \|\nabla_{\theta} \tilde{V}_{\pi_\theta}^{\pi_\theta}(\nu)\|_2 - \left\| \frac{\mathbf{d}_{\pi_\theta^\star, \rho}^{\pi_o^\star}}{\mathbf{d}_{\pi_\theta, \nu}^{\pi_o^\star}} \right\|_\infty \frac{\xi}{1-\gamma} + \frac{\lambda}{1-\gamma} (\log |\mathcal{A}| + 1) \\ &\quad + \frac{2}{1-\gamma} \left( L_r + \frac{\gamma}{1-\gamma} L_P (R_{\max} + \lambda \log |\mathcal{A}|) \right) \\ &\stackrel{(a)}{=} \sqrt{|\mathcal{S}||\mathcal{A}|} \left\| \frac{\mathbf{d}_{\pi_\theta^\star, \rho}^{\pi_o^\star}}{\mathbf{d}_{\pi_\theta, \nu}^{\pi_o^\star}} \right\|_\infty \|\nabla_{\theta} \tilde{V}_{\pi_\theta}^{\pi_\theta}(\nu)\|_2 - \left\| \frac{\mathbf{d}_{\pi_\theta^\star, \rho}^{\pi_o^\star}}{\mathbf{d}_{\pi_\theta, \nu}^{\pi_o^\star}} \right\|_\infty \frac{\xi}{1-\gamma} + \frac{\lambda}{1-\gamma} (1 + \log |\mathcal{A}|) \\ &\quad + \frac{2}{1-\gamma} \left( \xi + \frac{\gamma}{1-\gamma} \psi_{\max}(R_{\max} + \lambda \log |\mathcal{A}|) \right) \\ &\stackrel{(b)}{\leq} \sqrt{|\mathcal{S}||\mathcal{A}|} \left\| \frac{\mathbf{d}_{\pi_\theta^\star, \rho}^{\pi_o^\star}}{\mathbf{d}_{\pi_\theta, \nu}^{\pi_o^\star}} \right\|_\infty \|\nabla_{\theta} \tilde{V}_{\pi_\theta}^{\pi_\theta}(\nu)\|_2 + \frac{\xi}{1-\gamma} \\ &\quad + \frac{2\gamma}{1-\gamma} \psi_{\max}(R_{\max} + \lambda \log |\mathcal{A}|) + \frac{\lambda}{1-\gamma} (1 + \log |\mathcal{A}|) \\ &= \sqrt{|\mathcal{S}||\mathcal{A}|} \left\| \frac{\mathbf{d}_{\pi_\theta^\star, \rho}^{\pi_o^\star}}{\mathbf{d}_{\pi_\theta, \nu}^{\pi_o^\star}} \right\|_\infty \|\nabla_{\theta} \tilde{V}_{\pi_\theta}^{\pi_\theta}(\nu)\|_2 \\ &\quad + \frac{2}{1-\gamma} \left( \frac{\xi}{2} + \frac{\gamma}{1-\gamma} \psi_{\max}(R_{\max} + \lambda \log |\mathcal{A}|) \right) + \frac{\lambda}{1-\gamma} (1 + \log |\mathcal{A}|) \\ &\leq \sqrt{|\mathcal{S}||\mathcal{A}|} \left\| \frac{\mathbf{d}_{\pi_\theta^\star, \rho}^{\pi_o^\star}}{\mathbf{d}_{\pi_\theta, \nu}^{\pi_o^\star}} \right\|_\infty \|\nabla_{\theta} \tilde{V}_{\pi_\theta}^{\pi_\theta}(\nu)\|_2 \\ &\quad + \frac{R_{\max}}{1-\gamma} \left( 1 + \frac{2\gamma}{1-\gamma} \psi_{\max} \left( 1 + \frac{\lambda}{R_{\max}} \log |\mathcal{A}| \right) \right) + \frac{\lambda}{1-\gamma} (1 + \log |\mathcal{A}|) \end{aligned}$$

In (a), we substitute the values of  $L_r$  and  $L_P$  for softmax PeMDPs, and in (b), we use  $\left\| \frac{\mathbf{d}_{\pi_\theta^\star, \rho}^{\pi_o^\star}}{\mathbf{d}_{\pi_\theta, \nu}^{\pi_o^\star}} \right\|_\infty \geq 1$  (Lemma 9).

□

**Theorem 3** (Convergence of PePG in softmax PeMDPs – Part (b)). *Let  $\text{Cov} \triangleq \max_{\theta, \nu} \left\| \frac{\mathbf{d}_{\pi_\theta^\star, \rho}^{\pi_o^\star}}{\mathbf{d}_{\pi_\theta, \nu}^{\pi_o^\star}} \right\|_\infty$ . The gradient ascent algorithm on  $V_{\pi_\theta}^{\pi_\theta}(\rho)$  (Equation (9)) with step size  $\eta = \Omega \left( \frac{(1-\gamma)^2}{\gamma |\mathcal{A}|} \right)$  satisfies, for all distributions  $\rho \in \Delta(\mathcal{S})$ .*

(b) For entropy regularised case, if we set  $\lambda = \frac{(1-\gamma)R_{\max}}{1+\log |\mathcal{A}|}$ , we get

$$\min_{t < T} \left\{ \tilde{V}_{\pi_o^\star}^{\pi^\star}(\rho) - \tilde{V}_{\pi_\theta^{(t)}}^{\pi_\theta^{(t)}}(\rho) \right\} \leq \epsilon \text{ when } T = \Omega \left( \frac{R_{\max} |\mathcal{S}||\mathcal{A}|^2}{\epsilon^2 (1-\gamma)^3} \text{Cov}^2 \right), \text{ and } \epsilon = \Omega \left( \frac{1}{1-\gamma} \right).$$

1944 *Proof.* This proof follows similar steps as part (a) of Theorem 3 with two additional changes: (i) We have a  $\lambda$ , i.e. regularisation coefficient, dependent term due to the entropy regulariser. (ii) The maximum value of the soft value function is  
1945  $\frac{R_{\max} + \lambda \log |\mathcal{A}|}{1-\gamma}$  instead of  $\frac{R_{\max}}{1-\gamma}$  for the unregularised value function.  
1946  
1947

1948 **Step 1:** From Equation (31), we observe that the soft-value function  $\tilde{V}_{\pi_\theta}^{\pi_\theta^{(t)}}$  is  $L_\lambda$ -smooth.  
1949  
1950 Thus, following the Step 1 of Theorem 3, we get

$$\begin{aligned} \min_{t \in [T-1]} \|\nabla \tilde{V}_{\pi_\theta^{(t)}}^{\pi_\theta^{(t)}}(\rho)\|^2 &\leq \frac{1}{T\eta \left(1 - \frac{L_\lambda \eta}{2}\right)} \left( \tilde{V}_{\pi_\theta^{(0)}}^{\pi_\theta^{(0)}}(\rho) - \tilde{V}_{\pi_\theta^{(0)}}^{\pi_\theta^{(0)}}(\rho) \right) \\ &\leq \frac{R_{\max} + \lambda \log |\mathcal{A}|}{T\eta \left(1 - \frac{L_\lambda \eta}{2}\right) (1-\gamma)}. \end{aligned} \quad (54)$$

1955 The last inequality is true due to the fact that  $\tilde{V}_{\pi_\theta^{(0)}}^{\pi_\theta^{(0)}}(\rho) - \tilde{V}_{\pi_\theta^{(0)}}^{\pi_\theta^{(0)}}(\rho) \leq \tilde{V}_{\pi_\theta^{(0)}}^{\pi_\theta^{(0)}}(\rho) \leq \frac{R_{\max} + \lambda \log |\mathcal{A}|}{1-\gamma}$ .  
1956  
1957  
1958

1959 **Step 2:** Now, from Part (b) of Lemma 3, we obtain that  
1960  
1961  
1962

$$\begin{aligned} \min_{t \in [T-1]} \left( \tilde{V}_{\pi_\theta^{(0)}}^{\pi_\theta^{(0)}}(\rho) - \tilde{V}_{\pi_\theta^{(t)}}^{\pi_\theta^{(t)}}(\rho) \right)^2 &\leq \min_{t \in [T-1]} \left( \sqrt{|\mathcal{S}||\mathcal{A}|} \left\| \frac{\mathbf{d}_{\pi_\theta^{(t)}, \rho}^{\pi_\theta^{(t)}}}{\mathbf{d}_{\pi_\theta^{(t)}, \nu}^{\pi_\theta^{(t)}}} \right\|_\infty \|\nabla_\theta \tilde{V}_{\pi_\theta^{(t)}}^{\pi_\theta^{(t)}}(\nu)\|_2 + \frac{R_{\max}}{1-\gamma} \left( 1 + \frac{2\gamma}{1-\gamma} \psi_{\max} \left( 1 + \frac{\lambda}{R_{\max}} \log |\mathcal{A}| \right) \right) + \frac{\lambda}{1-\gamma} (1 + \log |\mathcal{A}|) \right)^2 \\ &\leq 2|\mathcal{S}||\mathcal{A}| \min_{t \in [T-1]} \left\| \frac{\mathbf{d}_{\pi_\theta^{(t)}, \rho}^{\pi_\theta^{(t)}}}{\mathbf{d}_{\pi_\theta^{(t)}, \nu}^{\pi_\theta^{(t)}}} \right\|_\infty^2 \|\nabla_\theta \tilde{V}_{\pi_\theta^{(t)}}^{\pi_\theta^{(t)}}(\nu)\|_2^2 + 2 \left( \frac{R_{\max}}{1-\gamma} \left( 1 + \frac{2\gamma}{1-\gamma} \psi_{\max} \left( 1 + \frac{\lambda}{R_{\max}} \log |\mathcal{A}| \right) \right) + \frac{\lambda}{1-\gamma} (1 + \log |\mathcal{A}|) \right)^2 \\ &\leq \frac{2|\mathcal{S}||\mathcal{A}| \text{Cov}^2(R_{\max} + \lambda \log |\mathcal{A}|)}{T\eta \left(1 - \frac{L_\lambda \eta}{2}\right) (1-\gamma)} + 2 \left( \frac{R_{\max}}{1-\gamma} \left( 1 + \frac{2\gamma}{1-\gamma} \psi_{\max} \left( 1 + \frac{\lambda}{R_{\max}} \log |\mathcal{A}| \right) \right) + \frac{\lambda}{1-\gamma} (1 + \log |\mathcal{A}|) \right)^2. \end{aligned}$$

1974 The last inequality is due to the upper bound on the minimum gradient norm as in Equation (54) and by definition of the  
1975 coverage parameter Cov.  
1976

1977 Thus, we conclude that  
1978  
1979

$$\begin{aligned} \min_{t \in [T-1]} \tilde{V}_{\pi_\theta^{(0)}}^{\pi_\theta^{(0)}}(\rho) - \tilde{V}_{\pi_\theta^{(t)}}^{\pi_\theta^{(t)}}(\rho) &\leq \sqrt{\frac{2|\mathcal{S}||\mathcal{A}| \text{Cov}^2(R_{\max} + \lambda \log |\mathcal{A}|)}{T\eta \left(1 - \frac{L_\lambda \eta}{2}\right) (1-\gamma)}} + \sqrt{2} \left( \frac{R_{\max}}{1-\gamma} \left( 1 + \frac{2\gamma}{1-\gamma} \psi_{\max} \left( 1 + \frac{\lambda}{R_{\max}} \log |\mathcal{A}| \right) \right) + \frac{\lambda}{1-\gamma} (1 + \log |\mathcal{A}|) \right). \end{aligned} \quad (55)$$

1985 **Step 4:** Now, by setting the  $T$ -dependent term in Equation (55) to  $\epsilon$ , we get  $T \geq \frac{2|\mathcal{S}||\mathcal{A}| \text{Cov}^2(R_{\max} + \lambda \log |\mathcal{A}|)}{\eta \left(1 - \frac{L_\lambda \eta}{2}\right) (1-\gamma) \epsilon^2}$ .  
1986

1987 Choosing  $\eta = \frac{1}{L_\lambda}$ ,  $\lambda = \frac{(1-\gamma)R_{\max}}{(1+\log |\mathcal{A}|)}$ , and  $\psi_{\max} = \mathcal{O}(\frac{1-\gamma}{\gamma})$ , we get the final expression  $T \geq \frac{8|\mathcal{S}||\mathcal{A}| \text{Cov}^2 L_\lambda R_{\max}}{(1-\gamma) \epsilon^2}$ , and  
1988

$$\min_{t \in [T-1]} \tilde{V}_{\pi_\theta^{(0)}}^{\pi_\theta^{(0)}}(\rho) - \tilde{V}_{\pi_\theta^{(t)}}^{\pi_\theta^{(t)}}(\rho) \leq \epsilon + \mathcal{O}\left(\frac{1}{1-\gamma}\right).$$

1991  
1992 Finally, noting that  $L_\lambda = \mathcal{O}\left(\max\left\{\frac{\gamma R_{\max} |\mathcal{A}| \psi_{\max}^2}{(1-\gamma)^2}, \frac{R_{\max} \psi_{\max}^2}{(1-\gamma)^2}\right\}\right)$ , we get  
1993  
1994

$$T = \Omega\left(\frac{|\mathcal{S}||\mathcal{A}|}{\epsilon^2 (1-\gamma)^3} \max\{1, \gamma |\mathcal{A}|\}\right).$$

□

1998

**H ABLATION STUDY ON ENTROPY REGULARISATION**

1999

2000

2001

2002

2003

2004

2005

2006

2007

2008

2009

2010

2011

2012

2013

2014

2015

2016

2017

2018

2019

2020

2021

2022

2023

2024

2025

2026

2027

2028

2029

2030

2031

2032

2033

2034

2035

2036

2037

2038

2039

2040

2041

2042

2043

2044

2045

2046

2047

2048

2049

2050

2051

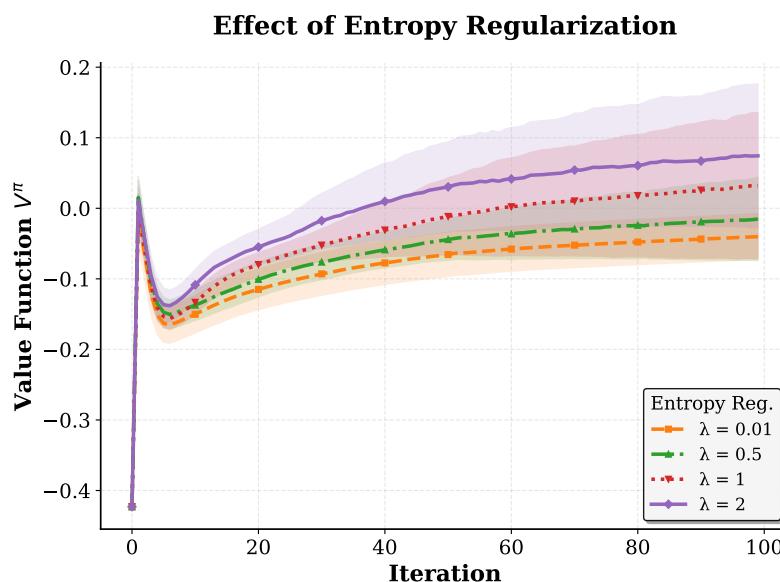


Figure 3: Ablation study for PePG for different values of regularised  $\lambda$  with 20 random seeds, each for 100 iterations

We conducted an ablation study across four entropy regularization strengths ( $\lambda \in \{0.01, 0.5, 1, 2\}$ ) to determine the optimal balance between exploration and convergence stability in RegPePG. The results demonstrate that  $\lambda = 2$  achieves the highest final performance (0.05), while smaller values ( $\lambda \leq 1$ ) converge to similar suboptimal levels around  $-0.01$  to 0, indicating that stronger entropy regularization enables more effective exploration of the policy space in performative settings.

2052    **I TECHNICAL LEMMAS**  
 2053  
 2054    **Lemma 9** (Lower Bound of Coverage). *For any  $\pi, \pi' \in \Pi(\Theta)$ , the following non-trivial lower bound holds,*  
 2055

$$\left\| \frac{\mathbf{d}_{\pi'}}{\mathbf{d}_{\pi}} \right\|_{\infty} \geq 1$$

2056    *Proof.*  
 2057  
 2058

$$\left\| \frac{\mathbf{d}_{\pi'}}{\mathbf{d}_{\pi}} \right\|_{\infty} = \max_{s,a} \frac{\mathbf{d}_{\pi'}(s,a)}{\mathbf{d}_{\pi}(s,a)} \geq \frac{1}{\sum_{s,a} w_{s,a}} \sum_{s,a} \frac{\mathbf{d}_{\pi'}(s,a)}{\mathbf{d}_{\pi}(s,a)} \cdot w_{s,a}$$

2064    Choose  $w_{s,a} = \mathbf{d}_{\pi}(s,a)$  Hence, we get,  
 2065

$$\max_{s,a} \frac{\mathbf{d}_{\pi'}(s,a)}{\mathbf{d}_{\pi}(s,a)} \geq \frac{\sum_{s,a} \mathbf{d}_{\pi'}(s,a)}{\sum_{s,a} \mathbf{d}_{\pi}(s,a)} = 1$$

2066    The last equality holds from the fact that the state-action occupancy measure is a distribution over  $\mathcal{S} \times \mathcal{A}$ . Hence,  
 2067     $\sum_{s,a} \mathbf{d}_{\pi'}(s,a) = \sum_{s,a} \mathbf{d}_{\pi}(s,a)$  □  
 2068

2069    **Lemma 10.** *The discounted state occupancy measure*  
 2070

$$\mathbf{d}_{\pi'}^{\pi}(s|s_0) \triangleq (1 - \gamma) \mathbb{E}_{\tau \sim \mathbb{P}_{\pi'}^{\pi}} \left[ \sum_{t=0}^{\infty} \gamma^t \mathbf{1}\{s_t = s\} \right]$$

2071    *is a probability mass function over the state-space  $\mathcal{S}$ .*  
 2072

2073    *Proof.* For each fixed  $s$  the integrand  $\sum_{t=0}^{\infty} \gamma^t \mathbf{1}\{s_t = s\} \geq 0$ , hence  $\mathbf{d}_{\pi'}^{\pi}(s|s_0) \geq 0$ .  
 2074

2075    To check normalization, we sum over all states and use Tonelli/Fubini (permitted because the summand is non-negative) to  
 2076    exchange sums and expectation:  
 2077

$$\sum_{s \in \mathcal{S}} \mathbf{d}_{\pi'}^{\pi}(s|s_0) = (1 - \gamma) \mathbb{E}_{\tau \sim \mathbb{P}_{\pi'}^{\pi}(\cdot|s_0)} \left[ \sum_{t=0}^{\infty} \gamma^t \sum_{s \in \mathcal{S}} \mathbf{1}\{s_t = s\} \right] = (1 - \gamma) \mathbb{E}_{\tau \sim \mathbb{P}_{\pi'}^{\pi}(\cdot|s_0)} \left[ \sum_{t=0}^{\infty} \gamma^t \cdot 1 \right] = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t = 1.$$

2078    Therefore  $\rho$  is a probability mass function on  $\mathcal{S}$ . □  
 2079

2080    A very similar argument holds for the discounted state-action occupancy measure  $\mathbf{d}_{\pi'}^{\pi}(s, a|s_0)$  as well.  
 2081

2082  
 2083  
 2084  
 2085  
 2086  
 2087  
 2088  
 2089  
 2090  
 2091  
 2092  
 2093  
 2094  
 2095  
 2096  
 2097  
 2098  
 2099  
 2100  
 2101  
 2102  
 2103  
 2104  
 2105