Independent Learning in Performative Markov Potential Games

Rilind Sahitaj[*]
RWTH Aachen & MPI-SWS

Paulius Sasnauskas[*]
MPI-SWS

Yiğit Yalın
MPI-SWS

Debmalya Mandal
University of Warwick

Goran Radanović
MPI-SWS

Abstract

Performative Reinforcement Learning (PRL) refers to a scenario in which the deployed policy changes the reward and transition dynamics of the underlying environment. In this work, we study multi-agent PRL by incorporating performative effects into Markov Potential Games (MPGs). We introduce the notion of a performatively stable equilibrium (PSE) and show that it always exists under a reasonable sensitivity assumption. We then provide convergence results for state-of-the-art algorithms used to solve MPGs. Specifically, we show that independent policy gradient ascent (IPGA) and independent natural policy gradient (INPG) converge to an approximate PSE in the *best-iterate* sense, with an additional term that accounts for the performative effects. Furthermore, we show that INPG asymptotically converges to a PSE in the *last-iterate* sense. As the performative effects vanish, we recover the convergence rates from prior work. For a special case of our game, we provide *finite-time last-iterate* convergence results for a repeated retraining approach, in which agents independently optimize a surrogate objective. We conduct extensive experiments to validate our theoretical findings.

1 INTRODUCTION

Multi-Agent Reinforcement Learning (MARL) is a framework for learning equilibrium policies in complex strategic environments. Although recent success stories in game playing (Silver et al., 2017; Brown and Sandholm, 2019; Vinyals et al., 2019) showcase the practical importance of scalable MARL algorithms, deploying them in real world settings often requires robustness properties that are not needed in games with well-defined rules. Specifically, it is important to account for higher-order effects that agents may have on the dynamics of the environment.

In MARL, we typically identify two types of feedback loops: a) one which is due to the RL nature of the problem setting, i.e., an agent's policy affecting future states, and b) one which is due to the multi-agent nature of the problem setting, i.e., an agent's transitions and rewards are affected by the other agents. However, the environment itself, i.e., reward function and transition probabilities, can be influenced by the policies that the agents deploy in the environment. We refer to this type of feedback loop as *performative* effects. Take, for example, AI-assistants, which interact with both users and other agents. As an AI-assistant interacts with its users, it may change their preferences, which in turn influence future interactions. This can be exploited by competing agents that may adapt their strategies accordingly.

Recently, *performative* effects of an agent on its environment have been studied in single-agent reinforcement learning (Mandal et al., 2023, *performative RL*), as well as in supervised learning (Perdomo et al., 2020, *performative prediction*). However, *performative* effects are unexplored in the context of MARL. Unlike prior work on performative prediction and performative RL, *performative* MARL has to simultaneously account for the three type of feedback loops from the previous paragraph.

Motivated by the practical importance of this setting, our goal is to understand the impact of *per-*

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

[*] This work was done as a part of an internship project at the Max Planck Institute for Software Systems.

Table 1: Iteration complexity of the independent policy gradient methods with gradient oracles. IPGA stands for independent policy gradient ascent and INPG abbreviates independent natural policy gradient. $\delta_{r,p} := \frac{1}{1-\gamma}\left(\omega_r + \frac{\gamma \cdot \omega_p \sqrt{S}}{1-\gamma}\right)$, where ω_r and ω_p are the sensitivity parameters modelling the *performative* effects, $\mathcal{W}_{r,p} := (n+1) \cdot n^2 \cdot S \cdot \delta_{r,p}$, S is the number of states, γ is the discount factor. $\kappa_\rho := \sup_t \sup_{\pi \in \Pi} \|d^{\pi}_{\rho,t}/\rho\|_\infty$ and its minimax-version $\widetilde{\kappa}_\rho := \inf_{\rho \in \Delta(\mathcal{S})} \kappa_\rho$, where A is the total number of actions over all agents, $A_{\max}$ is the largest action set over agents, n is the number of agents, C_Φ is the potential function difference upper bound, T is the number of rounds, $M_\rho := \sup_\pi \sup_{\pi'} \max_s \frac{1}{d^{\pi'}_{\rho,\pi}(s)} < \infty$, c is the lower bound for the probability of playing an optimal action.

Algorithm	**Theorem**	**Iteration Complexity**
IPGA	Theorem 3	$\mathcal{O}\left(\frac{\min\{\kappa_\rho, S\}^2 \sqrt{A \cdot n \cdot C_\Phi}}{\epsilon^2 \cdot (1-\gamma)^3} + \frac{8 \min\{\kappa_\rho, S\}^3 \cdot n \cdot A}{(1-\gamma)^{10}} \cdot \frac{\mathcal{W}_{r,p}}{\epsilon^2}\right)$
INPG	Theorem 4	$\mathcal{O}\left(\left(\frac{\tilde{\kappa}_\rho\left(\sqrt{n} + M \cdot \delta_{r,p}\right)}{\epsilon^2 \cdot c \cdot (1-\gamma)^{3.5}}\right)\right)$
INPG + log-barrier reg.	Theorem 5	$\mathcal{O}\left(\frac{n \cdot A_{\max} \cdot M^2}{\epsilon^2 \cdot (1-\gamma)^4} \cdot \max\{1, S \cdot \delta_{r,p}\}\right)$

formative effects on the convergence of well-known MARL algorithms. We base our formal framework on a class of Markov games, called Markov Potential Games (MPGs), and focus on independent policy gradient algorithms, which have been shown to converge in MPGs. Our contributions are as follows:

- **Framework:** We extend the MPG setting to incorporate *performative* effects. We introduce the notions of a performatively stable equilibrium (PSE) and performative regret (PReg).
- **Solution Concepts:** We show that a PSE exists under reasonable sensitivity assumptions. Furthermore, we prove that every PSE corresponds to an approximate Nash equilibrium.
- **Performative Regret Guarantees:** We study IPGA and two versions of INPG, with and without log-barrier regularization. We show that the algorithms achieve bounded performative regret, hence converging to a PSE in the best-iterate sense. Table 1 provides an overview of the convergence results.
- **Last-Iterate Convergence Guarantees:** We show that unregularized INPG achieves last-iterate convergence to an exact PSE asymptotically. However, compared to the regularized version, its guarantee depends on the probability of playing an optimal action. Furthermore, we show that the regularized version of INPG also achieves best-iterate convergence, but compared to the unregularized version, its guarantee does not require a lower bound on the probability of playing an optimal action. Furthermore, we generalize the repeated optimization approach from prior work on performative RL to multi-agent RL for a special class of *performative* MPGs. We show a *finite-time last-iterate* convergence guarantee for this method, which we prove by adapting the proof technique from Mandal et al. (2023) to our problem setting.
- **Experiments:** We evaluate the gradient-based algorithms on the safe-distancing game (Leonardos et al., 2022; Ding et al., 2022) and stochastic congestion games (Fox et al., 2022), showing that *performative* effects significantly impact convergence, affecting algorithms differently. According to our empirical results, the natural policy gradient methods are more robust against *performative* effects.

2 RELATED WORK

Markov Potential Games. Our work is related to Markov Games (Shapley, 1953) and to a subclass of games known as Markov Potential Games. Leonardos et al. (2022) show that in the infinite-horizon case, an independent policy gradient ascent algorithm converges to an ϵ-approximate Nash equilibrium after $\mathcal{O}(1/\epsilon^6)$ iterations. Ding et al. (2022) extend this to function approximation and improve to an iteration complexity of $\mathcal{O}(1/\epsilon^5)$ using a slightly different update rule. Given gradient-oracles, both methods achieve an $\mathcal{O}(1/\epsilon^2)$ iteration complexity. In the setting without gradient oracles, Mao et al. (2022) improve on the previous bounds with a $\mathcal{O}(1/\epsilon^{4.5})$ iteration complexity, by reducing the variance of the policy gradient algorithm. Furthermore, it is well known that independent natural policy gradient (INPG) converges in MPGs (Fox et al., 2022; Sun et al., 2023; Zhang et al., 2022; Alatur et al., 2024a) with sample complexity $\mathcal{O}(1/\epsilon^2)$. Under additional assumptions, Sun et al. (2023) improve this to an iteration complexity of $O(1/\epsilon)$, while Alatur et al. (2024a) reduce the linear dependence on the number of agents to a sublinear dependence, maintaining the $O(1/\epsilon^2)$ complexity.

In our setting, our guarantees additionally depend on an additive term that is dependent on the strength of the *performative* effects. We recover the guarantees from prior work for the considered algorithms as the *performative* effects become negligible.

Performative Prediction. Since the seminal work on performative prediction (Perdomo et al., 2020), various adaptations have been studied. We refer to the survey by Hardt and Mendler-Dünner (2023). A recent line of work studies performative prediction in a multi-agent setting (Li et al., 2022; Narang et al., 2023; Piliouras and Yu, 2023). More similar to reinforcement learning and related to our work, Brown et al. (2022) study a version of performative prediction in which the target population exhibits a state that captures historical information to account for gradual changes in the distribution. See also work by Li and Wai (2022); Ray et al. (2022); Izzo et al. (2022). For more details on the growing literature of performative prediction, we refer to Appendix C.

Performative Reinforcement Learning. Mandal et al. (2023) introduce the setting of performative reinforcement learning (PRL), where the rewards and transition function adapt to the policy. They propose a repeated retraining approach over the occupancy measure space that converges to an approximate performatively stable point. We generalize this technique to MPGs with performative effects and agent independent transitions. Rank et al. (2024) study an extension, where the environment additionally changes based on the past dynamics. Recently, Mandal and Radanovic (2025) extend PRL to the linear MDP setting, Cai et al. (2024) extend *performative* effects in linear dynamical systems, while Pollatos et al. (2025) study corruption-robustness in PRL.

3 FORMAL SETTING

First, we provide some basic notation that we will use throughout. We let $\Delta(X)$ denote the probability simplex over the finite set X. Further, given two vectors $a, b \in \mathbb{R}^k$ for a natural number $k \in \mathbb{N}$, we define a/b as the vector obtained by component-wise division.

3.1 Performative Markov Game

We define a n-player Game (MG) with *performative* effects induced by the adopted joint policy $\bar{\pi}$ as a tuple

$$\mathcal{G}(\bar{\pi}) = (\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^n, \{r_{i,\bar{\pi}}\}_{i=1}^n, P_{\bar{\pi}}, \gamma, \rho),$$

where $\mathcal{N} = \{1, \ldots, n\}$ is the set of agents, $\mathcal{S}$ is a (finite) state space, $\mathcal{A}_i$ is a finite action set of agent $i \in \mathcal{N}$ with the joint action space denoted as $\mathcal{A} := \prod_{i \in \mathcal{N}} \mathcal{A}_i$ for $i \in \mathcal{N}$. Under the adopted policy $\bar{\pi}$, $r_{i,\bar{\pi}} : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the i-th agent's reward function and $P_{\bar{\pi}}$ is the transition probability measure, characterized by a distribution $P_{\bar{\pi}}(\cdot \mid s, a)$ over $\mathcal{S}$, given an action $a \in \mathcal{A}$ and state $s \in \mathcal{S}$. Furthermore, $\gamma \in [0, 1)$ is the discount factor and $\rho \in \Delta(S)$ corresponds to the initial state distribution. We abbreviate the cardinalities of the action space and state space as $|\mathcal{A}| = A, |\mathcal{A}_i| = A_i$ and $|\mathcal{S}| = S$, respectively.

We assume that the joint policy $\pi = (\pi_i)_{i \in \mathcal{N}} \in \Pi := \Delta(\mathcal{A}_1)^{\mathcal{S}} \times \cdots \times \Delta(\mathcal{A}_n)^{\mathcal{S}}$ is a product of individual, stochastic policies. Specifically, we define agent's i policy $\pi_i : \mathcal{S} \to \Delta(\mathcal{A}_i)$ as a probability distribution $\pi_i(\cdot|s)$ over $\mathcal{A}_i$, given a state $s \in \mathcal{S}$. For brevity, we denote $\Pi_i = \Delta(\mathcal{A}_i)$. We emphasize that the probability transition function and the agent specific reward functions adapt to the underlying joint policy $\bar{\pi}$ specified in the game $\mathcal{G}(\bar{\pi})$.

We use superscript $t \in [T]$ to indicate dependence on the iterations of the methods we study. For example, the agents' joint policy at iteration t is denoted by π^t. When a function $f^{\pi^t}_{\pi^{t'}}$ depends on π^t and $\mathcal{G}(\pi^{t'})$, we use the shorthand notation $f^t_{t'}$.

We use superscript h to denote dependence on a single decision-making time step. More specifically, at a time step h, we are given a state $s^{h-1} \in \mathcal{S}$, an action profile $a^h = (a^h_1, \ldots, a^h_n) \in \mathcal{A}$ and observe rewards $r_{1,\bar{\pi}}(s^{h-1}, a^h), \ldots, r_{n,\bar{\pi}}(s^{h-1}, a^h)$ and transition to a state s^h.

The probability of a trajectory $\tau = (s^h, a^h, r^h)_{h=0}^{\infty}$ is given by: $\mathbb{P}^{\pi}_{\bar{\pi}}(\tau) = \rho(s^0) \cdot \Pi_{h \geq 0} P_{\bar{\pi}}(s^{h+1}|s^h, \pi(s^h))$, i.e., the trajectory τ is sampled by following policy π under the transition function $P_{\bar{\pi}}$. Furthermore, we define the *performative* value function of agent $i \in \mathcal{N}$ in $\mathcal{G}(\bar{\pi})$ under joint policy π as:

$$V^{\pi}_{i,\bar{\pi}}(\rho) = \mathbb{E}_{\tau \sim \mathbb{P}^{\pi}_{\bar{\pi}}} \Big[\sum_{h=0}^{\infty} \gamma^h r^h_{i,\bar{\pi}} \big| s^0 \sim \rho \Big]. \tag{1}$$

We denote by $Q^{\pi}_{i,\bar{\pi}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the action-value function of an agent $i \in \mathcal{N}$ in $\mathcal{G}(\bar{\pi})$ under joint policy π, that is

$$Q^{\pi}_{i,\bar{\pi}}(s, a) := \mathbb{E}_{\tau \sim \mathbb{P}^{\pi}_{\bar{\pi}}} \Big[\sum_{h=0}^{\infty} \gamma^h r^h_{i,\bar{\pi}} \big| s^0 = s, a^0 = a \Big].$$

Given initial state distribution ρ and policy π, the discounted state visitation distribution with respect to the underlying $\mathcal{G}(\bar{\pi})$ is given by:

$$d^{\pi}_{\rho,\bar{\pi}}(s) = (1 - \gamma) \mathbb{E}_{\tau \sim \mathbb{P}^{\pi}_{\bar{\pi}}} \left[\sum_{h=0}^{\infty} \gamma^h \mathbb{1}_{\{s^h = s, a^h = a\}} \middle| s^0 \sim \rho \right].$$

We adapt the definition of the distribution mismatch coefficient (see e.g., Ding et al. (2022)) to our setting. We use the shorthand notation $d^{\pi}_{\rho,t}$ for d^{π}_{ρ,π^t}. Furthermore, we define the distribution mismatch coefficient $\kappa_\rho \coloneqq \sup_t \sup_{\pi\in\Pi}\|d^{\pi}_{\rho,t}/\rho\|_\infty$ and the minimax-version $\tilde{\kappa}_\rho \coloneqq \inf_{\nu\in\Delta(\mathcal{S})}\sup_t \sup_{\pi\in\Pi}\|d^{\pi}_{\rho,t}/\nu\|_\infty$.

3.2 Performative Markov Potential Games

We are extending the Markov Potential Games (MPGs), which are special classes of MGs. MPGs assume the existence of a potential function, which, in our setting, would be policy-dependent. More formally, a MG with *performative* effects $\mathcal{G}$ is potential if for every $\bar{\pi}$, the induced game $\mathcal{G}(\bar{\pi})$ is a Markov Potential Game. This implies that for all $\bar{\pi}$ there exists a potential function $\Phi_{\bar{\pi}}$ such that

$$\Phi^{\pi}_{\bar{\pi}}(s) - \Phi^{\pi'_i,\pi_{-i}}_{\bar{\pi}}(s) = V^{\pi_i,\pi_{-i}}_{i,\bar{\pi}}(s) - V^{\pi'_i,\pi_{-i}}_{i,\bar{\pi}}(s) \qquad (2)$$

for all policies $\pi_i, \pi'_i \in \Pi_i, \pi_{-i} \in \Pi_{-i}$, states s and agents $i \in \mathcal{N}$. We set $\Phi(\nu) = \mathbb{E}_{s\sim\nu}[\Phi(s)]$. In the case of a static environment, i.e. $r_{i,\pi'} = r_{i,\pi''}$ for any $i \in \mathcal{N}$ and $P_{\pi'} = P_{\pi''}$ for any $\pi', \pi'' \in \Pi$, then the *performative* MPG simplifies to the standard MPG framework, introduced in Leonardos et al. (2022).

3.3 Solution Concepts

We generalize the solution concepts of a performatively optimal policy and a performatively stable policy from single-agent PRL (Mandal et al., 2023). The following notion of a Nash equilibrium (NE) generalizes the concept of a performatively optimal policy.

Definition 1 (ϵ-NE)**.** *A policy profile* $\pi = (\pi_i)_{i\in\mathcal{N}} \in \Pi$ *is an* ϵ*-NE if it satisfies*

$$V^{(\pi_i,\pi_{-i})}_{i,(\pi_i,\pi_{-i})}(\rho) \geq V^{(\pi'_i,\pi_{-i})}_{i,(\pi'_i,\pi_{-i})}(\rho) - \epsilon,$$

for all $i \in \mathcal{N}$*, all* $\pi'_i \in \Delta(\mathcal{A}_i)^{\mathcal{S}}$*.*

If ϵ is implicit, we alternatively say that π is an approximate NE. If the definition holds for $\epsilon = 0$, then π is called an exact NE. We extend the notion of a performatively stable policy to the multi-agent setting through the notion of a performatively stable equilibrium (PSE), defined as follows.

Definition 2 (ϵ-PSE)**.** *A policy profile* $\pi = (\pi_i)_{i\in\mathcal{N}} \in \Pi$ *is an* ϵ*-Performatively Stable Equilibrium if it satisfies*

$$V^{(\pi_i,\pi_{-i})}_{i,\pi}(\rho) \geq V^{(\pi'_i,\pi_{-i})}_{i,\pi}(\rho) - \epsilon,$$

for any $i \in \mathcal{N}$ *and for all* $\pi'_i \in \Delta(\mathcal{A}_i)^{\mathcal{S}}$*.*

If the definition holds for $\epsilon = 0$, we say that π is an exact PSE. In the latter solution concept, i.e., PSE, we essentially uncouple the deviation in policy π from the given game $\mathcal{G}(\pi)$. Hence, we emphasize that NE and PNE are two distinct equilibrium concepts in general *performative* Markov Potential Games.

3.4 Learning Objective

At each iteration t of the learning protocol, the agents deploy a joint policy π^t, which induces the game $\mathcal{G}(\pi^t)$. Then, each agent updates its policy π^t_i to π^{t+1}_i based on its performance (e.g., the return) in $\mathcal{G}(\pi^t)$. Quantities relevant for updating π^t_i can be estimated from the data obtained when deploying π^t.

Given a sequence of policies $\pi^1, \ldots, \pi^T$, we introduce a new regret notion called *performative* regret (PReg). Formally,

$$\mathrm{PReg}(T) \coloneqq \frac{1}{T}\sum_{t=1}^{T} \max_{i\in\mathcal{N}} \max_{\pi'_i} \left(V^{\pi'_i,\pi^t_{-i}}_{i,\pi^t}(\rho) - V^{\pi^t}_{i,\pi^t}(\rho) \right).$$

Intuitively, this notion captures the worst-case suboptimality gap across agents, averaged over the time horizon T, while accounting for the environment's response. If we bound $\mathrm{PReg}(T) \leq \epsilon$, then there is at least one iteration $t \in [T]$ such that π^t is an ϵ-PSE. Notably, the average policy $\pi_{avg} = \frac{1}{T}\sum_{t=1}^{T}\pi_t$ does not necessarily correspond to an ϵ-PSE in general, as this depends on the structure of the game $\mathcal{G}(\pi_{avg})$.

4 CHARACTERIZATION OF SOLUTION CONCEPTS

In this chapter, we argue that a PSE exists in every MPG with *performative* effects. Furthermore, a PSE corresponds to an approximate NE. We require a sensitivity assumption that bounds the magnitude of the *performative* effects, which is a standard assumption in the *performative* prediction/reinforcement learning literature. Unlike prior work in PRL (Mandal et al., 2023; Rank et al., 2024), our assumption is made with respect to the policy space.

Assumption 1 (Sensitivity)**.** *For any two policies* π *and* π'*, we have that*

$$\|r_{i,\pi}(\cdot,\cdot) - r_{i,\pi'}(\cdot,\cdot)\|_2 \leq \omega_r \cdot \|\pi - \pi'\|_2,$$
$$\|P_{\pi}(\cdot \mid \cdot,\cdot) - P_{\pi'}(\cdot \mid \cdot,\cdot)\|_2 \leq \omega_p \cdot \|\pi - \pi'\|_2,$$

for all $i \in \mathcal{N}$*.*

At a high-level idea, we prove the existence of a PSE by using the fact that $\Phi^{\pi'}_{\pi}$ is continuous in π' for a fixed underlying game induced by π and that $\Phi^{\pi'}_{\pi}$ is continuous in π for a fixed policy π'. More specifically, we show that the function $\widehat{\Phi} = \arg\max_{\pi'\in\Pi}\Phi^{\pi'}_{\pi}(\rho)$ is upper hemicontinuous, which allows us to use the

Kakutani fixed point theorem (Glicksberg, 1952) to show the existence of a fixed point that corresponds to a PSE.

Lemma 1. *For any state distribution $\rho \in \Delta(S)$, there exists a policy $\pi^* \in \Pi$ such that*

$$\Phi_{\pi^*}^{\pi^*}(\rho) - \Phi_{\pi^*}^{\pi_i, \pi^*_{-i}}(\rho) \geq 0, \tag{3}$$

for all $i \in \mathcal{N}$, $\pi_i \in \Pi_i$.

The existence result follows by a simple argument over the potential functions, see Appendix A.1.

Theorem 1. *Under Assumption 1, every performative MPG admits a PSE.*

It is important to understand the behavior at a PSE. We know that under Assumption 1, the difference between two value functions under two different $\mathcal{G}(\pi')$ and $\mathcal{G}(\pi'')$ is bounded, as formalized in the following lemma. Its proof is provided in Appendix A.3.

Lemma 2. *Under Assumption 1, in every performative MG, we have*

$$\left|V_{i,\pi'}^{\pi}(s) - V_{i,\pi''}^{\pi}(s)\right| \leq \delta_{r,p} \cdot \|\pi' - \pi''\|_2,$$

for all $\pi', \pi'' \in \Pi$, where $\delta_{r,p} \coloneqq \frac{1}{1-\gamma}\left(\omega_r + \frac{\gamma \cdot \omega_p \sqrt{S}}{1-\gamma}\right)$.

This result allows us to show that even though PSE and NE may differ, every PSE corresponds to an approximate NE, with an approximation factor that vanishes as the *performative* effects approach zero. Hence, the two solution concepts coincide as $\omega_r, \omega_p \to 0$. We provide the proof in Appendix A.2.

Theorem 2. *Given Assumption 1 in a performative MPG, every PSE is a $\delta_{r,p}$-NE.*

The next sections focus on algorithms for computing a PSE.

5 INDEPENDENT GRADIENT METHODS

In this section, we study gradient-based methods and provide convergence guarantees. First, in Section 5.1, we assume that we have access to a gradient oracle. In Section 5.2, we relax the latter assumption.

We show that Lemma 2 is critical to generalize the theoretical guarantees for independent policy gradient ascent (IPGA), independent natural policy gradient ascent (INPG), and regularized INPG (INPG reg.) to *performative* MPGs. Specifically, we leverage Lemma 2 to generalize the potential improvement argument from prior work (Ding et al., 2022; Zhang et al., 2022; Alatur et al., 2024a) to account for *performative* effects. We refer to Appendix A for the complete proofs of this section.

5.1 Independent Gradient Methods with Oracle

Next, we present the independent gradient methods and show the corresponding convergence guarantees.

5.1.1 Independent Policy Gradient Ascent

First, we consider the IPGA introduced by Ding et al. (2022). Each agent i updates its policy at iteration t according to the following update rule

$$\begin{aligned}\pi_i^{t+1}(\cdot|s) = \underset{\pi_i(\cdot|s)\in\Delta(\mathcal{A}_i)}{\arg\max} &\Big\{\big\langle \pi_i(\cdot|s), \bar{Q}_{i,t}^t(s,\cdot)\big\rangle_{\mathcal{A}_i} \\ &- \frac{1}{2\eta} \cdot \|\pi_i(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2\Big\},\end{aligned} \tag{4}$$

where η is the learning rate and $\bar{Q}_{i,t}^t(s,a_i) = \sum_{a_{-i}} \pi_{-i}^t(a_{-i} \mid s) \cdot Q_{i,t}^t(s,a_i,a_{-i})$ is the averaged Q_i-value at step t. Let $C_\Phi \in \mathbb{R}$ with $C_\Phi \geq |\Phi_{\bar{\pi}}^{\pi}(\mu) - \Phi_{\bar{\pi}}^{\pi'}(\mu)|$ for any $\pi, \pi', \bar{\pi} \in \Pi, \mu \in \Delta(\mathcal{S})$. Such a constant C_Φ always exists and is trivially upper-bounded by $\frac{n}{1-\gamma}$ (Ding et al., 2022, Lemma 18).

Theorem 3. *For $\eta = \frac{(1-\gamma)^4}{8\min\{\kappa_\rho, S\}^3 nA}$, running the IPGA algorithm (4) satisfies*

$$\begin{aligned}\mathrm{PReg}(T) \lesssim& \frac{\min\{\kappa_\rho, S\}^2\sqrt{AnC_\Phi}}{\sqrt{T}\cdot(1-\gamma)^3} \\ &+ \frac{\sqrt{8\min\{\kappa_\rho, S\}^3 nA}}{(1-\gamma)^5} \cdot \sqrt{\frac{\mathcal{W}_{r,p}}{T}},\end{aligned}$$

where $\mathcal{W}_{r,p} \coloneqq T \cdot (n+1) \cdot n^2 \cdot S \cdot \delta_{r,p}$.

Thus, the IPGA algorithm converges to an approximate PSE in the best-iterate sense.

5.1.2 Independent Natural Policy Gradients

Here, we consider total potential functions of the form

$$\Phi_{\pi'}^{\pi}(\rho) = \mathbb{E}_{s^0\sim\rho}\left[\sum_{h=0}^{\infty} \gamma^h \phi_{\pi'}^{\pi}(s^h, a^h)\Big|\pi\right],$$

induced by functions $\phi_{\pi'}^{\pi} : \mathcal{S} \times \mathcal{A} \to [0,1]$ for $\pi, \pi' \in \Pi$. For every iteration t, INPG dynamics have the following multiplicative update rule under the softmax parameterization:

$$\pi_i^{t+1}(a_i|s) = \pi_i^t(a_i|s)\frac{\exp\left(\frac{\eta}{1-\gamma}\bar{A}_{i,t}^t(s,a_i)\right)}{Z_i^t(s)}, \tag{5}$$

for every $\forall i \in \mathcal{N}, s \in \mathcal{S}, a_i \in \mathcal{A}_i$, where $\bar{A}_{i,t}^t(s,a_i) = \sum_{a_{-i}} \pi_{-i}^t(a_{-i}|s)\left(Q_{i,t}^t(s,a) - V_{i,t}^t(s)\right)$ is the marginalized advantage function and $Z_i^t(s)$ is the normalization term given as

$$Z_i^t(s) = \sum_{a_i} \pi_i^t(a_i|s)\exp\left(\frac{\eta}{1-\gamma}\bar{A}_{i,t}^t(s,a_i)\right) \geq 1. \tag{6}$$

We now introduce the following standard assumption in the analysis of INPG dynamics.

Assumption 2. *For any initial state distribution ρ, $\inf_{\pi}\inf_{\pi'}\min_s d^{\pi'}_{\rho,\pi}(s) > 0$.*

Based on this assumption, for any initial distribution ρ, we also define

$$M_\rho := \sup_{\pi}\sup_{\pi'}\max_{s}\frac{1}{d^{\pi'}_{\rho,\pi}(s)} < \infty,$$

and we drop the dependence on ρ when it is clear from the context. Moreover, we denote the lower bound for the probability of playing the optimal action by

$$c := \min_i \min_t \min_s \sum_{a_i^*\in\arg\max_{a_i\in A_i}\bar{Q}^{\pi^t}_{i,t}(s,a_i)} \pi^t_i(a^*_i|s) > 0.$$

Unregularized INPG Dynamics In contrast to IPGA, the regret introduced by the *performative* effects vanish as $T \to \infty$ under INPG dynamics.

Theorem 4. *Suppose that Assumption 1 and Assumption 2 hold. For $T \geq 1$ and $\eta \leq \frac{(1-\gamma)^2}{\sqrt{n}} + \frac{\sqrt{2}(1-\gamma)}{M\delta_{r,p}}$, the INPG dynamics (5) satisfy*

$$\mathrm{PReg}(T) \leq \sqrt{\frac{1}{T}\frac{3\tilde{\kappa}\left(\frac{\sqrt{n}}{1-\gamma}+\sqrt{2}M\delta_{r,p}\right)}{c(1-\gamma)^3}}.$$

Hence, INPG dynamics converge to an approximate-PSE in the best-iterate sense. In order to show asymptotic last-iterate convergence of INPG dynamics, we require the following standard assumption for the analysis of the INPG dynamics.

Assumption 3. *The stationary points of Eq. (5) are isolated.*

Given Assumption 3, we obtain that the INPG dynamics converges to an exact PSE in the last-iterate sense as $T \to \infty$. This holds because Lemma 8 and the boundedness of $\phi^{\pi}_{\pi'}$ implies $\lim_{t\to\infty} Z^t_i(s) = 1 \implies \lim_{t\to\infty}\pi^t(a_i|s)\bar{A}^t_{i,t}(s,a_i) = 0$, in which case the gradient norm of the potential function with respect to the policy approaches 0 in the limit. Along with Assumption 3, this observation implies that the sequence of policies π^t converges to some stationary policy π^∞. The rest of the proof follows the same lines as in Zhang et al. (2022, Section 12.0.2).

Regularized INPG Dynamics. Note that c can be arbitrarily small for bad initializations of the policy. This can slow down the convergence of INPG dynamics due to its $\frac{1}{c}$ dependence. We consider the INPG dynamics with log-barrier regularization to overcome this. Following Zhang et al. (2022), we define the regularized objective and potential function as follows:

$$\tilde{V}^{\pi'}_{i,\pi} = V^{\pi'}_{i,\pi}(\rho) + \lambda\sum_{s,a_i}\log\pi'_i(a_i|s),$$

$$\tilde{\Phi}^{\pi'}_{\pi}(\rho) = \Phi^{\pi'}_{\pi}(\rho) + \lambda\sum_i\sum_{s,a_i}\log\pi'_i(a_i|s).$$

We refer the reader to the work by Zhang et al. (2022) for further discussion on the choice of the regularizer. Under log-barrier regularization the INPG dynamics have the following update rule:

$$\pi^{t+1}_i(a_i|s) = \pi^t_i(a_i|s)\cdot\frac{\eta f^t_i(a_i|s)}{Z^t_i(s)}, \tag{7}$$

where $f^t_i(s,a_i)$ and $\tilde{Z}^t_i(s)$ are defined as

$$f^t_i(s,a_i) = \frac{1}{1-\gamma}\bar{A}^{\pi^t}_{i,t}(s,a_i) + \frac{\lambda}{d^t_{\rho,t}(s)\pi^t_i(a_i|s)} - \frac{\lambda\,|\mathcal{A}_i|}{d^t_{\rho,t}(s)},$$

$$\tilde{Z}^t_i(s) = \sum_{a_i}\pi^t_i(a_i|s)\exp\left(\eta f^t_i(s,a_i)\right).$$

Theorem 5. *Suppose that Assumption 1 and Assumption 2 hold. Then, for all $T \geq 1$, it holds that*

$$\text{Perform-Regret}(T) \leq \frac{9\sqrt{2}}{\eta\lambda(1-\gamma)T} + \lambda M A_{\max} + \eta n S\delta_{r,p}\left(\frac{1}{(1-\gamma)^2} + 4\lambda^2 A_{\max}M^2 + \frac{4\lambda M}{1-\gamma}\right),$$

Moreover, for any $\epsilon > 0$, set $\lambda = \frac{\epsilon}{3MA_{\max}}$ and

$$\eta = \min\Bigg\{\left(\frac{16\epsilon M}{3} + \frac{16M}{(1-\gamma)^2} + \frac{12nM}{(1-\gamma)^3}\right)^{-1}, \left(\frac{1}{(1-\gamma)^2} + \frac{4\epsilon^2}{9A_{\max}} + \frac{4\epsilon}{3A_{\max}(1-\gamma)^2}\right)^{-1} \times (3nS\delta_{r,p})^{-1}, \left(\frac{15}{(1-\gamma)^2} + 5\epsilon\right)^{-1}\Bigg\}.$$

Then, for $T \geq \mathcal{O}\left(\frac{nA_{\max}M^2}{\epsilon^2(1-\gamma)^4}\max\{1, S\delta_{r,p}\}\right)$, the regularized INPG dynamics (7) satisfy

$$\mathrm{PReg}(T) \leq \epsilon.$$

Note that the result no longer depends on the constant c, as highlighted in Zhang et al. (2022).

5.2 Independent Learning without Oracle

In this section, we remove the exact gradient requirement for the IPGA algorithm. In this case, the gradient can be estimated using offline data. At iteration t, we generate K trajectories by rolling out π^t in $\mathcal{G}(\pi^t)$ K times.

Each trajectory is a sequence of state-action-reward tuples $(\bar{s}^h, \bar{a}^h, \bar{r}^h)_{h=0}^{H-1}$ of length H, where $\bar{s}^0 \sim \rho$ and the H is the horizon. Horizon H is a random variable $H = \max_{i \in \mathcal{N}}(h_i + h_i')$, where h_i and h_i' are sampled from a geometric distribution with parameter $1-\gamma$. Time steps h_i to $h_i + h_i'$ are used for estimating Q-values of agent i. We follow the approach from Ding et al. (2022) to obtain an unbiased estimate of $\bar{Q}_{i,t}^t$, denoted as $R_{i,t}^k \coloneqq \sum_{h=h_i}^{h_i+h_i'-1} \bar{r}_i^h$. In the policy update phase, we obtain an estimate for $\bar{Q}_{i,t}^t$ by minimizing the expected regression loss:

$$\widehat{Q}_i^t(\cdot,\cdot) \approx \underset{\|\bar{Q}_{i,t}^t(\cdot,\cdot)\|_2 \leq \frac{\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}}{1-\gamma}}{\arg\min} \underbrace{\sum_{k=1}^{K} \left(R_{i,t}^k - \widehat{Q}_{i,t}^t(\bar{s}_i, \bar{a}_i^k)\right)^2}_{=:L_i^t(\widehat{Q}_{i,t}^t)}.$$

We assume that $\mathbb{E}\left[L_i^t(\widehat{Q}_{i,t}^t)\right] \leq \delta_{stat}$, where the expectation is taken over the randomness induced by constructing $\widehat{Q}_{i,t}^t$. As shown by Audibert and Catoni (2011), the statistical error can be bounded by $\delta_{stat} \in \mathcal{O}(\frac{S^2 A^2}{(1-\gamma)^4 \cdot K})$.

We take the estimated $\widehat{Q}$ to update the policy of each agent $i \in \mathcal{N}$:

$$\begin{aligned} \pi_i^{t+1}(\cdot|s) = \underset{\pi(\cdot|s) \in \Delta_\xi(\mathcal{A}_i)}{\arg\max} \Big\{ &\left\langle \pi_i(\cdot|s), \widehat{Q}_{i,t}^t(s,\cdot) \right\rangle_{\mathcal{A}_i} \\ &- \frac{1}{2\eta} \cdot \|\pi_i(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2 \Big\}. \end{aligned} \tag{8}$$

The projection is taken over $\Delta_\xi(\mathcal{A}_i) \coloneqq \{(1-\xi)\pi_i(\cdot \mid s) + \frac{\xi}{|\mathcal{A}_i|}, \forall \pi_i(\cdot|s)\}$, forcing every policy to establish a greedy-exploration policy with probability ξ.

We bound changes compared to Theorem 3, by taking account for estimation errors.

Theorem 6. *Running algorithm IPGA without gradient oracles (8) and setting $\eta = \frac{(1-\gamma)^4}{16\kappa_\rho^3 nA}$, we obtain,*

$$\mathbb{E}[\mathrm{PReg}(T)] \lesssim \mathcal{R}(\eta) + \left(\frac{\sqrt{A} \cdot (\kappa_\rho^2 \cdot n \cdot \delta_{\mathrm{stat}})^{1/3}}{(1-\gamma)^2}\right)$$

where we have that

$$\begin{aligned} \mathcal{R}(\eta) = &\frac{\min\{\kappa_\rho, S\}^2 \sqrt{AnC_\Phi}}{\sqrt{T} \cdot (1-\gamma)^3} \\ &+ \frac{\sqrt{8\min\{\kappa_\rho, S\}^3 nA\mathcal{W}_{r,p}}}{(1-\gamma)^5 \sqrt{T}} \end{aligned}$$

and $\mathcal{W}_{r,p} \coloneqq (n+1) \cdot n^2 \cdot S \cdot \delta_{r,p}$.

Extending the results of the INPG algorithm is not straightforward, even in standard MPG; we leave that for future work.

6 REPEATED OCCUPANCY MEASURE OPTIMIZATION

The results in the previous section do not provide *finite-time* last-iterate convergence guarantees. Instead, last-iterate convergence is established only for INPG, and this is an asymptotic convergence result. In this section, we study *finite-time* last-iterate convergence guarantees for a special case of *performative* MPGs. In particular, we focus on MPGs with *performative* effects and agent independent transitions.

Assumption 4. *For any MPG with performative effects $\mathcal{G}(\bar{\pi})$, it holds that for any $s, s' \in \mathcal{S}$ and any action $a \in \mathcal{A}$,*

$$P_{\bar{\pi}}(s' \mid s, a) = P_{\bar{\pi}}(s' \mid s).$$

As a method of interest, we consider the repeated optimization approach of Mandal et al. (2023), developed for single agent *performative* RL, and extend it to MPGs with *performative* effects. There are two critical aspects that we change in this method: i) our approach aims to optimize a surrogate objective instead of the regularized objective in Mandal et al. (2023), ii) our approach can be implemented using multi-agent optimization, where each agent independently optimizes a surrogate objective over the state-action occupancy measures. Note that this is in contrast to the previous sections, where agents were directly updating their policies. The policy of agent i can be obtained using the following expression:

$$\pi_i(a_i \mid s) \begin{cases} \frac{\mu_{i,\bar{\pi}}^\pi(s,a_i)}{\sum_{a_i} \mu_{i,\bar{\pi}}^\pi(s,a_i)} & \text{if } \sum_a \mu_{i,\bar{\pi}}^\pi(s,a_i) > 0, \\ \frac{1}{A} & \text{otherwise}, \end{cases} \tag{9}$$

where $\mu_{i,\bar{\pi}}^\pi$ is the state-action occupancy measure of agent $i \in \mathcal{N}$ over $\mathcal{S} \times \mathcal{A}_i$ for the joint policy π, given the underlying game $\mathcal{G}(\bar{\pi})$. We denote by D_i the set of feasible occupancy measures over $\mathcal{S} \times \mathcal{A}_i$ given game $\mathcal{G}(\bar{\pi})$. Furthermore, $\mu_{\bar{\pi}}^\pi = (\mu_{1,\bar{\pi}}^\pi, \ldots, \mu_{n,\bar{\pi}}^\pi)$ is the joint occupancy measure. When it is clear from context, we simplify the notation by writing $\mu_\pi^\pi = \mu$ and $\mu_{\pi'}^{\pi'} = \mu'$, respectively. Similarly, for each $i \in \mathcal{N}$, we write $\mu_{i,\pi}^\pi = \mu_i$, $\mu_{i,\pi'}^{\pi'} = \mu_i'$ for all $i \in \mathcal{N}$. More specifically, to update the current policy π^t, we aim to solve the following optimization problem:

$$\underset{\mu \in \mathcal{D}}{\arg\max} \left\langle \nabla_\pi \Phi_t^t(\rho), \mu \right\rangle - \frac{\lambda}{2} \|\mu\|_2^2, \tag{10}$$

where $\mathcal{D}$ is the set of valid occupancy measures in $\mathcal{G}(\pi^t)$. Here, the gradient of the potential is evaluated at the current policy and game $\mathcal{G}(\pi^t)$, specifically: $\nabla_\pi \Phi_t^t(\rho) = \nabla_\pi \Phi_{\pi'}^\pi(\rho)\big|_{\pi=\pi_t, \pi'=\pi_t}$. To obtain agents' policies from the solution to this optimization problem, we use the relation considered in Eq. (9). One can show that the optimization problem is feasible without knowledge of Φ because it holds that $\nabla_{\pi_i}\Phi^{\pi}_{\pi'}(\rho) = \nabla_{\pi_i}V^{\pi}_{i,\pi'}(\rho)$ for any $i \in \mathcal{N}$ and any π (Leonardos et al., 2022). However, to establish the convergence of repeatedly optimizing (10), we require the following sensitivity assumption over state-action occupancy measures and the following smoothness assumption on the gradients.

Assumption 5. *For any two policies π and π', we have that*

$$\begin{aligned}
&\textit{i.} && \|r_{i,\pi}(\cdot,\cdot) - r_{i,\pi'}(\cdot,\cdot)\|_2 \le \zeta_r \cdot \|\mu - \mu'\|_2,\\
&\textit{ii.} && \|P_{\pi}(\cdot \mid \cdot,\cdot) - P_{\pi'}(\cdot \mid \cdot,\cdot)\|_2 \le \zeta_p \cdot \|\mu - \mu'\|_2,\\
&\textit{iii.} && \|\nabla_{\pi}\Phi^{\pi}_{\pi'}(\rho) - \nabla_{\pi}\Phi^{\pi}_{\pi'}(\rho)\|_2 \le \beta \cdot \|\pi - \pi'\|_2,
\end{aligned}$$

for all $i \in \mathcal{N}$.

For general potential functions, it holds that $\beta \le \frac{2n\gamma A_{\max}}{(1-\gamma)^3}$, (Leonardos et al., 2022).

Under Assumption 5, repeatedly optimizing (10) results in an occupancy measure μ^{λ} associated with an approximate PSE π^{λ} for a sufficiently large value of λ. Now, instead of optimizing for the joint occupancy measure, we can consider the following optimization problems over individual occupancy measures, which allow agents to independently perform the policy update step:

$$\arg\max_{\mu_i \in \mathcal{D}_i} \left\langle \nabla_{\pi_i}\Phi^{t}_{t}(\rho), \mu_i \right\rangle - \frac{\lambda}{2}\|\mu_i\|_2^2. \tag{11}$$

To obtain agent i's policy from the solution of this optimization problem, we can use the relation (9). We can show that the obtained policies are the same as the ones obtained from the solution of (10) for the special class of *performative* MPGs, which we consider in this section. This holds because $\sum_a \mu^{\pi}_{\pi'}(s,a) = \sum_a \mu^{\bar{\pi}}_{\pi'}(s,a) =: \alpha_{\pi}(s)$ for a fixed game $\mathcal{G}(\pi)$ and any two policies $\pi, \bar{\pi}$, i.e., the visitation distribution is independent of the played policy. Combining this with the convergence of repeatedly optimizing (10), we obtain the following theorem.

Theorem 7. *For the choice of $\lambda \ge O\left((\zeta_p + \zeta_r) \cdot \frac{S^2\sqrt{n}\gamma A^{9/4}}{(1-\gamma)^6} + \frac{S^{3/2}\gamma A^{5/4}\beta}{(1-\gamma)^3 \min_s \alpha(s)}\right)$, let μ^{λ} be the fixed point of the objective in Eq. (10). If agents $i \in \mathcal{N}$ repeatedly optimize (11) for $T \ge 2(1-\mu)^{-1}\ln(2/\delta(1-\gamma))$ iterations, it holds that $\|\mu^T - \mu^{\lambda}\|_2 \le \delta$ and the performative gap is bounded by*

$$\begin{aligned}
&\max_{i\in\mathcal{N}}\max_{\pi'_i}\left(V^{\pi'_i,\pi^{(T)}_{-i}}_{i,\pi^T}(\rho) - V^{\pi^{(T)}}_{i,\pi^T}(\rho)\right)\\
&\le \frac{\kappa_\rho}{\min_s \alpha_\lambda(s)(1-\gamma)} \cdot \left(\sqrt{\max_i A_i} \cdot \delta + \frac{\lambda}{2(1-\gamma)}\right).
\end{aligned}$$

Proof sketch.

- Following the proof argument by Mandal et al. (2023), we adopt the dual perspective on Eq. (10), allowing us to analyze a strongly convex program with a fixed feasible region. However, we additionally need to show that $\nabla_{\bar{\pi}}\Phi^{\pi}_{\pi'}$ is Lipschitz continuous in the state action occupancy measure μ' (associated with policy π'). We obtain the Lipschitz constant that depends on the smoothness of the gradient and the sensitivity parameters.
- Translating back the dual solution, using strong duality, we show that μ^T converges to a fixed point μ^{λ}.
- Finally, we construct policy π^T from μ^T and combine a gradient domination argument with the fact that μ^{λ} maximizes expression in Eq. (10). □

This theorem shows the last-iterate convergence of the novel repeated optimization approach to an approximate PSE for sufficiently small ζ_p, ζ_r and β.

7 EXPERIMENTS

We study the empirical performance of IPGA and INPG on two MPGs from the literature – the safe-distancing game (Leonardos et al., 2022; Ding et al., 2022), and stochastic congestion games (Fox et al., 2022). The algorithms we evaluate are non-oracle gradient algorithms, meaning the gradient is estimated based on the rollouts of the policies. We replicate each experiment across 10 seeds for the safe-distancing game and 5 seeds for the stochastic congestion game. The plots report the mean and one standard deviation over these seeds. Full details of practical implementations and hyperparameters can be seen in Appendix B.1.

Environments. We take the safe-distancing game from Leonardos et al. (2022) and extend this environment to adapt to *performative* effects: the agents' actions are replaced with probability α with those sampled from a perturbed environment's optimal Q-values, similar to Mandal et al. (2023). More details are presented in Appendix B.2.1, with an illustration of this environment in Figure 8. Additionally, we take the stochastic congestion game (Fox et al., 2022) and modify it to have a *performative* response. More precisely, the environment transition probabilities and reward function change according to the deployed policy based on the sensitivity Assumption 1: upon playing policy π' the transition probabilities become $P_{i,\pi'} = P_{i,\pi_0} + \frac{\omega_p}{(1-\gamma)\sqrt{|S|\,|\mathcal{A}_i|}}(\pi' - \pi_0)\frac{1}{|S|}$, and the rewards $r_{i,\pi'} = r_{i,\pi_0} + \frac{\omega_r}{(1-\gamma)\sqrt{|S|\,|\mathcal{A}_i|}}(\pi' - \pi_0)$, where ω_p and ω_r are the sensitivity parameters indicating

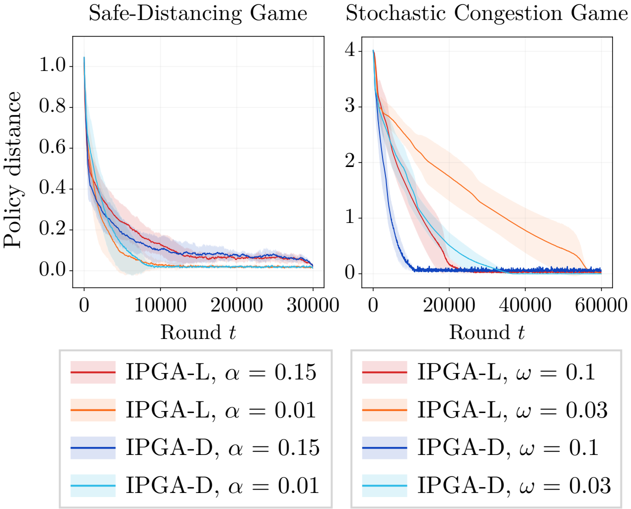

Figure 1: Comparison of IPGA-L and IPGA-D, showing the distance from the current policy to the average of the last 10 in that run: $\frac{1}{n}\sum_i^n \left\|\pi_i^t - \pi_i^{\text{last}}\right\|$, $\gamma = 0.99$. **Left**: IPGA-L $\eta = 0.00001$, IPGA-D $\eta = 0.0001$. **Right**: $\eta = 0.0003$.

the strength of the *performative* effects. In our experiments, we use the same value $\omega = \omega_p = \omega_r$ for both parameters. More details are presented in Appendix B.2.2, with an illustration of this environment in Figure 9.

Results. We start by showing how the *performative* effect influences convergence for IPGA in Figure 1. For completeness, we follow Ding et al. (2022) and include both the version of the algorithm we analyze (IPGA-D), and the variant proposed by Leonardos et al. (2022, IPGA-L). Generally, we notice that stronger *performative* effects, i.e., increasing α in the safe-distancing game or increasing ω in the stochastic congestion game, require a larger amount of iterations until convergence is observed. Additionally, in the safe-distancing game, higher *performative* strengths (e.g., $\alpha = 0.15$) lead to inexact convergence – the algorithms converge to a region of stable equilibria, exhibiting possible oscillations inside that region.

In the stochastic congestion game, we observe that IPGA-D converges faster than IPGA-L. A further study of the choice of the learning rate η can be seen in the appendix, Figure 7, which shows the performance of both methods.

We now shift our focus to the algorithms of interest that we analyze. A comparison of IPGA-D and both versions of INPG under *performative* effects can be seen in Figure 2. Typically, INPG converges faster than IPGA algorithms, and this is observed in our experiments as well. This reflects our theoretical guarantees, which indicate a better performance for natural policy gradient methods. Additionally, we observe that INPG (reg.) takes less rounds than INPG (unreg.) to converge, even under strong *performative* effects. However, in some cases the log-barrier regularization loses this advantage by having a large *performative* gap (e.g., Figure 2 right). We present more thorough experiments varying the learning rate η and performativity strength α, ω in Appendix B.1.

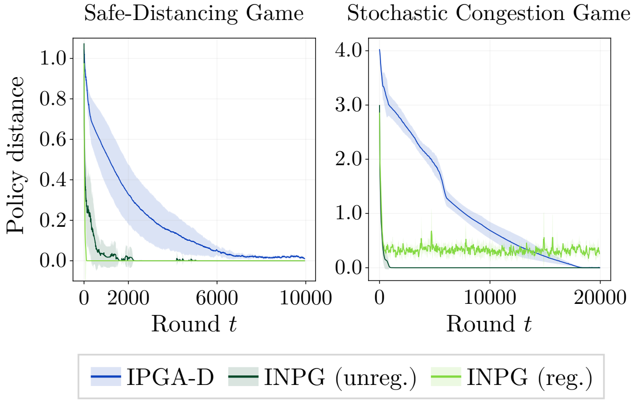

Figure 2: Comparison of IPGA-D, INPG (unreg.), INPG (reg.), showing the distance from the current policy to the average of the last 10 in that run: $\frac{1}{N}\sum_i^N \left\|\pi_i^t - \pi_i^{\text{last}}\right\|$, $\gamma = 0.99$. **Left**: $\alpha = 0.01$, $\eta = 0.0001$. **Right**: $\omega = 0.03$, $\eta = 0.0006$.

8 CONCLUSION

We provided a theoretical treatment of independent learning algorithms in performative Markov Potential Games. Our results establish best-iterate and asymptotic last-iterate convergence for different independent policy gradient algorithms, which showcases their robustness to performative effects. We further show that it is possible to obtain finite-time last-iterate convergence results for a special class of MPGs. It remains an open question whether our finite-time last-iterate convergence can be extended to the general case. For that, an interesting avenue may be to study weaker equilibria concepts.

Acknowledgements

This research was, in part, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 467367360. Additionally, this work was supported from the DFG under project number 514505843.

References

Alatur, P., Barakat, A., and He, N. (2024a). Independent policy mirror descent for markov potential games: Scaling to large number of players. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*.

Alatur, P., Ramponi, G., He, N., and Krause, A. (2024b). Provably learning Nash policies in constrained Markov potential games. In *AAMAS 2024*.

Anagnostides, I., Panageas, I., Farina, G., and Sandholm, T. (2023). On the convergence of no-regret learning dynamics in time-varying games. In *NeurIPS 2023*.

Audibert, J. and Catoni, O. (2011). Robust linear least squares regression. *Annals of Statistics*, 39(5).

Brown, G., Hod, S., and Kalemaj, I. (2022). Performative prediction in a stateful world. In *AISTATS 2022*.

Brown, N. and Sandholm, T. (2019). Superhuman AI for multiplayer poker. *Science*, 365(6456).

Cai, S., Han, F., and Cao, X. (2024). Performative control for linear dynamical systems. In *NeurIPS 2024*.

Cardoso, A. R., Abernethy, J. D., Wang, H., and Xu, H. (2019). Competing against nash equilibria in adversarially changing zero-sum games. In *ICML 2019*.

Ding, D., Wei, C., Zhang, K., and Jovanovic, M. R. (2022). Independent policy gradient for large-scale Markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *ICML 2022*.

Duvocelle, B., Mertikopoulos, P., Staudigl, M., and Vermeulen, D. (2023). Multiagent online learning in time-varying games. *Mathematics of Operations Research*, 48(2).

Fox, R., McAleer, S. M., Overman, W., and Panageas, I. (2022). Independent natural policy gradient always converges in Markov potential games. In *AISTATS 2022*.

Glicksberg, I. L. (1952). A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points. *Proceedings of the American Mathematical Society*, 3(1).

Guo, X., Li, X., Maheshwari, C., Sastry, S., and Wu, M. (2024). Markov α-potential games. *CoRR*, abs/2305.12553.

Hardt, M., Jagadeesan, M., and Mendler-Dünner, C. (2022). Performative power. In *NeurIPS 2022*.

Hardt, M. and Mendler-Dünner, C. (2023). Performative prediction: Past and future. *CoRR*, abs/2310.16608.

Izzo, Z., Zou, J., and Ying, L. (2022). How to learn when data gradually reacts to your model. In *AISTATS 2022*.

Jagadeesan, M., Zrnic, T., and Mendler-Dünner, C. (2022). Regret minimization with performative feedback. In *ICML 2022*.

Jiang, H., Cui, Q., Xiong, Z., Fazel, M., and Du, S. S. (2024). A black-box approach for non-stationary multi-agent reinforcement learning. In *ICLR 2024*.

Jordan, P., Barakat, A., and He, N. (2024). Independent learning in constrained Markov potential games. In *AISTATS 2024*.

Kulynych, B. (2022). Causal prediction can induce performative stability. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*.

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

Li, Q. and Wai, H. (2022). State dependent performative prediction with stochastic approximation. In *AISTATS 2022*.

Li, Q., Yau, C.-Y., and Wai, H.-T. (2022). Multi-agent performative prediction with greedy deployment and consensus seeking agents. In *NeurIPS 2022*.

Maheshwari, C., Wu, M., Pai, D., and Sastry, S. (2024). Independent and decentralized learning in markov potential games. *CoRR*, abs/2205.14590.

Mandal, D. and Radanovic, G. (2025). Performative reinforcement learning with linear markov decision processes. In *AISTATS 2025*.

Mandal, D., Triantafyllou, S., and Radanovic, G. (2023). Performative reinforcement learning. In *ICML 2023*.

Mao, W., Yang, L., Zhang, K., and Basar, T. (2022). On improving model-free algorithms for decentralized multi-agent reinforcement learning. In *ICML 2022*.

Mendler-Dünner, C., Ding, F., and Wang, Y. (2022). Anticipating performativity by predicting from predictions. In *NeurIPS 2022*.

Mendler-Dünner, C., Perdomo, J., Zrnic, T., and Hardt, M. (2020). Stochastic optimization for performative prediction. In *NeurIPS 2020*.

Mofakhami, M., Mitliagkas, I., and Gidel, G. (2023). Performative prediction with neural networks. In *AISTATS 2023*.

Narang, A., Faulkner, E., Drusvyatskiy, D., Fazel, M., and Ratliff, L. J. (2023). Multiplayer performative prediction: Learning in decision-dependent games. *Journal of Machine Learning Research*, 24(202).

Perdomo, J. C., Zrnic, T., Mendler-Dünner, C., and Hardt, M. (2020). Performative prediction. In *ICML 2020*.

Piliouras, G. and Yu, F.-Y. (2023). Multi-agent performative prediction: From global stability and optimality to chaos. In *EC 2023*.

Pollatos, V., Mandal, D., and Radanovic, G. (2025). On corruption-robustness in performative reinforcement learning. In *AAAI 25*.

Rank, B., Triantafyllou, S., Mandal, D., and Radanovic, G. (2024). Performative reinforcement learning in gradually shifting environments. In *UAI 2024*.

Ray, M., Ratliff, L. J., Drusvyatskiy, D., and Fazel, M. (2022). Decision-dependent risk minimization in geometrically decaying dynamic environments. In *AAAI 2022*.

Shapley, L. S. (1953). Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10).

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T. P., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nat.*, 550(7676).

Sun, Y., Liu, T., Zhou, R., Kumar, P. R., and Shahrampour, S. (2023). Provably fast convergence of independent natural policy gradient for Markov potential games. In *NeurIPS 2023*.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. In *NeurIPS 1999*.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. (2019). Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782).

Zhang, R., Mei, J., Dai, B., Schuurmans, D., and Li, N. (2022). On the global convergence rates of decentralized softmax gradient play in Markov potential games. In *NeurIPS 2022*.

Zhou, Z., Chen, Z., Lin, Y., and Wierman, A. (2023). Convergence rates for localized actor-critic in networked Markov potential games. In *UAI 2023*.

CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] URL: github.com/PauliusSasnauskas/performative-mpgs
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] We replicate the experiments on 5 different seeds for the safe-distancing game and 10 different seeds for the stochastic congestion game, and show the standard deviation across them.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator if your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Independent Learning in Performative Markov Potential Games Supplementary Materials

APPENDICES

A PROOFS AND DERIVATIONS

A.1 Existence of PSE

Proof of Lemma 1. We show that the function $\widehat{\Phi}(\pi) = \arg\max_{\pi' \in \Pi} \Phi_\pi^{\pi'}(\rho)$ has a fixed point. The idea is to show that $\widehat{\Phi}(\cdot)$ is hemicontinuous, so that a Kakutani fixed point theorem is applicable. Observe that $\widehat{\Phi}(\pi)$ is non-empty, as any policy π' can be deployed in the MPG $\mathcal{M}(\pi)$, so that the potential $\Phi_\pi^{\pi'}(\mu)$ is well-defined. Further, note that $\Phi_\pi^{\pi'}(\mu)$ has a global maximum with respect to parameter π', given a fixed π. This holds, because the set Π is compact and the function is continuous in π'. The latter holds, because for any agent $i \in \mathcal{N}$ and any deviation $\widehat{\pi}_i \in \Pi_i$, it holds that

$$\Phi_\pi^{\pi'}(\mu) - \Phi_\pi^{\widehat{\pi}_i, \pi'_{-i}}(\mu) = V_{i,\pi}^{\pi'}(\mu) - V_{i,\pi}^{\widehat{\pi}_i, \pi'_{-i}}(\mu)$$

and the value function $V_{i,\pi}^{\pi'}$ is continuous in π' for any $i \in \mathcal{N}$. From that, one can infer that $\Phi_\pi^{\pi'}$ is continuous in π' because we have for any $\widehat{\pi} \in \Pi$,

$$\Phi_\pi^{\pi'}(\mu) - \Phi_\pi^{\widehat{\pi}}(\mu) = \left(V_1^{\pi'}(\mu) - V_{1,\pi}^{\widehat{\pi}_1, \pi'_{-1}}(\mu)\right) + \left(V_{2,\pi}^{\widehat{\pi}_1, \pi'_{-1}} - V_{2,\pi}^{\widehat{\pi}_{\{1,2\}}, \pi'_{-\{1,2\}}}\right) + \cdots + \left(V_{n,\pi}^{\pi'_n, \widehat{\pi}_{-n}} - V_{n,\pi}^{\widehat{\pi}}(\mu)\right).$$

Further, one observes that $\Phi_\pi^{\pi'}$ is continuous in π due to (ω_r, ω_p)-sensitivity. So, we have the ingredients to apply Berge's maximum theorem. Finally, we can apply Berge's maximum theorem, this implies that $\widehat{\Phi}(\cdot)$ is upper hemicontinuous. We conclude that $\widehat{\Phi}$ has a fixed point π^*, by applying the Kakutani fixed point theorem (Glicksberg, 1952). Hence, it holds

$$\pi^* \in \arg\max_{\pi} \Phi_{\pi^*}^{\pi}(\mu).$$

We conclude that, $\Phi_{\pi^*}^{\pi^*}(\mu) \geq \Phi_{\pi^*}^{\pi_i, \pi^*_{-i}}(\mu)$ for any π_i. □

Using that auxiliary Lemma, we can show the existence of the solution concepts.

Proof of Theorem 1. Combining the definition in Eq. (2) and Lemma 1, the potential function Φ_*^* satisfies

$$\begin{aligned} 0 &\leq \Phi_{\pi^*}^{(\pi_i^*, \pi^*_{-i})}(\rho) - \Phi_{\pi^*}^{(\pi_i, \pi^*_{-i})}(\rho) \\ &= V_{i,\pi^*}^{(\pi_i^*, \pi^*_{-i})}(\rho) - V_{i,\pi^*}^{(\pi_i, \pi^*_{-i})}(\rho) \end{aligned}$$

for all policies π_i for all $i \in \mathcal{N}$. Therefore, we conclude that π^* is a performatively stable policy. □

A.2 Every PSE Is an Approximate NE

Proof. Suppose that π^* is a PSE. Then, for any agent $i \in \mathcal{N}$ and policy π_i,

$$V_{i,(\pi_i,\pi^*_{-i})}^{(\pi_i,\pi^*_{-i})}(\rho) - V_{i,\pi^*}^{\pi^*}(\rho) = V_{i,(\pi_i,\pi^*_{-i})}^{(\pi_i,\pi^*_{-i})}(\rho) - V_{i,\pi^*}^{(\pi_i,\pi^*_{-i})}(\rho) + V_{i,\pi^*}^{(\pi_i,\pi^*_{-i})}(\rho) - V_{i,\pi^*}^{\pi^*}(\rho) \leq \frac{1}{1-\gamma}\left(\omega_r + \frac{\gamma \omega_p \sqrt{S}}{1-\gamma}\right),$$

where the inequality follows from Definition 2 and Lemma 2. □

A.3 Bounding the Additional Costs Due to Performative Effects

We bound the sensitivity term, which occurs in our main analysis, Lemma 5. This boils down to bounding $|V_{i,x}^{\pi} - V_{i,x'}^{\pi}|$ for any two policies $x, x' \in \Pi$.

Proof of Lemma 2. First, for any $s \in \mathcal{S}$, it holds that for any two policies $x, x' \in \Pi$

$$V_{i,x}^{\pi}(s) = \sum_a \pi(a|s) \cdot Q_{i,x}^{\pi}(s,a).$$

Hence, we can bound

$$|V^\pi_{i,x}(s) - V^\pi_{i,x'}(s)| \le \max_{s,a}\left|Q^\pi_{i,x}(a,s) - Q^\pi_{i,x'}(s,a)\right|.$$

Rewriting over the Bellman equation, we have that for any $x \in \Pi$

$$Q^\pi_{i,x}(s,a) = r_{i,x}(s,a) + \gamma \sum_{s',a'} P_x(s'|s,a)\pi(a'|s')Q^\pi_{i,x}(s',a').$$

Exploiting that, We derive that for any $s \in \mathcal{S}, a \in \mathcal{A}$

$$\begin{aligned}
&\left|Q^\pi_{i,x}(s,a) - Q^\pi_{i,x'}(s,a)\right| \\
&\le |r_{i,x}(s,a) - r_{i,x'}(s,a)| + \gamma\left|\sum_{s',a'}\left(P_x(s'|s,a) - P_{x'}(s'|s,a)\right)\pi(a'|s')Q^\pi_{i,x}(s',a')\right| \\
&\quad + \gamma\left|\sum_{s',a'} P_{x'}(s'|s,a)\pi(a'|s')\left(Q^\pi_{i,x}(s',a') - Q^\pi_{i,x'}(s',a')\right)\right| \\
&\le |r_{i,x}(s,a) - r_{i,x'}(s,a)| + \frac{\gamma}{1-\gamma}\|P(\cdot|s,a) - P(\cdot|s,a)\|_1 + \gamma \max_{s',a'}\left|Q^\pi_x(s',a') - Q^\pi_{x'}(s',a')\right|.
\end{aligned}$$

Taking the maximum over s, a on the left-hand side implies that

$$\max_{s,a}\left|Q^\pi_{i,x}(s,a) - Q^\pi_{i,x'}(s,a)\right| \le \frac{1}{1-\gamma}\max_{s,a}|r_{i,x}(s,a) - r_{i,x'}(s,a)| + \frac{\gamma}{(1-\gamma)^2}\max_{s,a}\|P_x(\cdot|s,a) - P_{x'}(\cdot|s,a)\|_1.$$

By Assumption 1, we obtain that

$$\max_{s,a}|r_{i,x}(s,a) - r_{i,x'}(s,a)| = \|r_{i,x}(\cdot,\cdot) - r_{i,x'}(\cdot,\cdot)\|_\infty \le \|r_{i,x}(\cdot,\cdot) - r_{i,x'}(\cdot,\cdot)\|_2 \le \omega_r \cdot \|x - x'\|_2,$$

and that

$$\begin{aligned}
\max_{s,a}\|P_x(\cdot|s,a) - P_{x'}(\cdot|s,a)\|_1 &\le \sqrt{S}\cdot\max_{s,a}\|P_x(\cdot|s,a) - P_{x'}(\cdot|s,a)\|_2 \\
&\le \sqrt{S}\cdot\|P_x(\cdot|\cdot,\cdot) - P_{x'}(\cdot|\cdot,\cdot)\|_2 \\
&\le \sqrt{S}\cdot\omega_p\cdot\|x - x'\|_2.
\end{aligned}$$

In total, we obtain,

$$\begin{aligned}
\left|V^\pi_{i,x}(s) - V^\pi_{i,x'}(s)\right| &\le \frac{1}{1-\gamma}\|r_{i,x} - r_{i,x'}\|_\infty + \frac{\gamma}{(1-\gamma)^2}\max_{s,a}\|P_x(\cdot|s,a) - P_{x'}(\cdot|s,a)\|_1 \\
&\le \frac{1}{1-\gamma}\omega_r\cdot\|x - x'\|_2 + \frac{\gamma}{(1-\gamma)^2}\omega_p\cdot\sqrt{S}\cdot\|x - x'\|_2 \\
&\le \frac{1}{1-\gamma}\cdot\left(\omega_r + \frac{\gamma\cdot\omega_p\sqrt{S}}{1-\gamma}\right)\cdot\|x - x'\|_2.
\end{aligned}$$

□

A.4 Infinite Sample Case PGA – Proof of Theorem 3

We aim to bound the Nash regret. By the standard analysis of Ding et al. (2022, Theorem 1),

$$\begin{aligned}
&\sum_{t=1}^{T}\max_i\left(\max_{\pi'_i} V_{i,t}^{\pi'_i,\pi^t_{-i}}(\rho) - V_{i,t}^{\pi^t}(\rho)\right) \\
&\overset{(a)}{=} \frac{1}{1-\gamma}\sum_{t=1}^{T}\max_{\pi'_i}\sum_{s,a_i} d_{\rho,t}^{\pi'_i,\pi_{-i}}(s)\left(\pi'_i(a_i|s) - \pi_i^{(t)}(a_i|s)\right)\bar{Q}^t_{i,t}(s,a_i) \\
&\overset{(b)}{\le} \frac{3}{\eta(1-\gamma)}\sum_{t=1}^{T}\sum_{s} d_{\rho,t}^{\pi'_i,\pi^t_{-i}}(s)\cdot\left\|\pi_i^{t+1}(\cdot\mid s)) - \pi_i^t(\cdot\mid s)\right\|_2
\end{aligned}$$

$$\begin{aligned}
&\overset{(c)}{\leq} \frac{\sqrt{\sup_{t,\pi}\|d^{\pi}_{\rho,t}/\nu\|_\infty}}{\eta(1-\gamma)^{3/2}} \sum_{t=1}^{T}\sum_{s}\sqrt{d^{\pi'_i,\pi^t_{-i}}_{\rho,t}\cdot d^{\pi_i^{t+1},\pi^t_{-i}}_{\nu,t}}\cdot\left\|\pi_i^{t+1}(\cdot|s)-\pi_i^t(\cdot|s)\right\|_2 \\
&\overset{(d)}{\leq} \frac{\sqrt{\sup_{t,\pi}\|d^{\pi}_{\rho,t}/\nu\|_\infty}}{\eta(1-\gamma)^{3/2}} \sqrt{\sum_{t=1}^{T}\sum_{s} d^{\pi_i^{t+1},\pi^t_{-i}}_{\rho,t}(s)}\cdot\sqrt{\sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{s} d^{\pi_i^{t+1},\pi^{(t)}_{-i}}_{\nu,t}(s)\cdot\|\pi_i^{t+1}(\cdot|s)-\pi_i^t(\cdot|s))\|_2^2}, \\
&\overset{(e)}{\leq} \frac{\sqrt{\sup_{t,\pi}\|d^{\pi}_{\rho,t}/\nu\|_\infty}}{\eta(1-\gamma)^{3/2}}\cdot\sqrt{T}\cdot\sqrt{\sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{s} d^{\pi_i^{t+1},\pi^t_{-i}}_{v,t}(s)\cdot\|\pi_i^{t+1}(\cdot|s)-\pi_i^t(\cdot|s)\|^2},
\end{aligned}$$

where we have $\widetilde{\kappa}_\rho = \sup_t \inf_{v\in\Delta(\mathcal{S})}\sup_{\pi\in\Pi}\|d^{\pi}_{\rho,t}/v\|_\infty$. Note, that (a) follows by Lemma 3 and by abusing the notation i to represent $\arg\max_{\pi'_i}$. In (b), we use the observation by Ding et al. (2022) that by the optimality of π_i^{t+1}, it holds that

$$\left\langle \pi'_i(\cdot|s)-\pi_i^{t+1}(\cdot|s), \eta\bar{Q}^t_{i,t}(s,\cdot)-\pi_i^{t+1}(\cdot|s)+\pi_i^t(\cdot|s)\right\rangle_{\mathcal{A}_i}\leq 0 \quad \text{for any } \pi'_i\in\Pi_i,$$

which implies that,

$$\begin{aligned}
&\left\langle \pi'_i(\cdot|s)-\pi_i^t(\cdot|s),\bar{Q}^t_i(s,\cdot)\right\rangle_{\mathcal{A}_i} \\
&\leq \left\langle \pi'_i(\cdot|s)-\pi_i^{t+1}(\cdot|s),\pi_i^{t+1}(\cdot|s)-\pi_i^t(\cdot|s)\right\rangle_{\mathcal{A}_i}+\left\langle \pi_i^{t+1}(\cdot|s)-\pi_i^t(\cdot|s),\bar{Q}^t_i(s,\cdot)\right\rangle_{\mathcal{A}_i} \\
&\leq \frac{2}{\eta}\left\|\pi_i^{t+1}(\cdot|s)-\pi_i^t(\cdot|s)\right\|_2+\left\|\pi_i^{t+1}(\cdot|s)-\pi_i^t(\cdot|s)\right\|_2\left\|\bar{Q}^t_i(s,\cdot)\right\|_2 \\
&\leq \frac{3}{\eta}\left\|\pi_i^{t+1}(\cdot|s)-\pi_i^t(\cdot|s)\right\|_2,
\end{aligned}$$

where the line uses that $|\bar{Q}^t_{i,t}(s,\cdot)|_2\leq\frac{\sqrt{A}}{1-\gamma}$ and $\eta\leq\frac{1-\gamma}{\sqrt{A}}$. In (c), we choose an arbitrary $\nu\in\Delta(\mathcal{S})$ and exploit

$$\frac{d^{\pi'_i,\pi^t_{-i}}_{\rho,t}(s)}{d^{\pi_i^{t+1},\pi^t_{-i}}_{\nu,t}(s)}\leq\frac{d^{\pi'_i,\pi^t_{-i}}_{\rho,t}(s)}{(1-\gamma)\nu(s)}\leq\frac{\sup_{\pi,t}\|d^{\pi}_{\rho}/\nu\|_\infty}{(1-\gamma)}.$$

In (d), we exploit the Cauchy-Schwarz inequality, and we sum over all agents to replace the i from the $\arg\max_i$ in (a), and in (e) we proceed with using that the second term is equal to $\sqrt{T}$.

We apply the first guarantee in Lemma 5 to obtain:

$$\begin{aligned}
&\sum_{t=1}^{T}\max_i\left(\max_{\pi'_i,\pi^t_{-i}}V^{\pi'_i,\pi^t_{-i}}_{i,t}(\rho)-V^{\pi^t}_{i,t}(\rho)\right) \\
&\leq\frac{\sqrt{\widetilde{\kappa}_\rho}}{\eta(1-\gamma)^{3/2}}\cdot\sqrt{T}\cdot\sqrt{2\eta(1-\gamma)\left(\Phi^{T+1}_{T+1}(v)-\Phi^1_1(v)\right)+\frac{8\eta^3A^2n^2}{(1-\gamma)^4}T+\mathcal{W}_{r,p}\cdot2\eta(1-\gamma)} \\
&\leq\sqrt{\frac{\widetilde{\kappa}_\rho\cdot T\cdot C_\Phi}{\eta(1-\gamma)^2}}+\frac{\sqrt{\widetilde{\kappa}_\rho}}{\eta(1-\gamma)^{3/2}}\cdot\sqrt{T}\cdot\sqrt{\frac{4\eta^3A^2n^2}{(1-\gamma)^4}T}+\frac{\sqrt{\widetilde{\kappa}_\rho}}{\eta(1-\gamma)^{3/2}}\cdot\sqrt{T}\cdot\sqrt{2\eta(1-\gamma)\cdot\mathcal{W}_{r,p}} \\
&\lesssim\sqrt{\frac{\widetilde{\kappa}_\rho TC_\phi}{\eta(1-\gamma)^2}}+\sqrt{\frac{\widetilde{\kappa}_\rho\eta T^2A^2n^2}{(1-\gamma)^7}}+\frac{\sqrt{\widetilde{\kappa}_\rho\cdot T\cdot\mathcal{W}_{r,p}}}{\sqrt{\eta}(1-\gamma)^3}
\end{aligned}$$

where we use $\mathcal{W}_{r,p}=\frac{n(n+1)}{2}\cdot\delta_{r,p}\cdot\|\pi^{t+1}-\pi^t\|_2$ and that $C_\Phi=\max_{t,\pi,\pi',\mu}|\Phi^{\pi}_t(\mu)-\Phi^{\pi'}_t(\mu)|$. We complete the first guarantee by taking step size $\eta=\frac{(1-\gamma)^{5/2}\sqrt{C_\Phi}}{nA\sqrt{T}}$.

For completing the guarantee, we set $\eta\leq\frac{(1-\gamma)^4}{8\min\{\kappa_\nu,S\}^3nA}$ and we apply Lemma 5, for a large enough constant c, we obtain

$$\sum_{t=1}^{T}\max_i\left(\max_{\pi'_i,\pi^t_{-i}}V^{\pi'_i,\pi^t_{-i}}_{i,t}(\rho)-V^{\pi^t}_{i,t}(\rho)\right)$$

$$
\begin{aligned}
&\leq \frac{\sqrt{\sup_{t,\pi}\|d^{\pi}_{\rho,t}/\nu\|_\infty}}{\eta(1-\gamma)^{3/2}} \cdot \sqrt{T} \cdot \sqrt{2\eta(1-\gamma)\big(\Phi^{T+1}_T(v) - \Phi^1_T(v)\big) + \frac{8\eta^3 A^2 n^2}{(1-\gamma)^4} + \mathcal{W}_{r,p} \cdot 2\eta(1-\gamma)} \\
&\leq \sqrt{\frac{\sup_{t,\pi}\|d^{\pi}_{\rho,t}/\nu\|_\infty T C_\phi}{\eta(1-\gamma)^2}} + \sqrt{\frac{(\sup_{t,\pi}\|d^{\pi}_{\rho,t}/\nu\|_\infty) T C_\Phi}{\eta(1-\gamma)^2}} \sqrt{\frac{8\eta(\sup_{t,\pi}\|d^{\pi}_{\rho,t}/\nu\|_\infty) A^2 n^2}{(1-\gamma)^4}} + \frac{\sqrt{\widetilde{\kappa}_\rho \cdot T \cdot \mathcal{W}_{r,p}}}{\sqrt{\eta}(1-\gamma)} \\
&\leq c \cdot \sqrt{\frac{T C_\Phi \cdot \min\left\{\kappa_\nu, S\right\}^4 nA}{(1-\gamma)^6}} + \frac{\sqrt{8\min\{\kappa_\rho, S\}^3 nA}}{(1-\gamma)^5} \cdot \sqrt{\mathcal{W}_{r,p} T},
\end{aligned}
$$

where we use in the last inequality the following special cases: if $\nu = \rho$, then we have that $\sup_{t,\pi}\|d^{\pi}_{\rho,t}/\nu\|_\infty \leq \kappa_\nu$ and the first two terms are bounded by $O\left(\sqrt{\frac{\kappa_\rho^4 nATC_\Phi}{(1-\gamma)^6}}\right)$ for the choice $\eta \leq \frac{(1-\gamma)^4}{8\kappa_\rho^3 nA}$. Next, if we simply choose ν as the uniform distribution, we have that $\kappa_\nu \leq S$, $\eta = \frac{(1-\gamma)^4}{8S^3 nA} \leq \frac{(1-\gamma)^4}{8\kappa_\nu^3 nA}$ is a valid choice, finalizing the guarantee depending on S. Finally, we use the upper bound, given by Lemma 2, to show that

$$
\begin{aligned}
\mathcal{W}_{r,p} &= \frac{n(n+1)}{1-\gamma} \cdot \left(\omega_r + \frac{\gamma \cdot \omega_p \sqrt{S}}{1-\gamma}\right) \sum_t \|\pi^{t+1} - \pi^t\|_2 \\
&\leq \frac{n(n+1)}{1-\gamma} \cdot \left(\omega_r + \frac{\gamma \cdot \omega_p \sqrt{S}}{1-\gamma}\right) \sum_t \|\pi^{t+1} - \pi^t\|_1 \\
&\leq \frac{n(n+1)}{1-\gamma} \cdot \left(\omega_r + \frac{\gamma \cdot \omega_p \sqrt{S}}{1-\gamma}\right) \left\|\sum_t \pi^{t+1} - \pi^t\right\|_1 \\
&\leq \frac{n(n+1)}{1-\gamma} \cdot \left(\omega_r + \frac{\gamma \cdot \omega_p \sqrt{S}}{1-\gamma}\right) \cdot nS.
\end{aligned}
$$

Lemma 3 (Ding et al. (2022, Lemma 1))**.** *Given an underlying MPG(π') and a policy $\pi = (\pi_i, \pi_{-i})$, for agent i and any state distribution μ, it holds for any two policies $\hat{\pi}_i$ and $\bar{\pi}_i$*

$$
V^{\hat{\pi}_i,\pi_{-i}}_{i,\pi'}(\mu) - V^{\bar{\pi}_i,\pi_{-i}}_{i,\pi'}(\mu) = \frac{1}{1-\gamma} \sum_{s,a_i} d^{\hat{\pi}_i,\pi_{-i}}_{\mu,\pi'}(s) \cdot \big(\hat{\pi}_i - \bar{\pi}_i\big)(a_i|s) \cdot \bar{Q}^{\bar{\pi}_i,\pi_{-i}}_{i,\pi'}(s,a_i),
$$

where we have $\bar{Q}^{\bar{\pi}_i,\pi_{-i}}_{i,\pi'}(s,a_i) = \sum_{a_{-i}} \pi_{-i}(a_{-i}|s) \cdot Q^{\hat{\pi},\pi_{-i}}_{i,\pi'}(s,a_i)$.

Proof. Follows by e.g., C.1 by Leonardos et al. (2022) for a fixed underlying MPG(π'). □

We denote by $i \sim j$ the set of indices $\{\ell \mid i < \ell < j\}$, following the notation by Ding et al. (2022).

Lemma 4 (Ding et al. (2022, Lemma 2))**.** *For any function $\Psi^\pi : \Pi \to \mathbb{R}$ and any two policies $\pi, \pi' \in \Pi$,*

$$
\begin{aligned}
\Psi^{\pi'} - \Psi^{\pi} = &\sum_{i=1}^{n} \left(\Psi^{\pi'_i,\pi_{-i}} - \Psi^{\pi}\right) \\
&+ \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left(\Psi^{\pi_{<i,i\sim j},\pi'_{>j},\pi'_i,\pi'_j} - \Psi^{\pi_{<i,i\sim j},\pi'_{>j},\pi_i,\pi'_j} - \Psi^{\pi_{<i,i\sim j},\pi'_{>j},\pi'_i,\pi_j} + \Psi^{\pi_{<i,i\sim j},\pi'_{>j},\pi_i,\pi_j}\right).
\end{aligned}
$$

Lemma 5 (Policy Improvement)**.** *For an MPG according to Definition 2, for any state distribution μ and two consecutive policies π^{t+1} and π^t generated by the PGA Algorithm (4), we have:*

$$
\begin{aligned}
\Phi^{t+1}_{t+1}(\mu) - \Phi^t_t(\mu) \geq &\frac{1}{2\eta(1-\gamma)} \sum_{i=1}^{n} \sum_s d^{\pi^{t+1}_i,\pi_{-i}}_{\mu,t}(s) \cdot \left(1 - \frac{4\eta\kappa_\mu^3 An}{(1-\gamma)^4}\right) \cdot \left\|\pi^{t+1}_i(\cdot|s) - \pi^t_i(\cdot|s)\right\|^2 \\
&- \frac{n(n+1)}{2} \cdot \delta_{r,p} \cdot \|\pi^{t+1} - \pi^t\|_2,
\end{aligned}
$$

where we have $\kappa_\mu = \sup_t \sup_{\pi\in\Pi}\|d^{\pi}_{\mu,t}/\mu\|_\infty$ and $\delta_{r,p} := \frac{1}{1-\gamma}\left(\omega_r + \frac{\gamma\cdot\omega_p\sqrt{S}}{1-\gamma}\right)$.

Proof. By Lemma 4 with $\Psi^{\pi} = \Phi_t^{\pi}(\mu)$ and abbreviating $\pi' = \pi^{(t+1)}$, $\pi = \pi^{(t)}$, we have that

$$\Phi_t^{t+1}(\mu) - \Phi_t^{t}(\mu) = \mathbf{Diff}_{\alpha} + \mathbf{Diff}_{\beta},$$

where

$$\mathbf{Diff}_{\alpha} \coloneqq \sum_{i=1}^{n} \Phi_t^{\pi'_i,\pi_{-i}}(\mu) - \Phi_t^{\pi}(\mu),$$
$$\mathbf{Diff}_{\beta} \coloneqq \sum_{i=1}^{n} \sum_{j=i+1}^{n} \Big(\Phi_t^{\pi_{<i,i\sim j},\pi'_{>j},\pi'_i,\pi'_j}(\mu) - \Phi_t^{\pi_{<i,i\sim j},\pi'_{>j},\pi_i,\pi'_j}(\mu) - \ \Phi_t^{\pi_{<i,i\sim j},\pi'_{>j},\pi'_i,\pi_j}(\mu) + \Phi_t^{\pi_{<i,i\sim j},\pi'_{>j},\pi_i,\pi_j}(\mu) \Big).$$

According to the analysis Ding et al. (2022, Lemma 3.(*ii*)), this implies that,

$$\Phi_t^{t+1}(\mu) - \Phi_t^{t}(\mu) \geq \frac{1}{2\eta(1-\gamma)} \sum_{i=1}^{n} \sum_{s} d_{\mu,t}^{\pi_i^{t+1},\pi_{-i}}(s) \cdot \left(1 - \frac{4\eta\kappa_\mu^3 A n}{(1-\gamma)^4}\right) \cdot \left\|\pi_i^{t+1}(\cdot|s) - \pi_i^{t}(\cdot|s)\right\|^2 \tag{12}$$

Further, we derive that,

$$\Phi_{t+1}^{t+1}(\mu) - \Phi_t^{t}(\mu) = \Phi_{t+1}^{t+1} - \Phi_t^{t+1} + \Phi_t^{t+1}(\mu) - \Phi_t^{t}(\mu) \geq -|\Phi_{t+1}^{t+1} - \Phi_t^{t+1}| + \Phi_t^{t+1}(\mu) - \Phi_t^{t}(\mu).$$

By applying lemma 2 to upper bound the first term and by applying equation 12, we obtain our result. □

A.5 IPGA – Without Gradient Oracle

Similar as in section A.4, we extend the techniques by Ding et al. (2022) to obtain finite sample guarantees for our version of MPGs under performativity.

We bound the maximum occurring deviations over all time steps. The proof is based on Ding et al. (2022, Theorem 3) incorporating the additional costs due to performative effects.

Proof. We bound the performative regret that PGA Algorithm (4) achieves in a MPG with performative effects. Recall that $\overline{Q}_{i,t}$ corresponds to the exact Q-value averaged over the policies of agents $-i$ under the performative effect induced by π^t, while $\widehat{Q}_{i,t}$ corresponds to the computed estimation, and exploration rate $\xi \leq \frac{1}{2}$. In total, we have

$$
\begin{aligned}
&\sum_{t=1}^{T} \max_{i} \left(\max_{\pi'_i} V_{i,t}^{\pi'_i,\pi^t_{-i}}(\rho) - V_{i,t}^{\pi^t}(\rho) \right) \\
&\overset{(a)}{=} \frac{1}{1-\gamma} \sum_{t=1}^{T} \max_{\pi'_i} \sum_{s,a_i} d_{\rho,t}^{\pi'_i,\pi_{-i}}(s) \left(\pi'_i(a_i|s) - \pi_i^{(t)}(\cdot|s) \right) \bar{Q}^t_{i,t}(s,a_i) \\
&\overset{(b)}{\leq} \frac{1}{\eta(1-\gamma)} \sum_{t=1}^{T} \sum_{s} d_{\rho,t}^{\pi'_i,\pi^t_{-i}}(s) \left\| \pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s) \right\|_2 + \frac{\eta T \xi \sqrt{A}}{(1-\gamma)^2} \\
&\quad + \frac{1}{1-\gamma} \sum_{t=1}^{T} \sum_{s} d_{\rho,t}^{\pi'_i,\pi^t_{-i}}(s) \left\langle \pi'_i(\cdot|s) - \pi_i^t(\cdot|s), \bar{Q}^t_{i,t}(s,\cdot) - \widehat{Q}^t_{i,t}(s,\cdot) \right\rangle_{\mathcal{A}_i} \\
&\overset{(c)}{\leq} \frac{\sqrt{\kappa_\rho}}{\eta(1-\gamma)^{3/2}} \sum_{t=1}^{T} \sum_{s} \sqrt{d_{\rho,t}^{\pi_i^{t+1},\pi^t_{-i}}(s) \cdot d_{\rho,t}^{\pi'_i,\pi^t_{-i}}(s)} \left\| \pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s) \right\|_2 + \frac{\xi T \sqrt{A}}{(1-\gamma)^2} \\
&\quad + \frac{\kappa_\rho}{1-\gamma} \left| \sum_{t=1}^{T} \sum_{s} d_{\rho,t}^{\pi^t}(s) \cdot \left\langle \pi'_i(\cdot|s) - \pi_i^t(\cdot|s), \bar{Q}^t_{i,t}(s,\cdot) - \widehat{Q}^t_{i,t}(s,\cdot) \right\rangle_{\mathcal{A}_i} \right| \\
&\overset{(d)}{\leq} \frac{\sqrt{\kappa_\rho}}{\eta(1-\gamma)^{3/2}} \sqrt{\sum_{t=1}^{T} \sum_{s} d_{\rho,t}^{\pi'_i,\pi^t_{-i}}(s)} \cdot \sqrt{\sum_{t=1}^{T} \sum_{s} d_{\rho,t}^{\pi_i^{t+1},\pi^t_{-i}}(s) \left\| \pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s) \right\|_2^2} \\
&\quad + \frac{\xi T \sqrt{A}}{(1-\gamma)^2} + \frac{\kappa_\rho}{1-\gamma} \sum_{t=1}^{T} \sqrt{\frac{A L_i^t(\widehat{Q}^t_{i,t})}{\xi}} \\
&\overset{(e)}{\leq} \frac{\sqrt{\kappa_\rho}}{\eta(1-\gamma)^{3/2}} \sqrt{\sum_{t=1}^{T} \sum_{s} d_{\rho,t}^{\pi'_i,\pi^t_{-i}}(s)} \cdot \sqrt{\sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{s} d_{\rho,t}^{\pi_i^{t+1},\pi^t_{-i}}(s) \left\| \pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s) \right\|_2^2} \\
&\quad + \frac{\xi T \sqrt{A}}{(1-\gamma)^2} + \frac{\kappa_\rho}{1-\gamma} \sum_{t=1}^{T} \sqrt{\frac{A L_i^t(\widehat{Q}^t_{i,t})}{\xi}},
\end{aligned}
\tag{13}
$$

where we use the multi-agent performance difference Lemma 3 for (a), inequality (b) follows by abusing notation following Ding et al. (2022): policy π'_i represents the $\arg\max_{\pi'_i}$ and i captures the $\arg\max_i$ and by using the following property of the algorithm, (Ding et al., 2022, Equation 24) for $\xi \leq \frac{1}{2}$ and $\eta \leq \frac{1-\gamma}{\sqrt{A}}$:

$$
\left\langle \pi'_i(\cdot|s) - \pi_i^t(\cdot|s), \eta \widehat{Q}_i^t(s,\cdot) \right\rangle_{\mathcal{A}_i} \lesssim \frac{1}{\eta} \left\| \pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s) \right\|_2 + \frac{\xi \sqrt{A}}{1-\gamma} + \left\langle \pi'_i(\cdot|s) - \pi^t(\cdot|s), \overline{Q}_i^t(s,\cdot) - \widehat{Q}_i^t(s,\cdot) \right\rangle_{\mathcal{A}_i}.
$$

(c) follows by the definition of the distribution mismatch coefficient:

$$
\frac{d_{\rho,t}^{\pi'_i,\pi^t_{-i}}(s)}{d_{\rho,t}^{\pi_i^{t+1},\pi^t_{-i}}(s)} \leq \frac{d_{\rho,t}^{\pi'_i,\pi^t_{-i}}(s)}{(1-\gamma)\rho(s)} \leq \frac{\kappa_\rho}{1-\gamma}.
$$

(d) follows by applying the Cauchy-Schwartz inequality, Jensen inequality, and that

$$
\left| \sum_{t=1}^{T} \sum_{s} d_{\rho,t}^{\pi^t}(s) \left\langle \pi'_i(\cdot|s) - \pi_i^t(\cdot|s), \bar{Q}^t_{i,t}(s,\cdot) - \widehat{Q}^t_{i,t}(s,\cdot) \right\rangle_{\mathcal{A}_i} \right| \leq \sqrt{\frac{A L_i^t(\widehat{Q}^t_{i,t})}{\xi}}.
$$

In (e), we simply replace the $\arg\max_i$ by the sum over all players. The remaining part follows by applying the modified policy improvement Lemma 6 that takes into account performative effects.

We define $\mathcal{W}_{r,p} = \frac{n(n+1)}{2} \cdot \delta_{r,p} \cdot \|\pi^{t+1} - \pi^t\|_2$ and that $C_\Phi = \max_{t,\pi,\pi',\mu} |\Phi_t^\pi(\mu) - \Phi_t^{\pi'}(\mu)|$, which describes the performative costs. To obtain the guarantee, we apply Lemma 6 to Eq. (13), which leads to

$$
\mathbb{E}\left[\sum_{t=1}^{T} \max_{i} \left(\max_{\pi'_i} V_{i,t}^{\pi'_i,\pi^t_{-i}}(\rho) - V_{i,t}^{\pi^t}(\rho) \right) \right]
$$

$$
\begin{aligned}
&\lesssim \frac{\sqrt{\kappa_\rho}}{\eta(1-\gamma)^{3/2}}\sqrt{T}\sqrt{\eta(1-\gamma)(\Phi^{T+1}_{T+1}(\mu)-\Phi^1_1(\mu))+\frac{\eta^2\kappa_\rho A}{(1-\gamma)\xi}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\Big[L_i^t(\widehat{Q}^t_{i,t})\Big]+\mathcal{W}_{r,p}\cdot\eta\cdot(1-\gamma)} \\
&\quad+\frac{\xi T\sqrt{A}}{(1-\gamma)^2}+\frac{\kappa_\rho}{1-\gamma}\sum_{t=1}^{T}\sqrt{\frac{A\mathbb{E}\Big[L_i^t(\widehat{Q}^t_{i,t})\Big]}{\xi}} \\
&\lesssim \sqrt{\frac{\kappa_\rho T C_\Phi}{\eta(1-\gamma)^2}}+\frac{\kappa_\rho T}{(1-\gamma)^2}\sqrt{\frac{A\cdot n\cdot\delta_{stat}}{\xi}}+\frac{\xi T\sqrt{A}}{(1-\gamma)^2}+\frac{\sqrt{\kappa_\rho\cdot T\cdot\mathcal{W}_{r,p}}}{\sqrt{\eta}(1-\gamma)},
\end{aligned}
$$

where in the last line, we use that $C_\phi \geq \Phi^{T+1}_{T+1}-\Phi^1_1$, and that $\mathbb{E}[L_i^t(\widehat{Q}^t_{i,t})] \leq \delta_{\text{stat}}$. Finally, the guarantee follows by the choice $\eta = \frac{(1-\gamma)^4}{16\kappa_\rho^3 nA}$, and $\xi \leq \left(\kappa_\rho^2\cdot n\cdot\delta_{stat}\right)^{\frac{1}{3}}$. □

We obtain a similar policy improvement lemma as in Lemma 5, incurring additional costs for the estimation error in the expected regression loss $L_i^t(\widehat{Q}^t_{i,t})$.

Lemma 6 (Policy improvement)**.** *For an MPG with respect to a set D^* according to Definition 2, for any state distribution μ, the potential function $\Phi^\pi(\mu)$ and two consecutive policies π^{t+1} and π^t generated by the PGA Algorithm 4, we have:*

$$
\begin{aligned}
\Phi^{t+1}_{t+1}(\rho)-\Phi^t_t(\rho) \geq{}& \frac{1}{4\eta(1-\gamma)}\sum_{i=1}^{n}\sum_{s} d_{\rho,t}^{\pi_i^{t+1},\pi_{-i}^t}(s)\left(1-\frac{4\eta\kappa_\rho^3 nA}{(1-\gamma)^4}\right)\left(\left\|\pi^{t+1}(\cdot|s)-\pi^t(\cdot|s)\right\|_2^2\right) \\
&-\frac{\eta\kappa_\rho A}{(1-\gamma)^2\xi}\sum_{i=1}^{n}L_i^t(\widehat{Q}^t_{i,t})-\frac{n(n+1)}{2}\cdot\delta_{r,p}\cdot\|\pi^{t+1}-\pi^t\|_2,
\end{aligned}
$$

where we have $\kappa_\mu = \sup_t\sup_{\pi\in\Pi}\|d^\pi_{\mu,t}/\mu\|_\infty$ and $\delta_{r,p} \coloneqq \frac{1}{1-\gamma}\left(\omega_r+\frac{\gamma\cdot\omega_p\sqrt{S}}{1-\gamma}\right)$.

Proof. By Lemma 4 with $\Psi^\pi = \Phi^\pi_t(\mu)$ and abbreviating $\pi' = \pi^{(t+1)}$, $\pi = \pi^{(t)}$, we have that

$$
\Phi^{t+1}_t(\mu)-\Phi^t_t(\mu) = \mathbf{Diff}_\alpha + \mathbf{Diff}_\beta,
$$

where

$$
\begin{aligned}
\mathbf{Diff}_\alpha &\coloneqq \sum_{i=1}^{n}\Phi_t^{\pi'_i,\pi_{-i}}(\mu)-\Phi^\pi_t(\mu), \\
\mathbf{Diff}_\beta &\coloneqq \sum_{i=1}^{n}\sum_{j=i+1}^{n}\left(\Phi_t^{\pi_{<i,i\sim j},\pi'_{>j},\pi'_i,\pi'_j}(\mu)-\Phi_t^{\pi_{<i,i\sim j},\pi'_{>j},\pi_i,\pi'_j}(\mu)-\ \Phi_t^{\pi_{<i,i\sim j},\pi'_{>j},\pi'_i,\pi_j}(\mu)+\Phi_t^{\pi_{<i,i\sim j},\pi'_{>j},\pi_i,\pi_j}(\mu)\right).
\end{aligned}
$$

According to the analysis Ding et al. (2022, Lemma 6.(ii)), this implies that,

$$
\begin{aligned}
\Phi^{t+1}_t(\mu)-\Phi^t_t(\mu) \geq{}& \frac{1}{4\eta(1-\gamma)}\sum_{i=1}^{n}\sum_{s} d_{\rho,t}^{\pi_i^{t+1},\pi_{-i}^t}(s)\left(1-\frac{4\eta\kappa_\rho^3 nA}{(1-\gamma)^4}\right)\left(\left\|\pi^{t+1}(\cdot|s)-\pi^t(\cdot|s)\right\|_2^2\right) \\
&-\frac{\eta\kappa_\rho A}{(1-\gamma)^2\xi}\sum_{i=1}^{n}L_i^t(\widehat{Q}^t_{i,t}).
\end{aligned}
\tag{14}
$$

Further, we derive that,

$$
\Phi^{t+1}_{t+1}(\mu)-\Phi^t_t(\mu) = \Phi^{t+1}_{t+1}-\Phi^{t+1}_t+\Phi^{t+1}_t(\mu)-\Phi^t_t(\mu) \geq -|\Phi^{t+1}_{t+1}-\Phi^{t+1}_t|+\Phi^{t+1}_t(\mu)-\Phi^t_t(\mu).
$$

By applying lemma 2 to upper bound the first term and by applying equation 14, we obtain our result. □

A.6 Independent Natural Policy Gradient Ascent – Analysis

In general, this analysis of previous results draw on a performative difference lemma that can be adopted to the PRL setting by incorporating additional costs due to changes of the environment. Recall that we assume a total potential function,

We focus on the result by Zhang et al. (2022) with log-barrier regularization, because this allows us to have a technical assumption on stationary points under shifting environments. The natural generalization of the guarantees gives further justification for the robustness of NPG-type algorithms under performative effects in our experiments.

A.6.1 Unregularized Version

For a fixed MPG at time t, we have the following result by Alatur et al. (2024a, Proposition IX.2).

Lemma 7 (Policy Improvement)**.** *For MPG(π^t) and any initial distribution ρ, the following holds*

$$\Phi_t^{t+1}(\rho) - \Phi_t^t(\rho) \geq \left(\frac{1}{\eta} - \frac{\sqrt{n}}{(1-\gamma)^2}\right) \sum_s \mu_{\rho,t}^{\pi^{t+1}}(s) \mathrm{KL}\left(\pi^{t+1}(\cdot|s)||\pi^t(\cdot|s)\right) + \frac{1}{\eta} \sum_s \mu_{\rho,t}^{\pi^{t+1}}(s) \sum_i \log Z_i^t(s).$$

This result allows us to obtain the policy improvement lemma below under performative effects.

Lemma 8 (Policy Improvement under Performative Effects)**.** *For all $i \in \mathcal{N}$, it holds that*

$$\begin{aligned} V_{i,t+1}^{t+1}(\rho) - V_{i,t}^t(\rho) \geq &\left(\frac{1}{M}\left(\frac{1}{\eta} - \frac{\sqrt{n}}{(1-\gamma)^2}\right) - \frac{\sqrt{2}}{1-\gamma}\left(\omega_r + \frac{\gamma\omega_p\sqrt{S}}{1-\gamma}\right)\right) \mathrm{KL}\left(\pi^{t+1}||\pi^t\right) \\ &+ \frac{1}{\eta} \sum_s \mu_{\rho,t}^{\pi^{t+1}}(s) \sum_i \log Z_i^t(s). \end{aligned}$$

Moreover, for $\eta \leq (1-\gamma)\left(\frac{\sqrt{n}}{1-\gamma} + \sqrt{2}M\left(\omega_r + \frac{\gamma\omega_p\sqrt{S}}{1-\gamma}\right)\right)^{-1}$,

$$V_{i,t+1}^{t+1}(\rho) - V_{i,t}^t(\rho) \geq \frac{1}{\eta} \sum_s \mu_{\rho,t}^{\pi^{t+1}} \sum_i \log Z_i^t(s).$$

Proof. By the definition of the potential function, for all $i \in \mathcal{N}$:

$$\begin{aligned} V_{i,t+1}^{t+1}(\rho) - V_{i,t}^t(\rho) &= V_{i,t}^{t+1}(\rho) - V_{i,t}^t(\rho) - \left(V_{i,t}^{t+1}(\rho) - V_{i,t+1}^{t+1}(\rho)\right) \\ &= \Phi_t^{t+1}(\rho) - \Phi_t^t(\rho) - \left(V_{i,t}^{t+1}(\rho) - V_{i,t+1}^{t+1}(\rho)\right) \\ &\geq \Phi_t^{t+1}(\rho) - \Phi_t^t(\rho) - \left|V_{i,t+1}^{t+1}(\rho) - V_{i,t}^{t+1}(\rho)\right| \\ &\overset{(a)}{\geq} \left(\frac{1}{\eta} - \frac{\sqrt{n}}{(1-\gamma)^2}\right) \sum_s \mu_{\rho,t}^{\pi^{t+1}}(s) \mathrm{KL}\left(\pi^{t+1}(\cdot|s)||\pi^t(\cdot|s)\right) + \frac{1}{\eta} \sum_s \mu_{\rho,t}^{\pi^{t+1}}(s) \sum_i \log Z_i^t(s) \\ &\quad - \frac{1}{1-\gamma}\left(\omega_r + \frac{\gamma\omega_p\sqrt{S}}{1-\gamma}\right) \left\|\pi^{t+1} - \pi^t\right\|_2 \\ &\geq \left(\frac{1}{\eta} - \frac{\sqrt{n}}{(1-\gamma)^2}\right) \sum_s \mu_{\rho,t}^{\pi^{t+1}}(s) \mathrm{KL}\left(\pi^{t+1}(\cdot|s)||\pi^t(\cdot|s)\right) + \frac{1}{\eta} \sum_s \mu_{\rho,t}^{\pi^{t+1}}(s) \sum_i \log Z_i^t(s) \\ &\quad - \frac{1}{1-\gamma}\left(\omega_r + \frac{\gamma\omega_p\sqrt{S}}{1-\gamma}\right) \left\|\pi^{t+1} - \pi^t\right\|_1 \\ &\overset{(b)}{\geq} \left(\frac{1}{\eta} - \frac{\sqrt{n}}{(1-\gamma)^2}\right) \sum_s \mu_{\rho,t}^{\pi^{t+1}}(s) \mathrm{KL}\left(\pi^{t+1}(\cdot|s)||\pi^t(\cdot|s)\right) + \frac{1}{\eta} \sum_s \mu_{\rho,t}^{\pi^{t+1}}(s) \sum_i \log Z_i^t(s) \\ &\quad - \frac{\sqrt{2}}{1-\gamma}\left(\omega_r + \frac{\gamma\omega_p\sqrt{S}}{1-\gamma}\right) \mathrm{KL}(\pi^{t+1}||\pi^t) \end{aligned}$$

$$
\begin{aligned}
&\geq \sum_s \left(\left(\frac{1}{\eta} - \frac{\sqrt{n}}{(1-\gamma)^2} \right) \mu_{\rho,t}^{\pi^{t+1}}(s) - \frac{\sqrt{2}}{1-\gamma} \left(\omega_r + \frac{\gamma \omega_p \sqrt{S}}{1-\gamma} \right) \right) \mathrm{KL}\left(\pi^{t+1}(\cdot|s) || \pi^t(\cdot|s)\right) \\
&\quad + \frac{1}{\eta} \sum_s \mu_{\rho,t}^{\pi^{t+1}}(s) \sum_i \log Z_i^t(s) \\
&\overset{(c)}{\geq} \sum_s \left(\frac{1}{M} \left(\frac{1}{\eta} - \frac{\sqrt{n}}{(1-\gamma)^2} \right) - \frac{\sqrt{2}}{1-\gamma} \left(\omega_r + \frac{\gamma \omega_p \sqrt{S}}{1-\gamma} \right) \right) \mathrm{KL}\left(\pi^{t+1}(\cdot|s) || \pi^t(\cdot|s)\right) \\
&\quad + \frac{1}{\eta} \sum_s \mu_{\rho,t}^{\pi^{t+1}}(s) \sum_i \log Z_i^t(s) \\
&\geq \left(\frac{1}{M} \left(\frac{1}{\eta} - \frac{\sqrt{n}}{(1-\gamma)^2} \right) - \frac{\sqrt{2}}{1-\gamma} \left(\omega_r + \frac{\gamma \omega_p \sqrt{S}}{1-\gamma} \right) \right) \mathrm{KL}\left(\pi^{t+1} || \pi^t\right) \\
&\quad + \frac{1}{\eta} \sum_s \mu_{\rho,t}^{\pi^{t+1}}(s) \sum_i \log Z_i^t(s),
\end{aligned}
$$

where (a) follows from Lemma 2 and Lemma 7, (b) follows from Pinsker's inequality, and (c) follows from Assumption 2.

Thus, for $\eta \leq (1-\gamma) \left(\frac{\sqrt{n}}{1-\gamma} + \sqrt{2} M \left(\omega_r + \frac{\gamma \omega_p \sqrt{S}}{1-\gamma} \right) \right)^{-1}$,

$$
V_{i,t+1}^{t+1}(\rho) - V_{i,t}^t(\rho) \geq \frac{1}{\eta} \sum_s \mu_{\rho,t}^{\pi^{t+1}} \sum_i \log Z_i^t(s). \qquad \square
$$

Moreover, we borrow the following lemma that bounds the performative gap for the fixed MPG at time t from Alatur et al. (2024a, Lemma IX.3).

Lemma 9. *For $\eta \leq (1-\gamma)^2$ and any initial distribution ρ, it holds that*

$$
\left(\max_i \max_{\pi_i'} V_{i,t}^{\pi_i', \pi_{-i}^t}(\rho) - V_{i,t}^{\pi^t}(\rho) \right)^2 \leq \frac{3 \tilde{\kappa}_\rho}{c \eta^2 (1-\gamma)} \sum_i \sum_s \mu_{\rho,t}^{\pi^{t+1}}(s) \log Z_i^t(s),
$$

where $c := \min_i \min_t \min_s \sum_{a_i^* \in \arg\max_{a_i \in A_i} \bar{Q}_{i,t}^{\pi^t}(s,a_i)} \pi_i^t(a_i^*|s) > 0$.

We now have all the ingredients for the convergence proofs.

Proof of Theorem 4. By Jensen's inequality, Lemma 8 and Lemma 9:

$$
\begin{aligned}
\text{Perform-Regret}(T) &\leq \sqrt{\frac{1}{T} \sum_{t=1}^T \left(\max_i \max_{\pi_i'} V_{i,t}^{\pi_i', \pi_{-i}^t} - V_{i,t}^{\pi^t} \right)^2} \\
&\leq \sqrt{\frac{1}{T} \sum_{t=1}^T \frac{3 \tilde{\kappa}_\rho}{c \eta^2 (1-\gamma)} \sum_i \sum_s \mu_{\rho,t}^{\pi^{t+1}}(s) \log Z_i^t(s)} \\
&\leq \sqrt{\frac{1}{T} \sum_{t=1}^T \frac{3 \tilde{\kappa}_\rho}{c \eta (1-\gamma)} \left(V_{i,t+1}^{t+1}(\rho) - V_{i,t}^t(\rho) \right)} \\
&\leq \sqrt{\frac{1}{T} \frac{3 \tilde{\kappa}_\rho \left(\frac{\sqrt{n}}{1-\gamma} + \sqrt{2} M \left(\omega_r + \frac{\gamma \omega_p \sqrt{S}}{1-\gamma} \right) \right)}{c (1-\gamma)^3}},
\end{aligned}
$$

where we set $\eta \leq (1-\gamma) \left(\frac{\sqrt{n}}{1-\gamma} + \sqrt{2} M \left(\omega_r + \frac{\gamma \omega_p \sqrt{S}}{1-\gamma} \right) \right)^{-1}$ for the last inequality. $\square$

A.6.2 Regularized Version

We follow the proof of Zhang et al. (2022) and define the following for notational simplicity.

$$\Delta_i^t(s, a_i) \coloneqq \frac{\pi_i^{t+1}(a_i|s)}{\pi_i^t(a_i|s)} - 1.$$

We borrow the following four lemmas from Zhang et al. (2022). Their proofs follow from the same lines as in the original lemmas for the fixed MPG at time step t.

Lemma 10 (Lemma 24 by Zhang et al. (2022))**.** *For $\eta \le \frac{1}{15\left(\frac{1}{(1-\gamma)^2} + \lambda A_i M\right)}$ and $\theta_i^0 = 0$, the following is satisfied by the regularized INPG for all $t \ge 1$:*

$$\pi_{\theta_i^t}(a_i|s) \ge \frac{\lambda}{4\left(\lambda A_i M + (1-\gamma)^{-2}\right)}.$$

Lemma 11 (Lemma 26 by Zhang et al. (2022))**.** *For MPG(π^t), any initial distribution ρ, $\eta \le \frac{1}{15\left(\frac{1}{(1-\gamma)^2} + \lambda A_i M\right)}$, and $\theta_i^0 = 0$, the following is satisfied by the regularized INPG dynamics:*

$$\tilde{\Phi}_t^{t+1}(\rho) - \tilde{\Phi}_t^t(\rho) \ge \left(\frac{1}{2\eta} - 4\lambda A_{\max} M^2 \frac{4M}{(1-\gamma)^2} - \frac{3nM}{(1-\gamma)^3}\right) \sum_i \sum_{s,a_i} \mu_{\rho,t}^{\pi^t}(s) \pi_i^t(a_i|s) \Delta_i^t(s, a_i)^2.$$

Lemma 12 (Lemma 27 by Zhang et al. (2022))**.** *For MPG(π^t), any initial distribution ρ, $\eta \le \frac{1}{15\left(\frac{1}{(1-\gamma)^2} + \lambda A_i M\right)}$, and $\theta_i^0 = 0$, the following is satisfied by the regularized INPG dynamics:*

$$\sum_i \sum_{s,a_i} \mu_{\rho,t}^{\pi^t}(s) \pi_i^t(a_i|s) \Delta_i^t(s, a_i)^2 \ge \frac{\eta^2}{9} \sum_i \sum_{s,a_i} \mu_{\rho,t}^{\pi^t}(s) \pi_i^t(a_i|s) f_i^t(s, a_i)^2.$$

Lemma 13 (Lemma 28 by Zhang et al. (2022))**.** *For MPG(π^t), any initial distribution ρ, $\eta \le \frac{1}{15\left(\frac{1}{(1-\gamma)^2} + \lambda A_i M\right)}$, and $\theta_i^0 = 0$, the following inequality is satisfied by the regularized INPG dynamics:*

$$\max_i \max_{\pi_i'} V_{i,t}^{\pi_i', \pi_{-i}^t} - V_{i,t}^{\pi^t} \le \frac{\sum_i \sum_{s,a_i} \mu_{\rho,t}^{\pi^t}(s) \pi_i^t(a_i|s) f_i^t(s, a_i)^2}{4\lambda} + \lambda M A_{\max}.$$

The following lemma provides an upper bound on the KL divergence of the consecutive policies generated by the regularized INPG dynamics in terms of the regularization parameter. By Pinsker's inequality, this also allows us to bound the performative effects.

Lemma 14. *For MPG(π^t), any initial distribution ρ, $\eta \le \frac{1}{15\left(\frac{1}{(1-\gamma)^2} + \lambda A_i M\right)}$, and $\theta_i^0 = 0$, the following inquality is satisfied by the regularized INPG dynamics:*

$$\mathrm{KL}(\pi^{t+1} || \pi^t) \le \eta n S \left(\frac{1}{(1-\gamma)^2} + 4\lambda M \left(\lambda A_{\max} M + \frac{1}{(1-\gamma)^2}\right)\right).$$

Proof. By the regularized INPG dynamics and Lemma 10,

$$\begin{aligned}
\log\left(\frac{\pi_i^{t+1}(a_i|s)}{\pi_i^t(a_i|s)}\right) &= \frac{\eta}{1-\gamma} \bar{A}_{i,t}^{\pi^t}(s, a_i) + \frac{\eta\lambda}{\mu_{\rho,t}^{\pi^t}(s) \pi_i^t(a_i|s)} - \frac{\eta\lambda A_i}{\mu_{\rho,t}^{\pi^t}} - \log Z_i^t(s) \\
&\le \frac{\eta}{1-\gamma} \bar{A}_{i,t}^{\pi^t}(s, a_i) + \frac{\eta\lambda}{\mu_{\rho,t}^{\pi^t}(s) \pi_i^t(a_i|s)} \\
&\le \frac{\eta}{(1-\gamma)^2} + 4\eta\lambda M \left(\lambda A_{\max} M + \frac{1}{(1-\gamma)^2}\right).
\end{aligned}$$

Thus,

$$\begin{aligned}\mathrm{KL}(\pi^{t+1}||\pi^t) &= \sum_i \sum_s \sum_{a_i} \pi_i^{t+1}(a_i|s) \log\left(\frac{\pi_i^{t+1}(a_i|s)}{\pi_i^t(a_i|s)}\right) \\ &\leq \sum_i \sum_s \sum_{a_i} \pi_i^{t+1}(a_i|s)\left(\frac{\eta}{(1-\gamma)^2} + 4\eta\lambda M\left(\lambda A_{\max} M + \frac{1}{(1-\gamma)^2}\right)\right) \\ &= \eta n S\left(\frac{1}{(1-\gamma)^2} + 4\lambda M\left(\lambda A_{\max} M + \frac{1}{(1-\gamma)^2}\right)\right). \qquad \square\end{aligned}$$

We can now provide the proof of Theorem 5. We follow the proof by Zhang et al. (2022, Theorem 7) while taking the performative effects into account.

Proof of Theorem 5. By the definition of $\tilde{\Phi}$, Lemma 2, Lemma 11 and Lemma 12,

$$\begin{aligned}\tilde{V}_{i,t+1}^{t+1}(\rho) - \tilde{V}_{i,t}^t(\rho) &\geq \tilde{\Phi}_t^{t+1}(\rho) - \tilde{\Phi}_t^t(\rho) - \left|\tilde{V}_{i,t+1}^{t+1}(\rho) - \tilde{V}_{i,t}^{t+1}(\rho)\right| \\ &\geq \frac{1}{4\eta}\sum_i \sum_{s,a_i} \mu_{\rho,t}^{\pi^t}(s)\pi_i^t(a_i|s)\Delta_i^t(s,a_i)^2 - \frac{1}{1-\gamma}\left(\omega_r + \frac{\gamma\omega_p\sqrt{S}}{1-\gamma}\right)\left\|\pi^{t+1}-\pi^t\right\|_2 \\ &\geq \frac{\eta}{36}\sum_i \sum_{s,a_i} \mu_{\rho,t}^{\pi^t}(s)\pi_i^t(a_i|s)f_i^t(s,a_i)^2 - \frac{1}{1-\gamma}\left(\omega_r + \frac{\gamma\omega_p\sqrt{S}}{1-\gamma}\right)\left\|\pi^{t+1}-\pi^t\right\|_2.\end{aligned}$$

Then,

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} \mu_{\rho,t}^{\pi^t}(s)\pi_i^t(a_i|s)f_i^t(s,a_i)^2 &\leq \frac{1}{T}\frac{36\left(\tilde{V}_T^T(\rho) - \tilde{V}_0^0(\rho)\right)}{\eta} + \frac{1}{1-\gamma}\left(\omega_r + \frac{\gamma\omega_p\sqrt{S}}{1-\gamma}\right)\left\|\pi^{t+1}-\pi^t\right\|_2 \\ &\leq \frac{1}{T}\frac{36\left(\tilde{V}_T^T(\rho) - \tilde{V}_0^0(\rho)\right)}{\eta} + \frac{1}{1-\gamma}\left(\omega_r + \frac{\gamma\omega_p\sqrt{S}}{1-\gamma}\right)\left\|\pi^{t+1}-\pi^t\right\|_1 \\ &\leq \frac{1}{T}\frac{36\left(\tilde{V}_T^T(\rho) - \tilde{V}_0^0(\rho)\right)}{\eta} + \frac{1}{1-\gamma}\left(\omega_r + \frac{\gamma\omega_p\sqrt{S}}{1-\gamma}\right)\mathrm{KL}\left(\pi^{t+1}||\pi^t\right) \\ &\leq \frac{1}{T}\frac{36\sqrt{2}\left(\tilde{V}_T^T(\rho) - \tilde{V}_0^0(\rho)\right)}{\eta} \\ &\quad + \frac{\eta n S}{1-\gamma}\left(\omega_r + \frac{\gamma\omega_p\sqrt{S}}{1-\gamma}\right)\left(\frac{1}{(1-\gamma)^2} + 4\lambda M\left(\lambda A_i M + \frac{1}{(1-\gamma)^2}\right)\right),\end{aligned}$$

where we use Pinsker's inequality for the third inequality, and the last step follows from Lemma 14. Thus, by Lemma 13:

$$\begin{aligned}\text{Perform-Regret}(T) &\leq \frac{1}{T}\sum_{t=0}^{T-1} \mu_{\rho,t}^{\pi^t}(s)\pi_i^t(a_i|s)f_i^t(s,a_i)^2 + \lambda M A_{\max} \\ &\leq \frac{9\sqrt{2}\left(\tilde{V}_T^T(\rho) - \tilde{V}_0^0(\rho)\right)}{\eta\lambda T} + \lambda M A_{\max} \\ &\quad + \frac{\eta n S}{1-\gamma}\left(\omega_r + \frac{\gamma\omega_p\sqrt{S}}{1-\gamma}\right)\left(\frac{1}{(1-\gamma)^2} + 4\lambda M\left(\lambda A_i M + \frac{1}{1-\gamma}\right)\right) \\ &\leq \frac{9\sqrt{2}}{\eta\lambda(1-\gamma)T} + \lambda M A_{\max} \\ &\quad + \frac{\eta n S}{1-\gamma}\left(\omega_r + \frac{\gamma\omega_p\sqrt{S}}{1-\gamma}\right)\left(\frac{1}{(1-\gamma)^2} + 4\lambda M\left(\lambda A_{\max} M + \frac{1}{1-\gamma}\right)\right),\end{aligned}$$

where $A_i \coloneqq |\mathcal{A}_i|$ and $A_{\max} \coloneqq \max_{i\in\mathcal{I}} A_i$. Moreover, for any $\epsilon > 0$, by setting $\lambda = \frac{\epsilon}{3MA_{\max}}$,

$$\begin{aligned}\eta = \min\Bigg\{ & \frac{1}{15\left(\frac{1}{(1-\gamma)^2} + \lambda A_{\max} M\right)}, \frac{1}{4\left(4\lambda A_{\max} M^2 + \frac{4M}{(1-\gamma)^2} + \frac{3nM}{(1-\gamma)^3}\right)}, \\ & \frac{1-\gamma}{3nS}\left(\omega_r + \frac{\gamma\omega_p\sqrt{S}}{1-\gamma}\right)^{-1}\left(\frac{1}{(1-\gamma)^2} + 4\lambda M\left(\lambda A_{\max} M + \frac{1}{(1-\gamma)^2}\right)\right)^{-1}\Bigg\} \\ = \min\Bigg\{ & \left(\frac{15}{(1-\gamma)^2} + 5\epsilon\right)^{-1}, \left(\frac{16\epsilon M}{3} + \frac{16M}{(1-\gamma)^2} + \frac{12nM}{(1-\gamma)^3}\right)^{-1}, \\ & \frac{1-\gamma}{3nS}\left(\omega_r + \frac{\gamma\omega_p\sqrt{S}}{1-\gamma}\right)^{-1}\left(\frac{1}{(1-\gamma)^2} + \frac{4\epsilon^2}{9A_{\max}} + \frac{4\epsilon}{3A_{\max}(1-\gamma)^2}\right)^{-1}\Bigg\},\end{aligned}$$

and

$$T \geq \mathcal{O}\left(\frac{\tilde{V}_T^T(\rho) - \tilde{V}_0^0(\rho)}{\eta\lambda\epsilon}\right) \geq \mathcal{O}\left(\frac{nA_{\max}M^2}{\epsilon^2(1-\gamma)^4}\max\left\{1, S\left(\omega_r + \frac{\gamma\omega_p\sqrt{S}}{1-\gamma}\right)\right\}\right),$$

we obtain

$$\text{Perform-Regret}(T) \leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} \leq \epsilon. \qquad \square$$

A.7 Last-Iterate Convergence in MPGs with Performative Effects and Agent Independent Transitions

Recall the sensitivity assumption.

Assumption 6 (Sensitivity)**.** *For any two policies π and π', we have that for all $i \in \mathcal{N}$,*

$$\begin{aligned}\|r_{i,\pi}(\cdot,\cdot) - r_{i,\pi'}(\cdot,\cdot)\|_2 &\leq \zeta_r \cdot \|\mu - \mu'\|_2, \\ \|P_\pi(\cdot\mid\cdot,\cdot) - P_{\pi'}(\cdot\mid\cdot,\cdot)\|_2 &\leq \zeta_p \cdot \|\mu - \mu'\|_2.\end{aligned}$$

We provide a stronger variant of Theorem 7, by given a tighter choice for the regularization parameter λ which only provides a sublinear dependence on the number of agents.

Theorem 8. *Let $\alpha_{\min} = \min_{s,\pi}\alpha_\pi(s)$, let*

$$\lambda > \frac{\sqrt{A_{\max}}}{A_{\min}} \cdot O\left(\zeta_p \cdot \frac{S^2\sqrt{n}\gamma A_{\max}^{9/4}}{(1-\gamma)^6} + \zeta_r \cdot \frac{S^{3/2}\gamma\sqrt{n}A_{\max}^{9/4}}{(1-\gamma)^4} + \frac{S^{3/2}\gamma A^{5/4}\beta}{(1-\gamma)^3\alpha_{\min}}\right)$$

be the fixed point of the objective in Eq. (10). It holds that, $\|\mu^T - \mu^\lambda\|_2 \leq \delta$ for $T \geq 2(1-\mu)^{-1}\ln(2/\delta(1-\gamma))$ and the performative gap is bounded:

$$\max_{i\in\mathcal{N}}\max_{\pi_i'}\left(V_{i,\pi^T}^{\pi_i',\pi_{-i}^{(T)}}(\rho) - V_{i,\pi^T}^{\pi^{(T)}}(\rho)\right) \leq \frac{\kappa_\rho}{\min_s \alpha_\lambda(s)(1-\gamma)} \cdot \left(\sqrt{A_{\max}} \cdot \delta + \frac{\lambda}{2(1-\gamma)}\right).$$

Proof of Theorem 7. Given policy π^λ induced by the computed μ^λ, given Lemma 15, suppose that we bound the performative-gap

$$\begin{aligned}\max_i\max_{\pi_i'} V_{i,\pi}^{\pi_i',\pi_{-i}}(\rho) - V_{i,\pi}^{\pi}(\rho) &= \max_i\max_{\pi_i'} \Phi_\pi^{\pi_i',\pi_{-i}}(\rho) - \Phi_\pi^\pi(\rho) \\ &\leq \frac{\kappa_\rho}{1-\gamma}\max_i\max_{\pi_i'}\langle \pi_i' - \pi_i, \nabla_{\pi_i}\Phi_\pi^\pi(\rho)\rangle \\ &\leq \frac{\kappa_\rho}{1-\gamma}\cdot\langle \pi' - \pi, \nabla_\pi\Phi_\pi^\pi(\rho)\rangle \\ &\leq \kappa_\rho \cdot \left\langle \frac{\mu'}{d_\lambda'} - \frac{\mu}{d_\lambda}, \nabla_\pi\Phi_\pi^\pi(\rho)\right\rangle,\end{aligned}$$

where we apply the gradient domination property in the first inequality. Further, we exploit the correspondence between policies and occupancy measures due to agent independent transitions in the third inequality. Further, we have

$$\begin{aligned}
\kappa_\rho \cdot \left\langle \frac{\mu'}{d'_\lambda} - \frac{\mu}{d_\lambda}, \nabla_\pi \Phi_\pi^\pi(\rho) \right\rangle &\leq \frac{\kappa_\rho}{\min_s \alpha_\lambda(s)} \langle \mu' - \mu, \nabla_\pi \Phi_\pi^\pi(\rho) \rangle \\
&\leq \frac{\kappa_\rho}{\min_s \alpha_\lambda(s)} \cdot \left[\langle \mu' - \mu^\lambda, \nabla_\pi \Phi_\pi^\pi(\rho) \rangle + \langle \mu^\lambda - \mu, \nabla_\pi \Phi_\pi^\pi(\rho) \rangle \right] \\
&\leq \frac{\kappa_\rho}{\min_s \alpha_\lambda(s)} \cdot \left[\frac{\sqrt{A_{\max}} \cdot \delta}{1-\gamma} + \frac{\lambda}{2} \left(\|\mu'\|_2^2 - \|\mu\|_2^2 \right) \right] \\
&\leq \frac{\kappa_\rho}{\min_s \alpha_\lambda(s)(1-\gamma)} \cdot \left(\sqrt{A_{\max}} \cdot \delta + \frac{\lambda}{2(1-\gamma)} \right). \qquad \square
\end{aligned}$$

First, we show the convergence.

Lemma 15. *Repeatedly optimizing Eq. (10) converges to a fixed point μ^λ, more precisely, given that*

$$\lambda > \frac{\sqrt{A_{\max}}}{A_{\min}} \cdot O\left(\zeta_p \cdot \frac{S^2 \sqrt{n} \gamma A_{\max}^{9/4}}{(1-\gamma)^6} + \zeta_r \cdot \frac{S^{3/2} \gamma \sqrt{n} A_{\max}^{9/4}}{(1-\gamma)^4} + \frac{S^{3/2} \gamma A^{5/4} \beta}{(1-\gamma)^3 \alpha_{\min}} \right)$$

chosen up to a sufficiently large constant, it holds that $\|\mu^T - \mu^\lambda\|_2 \leq \delta$ for $T \geq 2(1-\mu)^{-1} \ln(2/\delta(1-\gamma))$.

Let us recall the primal objective in Eq. (10) and explicitly formulate the constraint that $\mu = (\mu_1, \dots, \mu_n)$ where μ_i is a state-action occupancy measure over $\mathcal{S} \times \mathcal{A}_i$ (add flow constraints and $\mu \geq 0$). Hence, we consider the following objective:

$$\begin{aligned}
\max_{\mu \geq 0} \quad & \left\langle \nabla \Phi_t^t(\rho), \mu \right\rangle - \frac{\lambda}{2} \cdot \|\mu\|_2^2 \\
\text{s.t.} \quad & \sum_{a_i} \mu_i(s, a_i) = \rho(s) + \gamma \cdot \sum_{s'} P_t(s \mid s') \sum_{a_i} \mu_i(s', a_i) \quad \text{for all} \quad s \in \mathcal{S}, i \in \mathcal{N}.
\end{aligned} \tag{15}$$

The corresponding Lagrangian is formulated as:

$$\begin{aligned}
\mathcal{L}(\mu, h) =& \langle \mu, \nabla \Phi_t^t(\rho) \rangle - \frac{\lambda}{2} \|\mu\|_2^2 \\
&+ \sum_{i \in \mathcal{N}} \sum_s h_i(s) \left(- \sum_{a_i} \mu_i(s, a_i) + \rho(s) + \gamma \cdot \sum_{s'_i, a_i} \mu_i(s', a_i) P_t(s \mid s') \right).
\end{aligned} \tag{16}$$

To find an optimal μ, we take the gradient $\nabla_\mu \mathcal{L}$ and set it to zero:

$$\partial_{x_{i,s,a_i}} \Phi_t^t(\rho) - \lambda \cdot \mu_i(s, a_i) - h_i(s) + \gamma \cdot \sum_{\widetilde{s}} h_i(\widetilde{s}) P_t(\widetilde{s} \mid s) = 0.$$

Recall, that with $\partial_{x_{i,s,a_i}} \Phi_t^t(\rho)$, we refer to the partial derivative with respect to the played occupancy measure. Solving for $\mu_i(s, a_i)$, we obtain that

$$\mu_i(s, a_i) = \frac{1}{\lambda} \cdot \left(\partial_{x_{i,s,a_i}} \Phi_t^t(\rho) - h_i(s) + \gamma \cdot \sum_{\widetilde{s}} h_i(\widetilde{s}) P_t(\widetilde{s} \mid s) \right).$$

We substitute this value back to obtain the Lagrangian dual formulation:

$$\begin{aligned}
\min_{h \in \mathbb{R}^{n \times S}} & - \frac{1}{\lambda} \sum_i \sum_{s, a_i} h_i(s) \cdot \partial_{x_{i,s,a_i}} \Phi_t^t(\rho) + \frac{\gamma}{\lambda} \sum_i \sum_s \sum_{s'.a_i} \partial_{\mu_i, s', a_i} \Phi(s', a_i) \cdot h_i(s) \cdot P_t(s \mid s') \\
&+ \sum_i \sum_s h_i(s) \rho(s) + \frac{1}{2\lambda} \sum_i A_i \sum_s h_i(s)^2 - \frac{\gamma}{\lambda} \sum_i \sum_{s, a_i} h_i(s) \sum_{s'_i} h_i(s') P_t(s \mid s') \\
&+ \frac{\gamma^2}{2\lambda} \sum_i \sum_{s, a_i} \sum_{\widetilde{s}, \widehat{s}} h_i(\widetilde{s}) h_i(\widehat{s}) P_t(\widehat{s} \mid s) P_t(\widetilde{s} \mid s).
\end{aligned} \tag{17}$$

The dual is objective is parameterized with $\nabla\Phi_t^t$ and probability transition function P_t, which illustrates the performative effect given the underlying game $\mathcal{G}_t$ (shorthand notation for $\mathcal{G}(\pi^t)$) induced by the state occupancy measure μ^t. Further, we have that the gradient is dependent on the played policy π^t. These parameters capture indeed the influence of μ^t and we denote the dual objective as $\mathcal{L}(\cdot, \mathcal{G}_t, \pi^t)$ to express that.

Let $\mathrm{GD}(\mu^t)$ be the optimal solution to the primal problem, given that $\mathcal{G}_t$ is the underlying game. We show that $\mathrm{GD}(\cdot)$ corresponds to a contraction mapping. Given two occupancy measures $\mu, \widehat{\mu}$, we denote $\nabla\Phi = \nabla\Phi_\pi^\pi$ (resp. $\nabla\widehat{\Phi} = \nabla\Phi_{\widehat{\pi}}^{\widehat{\pi}}$) and P (resp. $\widehat{P}$) and π (respectively $\widehat{\pi}$) as the implemented policy. Further, let h respectively $\widehat{h}$ be the associated optimal solutions for the dual objective in Eq. (17). By the strong-convexity property, see Lemma 16, the following both inequalities hold:

$$\mathcal{L}(h, M, \pi) - \mathcal{L}(\widehat{h}, M, \pi) \geq \left\langle h - \widehat{h}, \nabla\mathcal{L}(\widehat{h}, M)\right\rangle + \frac{A_{\min}(1-\gamma)^2}{2\lambda}\left\|h - \widehat{h}\right\|_2^2$$

$$\mathcal{L}(\widehat{h}, M, \pi) - \mathcal{L}(h, M, \pi) \geq \frac{A_{\min}(1-\gamma)^2}{2\lambda}\left\|h - \widehat{h}\right\|_2^2,$$

which implies that

$$-\frac{A_{\min}(1-\gamma)^2}{\lambda}\left\|h - \widehat{h}\right\|_2^2 \geq \left\langle h - \widehat{h}, \nabla\mathcal{L}(\widehat{h}, M)\right\rangle = \left\langle h - \widehat{h}, \nabla\mathcal{L}(\widehat{h}, M) - \nabla\mathcal{L}(\widehat{h}, \widehat{M})\right\rangle, \tag{18}$$

where the last inequality uses that $\widehat{h}$ is optimal for $\mathcal{L}(\cdot, \widehat{M})$. Further, we can apply Lemma 17 to show that:

$$\begin{aligned}\left\|\nabla\mathcal{L}(\widehat{h}, M) - \nabla\mathcal{L}(\widehat{h}, \widehat{M})\right\|_2 \leq\ & \frac{\gamma S\sqrt{10A_{\max}\left(1 + \|\nabla\Phi(\rho)\|_\infty\right)}}{\lambda}\left\|\nabla\Phi(\rho) - \nabla\widehat{\Phi}(\rho)\right\|_2 \\ & + \frac{5\gamma S\sqrt{A_{\max}\left(1 + \|\nabla\Phi(\rho)\|_\infty\right)}}{\lambda}\|h\|_2\left\|P - \widehat{P}\right\|_2.\end{aligned}$$

Further, observe that using Lemma 20 and Lemma 21, we obtain that, (recall that $\Phi = \Phi_\pi^\pi$ and $\widehat{\Phi} = \Phi_{\widehat{\pi}}^{\widehat{\pi}}$)

$$\begin{aligned}\|\nabla\Phi - \nabla\widehat{\Phi}(\rho)\|_2 &\leq \|\nabla\Phi - \nabla\Phi_\pi^{\widehat{\pi}}\|_2 + \|\nabla\Phi_\pi^{\widehat{\pi}} - \nabla\widehat{\Phi}\|_2 \\ &\leq \left(\frac{\beta}{\alpha_{\min}} + \frac{2\gamma\sqrt{nSA_{\max}}}{(1-\gamma)^3}\zeta_p + \frac{\sqrt{nA_{\max}}}{(1-\gamma)^2}\zeta_r\right)\|\mu - \widehat{\mu}\|_2,\end{aligned}$$

where we denote $\alpha_{\min} = \min_{\pi,s}\alpha_\pi(s)$. By combining Lemma 19 with the observation that $\|\nabla\Phi(\rho)\|_\infty \leq \frac{\sqrt{A_{\max}}}{(1-\gamma)^2}$, which holds independent of the choice of the underlying environment $\mathcal{M}$ and the played policy π, we get that $\|h\|_2 \leq \frac{\sqrt{9nS}}{(1-\gamma)^2}\|\nabla\Phi_t^t(\rho)\|_\infty \leq \frac{3\sqrt{nSA_{\max}}}{(1-\gamma)^4}$. Moreover, we apply the Sensitivity Assumption 6 to bound $\|P - \widehat{P}\|_2 \leq \omega_p \cdot \|\mu - \widehat{\mu}\|_2$. This leads to

$$\begin{aligned}\left\|\nabla\mathcal{L}(\widehat{h}, M) - \nabla\mathcal{L}(\widehat{h}, \widehat{M})\right\|_2 \leq & \left(\frac{S\gamma A_{\max}^{3/4}\beta}{\lambda(1-\gamma)\alpha_{\min}}\right)\cdot\|\mu - \widehat{\mu}\|_2 \\ & + \frac{38\gamma\sqrt{n}S^{3/2}A_{\max}^{5/4}}{\lambda(1-\gamma)^4}\cdot\zeta_p\cdot\|\mu - \widehat{\mu}\|_2 + \frac{\zeta_r\gamma S\sqrt{n}A_{\max}^{5/4}}{(1-\gamma)^2}\cdot\|\mu - \widehat{\mu}\|_2.\end{aligned}$$

We substitute the latter bound into Eq. (18) to derive

$$\begin{aligned}-\frac{A_{\min}(1-\gamma)^2}{\lambda}\|h - \widehat{h}\|_2^2 &\geq -\|h - \widehat{h}\|_2\cdot\|\nabla\mathcal{L}(\widehat{h}, M) - \nabla\mathcal{L}(\widehat{h}, \widehat{M})\|_2 \\ &\geq -\|h - \widehat{h}\|_2\left(\frac{38\gamma S^{3/2}\cdot\sqrt{n}A_{\max}^{5/4}}{\lambda(1-\gamma)^4}\cdot\zeta_p + \frac{\zeta_r\gamma S\sqrt{n}A_{\max}^{5/4}}{(1-\gamma)^2} + \frac{S\gamma A^{3/4}\beta}{\lambda(1-\gamma)\alpha_{\min}}\right)\cdot\|\mu - \widehat{\mu}\|_2,\end{aligned}$$

so that

$$\|h - \widehat{h}\|_2 \leq \frac{\lambda}{A_{\min}(1-\gamma)^2}\left(\frac{38\gamma S^{3/2}\cdot\sqrt{n}A_{\max}^{5/4}}{\lambda(1-\gamma)^4}\cdot\zeta_p + \frac{\zeta_r\gamma S\sqrt{n}A_{\max}^{5/4}}{(1-\gamma)^2} + \frac{S\gamma A^{3/4}\beta}{\lambda(1-\gamma)\alpha_{\min}}\right)\|\mu - \widehat{\mu}\|_2.$$

Finally, we have the ingredients to bound the difference between the optimal primal solution ($\mathrm{GD}(\mu)$) when the deployed occupancy measure is μ with $\mathrm{GD}(\widehat{\mu})$. Let us define

$$
\begin{aligned}
&4\left(\frac{\sqrt{nA_{\max}}}{(1-\gamma)^2}\zeta_r + 2\frac{\sqrt{SnA_{\max}}\zeta_p\gamma}{(1-\gamma)^3}\right) + 6\gamma\zeta_p \left\|h\right\|_2 \\
&\leq 4\left(\frac{\sqrt{nA_{\max}}}{(1-\gamma)^2}\zeta_r + 2\frac{\sqrt{SnA_{\max}}\zeta_p\gamma}{(1-\gamma)^3}\right) + 6\gamma\zeta_p\sqrt{nSA_{\max}}/(1-\gamma)^4 =: K.
\end{aligned}
$$

By Lemma 18, we have that,

$$
\begin{aligned}
&\|\mathrm{GD}(\mu) - \mathrm{GD}(\widehat{\mu})\|_2 \leq \left(1 + \frac{K}{\lambda}\right)\frac{3\sqrt{SA_{\max}}}{\lambda A_{\min}(1-\gamma)^2}\left\|h - \widehat{h}\right\|_2 \\
&\leq \underbrace{\left(1+\frac{K}{\lambda}\right)\frac{\sqrt{SA_{\max}}}{(1-\gamma)^2 A_{\min}}\left(\frac{38\gamma S^{3/2}\cdot\sqrt{n}A_{\max}^{5/4}}{\lambda(1-\gamma)^4}\cdot\zeta_p + \frac{\zeta_r\gamma S\sqrt{n}A_{\max}^{5/4}}{(1-\gamma)^2} + \frac{S\gamma A^{3/4}\beta}{\lambda(1-\gamma)\alpha_{\min}}\right)}_{\ell}\|\mu - \widehat{\mu}\|_2\,.
\end{aligned}
$$

So, we obtain a bound

$$\|\mathrm{GD}(\mu) - \mathrm{GD}(\widehat{\mu})\|_2 \leq \xi\cdot\|\mu - \widehat{\mu}\|_2.$$

For the choice

$$\lambda > \frac{\sqrt{A_{\max}}}{A_{\min}}\cdot O\left(\zeta_p\cdot\frac{S^2\sqrt{n}\gamma A_{\max}^{9/4}}{(1-\gamma)^6} + \zeta_r\cdot\frac{S^{3/2}\gamma\sqrt{n}A_{\max}^{9/4}}{(1-\gamma)^4} + \frac{S^{3/2}\gamma A^{5/4}\beta}{(1-\gamma)^3\alpha_{\min}}\right),$$

we have that $\xi < 1$ for selecting λ up to sufficiently large constants. So that the operator GD is indeed a contraction map and converges against a fixed point μ^λ. This implies, that for $t \geq \ln\left(\|d_0 - d_\lambda\|_2)/\delta\right)\ln 1/\xi$, it is guaranteed that $\|\mu^t - \mu^\lambda\|_2 \leq \delta$.

Lemma 16. *The dual objective $\mathcal{L}_\mu(\cdot, M)$ (defined in Eq. (17)) is $\left[A_{\min}\frac{(1-\gamma)^2}{\lambda}\right]$-strongly convex with respect to h given a fixed M.*

Proof. The partial derivative of the objective in Eq. (17) is given by

$$
\begin{aligned}
\frac{\partial\mathcal{L}_\mu(h)}{\partial h_i(s)} &= -\frac{1}{\lambda}\sum_{a_i}\partial_{x_{i,s,a_i}}\Phi_t^t(\rho) + \rho(s) + \frac{\gamma}{\lambda}\sum_{s',a_i}\partial_{x_{i,s,a_i}}\Phi_t^t(\rho)P_t(s \mid s') \\
&\quad + \frac{A_i}{\lambda}h_i(s) - \frac{\gamma}{\lambda}\sum_{s'}h_i(s')\sum_{a_i}\left(P_t(s' \mid s) + P_t(s \mid s')\right) \\
&\quad + \frac{\gamma^2}{\lambda}\sum_{\widetilde{s}}h_i(\widetilde{s})\sum_{s',a_i}P_t(\widetilde{s} \mid s')P_t(s \mid s').
\end{aligned} \tag{19}
$$

Let $A_{\min} := \min_i A_i$, and I denote the $S\times S$ identity matrix. Moreover, define $M \in \mathbb{R}^{S\times S}$ such that $M(s,s') = P_t(s' \mid s)$ for all $s, \widetilde{s} \in \mathcal{S}$. We obtain the following inequality:

$$
\begin{aligned}
&\left\langle\nabla\mathcal{L}_\mu(h, M, \pi) - \nabla\mathcal{L}_\mu(\widehat{h}, M, \pi), h - \widehat{h}\right\rangle \\
&= \frac{1}{\lambda}\sum_{i,s}A_i\left(h_i(s) - \widehat{h}_i(s)\right)^2 \\
&\quad - \frac{\gamma}{\lambda}\sum_{i,s}\left(h_i(s) - \widehat{h}_i(s)\right)\sum_{s'}\left(h_i(s') - \widehat{h}_i(s')\right)\sum_{a_i}\left(P_t(s' \mid s) + P_t(s \mid s')\right) \\
&\quad + \frac{\gamma^2}{\lambda}\sum_{i,s}\left(h_i(s) - \widehat{h}_i(s)\right)\sum_{\widetilde{s}}\left(h_i(\widetilde{s}) - \widehat{h}_i(\widetilde{s})\right)\sum_{s',a_i}P_t(\widetilde{s} \mid s')P_t(s \mid s') \\
&= \frac{1}{\lambda}\sum_i\sum_{a_i}\left(h_i - \widehat{h}_i\right)^T\left(I - \gamma M - \gamma^2 M^T M\right)\left(h_i - \widehat{h}_i\right)
\end{aligned}
$$

$$
\begin{aligned}
&\geq \frac{(1-\gamma)^2}{\lambda} \sum_i A_i \left\| h_i - \widehat{h}_i \right\|_2^2 \\
&\geq A_{\min} \frac{(1-\gamma)^2}{\lambda} \left\| h - \widehat{h} \right\|_2^2,
\end{aligned}
$$

where the first inequality follows from Mandal et al. (2023, Lemma 5). □

Lemma 17. *The dual objective $\mathcal{L}$ (defined in Eq. (17)) satisfies the following bound for any h and MPGs $M, \widehat{M}$:*

$$
\begin{aligned}
\left\| \nabla \mathcal{L}(h, M, \pi) - \nabla \mathcal{L}(h, \widehat{M}, \widehat{\pi}) \right\|_2 \leq & \frac{\gamma S \sqrt{10 A_{\max} \left(1 + \|\nabla \Phi(\rho)\|_\infty\right)}}{\lambda} \left\| \nabla \Phi(\rho) - \nabla \widehat{\Phi}(\rho) \right\|_2 \\
& + \frac{5 \gamma S \sqrt{A_{\max} \left(1 + \|\nabla \Phi(\rho)\|_\infty\right)}}{\lambda} \|h\|_2 \left\| P - \widehat{P} \right\|_2 .
\end{aligned}
$$

Proof. Let $A_{\max} \coloneqq \max_i A_i$. By the partial derivative of the dual objective in Eq. (19), we obtain

$$
\begin{aligned}
& \left\| \nabla \mathcal{L}_\mu(h, M, \pi) - \nabla \mathcal{L}_\mu(h, \widehat{M}, \widehat{\pi}) \right\|_2^2 \\
& \leq \frac{5}{\lambda^2} \sum_i \sum_s \sum_{a_i} \left(\partial_{x_{i,s,a_i}} \Phi(\rho) - \partial_{x_{i,s,a_i}} \widehat{\Phi}(\rho) \right)^2 \\
& \quad + \frac{5\gamma^2}{\lambda^2} \sum_i \sum_s \sum_{s', a_i} \left(\partial_{x_{i,s,a_i}} \Phi(\rho) P(s' \mid s) - \partial_{x_{i,s,a_i}} \widehat{\Phi}(\rho) \widehat{P}(s' \mid s) \right)^2 \\
& \quad + \frac{5\gamma^2}{\lambda^2} \sum_i \sum_s \sum_{s', a_i} h_i(s') \left(P(s' \mid s) - \widehat{P}(s' \mid s) \right)^2 \\
& \quad + \frac{5\gamma^2}{\lambda^2} \sum_i \sum_s \sum_{s', a_i} h_i(s') \left(P(s \mid s') - \widehat{P}_i(s \mid s') \right)^2 \\
& \quad + \frac{5\gamma^4}{\lambda^2} \sum_i \sum_s \sum_{\widetilde{s}} h_i(\widetilde{s}) \sum_{s', a_i} \left(P(\widetilde{s} \mid s') P(s \mid s') - \widehat{P}(\widehat{s} \mid s') \widehat{P}(s \mid s') \right)^2 \\
& \leq \frac{5}{\lambda^2} \left\| \nabla \Phi(\rho) - \nabla \widehat{\Phi}(\rho) \right\|_2^2 + \frac{5\gamma^2}{\lambda^2} \left(1 + \|\nabla \Phi(\rho)\|_\infty\right) S^2 A_{\max} \left\| \nabla \Phi(\rho) - \nabla \widehat{\Phi}(\rho) \right\|_2^2 \\
& \quad + \frac{5\gamma^2}{\lambda^2} \left(1 + \|\nabla \Phi(\rho)\|_\infty\right) \|\nabla \Phi(\rho)\|_\infty S A_{\max} \left\| P - \widehat{P} \right\| \\
& \quad + \frac{10\gamma^2}{\lambda^2} A_{\max} \|h\|_2^2 \left\| P - \widehat{P} \right\|_2^2 + \frac{20\gamma^4}{\lambda^2} S^2 A_{\max} \|h\|_2^2 \left\| P - \widehat{P} \right\|_2^2 \\
& = \left(\frac{5}{\lambda^2} + \frac{5\gamma^2}{\lambda^2} \left(1 + \|\nabla \Phi(\rho)\|_\infty\right) S^2 A_{\max} \right) \left\| \nabla \Phi(\rho) - \nabla \widehat{\Phi}(\rho) \right\|_2^2 \\
& \quad + \left(\frac{5\gamma^2}{\lambda^2} \left(1 + \|\nabla \Phi(\rho)\|_\infty\right) S^2 A_{\max} + \frac{10\gamma^2}{\lambda^2} A_{\max} \|h\|_2^2 + \frac{20\gamma^4}{\lambda^2} S^2 A_{\max} \|h\|_2^2 \right) \left\| P - \widehat{P} \right\|_2^2,
\end{aligned}
$$

where the first inequality follows from Jensen's inequality and the second inequality uses the following inequalities:

$$
\begin{aligned}
& \sum_{s', a_i} \left(\partial_{x_{i,s,a_i}} \Phi(\rho) P(s' \mid s) - \partial_{x_{i,s,a_i}} \widehat{\Phi}(\rho) \widehat{P}(s' \mid s) \right)^2 \\
& \leq \sum_{s', a_i} \left(\left| \partial_{x_{i,s,a_i}} \Phi(\rho) - \partial_{x_{i,s,a_i}} \widehat{\Phi}(\rho) \right| + \|\nabla \Phi(\rho)\|_\infty \left| P(s \mid s') - \widehat{P}(s \mid s') \right| \right)^2 \\
& \leq \left(1 + \|\nabla \Phi(\rho)\|_\infty\right) \left(\sum_{s', a_i} \left| \partial_{x_{i,s,a_i}} \Phi(\rho) - \partial_{x_{i,s,a_i}} \widehat{\Phi}(\rho) \right| \right)^2 \\
& \quad + \left(1 + \|\nabla \Phi(\rho)\|_\infty\right) \|\nabla \Phi(\rho)\|_\infty \left(\sum_{s', a_i} \left| P(s \mid s') - \widehat{P}(s \mid s') \right| \right)^2
\end{aligned}
$$

$$\begin{aligned}
&\leq \left(1 + \left\|\nabla\Phi(\rho)\right\|_\infty\right) S^2 A_i \sum_{a_i} \left(\partial_{x_{i,s,a_i}} \ \Phi(\rho) - \partial_{x_{i,s,a_i}} \widehat{\Phi}(\rho)\right)^2 \\
&\quad + \left(1 + \left\|\nabla\Phi(\rho)\right\|_\infty\right) \left\|\nabla\Phi(\rho)\right\|_\infty S A_i \sum_{s',a_i} \left(P(s,s') - \widehat{P}(s \mid s')\right)^2;
\end{aligned}$$

$$\begin{aligned}
&\sum_i \sum_s \sum_{s',a_i} h_i(s') \left(P(s' \mid s) - \widehat{P}(s' \mid s)\right)^2 \\
&\leq \sum_i \sum_s \left(\sum_{s',a_i} h_i^2(s')\right) \left(\sum_{s',a_i} \left(P(s' \mid s) - \widehat{P}(s' \mid s)\right)^2\right) \\
&\leq \sum_i A_i \left\|h_i\right\|_2^2 \left\|P - \widehat{P}\right\|_2^2 \\
&\leq n A_{\max} \left\|h\right\|_2^2 \left\|P - \widehat{P}\right\|_2^2;
\end{aligned}$$

$$\begin{aligned}
&\sum_i \sum_s \sum_{s',a_i} h_i(s') \left(P(s \mid s') - \widehat{P}(s \mid s')\right)^2 \\
&\leq \sum_i \sum_s \left(\sum_{s',a_i} h_i^2(s')\right) \left(\sum_{s',a_i} \left(P(s \mid s') - \widehat{P}(s \mid s')\right)^2\right) \\
&\leq n A_{\max} \left\|h_i\right\|_2^2 \left\|P - \widehat{P}\right\|_2^2 \\
&\leq A_{\max} \left\|h\right\|_2^2 \left\|P - \widehat{P}\right\|_2^2;
\end{aligned}$$

$$\begin{aligned}
&\sum_i \sum_s \sum_{\widetilde{s},a_i,s'} h_i(\widetilde{s}) \left(P(\widetilde{s} \mid s') P(s \mid s') - \widehat{P}(\widehat{s} \mid s') \widehat{P}(s \mid s')\right)^2 \\
&\leq \sum_i \sum_s \left(\sum_{s',\widetilde{s}} h_i^2(\widetilde{s})\right) \left(\sum_{s',\widetilde{s}} \left(P(\widetilde{s}, s') P(s \mid s') - \widehat{P}(\widehat{s} \mid s') \widehat{P}(s \mid s')\right)^2\right) \\
&\leq \sum_i S A_i \left\|h_i\right\|_2^2 \left(\sum_{s,\widetilde{s}} \sum_{s',a_i} \left(P(\widetilde{s} \mid s') P(s \mid s') - \widehat{P}(\widehat{s} \mid s') \widehat{P}(s \mid s')\right)^2\right) \\
&\leq \sum_i S A_i \left\|h_i\right\|_2^2 \sum_{s,\widetilde{s}} \sum_{s',a_i} \left(\left|P(\widetilde{s} \mid s') - P(s \mid s')\right| + \left|\widehat{P}(\widehat{s} \mid s') - \widehat{P}(s \mid s')\right|\right)^2 \\
&\leq 4 \sum_i S A_i \left\|h_i\right\|_2^2 \sum_{s,\widetilde{s}} \sum_{s',a_i} \left(\left(P(\widetilde{s} \mid s') - P(s \mid s')\right)^2 + \left(\widehat{P}(\widehat{s} \mid s') - \widehat{P}(s \mid s')\right)^2\right) \\
&\leq 4 S^2 A_{\max} \sum_i \left\|h_i\right\|_2^2 \left\|P - \widehat{P}\right\|_2^2 \\
&\leq 4 S^2 A_{\max} \left\|h\right\|_2^2 \left\|P - \widehat{P}\right\|_2^2.
\end{aligned}$$

□

Lemma 18. *Consider two state-action occupancy measures μ and $\widehat{\mu}$. For $\lambda \geq 4n\zeta_r\sqrt{S} + 6\gamma\omega_p \left\|h\right\|_2$, the following bound holds:*

$$\left\|\mu - \widehat{\mu}\right\|_2 \leq \left(1 + \frac{4\left(\frac{\sqrt{nA_{\max}}}{(1-\gamma)^2}\zeta_r + 2\frac{\sqrt{SnA_{\max}}\zeta_p\gamma}{(1-\gamma)^3}\right) + 6\gamma\zeta_p \left\|h\right\|_2}{\lambda}\right) \frac{3\sqrt{SA_{\max}}}{\lambda} \left\|h - \widehat{h}\right\|_2^2.$$

Proof. By the partial derivative of the dual objective in Eq. (19), we obtain

$$\begin{aligned}
(\mu_i(s,a_i) - \widehat{\mu}_i(s,a_i))^2 &= \frac{1}{\lambda^2}\Bigg(\Big(\partial_{x_{i,s,a_i}}\Phi(\rho) - \partial_{x_{i,s,a_i}}\widehat{\Phi}(\rho)\Big) + \Big(-h_i(s) + \widehat{h}_i(s)\Big) \\
&\quad + \gamma\Bigg(\sum_{\widetilde{s}} h_i(\widetilde{s})P(\widetilde{s} \mid s) - \sum_{\widetilde{s}} \widehat{h}_i(\widetilde{s})\widehat{P}(\widetilde{s} \mid s)\Bigg)\Bigg)^2 \\
&\leq \frac{3}{\lambda^2}\Bigg(\Big(\partial_{x_{i,s,a_i}}\Phi(\rho) - \partial_{x_{i,s,a_i}}\widehat{\Phi}(\rho)\Big)^2 + \Big(-h_i(s) + \widehat{h}_i(s)\Big)^2 \\
&\quad + \gamma^2\Bigg(\sum_{\widetilde{s}} h_i(\widetilde{s}_i)P(\widetilde{s} \mid s) - \sum_{\widetilde{s}} \widehat{h}_i(\widetilde{s}_i)\widehat{P}(\widetilde{s} \mid s)\Bigg)^2\Bigg) \\
&\leq \frac{3}{\lambda^2}\Bigg(\Big(\partial_{x_{i,s,a_i}}\Phi(\rho) - \partial_{x_{i,s,a_i}}\widehat{\Phi}(\rho)\Big)^2 + \Big(-h_i(s) + \widehat{h}_i(s)\Big)^2 \\
&\quad + 2\gamma\Bigg(\sum_{\widetilde{s}} \Big(h_i(\widetilde{s}) - \widehat{h}_i(\widetilde{s})\Big)P(\widetilde{s} \mid s)\Bigg)^2 \\
&\quad + 2\gamma\Bigg(\sum_{\widetilde{s}} \widehat{h}_i(\widetilde{s}_i)\Big(P(\widetilde{s} \mid s) - \widehat{P}(\widetilde{s} \mid s)\Big)\Bigg)\Bigg)^2 \\
&\leq \frac{3}{\lambda^2}\Bigg(\Big(\partial_{x_{i,s,a_i}}\Phi(\rho) - \partial_{x_{i,s,a_i}}\widehat{\Phi}(\rho)\Big)^2 + \Big(-h_i(s) + \widehat{h}_i(s)\Big)^2 \\
&\quad + 2\gamma\Bigg(\sum_{\widetilde{s}} \Big(h_i(\widetilde{s}) - \widehat{h}_i(\widetilde{s})\Big)P(\widetilde{s} \mid s)\Bigg)^2 \\
&\quad + 2\gamma\Bigg(\sum_{\widetilde{s}} \widehat{h}_i(\widetilde{s})\Big(P(\widetilde{s} \mid s) - \widehat{P}(\widetilde{s} \mid s)\Big)\Bigg)\Bigg)^2 \\
&\leq \frac{3}{\lambda^2}\Bigg(\Big(\partial_{x_{i,s,a_i}}\Phi(\rho) - \partial_{x_{i,s,a_i}}\widehat{\Phi}(\rho)\Big)^2 + \Big(-h_i(s) + \widehat{h}_i(s)\Big)^2 \\
&\quad + 2\gamma\Bigg(\sum_{\widetilde{s}} \Big(h_i(\widetilde{s}) - \widehat{h}_i(\widetilde{s})\Big)\Bigg)^2 \\
&\quad + 2\gamma\Bigg(\sum_{\widetilde{s}} \widehat{h}_i^2(\widetilde{s})\Bigg)\Bigg(\sum_{\widetilde{s}} \Big(P(\widetilde{s} \mid s) - \widehat{P}(\widetilde{s} \mid s)\Big)^2\Bigg).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\mu - \widehat{\mu}\|_2^2 &\leq \frac{3}{\lambda^2}\left\|\nabla\Phi(\rho) - \nabla\widehat{\Phi}(\rho)\right\|_2^2 + \frac{3}{\lambda^2}\left\|h - \widehat{h}\right\|_2^2 + \frac{6\gamma^2}{\lambda^2}\sum_i\sum_{s,a_i}\left\|h_i - \widehat{h}_i\right\|_2^2 \\
&\quad + \frac{6\gamma^2}{\lambda^2}\sum_i\sum_{s,a_i}\|h_i\|_2^2\left\|P(\cdot \mid s) - \widehat{P}(\cdot \mid s)\right\|_2^2 \\
&\leq \frac{3}{\lambda^2}\left\|\nabla\Phi(\rho) - \nabla\widehat{\Phi}(\rho)\right\|_2^2 + \frac{9}{\lambda^2}S \cdot \sum_i A_i \cdot \left\|h - \widehat{h}\right\|_2^2 + \frac{6\gamma^2}{\lambda^2}\|h\|_2^2\left\|P - \widehat{P}\right\|_2^2.
\end{aligned}$$

By Lemma 20 and Assumption 6:

$$\|\mu - \widehat{\mu}\|_2 \leq \frac{2}{\lambda}\left(\frac{\sqrt{nA_{\max}}}{(1-\gamma)^2}\zeta_r + 2\frac{\sqrt{SnA_{\max}}\zeta_p\gamma}{(1-\gamma)^3}\right)\|\mu - \widehat{\mu}\|_2 + \frac{3}{\lambda}\sqrt{SA_{\max}}\left\|h - \widehat{h}\right\|_2 + \frac{3\gamma}{\lambda}\zeta_p\|h\|_2\|\mu - \widehat{\mu}\|_2 .$$

Rearranging the terms yields the following bound:

$$\|\mu - \widehat{\mu}\|_2 \leq \left(1 - \frac{2\left(\frac{\sqrt{nA_{\max}}}{(1-\gamma)^2}\zeta_r + 2\frac{\sqrt{SnA_{\max}}\zeta_p\gamma}{(1-\gamma)^3}\right) + 3\gamma\zeta_p\|h\|_2}{\lambda}\right)^{-1}\frac{3\sqrt{SA_{\max n}}}{\lambda}\left\|h - \widehat{h}\right\|_2^2$$

$$\leq \left(1 + \frac{4\left(\frac{\sqrt{nA_{\max}}}{(1-\gamma)^2}\zeta_r + 2\frac{\sqrt{SnA_{\max}}\zeta_p\gamma}{(1-\gamma)^3}\right) + 6\gamma\zeta_p \left\|h\right\|_2}{\lambda}\right) \frac{3\sqrt{SA_{\max}}}{\lambda} \left\|h - \widehat{h}\right\|_2^2$$

for $\lambda \geq 4\left(\frac{\sqrt{nA_{\max}}}{(1-\gamma)^2}\zeta_r + 2\frac{\sqrt{SnA_{\max}}\zeta_p\gamma}{(1-\gamma)^3}\right)$. □

Lemma 19. *The norm of the optimal solution to the dual objective (defined in Eq. (17)) is bounded by* $\frac{\sqrt{9nS}}{(1-\gamma)^2}\left\|\nabla\Phi_t^t(\rho)\right\|_\infty$ *for any MPG* $\mathcal{M}$.

Proof. Rearranging the terms of the partial derivative of the dual objective in Eq. (19), we have

$$\begin{aligned}
&h_i(s)\left[\frac{A_i}{\lambda} - \frac{2\gamma}{\lambda}\sum_{a_i} P_t(s' \mid s) + \frac{\gamma^2}{\lambda}\sum_{s',a_i} P_t^2(s \mid s')\right] \\
&+ \sum_{\widetilde{s}\neq s} h_i(\widetilde{s})\left[-\frac{\gamma}{\lambda}\sum_{a_i} P_t(\widetilde{s} \mid s) - \frac{\gamma}{\lambda}\sum_{a_i} P_t(s \mid \widetilde{s}) + \frac{\gamma^2}{\lambda}\sum_{s',a_i} P_t(\widetilde{s} \mid s')P_t(s \mid s')\right] \\
&= \frac{1}{\lambda}\sum_{a_i}\partial_{x_{i,s,a_i}}\Phi_t^t(\rho) - \rho(s) - \frac{\gamma}{\lambda}\sum_{s',a_i}\partial_{x_{i,s,a_i}}\Phi_t^t(\rho)P_t(s \mid s').
\end{aligned}$$

For each $i \in \mathcal{N}$, define $B_i \in \mathbb{R}^{S\times S}$ and $b_i \in \mathbb{R}^S$, where

$$B_i(s, \widetilde{s}_i) = \begin{cases} \frac{A_i}{\lambda} - \frac{2\gamma}{\lambda}\sum_{a_i} P_t(s' \mid s, a_i) + \frac{\gamma^2}{\lambda}\sum_{s',a_i} P_t^2(s \mid s', a_i) & s = \widetilde{s} \\ -\frac{\gamma}{\lambda}\sum_{a_i} P_t(\widetilde{s} \mid s, a_i) - \frac{\gamma}{\lambda}\sum_{a_i} P_t(\widetilde{s}, a_i, s) + \frac{\gamma^2}{\lambda}\sum_{s',a_i} P_t(\widetilde{s} \mid s', a_i)P_t(s \mid s', a_i) & \mathrm{s} \neq \widetilde{\mathrm{s}} \end{cases}$$

and

$$b_i(s) = \frac{1}{\lambda}\sum_{a_i}\partial_{\mu_i,s,a_i}\Phi_t^t(\rho) - \rho(s) - \frac{\gamma}{\lambda}\sum_{s',a_i}\partial_{\mu_i,s,a_i}\Phi_t^t(\rho)P_t(s \mid s', a_i).$$

Then, the optimal solution of the system is the solution of $B_i h_i = b_i$ for all agents $i \in \mathcal{N}$. For each $i \in \mathcal{N}$ and $a_i \in \mathcal{A}_i$, define $M_{i,a_i} \in \mathbb{R}^{S\times S}$ such that $M_{t,i,a_i}(s, s') = P_t(s' \mid s, a_i)$ for all $s, \widetilde{s} \in \mathcal{S}_i$. Let I_i denote an $S \times S$ identity matrix. We have

$$B_i = \frac{A_i}{\lambda}I_i - \frac{\gamma}{\lambda}\sum_{a_i}\left(M_{t,i,a_i} + M_{t,i,a_i}^T\right) + \frac{\gamma^2}{\lambda}M_{t,i,a_i}^T M_{t,i,a_i} \leq \frac{A_i(1-\gamma)^2}{\lambda}I_i,$$

where the inequality follows from Mandal et al. (2023, Lemma 5). Moreover,

$$\begin{aligned}
&\left(\frac{1}{\lambda}\sum_{a_i}\partial_{\mu_i,s,a_i}\Phi_t^t(\rho) - \rho(s) - \frac{\gamma}{\lambda}\sum_{s',a_i}\partial_{\mu_i,s,a_i}\Phi_t^t(\rho)P_t(s \mid s', a_i)\right)^2 \\
&\leq \frac{3}{\lambda^2}\left(\sum_{a_i}\partial_{\mu_i,s,a_i}\Phi_t^t(\rho)\right)^2 + 3\rho^2(s) + \frac{3\gamma^2}{\lambda}\left(\sum_{s',a_i}\partial_{\mu_i,s,a_i}\Phi_t^t(\rho)P_t(s \mid s', a_i)\right)^2 \\
&\leq \frac{3}{\lambda^2}A_i^2\left\|\nabla\Phi_t^t(\rho)\right\|_\infty^2 + 3\rho^2(s) + \frac{3\gamma^2}{\lambda^2}S\left(\sum_{a_i}\left(\partial_{\mu_i,s,a_i}\Phi_t^t(\rho)\right)^2\right)\left(\sum_{s',a_i}\left(P_t(s \mid s', a_i)\right)^2\right) \\
&\leq \frac{3}{\lambda^2}A^2\left\|\nabla\Phi_t^t(\rho)\right\|_\infty^2 + 3\rho^2(s) + \frac{3\gamma^2}{\lambda^2}S\left(\sum_{a_i}\left(\partial_{\mu_i,s,a_i}\Phi_t^t(\rho)\right)^2\right)\left(\sum_{s',a_i}\left(P_t(s \mid s', a_i)\right)^2\right).
\end{aligned}$$

Thus, we obtain

$$\left\|h\right\|_2^2 \leq \sum_i \frac{\left\|b_i\right\|_2^2}{\left(\lambda_{\min}(B_i)\right)^2}$$

$$
\begin{aligned}
&\leq \frac{\lambda^2}{(1-\gamma)^4 A_{\max}^2} \sum_{i,s} \left(\frac{3}{\lambda^2} A_i^2 \left\| \nabla \Phi_t^t(\rho) \right\|_\infty^2 + 3\rho^2(s) \right) \\
&\quad + \frac{\lambda^2}{(1-\gamma)^4 A_{\max}^2} \sum_{i,s} \frac{3\gamma^2}{\lambda^2} S \left(\sum_{a_i} \left(\partial_{\mu_{i,s,a_i}} \Phi_t^t(\rho) \right)^2 \right) \left(\sum_{s',a_i} \left(P_t(s \mid s', a_i) \right)^2 \right) \\
&\leq \frac{\lambda^2}{(1-\gamma)^4 A_{\max}^2} \left(\frac{3}{\lambda^2} n S A_{\max}^2 \left\| \nabla \Phi_t^t(\rho) \right\|_\infty^2 + 3 + \frac{3\gamma^2}{\lambda^2} S A_{\max} \left\| \nabla \Phi_t^t(\rho) \right\|_\infty^2 \right).
\end{aligned}
$$

For $\lambda < \sqrt{nS} A_{\max} \left\| \nabla \Phi_t^t(\rho) \right\|_\infty$, we can further simplify the bound to obtain

$$
\|h\|_2^2 \leq \frac{\lambda^2}{(1-\gamma)^4 A_{\max}^2} \frac{9nSA_{\max}^2}{\lambda^2} \left\| \nabla \Phi_t^t(\rho) \right\|_\infty^2 \leq \frac{9nS}{(1-\gamma)^4} \left\| \nabla \Phi_t^t(\rho) \right\|_\infty^2 \qquad \square
$$

Lemma 20. *Let M and $\widehat{M}$ be two different underlying MPGs associated with state-action occupancy measures, μ and $\widehat{\mu}$, given policy $\widehat{\pi}$, it holds that*

$$
\left\| \nabla \Phi_M^{\tilde{\pi}} - \nabla \Phi_{\widehat{M}}^{\tilde{\pi}} \right\|_2 \leq \|\mu - \widehat{\mu}\|_2 \cdot \left(\frac{\sqrt{nA_{\max}}}{(1-\gamma)^2} \zeta_r + 2 \frac{\sqrt{SnA_{\max}} \zeta_p \gamma}{(1-\gamma)^3} \right).
$$

Proof. By the policy gradient theorem (Sutton et al., 1999), for an agent $i \in \mathcal{N}$ policy $\pi \in \Pi$:

$$
\frac{\partial \Phi_\pi^{\tilde{\pi}}}{\partial \pi_i(a_i|s)} = \frac{1}{1-\gamma} d_\pi^{\tilde{\pi}}(s) \bar{Q}_{i,\pi}^{\tilde{\pi}}(s, a_i).
$$

Then,

$$
\begin{aligned}
&\left\| \nabla \Phi_M^{\tilde{\pi}} - \nabla \Phi_{\widehat{M}}^{\tilde{\pi}} \right\|_2^2 \\
&= \frac{1}{(1-\gamma)^2} \sum_i \sum_{s,a_i} \left(d_{\rho,\pi}^{\tilde{\pi}}(s) \bar{Q}_{i,\pi}^{\tilde{\pi}}(s,a_i) - d_{\rho,\widehat{\pi}}^{\tilde{\pi}}(s) \bar{Q}_{i,\widehat{\pi}}^{\tilde{\pi}}(s,a_i) \right)^2 \\
&\leq \frac{1}{(1-\gamma)^2} \sum_i \sum_{s,a_i} \left(\left| d_{\rho,\pi}^{\tilde{\pi}}(s) - d_{\rho,\widehat{\pi}}^{\tilde{\pi}}(s) \right| \bar{Q}_{i,\pi}^{\tilde{\pi}}(s,a_i) + d_{\rho,\widehat{\pi}}^{\tilde{\pi}}(s) \left| \bar{Q}_{i,\pi}^{\tilde{\pi}}(s,a_i) - \bar{Q}_{i,\widehat{\pi}}^{\tilde{\pi}}(s,a_i) \right| \right)^2 \\
&\leq \frac{1}{(1-\gamma)^2} \sum_i \sum_{s,a_i} \left(\left| d_{\rho,\pi}^{\tilde{\pi}}(s) - d_{\rho,\widehat{\pi}}^{\tilde{\pi}}(s) \right| \bar{Q}_{i,\pi}^{\tilde{\pi}}(s,a_i) + d_{\rho,\widehat{\pi}}^{\tilde{\pi}}(s) \left| \bar{Q}_{i,\pi}^{\tilde{\pi}}(s,a_i) - \bar{Q}_{i,\widehat{\pi}}^{\tilde{\pi}}(s,a_i) \right| \right)^2 \\
&\leq \frac{2}{(1-\gamma)^2} \sum_i \sum_{s,a_i} \left[\left| d_{\rho,\pi}^{\tilde{\pi}}(s) - d_{\rho,\widehat{\pi}}^{\tilde{\pi}}(s) \right| \bar{Q}_{i,\pi}^{\tilde{\pi}}(s,a_i) \right]^2 + \left[d_{\rho,\widehat{\pi}}^{\tilde{\pi}}(s) \left| \bar{Q}_{i,\pi}^{\tilde{\pi}}(s,a_i) - \bar{Q}_{i,\widehat{\pi}}^{\tilde{\pi}}(s,a_i) \right| \right]^2 \\
&\leq \frac{2}{(1-\gamma)^4} \sum_i \sum_{s,a_i} \left[d_{\rho,\pi}^{\tilde{\pi}}(s) - d_{\rho,\widehat{\pi}}^{\tilde{\pi}}(s) \right]^2 + \frac{2}{(1-\gamma)^2} \sum_i \sum_{s,a_i} \left[d_{\rho,\widehat{\pi}}^{\tilde{\pi}}(s) \left| \bar{Q}_{i,\pi}^{\tilde{\pi}}(s,a_i) - \bar{Q}_{i,\widehat{\pi}}^{\tilde{\pi}}(s,a_i) \right| \right]^2 \\
&\lesssim \|\mu - \widehat{\mu}\|_2 \cdot \left(\frac{SnA_{\max} \cdot \zeta_p^2}{(1-\gamma)^4} + \left[\frac{\sqrt{nA_{\max}}}{(1-\gamma)^2} \left(\zeta_r + \frac{\gamma \cdot \zeta_p \sqrt{S}}{1-\gamma} \right) \right]^2 \right).
\end{aligned}
$$

For the last inequality to hold, we exploit the following bounds: first, we use that

$$
\begin{aligned}
\sum_{i,s,a_i} \left(d_{\rho,\pi}^{\tilde{\pi}}(s) - d_{\rho,\widehat{\pi}}^{\tilde{\pi}}(s) \right)^2 &\leq (1-\gamma)^2 \cdot nA_{\max} \cdot \|\tilde{\mu}_\pi - \tilde{\mu}_{\widehat{\pi}}\|_2^2 \\
&\leq 3SnA_{\max} \cdot \zeta_p^2 \cdot \|\mu - \widehat{\mu}\|_2^2,
\end{aligned}
$$

where the last inequality uses the same computation as in Lemma 21. Further, exploiting Lemma 2, we obtain

$$
\max_{s,a} \left| \bar{Q}_{i,\pi}^{\tilde{\pi}}(s,a_i) - \bar{Q}_{i,\widehat{\pi}}^{\tilde{\pi}}(s,a_i) \right| \leq \frac{1}{1-\gamma} \cdot \left(\zeta_r + \frac{\gamma \cdot \zeta_p \sqrt{S}}{1-\gamma} \right) \cdot \|\mu - \widehat{\mu}\|_2.
$$

By combining, we conclude the proof. $\square$

Lemma 21. *It holds that* $\|\nabla\Phi_\pi^{\pi'} - \Phi_\pi^{\pi''}\|_2 \lesssim \frac{\beta}{\min_{s,\pi}\alpha_\pi(s)}\left(1+\frac{\zeta_p\sqrt{S}}{1-\gamma}\right)\|\mu''-\mu'\|_2$.

Proof. First, we use the assumption that Φ_π^x is β-smooth in x, we have

$$\|\nabla\Phi_\pi^{\pi'} - \nabla\Phi_\pi^{\pi''}\|_2 \leq \beta\cdot\|\pi'-\pi''\|_2.$$

Due to agent-independent transitions, it holds that

$$\|\pi'-\pi''\|_2 \leq \frac{\|\mu_{\pi'}^{\pi'} - \mu_{\pi'}^{\pi''}\|_2}{\min_s \alpha_\pi(s)} \leq \frac{\|\mu_{\pi'}^{\pi'} - \mu_{\pi''}^{\pi''}\|_2 + \|\mu_{\pi''}^{\pi''} - \mu_{\pi'}^{\pi''}\|_2}{\min_s \alpha_\pi(s)}.$$

It remains to show that $\|\mu_{\pi''}^{\pi''} - \mu_{\pi'}^{\pi''}\|_2$ is bounded by a fraction of $\|\mu_{\pi''}^{\pi''} - \mu_{\pi'}^{\pi''}\|_2$. We rewrite the expression explicitly and abuse notation to express the probability that an action/state pair occurs, i.e.,

$$\begin{aligned}
\|\mu_{\pi''}^{\pi''} - \mu_{\pi'}^{\pi''}\|_2 &= \sum_{i\in\mathcal{N}}\sum_{s\in\mathcal{S}}\sum_{a_i}\left[\sum_{t=0}^{\infty}\left(P_{\pi''}(s^t=s, a_i^t=a_i \mid \pi'') - P_{\pi'}(s_t=s, a_i^t=a_i \mid \pi'')\right)\right]^2 \\
&\leq \frac{1}{(1-\gamma)^2}\sum_{i,s,a_i}\left(\max_t\left(P_{\pi''}(s^t=s, a_i^t=a_i \mid \pi'') - P_{\pi'}(s_t=s, a_i^t=a_i \mid \pi'')\right)\right)^2 \\
&= \frac{1}{(1-\gamma)^2}\sum_{i,s,a_i}\left(\sum_{s'}\left(\left|P_{\pi''}(s^t=s, a_i^t=a_i \mid \pi'', s') - P_{\pi'}(s_t=s, a_i^t=a_i \mid \pi'', s')\right|\cdot\max_{x\in\{\pi',\pi''\}}P_x(s_{t-1}=s)\right)\right)^2 \\
&\leq \frac{3S}{(1-\gamma)^2}\sum_{i,s,a_i}\sum_i\sum_{s,s',a}|P_{\pi''}(s,a\mid s') - P_{\pi'}(s,a\mid s')|^2 \\
&\leq \frac{3S}{(1-\gamma)^2}\|P_{\pi''}(\cdot\mid\cdot,\cdot) - P_{\pi'}(\cdot\mid\cdot,\cdot)\|_2^2 \\
&\leq \frac{3S\zeta_p^2}{(1-\gamma)^2}\|\mu''-\mu'\|_2^2,
\end{aligned}$$

where we especially make use of the Cauchy-Schwarz inequality in the second inequality and the sensitivity assumption in the last inequality. Putting the inequalities together, it holds that

$$\|\pi'-\pi''\|_2 \lesssim \frac{1}{\min_{s,\pi}\alpha_\pi(s)}\left(1+\frac{\zeta_p\sqrt{S}}{1-\gamma}\right)\|\mu''-\mu'\|_2 \qquad \square$$

The next statement can be proved as Lemma 2, by exchanging the definition of sensitivity.

Lemma 22. *We have for all* $\pi, \widehat{\pi}$*, given a fixed* $\tilde{\pi}$*, it holds that*

$$\max_{s,a}\left|\bar{Q}_{i,\pi}^{\tilde{\pi}}(s,a_i) - \bar{Q}_{i,\widehat{\pi}}^{\tilde{\pi}}(s,a_i)\right| \leq \frac{1}{1-\gamma}\cdot\left(\zeta_r + \frac{\gamma\cdot\zeta_p\sqrt{S}}{1-\gamma}\right)\cdot\|\mu-\widehat{\mu}\|_2\,.$$

B EXPERIMENTAL SETUP

This section describes details about the experiments presented in Section 7. The code can be found in https://github.com/PauliusSasnauskas/performative-mpgs.

B.1 Algorithms

The policy distance presented in Figures 3, 4, 5, 6 is the distance from the current policy to the average of the last 10 in that run:

$$\text{Policy distance }(t) = \frac{1}{N}\sum_i^N\left\|\pi_i^t - \pi_i^{\text{last}}\right\|,$$

where π_i^{last} is the average of the last 10 policies in that run.

B.1.1 Independent Projected Gradient Ascent (IPGA)

We improve upon the code presented by Leonardos et al. (2022) for the IPGA algorithm, which is shown in Algorithm 1. We set the hyperparameters as shown in Table 2. We name the regularized IPGA version in the name of Leonardos et al. (2022) – *IPGA-L*, and the unregularized version in the name of Ding et al. (2022) – *IPGA-D*. The comparison of performance for different performativity strengths and different learning rates can be seen in Figure 3 for the safe-distancing game, and in Figure 4 for the stochastic congestion game.

Table 2: Hyperparameters used for the IPGA algorithm (unless noted otherwise).

Parameter	**Value**
Learning rate η	0.0001
Discount factor γ	0.99
Number of episodes per round	20

Algorithm 1 Independent Projected Gradient Ascent Practical Implementation

```
Input: $\eta$ – step size, $T$ – number of rounds, $K$ – episodes per round
Init.: for all $i \in \mathcal{N}, a_i \in \mathcal{A}_i, s \in S$, let $\pi_i^t(a_i|s) = 1/|\mathcal{A}_i|$.
for $t = 1$ to $T$ do
    Roll out policies $\pi^t$ for $K$ episodes to get trajectories $\tau$
    Compute state visitation distribution $\mu(s)$ from trajectories $\tau$
    Compute state-action value function $Q_i(s,a)$ from trajectories $\tau$ for each agent $i \in \mathcal{N}$
    for $i = 1$ to $n$ (simultaneously) do
        Compute $g_i(s,a) = \mu(s)\, Q_i(s,a) \quad \forall s \in S, a \in \mathcal{A}_i$        ▷ Set $\mu(s) = 1\ \forall s \in S$ for IPGA-D version
        Update $\pi_i^{t+1}(a|s) = \mathrm{Proj}_{\Delta_{\mathcal{A}_i}}(\pi_i^t(a|s) + \eta g(s,a)) \quad \forall s \in S, a \in \mathcal{A}_i$
    end for
end for
```

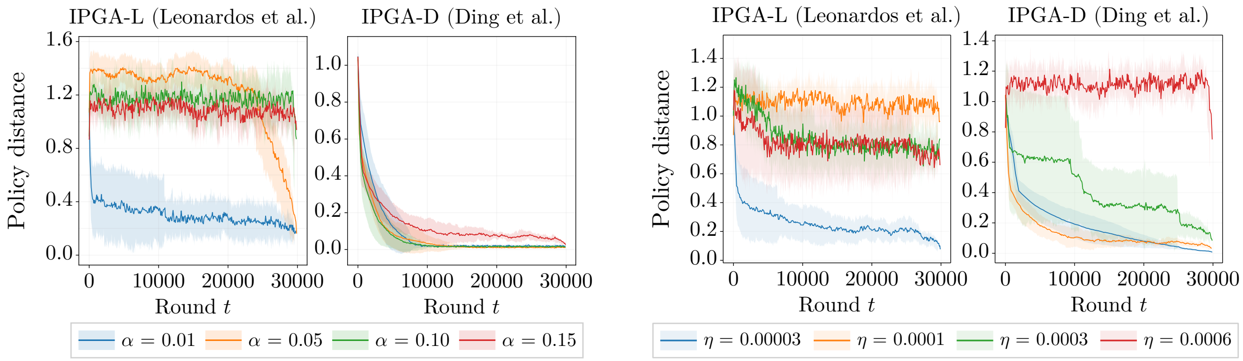

Figure 3: Comparison of IPGA-L and IPGA-D in the safe-distancing game varying the performativity strength α (left two plots, $\eta = 0.0001$), and learning rate η (right two plots, $\alpha = 0.15$). Mean and standard deviation over 10 experiment replications.

B.1.2 Independent Natural Policy Gradient (INPG)

We implement our own version of the INPG algorithm shown in Algorithm 2, based on the variant by Fox et al. (2022). We name this algorithm *INPG (unregularized)* in the plots. We name the log-barrier regularized version *INPG (regularized)*. We set the hyperparameters as shown in Table 3.

The comparison of performance for different performativity strengths and different learning rates in the safe-distancing game can be seen in Figure 5, in the stochastic congestion game in Figure 6.

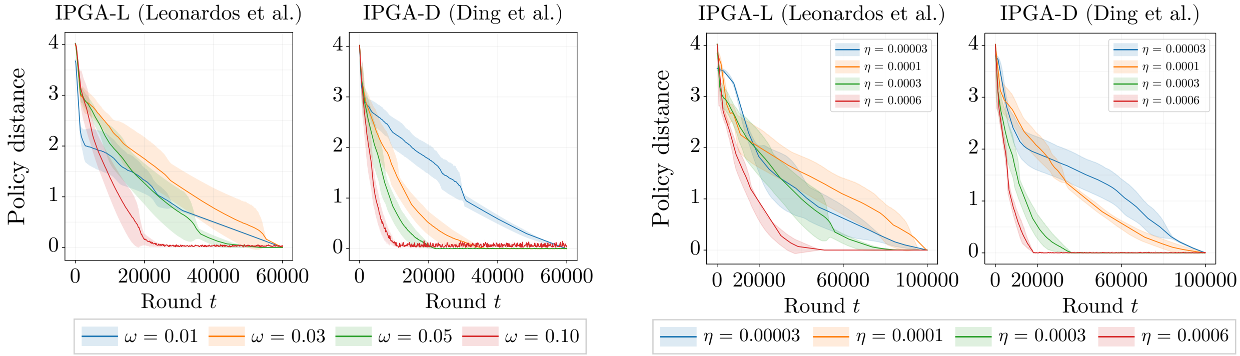

Figure 4: Comparison of IPGA-L and IPGA-D in the stochastic congestion game varying the performativity strength $\omega = \omega_r = \omega_p$ (left two plots, $\eta = 0.0003$), and learning rate η (right two plots, $\omega = 0.03$). Mean and standard deviation over 5 experiment replications.

Table 3: Hyperparameters used for the INPG algorithm (unless noted otherwise).

Parameter	**Value**
Learning rate η	0.0001
Discount factor γ	0.99
Number of episodes per round	20
Regularizer strength λ (only in regularized version)	0.003

Figure 7 shows the importance of selecting an appropriate learning rate for the PGA algorithms. Some learning rates are more stable for a larger set of performativity strengths α, as seen in the plot ($\eta = 0.00001$ for IPGA-L and $\eta = 0.0001$ for IPGA-D).

Algorithm 2 Independent Natural Policy Gradient Practical Implementation

```
Input: $\eta$ – step size, $T$ – number of rounds, $K$ – episodes per round
Init.: for all $i \in \mathcal{N}, a_i \in \mathcal{A}_i, s \in S$, let $\pi_i^t(a_i|s) = 1/|\mathcal{A}_i|$.
for $t = 1$ to $T$ do
    Roll out policies $\pi^t$ for $K$ episodes to get trajectories $\tau$
    Compute state visitation distribution $\mu(s)$ from trajectories $\tau$
    Compute value functions $V_i(s)$ and $Q_i(s,a)$ from trajectories $\tau$ for each agent $i \in \mathcal{N}$
    for $i = 1$ to $n$ (simultaneously) do
        Compute $A_i(s,a) = Q_i(s,a) - V_i(s) \quad \forall s \in S, a \in \mathcal{A}_i$
        Update $\pi_i^{t+1}(a|s) = \pi_i^t(a|s) \exp\left(\frac{\eta}{1-\gamma} A_i(s,a) + \frac{\lambda}{\mu(s)\pi_i^t(a|s)} - \frac{\lambda|\mathcal{A}_i|}{\mu(s)}\right) \frac{1}{Z} \quad \forall s \in S, a \in \mathcal{A}_i$
                                        ▷ $Z$ is the renormalization term. For INPG (unregularized) we set $\lambda = 0$.
    end for
end for
```

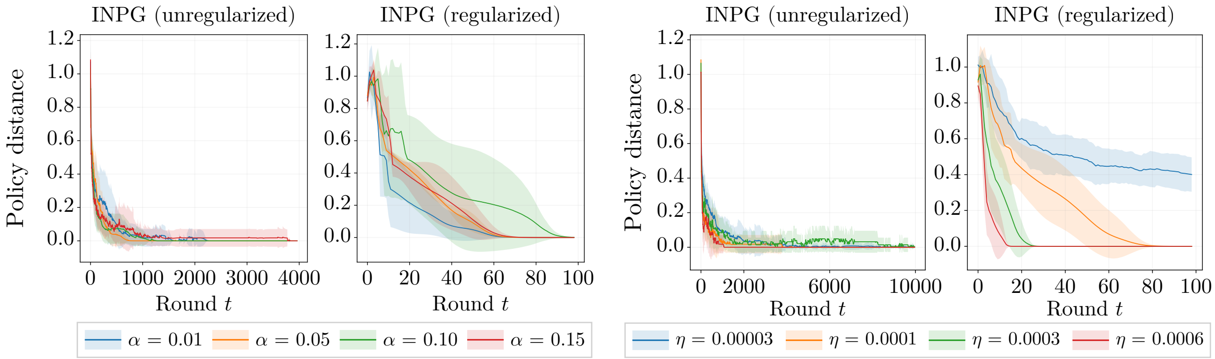

Figure 5: Comparison of INPG regularized vs. unregularized version in the safe-distancing game varying the performativity strength α (left two plots, $\eta = 0.0001$) and learning rate η (right two plots, $\alpha = 0.15$). Mean and standard deviation over 10 experiment replications. In the INPG (regularized) version with $\eta = 0.00003$ (blue line) in the rightmost plot converges after approx. 3000 rounds (not seen in the plot). (Note the stark difference in the number of rounds.)

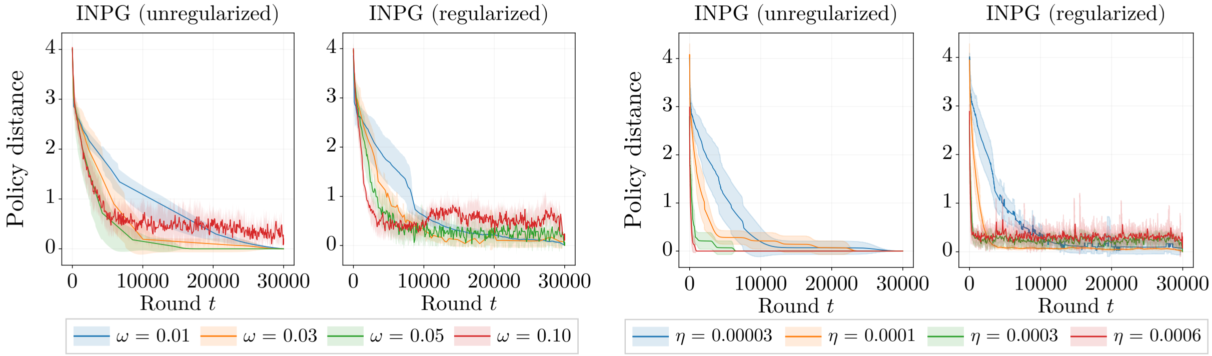

Figure 6: Comparison of INPG regularized vs. unregularized version in the stochastic congestion game varying the performativity strength $\omega = \omega_r = \omega_p$ (left two plots, $\eta = 0.00003$) and learning rate η (right two plots, $\omega = 0.03$). Mean and standard deviation over 5 experiment replications.

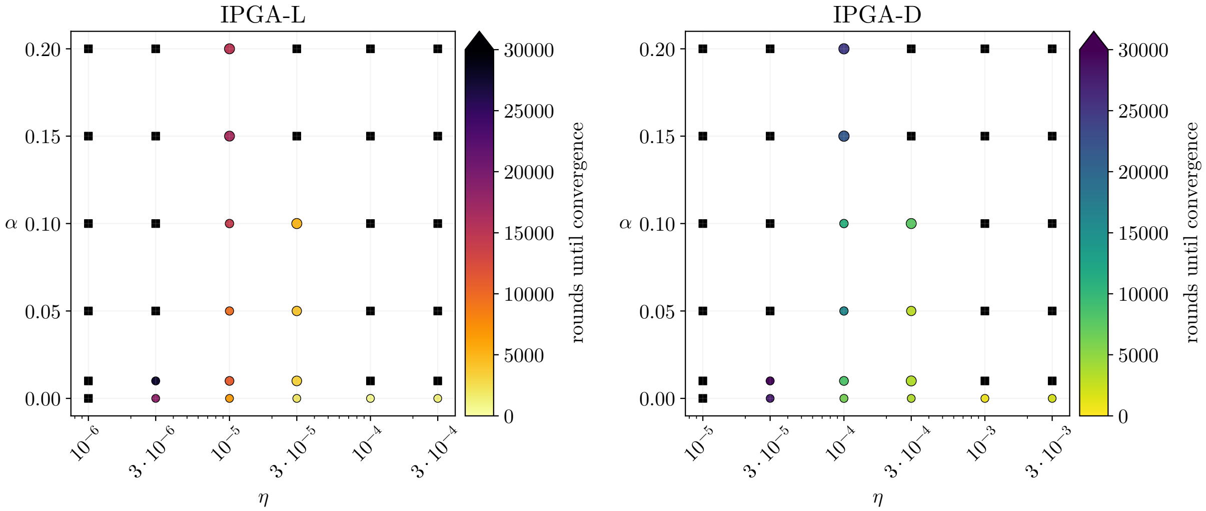

Figure 7: Comparison of different values of α (linear scale) and η (log scale) for IPGA-L and IPGA-D in the safe-distancing environment. Circles indicate the runs converged in under 30000 rounds with the number of rounds indicated by the color scale shown on the right. Black squares indicate the runs did not converge in 30000 rounds. Values shown are from the mean over 10 runs.

B.2 Environments

B.2.1 Safe-Distancing Game

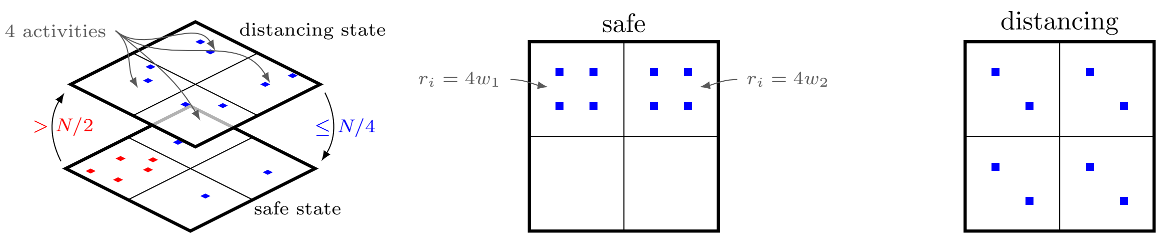

Figure 8: An illustration of the Safe-Distancing environment that is used in the experiments. **Left:** illustration of the two states. When more than $\frac{N}{2} = 4$ agents in the *safe* state are performing the same activity (5 red squares) all agents are transitioned to the *distancing* state, where they have to spread out, and no more than $\frac{N}{4} = 2$ agents may perform an activity to transition back to the *safe* state. **Center:** illustration of the optimal joint policy in the *safe* state (w_1 and w_2 being the two highest weighted rewards). **Right:** illustration of the optimal joint policy in the *distancing* state. Figure adapted from Leonardos et al. (2022).

We use one environment setup defined by Leonardos et al. (2022) – the safe-distancing game. We consider an MDP with two states, one state is called *safe*, the other – *distancing*. There are $N = 8$ agents, $|\mathcal{A}_i| = 4$ activities the agents can perform. In both states the reward each agent receives for performing activity $a_i = k$ is equal to a weight w_k multiplied with the number of agents performing that activity. The weights satisfy $w_1 > w_2 > w_3 > w_4$, i.e., activity 1 is the most preferable. If more than $\frac{N}{2} = 4$ agents are performing the same activity, all agents are transitioned to the *distancing* state. At the *distancing* state the reward weights are the same, except the reward is reduced by a (considerably large) constant $c = 100$. To transition back to the *safe* state the agents have to distribute themselves evenly among the activities, i.e., no more than $\frac{N}{4} = 2$ agents may perform the same activity. A visualization of the game and example policies can be seen in Figure 8. In our experiments we set $(w_1, w_2, w_3, w_4) = (4, 3, 2, 1)$.

Performative Effect. To model the performative response we modify the environment by taking inspiration from Mandal et al. (2023), and do as follows. Each agent is controlled by a principal agent (the learning algorithm), and an influencer agent. The influencer agent may override some of the actions taken by the principal agent. Therefore, the principal agent's effective environment is performative.

For example, the principal agent selects one of $|\mathcal{A}_i| = 4$ actions. The influencer agent may choose to keep this action the same, or intervene, by overriding the action and choosing a different activity. The influencer agent maintains Q-values of $|\mathcal{A}_i| + 1 = 5$ actions – four for intervening by changing into one of $|\mathcal{A}_i| = 4$ actions, and an additional one for no intervention. Parameter α controls the performative strength. With a probability of $1 - \alpha$ the original principal agent action is selected, and with probability α (e.g., $\alpha = 0.15$) the influencer agent action gets activated. Its Q-values are computed on a perturbed environment, as described by Mandal et al. (2023). The action selected by the influencer agent is sampled from:

$$\pi_2(a_i|s) = \frac{\exp(Q^{*|\pi_1}(s|a_i)}{\sum_j \exp(Q^{*|\pi_1}(s, a_j))}.$$

In our experiments, we use the default $\alpha = 0.15$, unless noted otherwise.

B.2.2 Stochastic Congestion Game

We use the experiment setup defined by Fox et al. (2022) – the stochastic congestion game. We consider an MDP with 5 states, $N = 4$ agents. The states and actions transition as shown in Figure 9. In every state each

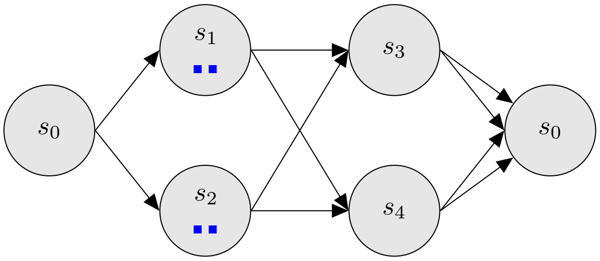

Figure 9: An illustration of the stochastic congestion game. From states s_3 and s_4 the game wraps around and state s_0 is reached. The agents (blue squares) in state s_1 can choose the same action (e.g., the edge to s_3), yielding a reward of 15 each, or can choose different actions (edges to s_3 and s_4), yielding a reward of 50 each. Figure adapted from Fox et al. (2022).

agent can perform one of $|\mathcal{A}_i| = 2$ actions. The reward received by each agent is based on the number of agents choosing the same action at that same state. In our experiments, if only one agent is choosing that action, the reward is $r_i = 50$, if two agents are choosing the same action $r_i = 15$, three $r_i = 5$, four $r_i = 1$. The agents start at state s_0.

Performative Effect. To model the performative effect in this environment, we change the rewards and transition probabilities as follows:

$$r_{i,\pi'} = r_{i,\pi_0} + \frac{\omega_r}{(1-\gamma)\sqrt{|\mathcal{S}|\,|\mathcal{A}_i|}}(\pi' - \pi_0)\,, \tag{20}$$

$$P_{i,\pi'} = P_{i,\pi_0} + \frac{\omega_p}{(1-\gamma)\sqrt{|\mathcal{S}|\,|\mathcal{A}_i|}}(\pi' - \pi_0)\frac{1}{|\mathcal{S}|}\,, \tag{21}$$

varying the strength via ω_r and ω_p. We set π_0 to the initial uniform-random policy.

We restrict the changes the transition kernel in $P_{i,\pi'}$ to be valid for the game, (for example, following the naming in Figure 9, in state s_2 the kernel is restricted to only transition to s_3 or to s_4, and not to, for example, s_0).

In our experiments, we use the defaults $\omega_r = \omega_p = 0.03$, unless noted otherwise.

B.3 Computing Infrastructure

We ran the experiments on an internal computing cluster with NVIDIA A100 80 GB GPUs. Running 10 experiment replications in parallel, a single round in the safe-distancing environment takes approx. 0.7 s, 10000 rounds – approx. 2 h. Running 5 experiment replications in parallel, a single round in the stochastic congestion game takes approx. 1.6 s, 10000 rounds – approx. 4.5 h.

C ADDITIONAL RELATED WORK

Performative Prediction. Since the seminal work of Performative Prediction (Perdomo et al., 2020), various adaptations has been studied, we refer to the survey by Hardt and Mendler-Dünner (2023). There are variations considering stochastic optimization (Mendler-Dünner et al., 2020) for finding performatively stable points. Performative power – a notion that measures the potential of a firm to influence the population distribution of participants on an online platform (Hardt et al., 2022), performative prediction with neural networks (Mofakhami et al., 2023) considers a setting, which allows weaker assumptions on the loss function, regret minimization with performative feedback (Jagadeesan et al., 2022) and the connection between performativity and causality has been studied by Mendler-Dünner et al. (2022); Kulynych (2022).

Multi-Agent Performative Prediction A recent line of work studies performative prediction in a multi-agent setting with slightly different frameworks, e.g., consensus seeking agents where an agent i has a local distribution that is not effected by the other agents' decisions (Li et al., 2022), smooth games where the local distributions are affected by the joint decision (Narang et al., 2023), global distributions (Piliouras and Yu, 2023). Among these, our setting is conceptually closest to Narang et al. (2023), which provide a game theoretic notion of multi-agent performative prediction. In this work, they provide methods to converge to a PSE in (stateless) strongly-monotone games.

Variations of MPGs. MPGs have also been considered in different variations, e.g., α-approximate MPGs (Guo et al., 2024), Networked MPGs (Zhou et al., 2023), fully decentralized settings (agents may not require to know if other agents exist) (Maheshwari et al., 2024). Another recent line of work considers constrained MPGs to study MARL under safety constraints (Alatur et al., 2024b; Jordan et al., 2024).

Non-stationary Multi-agent Reinforcement Learning. Our work is also related non-stationary MARL. There is work on full-information settings (e.g. gradients are known), see e.g., Cardoso et al. (2019); Anagnostides et al. (2023); Duvocelle et al. (2023), bandit feedback (gradient estimations are required) (Jiang et al., 2024).