Independent Natural Policy Gradient Always Converges in Markov Potential Games

Roy Fox
UC Irvine

Stephen McAleer
CMU

William Overman
UC Irvine

Ioannis Panageas
UC Irvine

Abstract

Natural policy gradient has emerged as one of the most successful algorithms for computing optimal policies in challenging Reinforcement Learning (RL) tasks, but very little was known about its convergence properties until recently. The picture becomes more blurry when it comes to multi-agent RL (MARL), where only few works have theoretical guarantees for convergence to Nash policies. In this paper, we focus on a particular class of multi-agent stochastic games called Markov Potential Games and prove that Independent Natural Policy Gradient always converges using constant learning rates. The proof deviates from existing approaches and overcomes the challenge that Markov potential Games do not have unique optimal values (as single-agent settings exhibit), leading different initializations to different limit point values. We complement our theoretical results with experiments that indicate that Natural Policy Gradient outperforms Policy Gradient in multi-state congestion games.

1 INTRODUCTION

Reinforcement learning has proven successful on a number of sequential decision making tasks (Mnih et al., 2015; Akkaya et al., 2019; Arulkumaran et al., 2017). However, much of this recent progress has been in single-agent domains. In multi-agent reinforcement learning domains, many of the assumptions that are useful for understanding single agent reinforcement learning do not hold. For example, from the perspective of a single agent, the MDP is not stationary over time as other agents change their policies. The complications arising in the multi-agent reinforcement learning setting make it much less understood than the single-agent setting.

Despite the difficulty of multi-agent reinforcement learning (RL), recent RL algorithms have been developed for the fully-competitive two-player setting (Lanctot et al., 2017; Heinrich and Silver, 2016; Brown et al., 2019; McAleer et al., 2020, 2021; Hennes et al., 2019; Daskalakis et al., 2020), achieving superhuman performance on StarCraft (Vinyals et al., 2019), Go and Chess (Silver et al., 2017), and Poker (Brown and Sandholm, 2018; Moravčík et al., 2017). Recent multi-agent RL algorithms have also achieved success on the multi-agent fully cooperative reward setting (Foerster et al., 2017; Lowe et al., 2017; Sunehag et al., 2017; Foerster et al., 2018; OroojlooyJadid and Hajinezhad, 2019; Kuba et al., 2021). Very recently, Markov Potential Games (MPGs) have emerged as a generalization of fully cooperative settings (Leonardos et al., 2021; Zhang et al., 2021; Mguni et al., 2021). A Markov game is a MPG if there exists a potential function such that if one agent changes their policy, the difference in their value function is the same as the difference in the potential function in all states. MPGs generalize normal-form potential games to multi-step Markov games and include many important examples such as routing games (described below) and cooperative games.

In multi-agent games, the most common solution concept is a Nash equilibrium policy (Nash et al., 1950). All agents are said to be in a Nash policy if no agent can improve their reward by deviating to another policy provided that the other agents do not deviate. Since Nash equilibria exist in every finite game, Nash equilibria serve as a desriptive solution concept, because if rational agents are not in a Nash equilibrium, they will change their policy. In two-player zero sum games and cooperative games, Nash equilibria also serve as a prescriptive solution concept, and are in a sense optimal.

Calculating Nash policies in MPGs is of interest for a variety of reasons. First, Nash policy outcomes can be compared with max-welfare outcomes to determine the price of anarchy, a measure of the inefficiency of

individual selfish behavior (Roughgarden, 2009). Second, MPGs have the property that pure Nash policies always exist, and can be found by *locally* maximizing the potential function. Therefore, MPGs allow easier analysis of equilibrium behaviour than is known for generic general-sum Markov Games.

A natural question to ask in this setting is whether independent learners can converge to a Nash policy in the last iterate. The only existing algorithm guaranteed to converge in MPGs in the last iterate is independent policy gradient (Leonardos et al., 2021; Zhang et al., 2021). Although independent policy gradient will converge to a Nash policy, it can converge very slowly. This is a common problem in single-agent policy gradient (PG) as well, where policies can often be on the boundary of the action-distribution simplex, giving a small gradient signal.

Conversely, Natural Policy Gradient (NPG) has been shown to converge much faster than PG in the single-agent setting because NPG is able to modify the gradient to induce a large update even when the gradient is very small (Agarwal et al., 2020). The main result of this paper is to show that when n agents independently run Natural Policy Gradient in a Markov Potential Game they will converge asymptotically to a Nash equilibrium. We also show experimentally that Independent Natural Policy Gradient indeed converges faster than Independent Policy Gradient in two separate settings.

If one is satisfied with time averaged convergence, then it has been shown that $O(\frac{1}{\varepsilon^2})$ convergence to correlated and coarse correlated equilibria can be achieved (Song et al., 2021; Jin et al., 2021). Obtaining results for last iterate convergence, which we do here, is a much more challenging task.

Our results. Our main technical result is summarized by the following informal statement of Theorem 3.1:

Theorem 1.1 (Informal). *Consider a Markov Potential Game in which all agents are updating their policies according to Independent Natural Policy Gradient. For small enough stepsize η, Independent Natural Policy Gradient exhibits last-iterate (asymptotic) convergence to Nash-equilibrium policies.*

Remark 1 (Stepsize needs to be small). The stepsize cannot be arbitrarily large, as it has been shown that the Multiplicative Weights Update Algorithm in potential games (non-sequential MPGs) exhibits chaotic behavior for large stepsizes (Palaiopanos et al., 2017). Please see Remark 2 that argues that Multiplicative Weights Update is effectively Natural Policy Gradient on the policy space.

Technical Overview. Our main technical result is to show that Indepedent Natural Policy Gradient (INPG) globally converges to equilibrium policies for a fixed stepsize in Markov Potential Games. The proof can be broken down into three main steps. The first step is to show that INPG is equivalent to Natural Gradient Ascent on the potential function Φ induced by the Markov Potential Game. Having shown this fact, in the second step we show that Φ is non-decreasing in each INPG iteration. This step is rather technical as it involves showing that the gradient of Φ is locally Lipschitz in the softmax parametrization, with respect to a Mahalanobis norm[1] with a positive diagonal conditioning matrix induced by the Fisher information matrix. The last step is to show that the limit points of Independent Natural Policy Gradient are indeed equilibrium policies. Note that in the softmax parametrization this does not follow immediately from the monotonic non-decrease of the potential. We prove this step by exploiting the structure of the Multiplicative Weights Update Algorithm, which is essentially the same as Natural Policy Gradient on the policy space (see Agarwal et al. (2020) and Remark 2).

2 PRELIMINARIES

Markov Decision Process (MDP). The notation we use is standard and largely follows Agarwal et al. (2020) and Leonardos et al. (2021). We consider a setting with n agents who select actions in a Markov Decision Process (MDP). A MDP is a tuple $(\mathcal{S}, N, \{\mathcal{A}_i, r_i\}_{i \in N}, P, \gamma, \mu)$ in which:

- $\mathcal{S}$ is a finite state space. We write $\Delta(\mathcal{S})$ to denote the set of all probability distributions over the set $\mathcal{S}$ and $|\mathcal{S}|$ for its cardinality.

- $N = \{1, 2, \ldots, n\}$ is the set of agents in the game.

- $\mathcal{A}_i$ denotes a finite set of actions for agent i. Using common conventions in Game Theory, we write $\mathcal{A} = \times_{i \in N} \mathcal{A}_i$ and $\mathcal{A}_{-i} = \times_{j \neq i} \mathcal{A}_j$ to denote the joint action space of all agents and of all agents but i, respectively. We also have that $\mathbf{a} = (a_i, \mathbf{a_{-i}}) \in \mathcal{A}$ in which $a_i \in \mathcal{A}_i$ and $\mathbf{a}_{-i} \in \mathcal{A}_{-i}$.

- $r_i : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the individual reward function of agent $i \in N$, i.e., $r_i(s, a_i, \mathbf{a}_{-i})$ is the instantaneous reward of agent i when agent i takes action a_i and the other agents choose $\mathbf{a}_{-i}$ at state $s \in \mathcal{S}$.

- P is the transition probability matrix, for which $P(s' \mid s, \mathbf{a})$ is the probability of transitioning from s to s' given that agents choose joint action $\mathbf{a} \in \mathcal{A}$.

---

[1]For a positive definite matrix A, the Mahalanobis norm is $\|x\|_A = \sqrt{x^\top A x}$.

- $\gamma \in [0, 1)$ is a discount factor, same for all agents. For finite-horizon MDPs we take $\gamma = 1$ and in the proofs we substitute $\frac{1}{1-\gamma}$ with the horizon.

- $\mu \in \Delta(\mathcal{S})$ is the distribution of the initial state.

Policies and Value Functions. For each agent $i \in N$, a deterministic, stationary policy $\pi_i : \mathcal{S} \to \mathcal{A}_i$ specifies the action of agent i at each state $s \in \mathcal{S}$. A stochastic policy $\pi_i : \mathcal{S} \to \Delta(\mathcal{A}_i)$ denotes a probability distribution over the actions of agent i for each state $s \in \mathcal{S}$. We denote by $\Delta(\mathcal{A}_i)^{|\mathcal{S}|}$, $\Delta(\mathcal{A})^{|\mathcal{S}|}$, and $\Delta(\mathcal{A}_{-i})^{|\mathcal{S}|}$ the set of all stochastic policies for agent i, the set of all joint (product distribution) stochastic policies for all agents, and for all agents but i, respectively. A joint policy π induces a distribution over trajectories $\tau = (s_t, \mathbf{a}_t, \mathbf{r}_t)_{t \geq 0}$, where s_0 is drawn from the initial state distribution μ, $a_{i,t}$ is drawn from $\pi_i(\cdot \mid s_t)$ for all agents $i \in N$ and $\mathbf{r}_t$ depends on $s_t, \mathbf{a}_t$.

In the multi-agent setting, each agent is aiming to maximize her respective value function. The value function $V_i^\pi(s) : \Delta(\mathcal{A})^{|\mathcal{S}|} \to \mathbb{R}$ gives the expected reward of agent $i \in N$ when the MDP starts from $s_0 = s$ and the agents use joint policy π

$$V_i^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, \mathbf{a}_t) \,\middle|\, s_0 = s\right]. \qquad (1)$$

We also denote $V_i^\pi(\mu) = \mathbb{E}_{s \sim \mu}[V_i^\pi(s)]$ if the initial state follows distribution is μ. Analogously, one can define for each agent i the Q-function $Q_i^\pi(s, \mathbf{a}) = r_i(s, \mathbf{a}) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, \mathbf{a})}[V_i^\pi(s')]$ and advantage function $A_i^\pi(s, \mathbf{a}) = Q_i^\pi(s, \mathbf{a}) - V_i^\pi(s)$. The solution concept that we will be focusing on in this paper is the Nash equilibrium joint policy.

Definition 1 (Equilibrium Joint Policy). A joint policy, $\pi^* = (\pi_i^*)_{i \in N} \in \Delta(\mathcal{A})^{|\mathcal{S}|}$, is an equilibrium policy if for each agent i,

$$V_i^{(\pi_i^*, \pi_{-i}^*)}(s) \geq V_i^{(\pi_i, \pi_{-i}^*)}(s),$$

for all $\pi_i \in \Delta(\mathcal{A}_i)^{|\mathcal{S}|}$, and all $s \in \mathcal{S}$, i.e., if agent i's policy π_i^* maximizes her value function for each starting state $s \in \mathcal{S}$ given the joint policy of the other agents π_{-i}^*, and this is true for all agents i. Similarly, a joint policy $\pi^* = (\pi_i^*)_{i \in N}$ is an ϵ-equilibrium policy if there exists an $\epsilon > 0$ so that for each agent i

$$V_i^{(\pi_i^*, \pi_{-i}^*)}(s) \geq V_i^{(\pi_i, \pi_{-i}^*)}(s) - \epsilon,$$

for all $\pi_i \in \Delta(\mathcal{A}_i)^{|\mathcal{S}|}$ and all $s \in \mathcal{S}$.

Markov Potential Games. For the rest of the paper we will focus on a particular class of multi-agent MDPs called Markov Potential Games.

Definition 2 (Markov Potential Game (Mguni, 2020; Mguni et al., 2021; Leonardos et al., 2021; Zhang et al., 2021)). A Markov Decision Process is called a *Markov Potential Game (MPG)* if there exists a (state-dependent) function $\Phi^\pi(s) : \Delta(\mathcal{A})^{|\mathcal{S}|} \to \mathbb{R}$ so that

$$\Phi^{(\pi_i, \pi_{-i})}(s) - \Phi^{(\pi_i', \pi_{-i})}(s) = V_i^{(\pi_i, \pi_{-i})}(s) - V_i^{(\pi_i', \pi_{-i})}(s), \qquad (2)$$

for all agents $i \in N$, all states $s \in \mathcal{S}$ and all policies $\pi_i, \pi_i' \in \Delta(\mathcal{A}_i)^{|\mathcal{S}|}$ and $\pi_{-i} \in \Delta(A_{-i})^{|\mathcal{S}|}$. By linearity of expectation, the same is true if $s \sim \mu$

$$\Phi^{(\pi_i, \pi_{-i})}(\mu) - \Phi^{(\pi_i', \pi_{-i})}(\mu) = V_i^{(\pi_i, \pi_{-i})}(\mu) - V_i^{(\pi_i', \pi_{-i})}(\mu),$$

where $\Phi^\pi(\mu) := \mathbb{E}_{s \sim \mu}[\Phi^\pi(s)]$.

We conclude this subsection with the definition of the mismatch coefficient.

Discounted State Distribution. It is useful to define the discounted state visitation distribution $d_{s_0}^\pi(s)$ for $s \in \mathcal{S}$ that is induced by a policy π

$$d_{s_0}^\pi(s) := (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathrm{Pr}^\pi(s_t = s \mid s_0), \qquad \text{for all } s \in \mathcal{S}. \qquad (3)$$

We write $d_\mu^\pi(s) = \mathbb{E}_{s_0 \sim \mu}[d_{s_0}^\pi(s)]$ to denote the discounted state visitation distribution when the initial state distribution is μ.

Definition 3 (Distribution Mismatch coefficient (Agarwal et al., 2020)). Let μ be any distribution in $\Delta(\mathcal{S})$ and let $\mathcal{O}$ be the set of policies $\pi \in \Delta(\mathcal{A})^\mathcal{S}$. We call

$$M := \max_{\pi, \tilde{\pi} \in \mathcal{O}} \left\| \frac{d_\mu^\pi}{d_\mu^{\tilde{\pi}}} \right\|_\infty$$

the *distribution mismatch coefficient*, where d_μ^π is the discounted state distribution (3).

Example 1 (**Stochastic Congestion Games**). We are given a directed acyclic graph G, with a source s and a destination node t and N agents (Figure 1). Every state of the finite-horizon MDP consists of the positions of all agents on the graph. The MDP terminates when they all reach t. Each agent chooses an action (an edge e which is the node to go next) and the cost he experiences is a function of the load l_e of the edge e that he chooses $c_e(l_e)$, where the load is the total number of agents choosing the same edge. This setting is a MPG with the following potential function. Suppose we are at state s and let $\pi := (\pi_1, ..., \pi_N)$, then we have

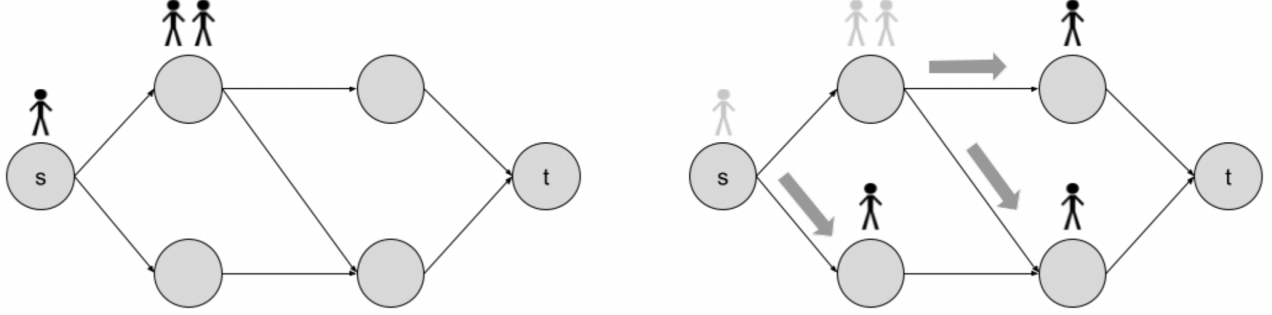$$\Phi^\pi(s) = \sum_{e \text{ reachable from } s} \mathbb{E}_\pi\left[\sum_{k=1}^{l_e} c_e(k)\right].$$

Figure 1: Two different configurations of the stochastic congestion game on 6 vertices. These correspond to two different states in the associated MDP. Arrows in the right figure show the actions chosen by the 3 agents to move from the MDP state on the left to MDP state on the right.

Consider any player i. With l_e^{-i} the number of agents other than i choosing edge e, we have that

$$
\begin{aligned}
\Phi^{(\pi_i, \pi_{-i})}(s) &= \sum_{e \text{ reachable from } s} \pi_i(e|s) \mathbb{E}_{\pi_{-i}} \left[\sum_{k=1}^{l_e^{-i}+1} c_e(k) \right] \\
&\quad + (1 - \pi_i(e|s)) \mathbb{E}_{\pi_{-i}} \left[\sum_{k=1}^{l_e^{-i}} c_e(k) \right] \\
&= \sum_{e \text{ reachable from } s} \pi_i(e|s) \mathbb{E}_{\pi_{-i}} \left[c_e(l_e + 1) \right] \\
&\quad + \sum_{e \text{ reachable from } s} \mathbb{E}_{\pi_{-i}} \left[\sum_{k=1}^{l_e} c_e(k) \right].
\end{aligned}
$$

So if player i deviates from π_i then the change in the potential will equal the change in their individual utility since the common term in the second sum cancels as it does not depend on π_i. We refer to this illustrative class of MPGs zas stochastic congestion games.

2.1 Natural Policy Gradient

Softmax parametrization. We assume that each agent i is represented by its logits — a vector θ_i of size $\mathcal{S} \times \mathcal{A}_i$ such that the policy is given by

$$
\pi_{\theta_i}(a|s) = \frac{\exp(\theta_{i,s,a})}{\sum_{a' \in \mathcal{A}_i} \exp(\theta_{i,s,a'})} \text{ for all } a \in \mathcal{A}_i, s \in \mathcal{S}. \tag{4}
$$

Independent Natural Policy Gradient (INPG). Given the softmax parametrization, we can define Independent Natural Policy Gradient:

$$
\theta_i^{(t+1)} = \theta_i^{(t)} + \eta F_{i,\mu}^+(\theta^{(t)}) \nabla_{\theta_i} V_i^{\pi_{\theta^{(t)}}} \text{ for each agent } i, \tag{5}
$$

where $F_{i,\mu}^+(\theta)$ is the pseudo-inverse of the Fischer matrix

$$
\begin{aligned}
F_{i,\mu}(\theta) = \mathbb{E}_{s \sim d_\mu^{\pi_\theta}} \mathbb{E}_{a_i \sim \pi_{\theta_i}(.|s)} \\
\left[\nabla_{\theta_i} \log \pi_{\theta_i}(a|s) \nabla_{\theta_i} \log \pi_{\theta_i}(a|s)^\top \right]
\end{aligned}
$$

and η is the step-size.

NPG equations can be simplified using the advantage function:

Lemma 2.1 (Simplified equations of INPG (Agarwal et al., 2020)). *For each agent i, state s, and action a, we have that*

$$
\theta_{i,s,a}^{(t+1)} = \theta_{i,s,a}^{(t)} + \frac{\eta}{1-\gamma} A_i^{(t)}(s,a), \tag{INPG}
$$

where $A_i^\theta(s,a) = \mathbb{E}_{\mathbf{a}_{-i} \sim \pi_{\theta_{-i}}}[A_i^{\pi_\theta}(s, a, \mathbf{a}_{-i})]$ is the expectation of the advantage function of i over the actions of all agents but i who use joint policy $\pi_{\theta_{-i}}$, and $A_i^{(t)}$ is shorthand for $A_i^{\theta^{(t)}}$.

Remark 2. If we make a change of variables from θ to probabilities π, the equations of INPG become the standard Multiplicative Weights Update algorithm (MWU), that is

$$
\pi_i^{(t+1)}(a|s) = \pi_i^{(t)}(a|s) \frac{\exp(\frac{\eta}{1-\gamma} A_i^\pi(s,a))}{Z_{i,s}^{(t)}}, \quad \text{(MWU)}
$$

where $Z_{i,s}^{(t)}$ is the renormalization term. Observe that the fixed points π of (MWU) satisfy either $\pi_i(a|s) = 0$ or $A_i^\pi(s,a) = 0$ for all $a \in \mathcal{A}_i, s \in \mathcal{S}$ and agents i.

3 CONVERGENCE RESULTS

In this section we prove that INPG converges. The main theorem is stated formally as follows:

Theorem 3.1. *Given a Markov Potential Game, assume that all the agents use Independent Natural Policy Gradient (INPG) with stepsize $\eta < \frac{(1-\gamma)^3}{27n^2 A_{\max}^2 M}$ and that the fixed points of (MWU) are isolated. It holds that the dynamics (INPG) converge pointwise (last iterate convergence) to equilibrium policies.*

The first step towards our result is to show that INPG is equivalent to running Natural Policy Gradient (NPG) on the potential function.

Lemma 3.2. *We consider Natural Policy Gradient dynamics on function Φ^{π_θ}, that gives*

$$\theta^{(t+1)} = \theta^{(t)} + \eta F_\mu^+(\theta^{(t)})\nabla_\theta \Phi^{\pi_{\theta^{(t)}}} \qquad (6)$$

where $F_\mu^+(\theta)$ is the pseudo-inverse of the Fischer matrix

$$F_\mu(\theta) = \mathbb{E}_{s\sim d_\mu^{\pi_\theta}} \mathbb{E}_{\mathbf{a}\sim\pi_\theta(.|s)}$$
$$\left[\nabla_\theta \log \pi_\theta(\mathbf{a}|s)\nabla_\theta \log \pi_\theta(\mathbf{a}|s)^\top\right].$$

It is true that (INPG) is the same dynamics as (6).

Proof. The first fact we are going to use is that

$$\nabla_{\theta_i}\Phi^{\pi_\theta} = \nabla_{\theta_i}V_i^{\pi_\theta}. \qquad (7)$$

This is true since $\nabla_{\pi_{\theta_i}}\Phi^{\pi_\theta} = \nabla_{\pi_{\theta_i}}V_i^{\pi_\theta}$ (Proposition B.1 in Leonardos et al. (2021)) and by applying the chain rule. Moreover, since π_θ is a product distribution, we have

$$\log \pi_\theta(\mathbf{a}|s) = \sum_i \log \pi_{\theta_i}(a_i|s), \qquad (8)$$

where $\mathbf{a} = (a_1, ..., a_n)$. Assuming $i \neq j$, we have

$$\mathbb{E}_{\mathbf{a}\sim\pi_\theta}[\nabla_\theta \pi_{\theta_i}(a_i|s)\nabla_\theta \pi_{\theta_j}(a_j|s)^\top] =$$
$$\mathbb{E}_{\mathbf{a}\sim\pi_\theta}[\nabla_{\theta_i}\pi_{\theta_i}(a_i|s)]\mathbb{E}_{\mathbf{a}\sim\pi_\theta}\left[\nabla_\theta\pi_{\theta_j}(a_j|s)\right]^\top = 0, \qquad (9)$$

where the first equality comes from independence and the last equality from the fact that the expectation of the derivative of the log-likelihood is zero. We conclude that

$$F_\mu(\theta) = \mathbb{E}_{s\sim d_\mu^{\pi_\theta}} \mathbb{E}_{\mathbf{a}\sim\pi_\theta(.|s)}$$
$$\left[\nabla_\theta \log \pi_\theta(a|s)\nabla_\theta \log \pi_\theta(a|s)^\top\right]$$
$$\stackrel{(8),(9)}{=} \sum_{i\in N} \mathbb{E}_{s\sim d_\mu^{\pi_\theta}} \mathbb{E}_{\mathbf{a}\sim\pi_\theta(.|s)}$$
$$\left[\nabla_\theta \log \pi_{\theta_i}(a_i|s)\nabla_\theta \log \pi_{\theta_i}(a_i|s)^\top\right]$$
$$\stackrel{\text{prod.distr.}}{=} \sum_{i\in N} \mathbb{E}_{s\sim d_\mu^{\pi_\theta}} \mathbb{E}_{\mathbf{a}_i\sim\pi_{\theta_i}(.|s)}$$
$$\left[\nabla_\theta \log \pi_{\theta_i}(a_i|s)\nabla_\theta \log \pi_{\theta_i}(a_i|s)^\top\right],$$

i.e., the Fisher matrix $F_\mu(\theta)$ is block-diagonal with each block being the Fisher matrix of each individual

player. Finally, from Lemma 5.1 in (Agarwal et al., 2020) it holds that

$$F_{i,\mu}^+(\theta)\nabla_{\theta_i}V_i^{\pi_\theta} = \frac{1}{1-\gamma}\mathbb{E}_{\mathbf{a}_{-i}\sim\pi_{\theta_{-i}}}A_i^{\pi_\theta}(s, a, \mathbf{a}_{-i}). \quad (10)$$

The proof is now concluded by (7), (10), and the fact that the pseudo-inverse of a block-diagonal matrix is block-diagonal with blocks that are the pseudo-inverse of each block of the original matrix. $\square$

With Lemma 3.2 established, we next show that the potential function is non-decreasing along the dynamics (INPG). This requires the following proposition, the proof of which can be found in the appendix:

Proposition 3.3 (Smoothness — Mahalanobis)**.** *Let $\theta, \tilde{\theta}$ be such that $\left\|\theta - \tilde{\theta}\right\|_\infty \leq \frac{\eta}{1-\gamma}$. There exists a constant $L = \frac{27n^2 A_{\max}^2 M}{(1-\gamma)^3}$ so that*

$$-\Phi^{\pi_{\tilde{\theta}}}(\mu) \leq -\Phi^{\pi_\theta}(\mu) + \langle -\nabla_\theta\Phi^{\pi_\theta}(\mu), \tilde{\theta} - \theta\rangle +$$
$$\frac{L}{2}\left\|\tilde{\theta} - \theta\right\|_{D^\theta}^2,$$

where D^θ is a diagonal matrix of size $(\sum_i |S| \cdot |A_i|) \times (\sum_i |S| \cdot |A_i|)$ with diagonal entry $(i, s, a), (i, s, a)$ to be $d_\mu^{\pi_\theta}(s)\pi_{\theta_i}(a|s)$ and $A_{\max} = \max_{i\in N}|A_i|$.

Lemma 3.4 ($\Phi^{(t)}$ is non-decreasing)**.** *Assume that the agents use the dynamics (INPG). It holds that*

$$\Phi^{(t+1)}(\mu) \geq \Phi^{(t)}(\mu) \text{ for all } t. \qquad (11)$$

Proof. Using Lemma 3.2, it suffices to focus on the dynamics (6), i.e., we will analyze Natural Policy Gradient on the potential function $\Phi^{\pi_\theta}(\mu)$. Using Proposition B1 of (Leonardos et al., 2021), it holds for agent i that,

$$\frac{\partial \Phi^{\pi_\theta}}{\partial \theta_{i,s,a}} = \frac{\partial V_i^{\pi_\theta}}{\partial \theta_{i,s,a}}$$
$$= d_\mu^{\pi_\theta}(s)\pi_{\theta_i}(a|s)\mathbb{E}_{\mathbf{a}_{-i}\sim\pi_{\theta_{-i}}}A_i^{\pi_\theta}(s, a, \mathbf{a}_{-i}),$$

hence by defining $D^{(t)}$ to be a diagonal matrix of size $(\sum_i |\mathcal{S}| \times |\mathcal{A}_i|)^2$ with diagonal entry $(i, s, a_i), (i, s, a_i)$ to be $d_\mu^{\pi_{\theta^{(t)}}}(s)\pi_{\theta_i^{(t)}}(a_i|s)$ we have that NPG on Φ is the same as

$$\theta^{(t+1)} = \theta^{(t)} + \eta\left(D^{(t)}\right)^{-1}\nabla_\theta\Phi^{(t)}(\mu). \qquad (12)$$

Since all rewards are in $[0, 1]$ it follows that $\left\|\theta^{(t+1)} - \theta^{(t)}\right\|_\infty \leq \frac{\eta}{1-\gamma}$. Given Proposition 3.3, it follows that

$$-\Phi^{(t+1)}(\mu) \leq -\Phi^{(t)}(\mu) + \langle -\nabla_\theta\Phi^{(t)}(\mu), \theta^{(t+1)} - \theta^{(t)}\rangle +$$
$$\frac{L}{2}\left\|\theta^{(t+1)} - \theta^{(t)}\right\|_{D^{(t)}}^2$$
$$= -\Phi^{(t)}(\mu) - \eta\left\|\nabla_\theta\Phi^{(t)}(\mu)\right\|_{\left(D^{(t)}\right)^{-1}}^2 +$$
$$\frac{\eta^2 L}{2}\left\|\nabla_\theta\Phi^{(t)}(\mu)\right\|_{\left(D^{(t)}\right)^{-1}}^2.$$

Therefore, by choosing $\eta < \frac{1}{L}$ we have that

$$\Phi^{(t+1)} \geq \Phi^{(t)} + \frac{1}{2L} \left\| \nabla_\theta \Phi^{(t)}(\mu) \right\|^2_{(D^{(t)})^{-1}} > \Phi^{(t)}, \quad (13)$$

and the proof is complete. $\qquad\square$

From Lemma 3.4 and the fact that Φ is defined on a compact domain and is continuous (hence attains a maximum) we conclude that if $\eta < \frac{(1-\gamma)^3}{27n^2 A_{\max}^2 M}$ then $\Phi^{(t)}$ converges to a limit $\Phi^{(\infty)}$. We shall show that the corresponding policies are indeed equilibrium policies. First, we prove that the limit $\Phi^{(\infty)}$ is a fixed point of the Independent Natural Policy Gradient dynamics.

Lemma 3.5 (Convergence to fixed points). *INPG dynamics (INPG) converges to fixed points.*

Proof. Let $\Omega \subset \Delta(\mathcal{A})^{|\mathcal{S}|}$ be the set of limit points of $\pi_{\theta^{(t)}}$. $\Phi^{(t)}$ is increasing with respect to time t by Lemma 3.4 and so, because Φ is bounded on $\Delta(\mathcal{A})^{|\mathcal{S}|}$, $\Phi^{(t)}$ converges as $t \to \infty$ to $\Phi^{(\infty)} = \inf_t \{\Phi^{(t)}\}$. By continuity of Φ we get that $\Phi^q = \lim_{t\to\infty} \Phi^{q_{\theta(t)}} = \Phi^{(\infty)}$ for all $q \in \Omega$. So Φ is constant on Ω. Also $q^{(t)} = \lim_{n\to\infty} q^{(t_n+t)}$ as $n \to \infty$ for some sequence of times $\{t_i\}$ and so $q^{(t)}$ lies in Ω, i.e. Ω is invariant. Thus, if $q \equiv q^{(0)} \in \Omega$ the trajectory $q^{(t)}$ lies in Ω and so $\Phi^{q^{(t)}} = \Phi^\infty$ on the trajectory. But Φ is strictly increasing except on fixed points and so Ω consists entirely of fixed points. $\quad\square$

We conclude the section by showing that if $\pi^{(0)}$ is in the interior of $\Delta(\mathcal{A})^{|\mathcal{S}|}$ (all θ are bounded at initialization) then we have convergence to equilibrium policies.

Lemma 3.6 (Convergence to equilibrium policies). *Assume that the fixed points of (INPG) are isolated. Let $\pi^{(0)}$ be a point in the interior of $\Delta(\mathcal{A})^{|\mathcal{S}|}$. It follows that $\lim_{t\to\infty} \pi^{(t)} = \pi^{(\infty)}$ is an equilibirum policy.*

Proof. We showed in Lemma 3.5 that INPG dynamics (INPG) converges, hence $\lim_{t\to\infty} \pi^{(t)}$ exists (under the assumption that the fixed points are isolated) and is equal to a fixed point of the dynamics $\pi^{(\infty)}$. Also it is clear from the dynamics that $\Delta(\mathcal{A})^{|\mathcal{S}|}$ is invariant, i.e., for all $t \geq 0$, we have $\sum_{a\in\mathcal{A}_i} \pi_i^{(t)}(a|s) = 1$ for all $s \in \mathcal{S}$, $i \in N$ and $\pi_i^{(t)}(a|s) > 0$ for all $a \in \mathcal{A}_i, s \in \mathcal{S}$, and $i \in N$ (since $\pi^{(0)}$ is in the interior of $\Delta(\mathcal{A})^{|\mathcal{S}|}$).

Assume that $\pi^{(\infty)}$ is not an equilibrium policy, then there exists an agent i, a state s and an action $a \in \mathcal{A}_i$ so that $A_i^{\pi^{(\infty)}}(s,a) > 0$ and $\pi_i^{(\infty)}(a|s) = 0$. Fix a $\zeta > 0$ and let $U_\zeta = \{\pi : \frac{\eta A_i^\pi(s,a)}{1-\gamma} > \zeta + \log Z_{i,s}^\pi\}$. By continuity we have that U_ζ is open. It is also true that $\pi^{(\infty)} \in U_\zeta$ for ζ small enough since $Z_{i,s}^{\pi^{(\infty)}} = 1$.

Since $\pi^{(t)}$ converges to $\pi^{(\infty)}$ as $t \to \infty$, there exists a time t_0 so that for all $t' \geq t_0$ we have that $\pi^{(t')} \in U_\zeta$.

However, from INPG dynamics (INPG) we get that if $\pi^{(t')} \in U_\zeta$ then $e^{\frac{\eta A_i^{(t')}(s,a)}{1-\gamma}} \geq e^\zeta Z_{i,s}^{(t')} \geq Z_{i,s}^{(t')}$. Hence

$$\pi_i^{(t'+1)}(a|s) = \pi_i^{(t')}(a|s) \frac{\exp\left(\frac{\eta A_i^{(t)}(s,a)}{1-\gamma}\right)}{Z_{i,s}^{(t')}} \geq \pi_i^{(t')}(a|s) > 0,$$

i.e., $\pi_i^{(t')}(a|s)$ is positive and increasing with $t' \geq t_0$. We reached a contradiction since $\pi_i^{(t)}(a|s) \to \pi_i^{(\infty)}(a|s) = 0$, thus $\pi^{(\infty)}$ is an equilibrium policy. $\quad\square$

The proof of Theorem 3.1 follows by Lemmas 3.5 and 3.6.

4 EXPERIMENTS

In this section we compare the empirical performance of Independent Natural Policy Gradient (INPG) to Independent Policy Gradient (IPG) in both the Stochastic Congestion Game (SCG) environment described in section 2 and the environment introduced by Leonardos et al. (2021), which we will call the *distancing game*. We compare INPG to IPG in the distancing game setting to show empirical improvement over previous results, but also include this new SCG setting for several reasons. The distancing game considered in Leonardos et al. (2021) had only 2 states and the required number of iterations for the experiments to empirically converge (for policies to not change by more than $1e-16$ between iterates) was far less than theoretically guaranteed. The SCG setting we use, which has more states yet less agents, required many more iterations of IPG, demonstrating a significant gap in the empirically observed convergence rates of IPG and INPG. Experiments were run on a 2021 MacBook Air.

4.1 Distancing Game

For completeness, we briefly reproduce the description of the distancing game as given in Leonardos et al. (2021).

Experimental setup: We consider a MDP in which every state is a congestion game (Rosenthal, 1973). There are 8 agents, 4 facilities that the agents can select from and 2 states: a *safe* state and a *spread* state. In both states, the agents prefer to share their facility with as many other agents as possible. Specifically, the reward of each agent for choosing the action corresponding to facility k is equal to a positive weight w_k^{safe} times the number of agents at $k = A, B, C, D$. The weights satisfy $w_D^{\text{safe}} > w_C^{\text{safe}} > w_B^{\text{safe}} > w_A^{\text{safe}}$, i.e., facility D is the most preferred facility by all agents.

However, if more than 4 agents choose the same facility, then the MDP transitions to the spread state. The weight function is the same for all agents at the spread state as well, but the reward is reduced by a large constant amount $c > 0$. To return to the safe state, the agents need to maximally spread out amongst the facilities, that is, no more than 2 agents may be in the same facility.

Independent Policy Gradient Parameters: We perform episodic updates with $T = 20$ steps. At each iteration, we estimate the Q-functions, the value function, the discounted visitation distributions and, hence, the policy gradients using the average of mini-batches of size 20. We use discount factor $\gamma = 0.99$ and learning rate $\eta = 0.0001$. Empirical convergence was determined to have occurred if the L1 distance between the consecutive policies of each agent was less than $10e - 16$.

Independent Natural Policy Gradient Parameters: We use the same parameters for the INPG implementation as the IPG implementation. Specifically, we perform episodic updates with $T = 20$ steps. We estimate the Q-functions and and value function at each iteration, which allows us to obtain an estimate for the advantage function that is used in the INPG update. This is done using mini-batches of size 20. Again the discount factor is $\gamma = 0.99$ and the learning rate is $\eta = 0.0001$. Empirical convergence was determined to have occurred if the L1 distance between the consecutive policies of each agent was less than $10e - 16$.

Results: Figure 2 and Figure 3 depict the L1-accuracy in the policy space at each iteration, i.e., L1-accuracy $= \frac{1}{N} \sum_{i \in \mathcal{N}} |\pi_i - \pi_i^{\text{final}}| = \frac{1}{N} \sum_{i \in \mathcal{N}} \sum_s \sum_a |\pi_i(a \mid s) - \pi_i^{\text{final}}(a \mid s)|$. This is the average distance between the current policy and the final policy of all agents.

We can see from Figure 2 and Figure 3 that on average and in almost every run, Independent Natural Policy Gradient converges faster than Independent Policy Gradient. This matches our intuitive expectation that NPG methods should require fewer iterations than PG methods to achieve convergence (Agarwal et al., 2020).

4.2 Stochastic Congestion Game

We now compare Independent Policy Gradient and Independent Natural Policy Gradient for the environment described in Example 1.

Experimental setup: We consider a MDP where each state corresponds to a configuration of n agents
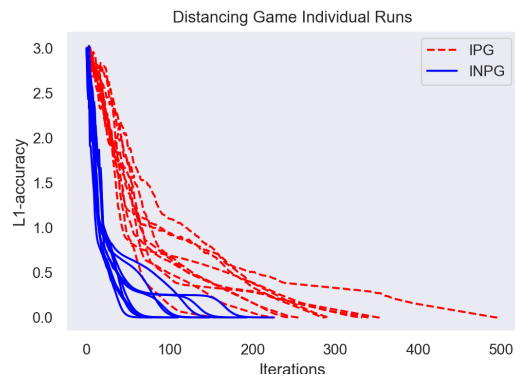

Figure 2: Trajectories of the L1-accuracy for 10 runs each of Independent Policy Gradient and Independent Natural Policy Gradient in the distancing game environment with 8 agents and learning rate of $\eta = .0001$.
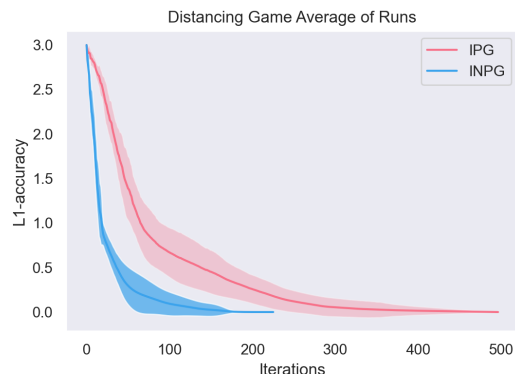

Figure 3: Mean L1-accuracy with shaded region of one standard deviation for the trajectories of Figure 2.

on a graph such as in Figure 1, except that the internal layers are fully connected. The graph is directed and the edges only move from left to right. It is important to note that the states of the MDP are not the vertices of this graph but the entire configuration, which is captured by the location of each agent. Thus there are $|V|^n$ possible states in the MDP where $|V|$ is the number of vertices in the DAG.

The actions of each agent at a given state of the MDP are to choose the edge from their current, corresponding vertex in the DAG that they want to move along. They then receive a reward inversely proportional to the number of agents who chose this edge and the MDP transitions to the next configuration state. If the agents reach t we can either send them along a constant reward edge back to s or consider the episode to be terminated.

The specific DAG on which the experiments were run

was with 2 internal fully-connected layers of 2 vertices each, giving the graph 6 total vertices consisting of s, t, and four internal vertices. Thus at each state each agent has two possible actions to choose from, corresponding to the two edges coming out of their vertex in the DAG. The specific value of n chosen for comparing independent policy gradient and independent natural policy gradient was 4. This value was chosen due to the fact that for $n \geq 5$ agents, IPG required more than 3000 iterations, each of which require a nontrivial amount of time due to needing to make tabular updates for each state-action pair.

Parameters: All parameters and methods were the same as for IPG and INPG in the distancing game setting of the previous subsection.

Results: Figure 4 and Figure 5 depict the L1-accuracy in the policy space at each iteration, i.e., L1-accuracy $= \frac{1}{N} \sum_{i \in \mathcal{N}} |\pi_i - \pi_i^{\text{final}}| = \frac{1}{N} \sum_{i \in \mathcal{N}} \sum_s \sum_a |\pi_i(a \mid s) - \pi_i^{\text{final}}(a \mid s)|$. This is the average distance between the current policy and the final policy of all agents.
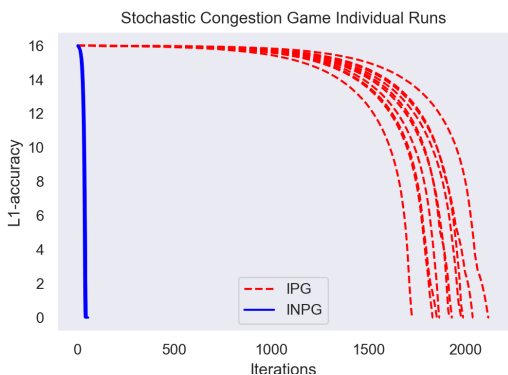

Figure 4: Trajectories of the L1-accuracy for 10 runs each of Independent Policy Gradient and Independent Natural Policy Gradient in the stochastic congestion game environment with 4 agents and learning rate of $\eta = .0001$.

These results show a dramatic difference in performance of IPG and INPG in the stochastic congestion game environment. The agents running IPG took about 2000 iterations to converge, while the agents running INPG took only about 50. This experiment also shows that a key distinguishing aspect between single-agent PG and NPG also can appear in the multi-agent case; IPG suffers from a long flat region of little improvement when far from a Nash policy, while INPG quickly moves the agents towards a Nash policy. This is in line with observed differences between single-agent PG and NPG
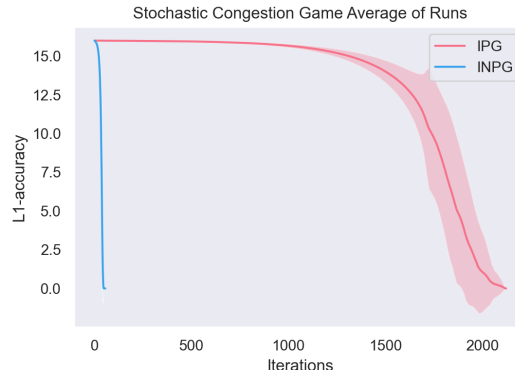

Figure 5: Mean L1-accuracy with shaded region of one standard deviation for the trajectories of Figure 4.

moving towards optimal policies (Kakade, 2001).

Furthermore, as mentioned above, IPG was not able to converge within 3000 iterations for $n \geq 5$ agents in the stochastic congestion game setting. In comparison, as see in Figure 6, INPG was able to converge in less than 100 iterations even with 8 agents.
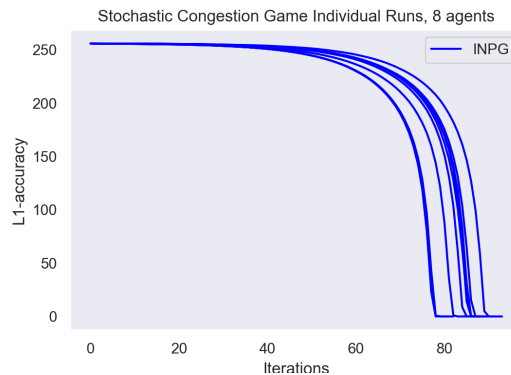

Figure 6: Trajectories of the L1-accuracy for 10 runs each of independent natural policy gradient in the stochastic congestion game environment with 8 agents and learning rate of $\eta = .0001$.

These experimental results highlight the key reason that guarantees regarding the convergence of independent natural policy gradient are necessary to obtain; INPG generally outperforms IPG in all of our experiments, in some cases by a significantly large margin. Again, this is in line with expectations based on the performance of single-agent policy gradient versus natural policy gradient (Agarwal et al., 2020).

5 CONCLUSION

In this paper we showed that independent Natural Policy Gradient will always converge to a Nash equilibrium in Markov Potential Games under mild conditions. We also show experimental results where NPG converges much faster than independent policy gradients.

In future work, we will prove rates of convergence for NPG in MPGs. We will also look into having each learner estimate advantages instead of receiving oracle advantages and will prove results with function approximation. It is also unknown whether trust region methods like TRPO will converge in MPGs which generalize cooperative games, for which TRPO convergence has been established (Kuba et al., 2021). Going beyond this work, a large open problem remain proving (or disproving) the convergence of independent RL agents to correlated equilibrium in general-sum Markov games.

References

A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. Optimality and Approximation with Policy Gradient Methods in Markov Decision Processes. In J. Abernethy and S. Agarwal, editors, *Proceedings of 33rd Conference on Learning Theory*, volume 125 of *PMLR*, pages 64–66, 2020. URL http://proceedings.mlr.press/v125/agarwal20a.html.

I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6): 26–38, 2017.

N. Brown and T. Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

N. Brown, A. Lerer, S. Gross, and T. Sandholm. Deep counterfactual regret minimization. In *International conference on machine learning*, pages 793–802. PMLR, 2019.

C. Daskalakis, D. J. Foster, and N. Golowich. Independent policy gradient methods for competitive reinforcement learning. *Neural Information Processing Systems (NeurIPS) 2020*, 2020.

J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

J. N. Foerster, R. Y. Chen, M. Al-Shedivat, S. Whiteson, P. Abbeel, and I. Mordatch. Learning with opponent-learning awareness. *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2017.

J. Heinrich and D. Silver. Deep reinforcement learning from self-play in imperfect-information games. *arXiv preprint arXiv:1603.01121*, 2016.

D. Hennes, D. Morrill, S. Omidshafiei, R. Munos, J. Perolat, M. Lanctot, A. Gruslys, J.-B. Lespiau, P. Parmas, E. Duenez-Guzman, et al. Neural replicator dynamics. *AAMAS 2020*, 2019.

C. Jin, Q. Liu, Y. Wang, and T. Yu. V-learning – a simple, efficient, decentralized algorithm for multiagent rl. 2021.

S. Kakade. A natural policy gradient. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, page 1531–1538, Cambridge, MA, USA, 2001. MIT Press.

J. G. Kuba, R. Chen, M. Wen, Y. Wen, F. Sun, J. Wang, and Y. Yang. Trust region policy optimisation in multi-agent reinforcement learning. *arXiv preprint arXiv:2109.11251*, 2021.

M. Lanctot, V. Zambaldi, A. Gruslys, A. Lazaridou, K. Tuyls, J. Pérolat, D. Silver, and T. Graepel. A unified game-theoretic approach to multiagent reinforcement learning. *Neural Information Processing Systems (NIPS)*, 2017.

S. Leonardos, W. Overman, I. Panageas, and G. Piliouras. Global convergence of multi-agent policy gradient in markov potential games. *CoRR*, abs/2106.01969, 2021. URL https://arxiv.org/abs/2106.01969.

R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Neural Information Processing Systems (NIPS)*, 2017.

S. McAleer, J. Lanier, R. Fox, and P. Baldi. Pipeline psro: A scalable approach for finding approximate nash equilibria in large games. *Neural Information Processing Systems (NeurIPS)*, 2020.

S. McAleer, J. Lanier, P. Baldi, and R. Fox. Xdo: A double oracle algorithm for extensive-form games. *arXiv preprint arXiv:2103.06426*, 2021.

D. Mguni. Stochastic Potential Games. *arXiv e-prints*, art. arXiv:2005.13527, May 2020.

D. Mguni, Y. Wu, Y. Du, Y. Yang, Z. Wang, M. Li, Y. Wen, J. Jennings, and J. Wang. Learning in nonzero-sum stochastic games with potentials. *International Conference on Machine Learning (ICML) 2021*, 2021.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, Feb 2015. ISSN 1476-4687. doi: 10.1038/nature14236.

M. Moravčík, M. Schmid, N. Burch, V. Lisỳ, D. Morrill, N. Bard, T. Davis, K. Waugh, M. Johanson, and M. Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356 (6337):508–513, 2017.

J. F. Nash et al. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36 (1):48–49, 1950.

A. OroojlooyJadid and D. Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *arXiv preprint arXiv:1908.03963*, 2019.

G. Palaiopanos, I. Panageas, and G. Piliouras. Multiplicative weights update with constant step-size in congestion games: Convergence, limit cycles and chaos. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5872–5882, 2017.

R. Rosenthal. A class of games possessing pure-strategy Nash equilibria. *International Journal of Game Theory*, 2(1):65–67, 1973.

T. Roughgarden. Intrinsic robustness of the price of anarchy. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 513–522, 2009.

D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Z. Song, S. Mei, and Y. Bai. When can we learn general-sum markov games with a large number of players sample-efficiently? 2021.

P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning. *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2017.

O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

R. Zhang, Z. Ren, and N. Li. Gradient play in multi-agent markov stochastic games: Stationary points and convergence. *arXiv preprint arXiv:2106.00198*, 2021.

Independent Natural Policy Gradient Always Converges in Markov Potential Games: Supplementary Materials

6 Proof of Proposition 3.3

Proof. Let $\left\|\tilde{\theta} - \theta\right\|_\infty \leq \frac{\eta}{1-\gamma}$. We show that

$$1 - \frac{2\eta}{1-\gamma} \leq \frac{\pi_{\theta_i}(a|s)}{\pi_{\tilde{\theta}_i}(a|s)} \leq 1 + \frac{4\eta}{1-\gamma} \tag{14}$$

for all agents i and $a \in A_i$. It holds that

$$\begin{aligned}
\pi_{\theta_i}(a|s) &= \frac{e^{\theta_{i,s,a}}}{\sum_{a' \in A_i} e^{\theta_{i,s,a'}}} \\
&\leq \frac{e^{\tilde{\theta}_{i,s,a} + \frac{\eta}{1-\gamma}}}{\sum_{a' \in A_i} e^{\tilde{\theta}_{i,s,a'} - \frac{\eta}{1-\gamma}}} \\
&= \pi_{\tilde{\theta}_i}(a|s) e^{\frac{2\eta}{1-\gamma}} \leq \pi_{\tilde{\theta}_i}(a|s)\left(1 + 4\frac{\eta}{1-\gamma}\right),
\end{aligned}$$

where the last inequality works as $e^x \leq 2x + 1$ for $x \in [0, 1/2]$, i.e., we assume $\eta \leq \frac{(1-\gamma)}{2}$. Similarly,

$$\begin{aligned}
\pi_{\theta_i}(a|s) &= \frac{e^{\theta_{i,s,a}}}{\sum_{a' \in A_i} e^{\theta_{i,s,a'}}} \\
&\geq \frac{e^{\tilde{\theta}_{i,s,a} - \frac{\eta}{1-\gamma}}}{\sum_{a' \in A_i} e^{\tilde{\theta}_{i,s,a'} + \frac{\eta}{1-\gamma}}} \\
&= \pi_{\tilde{\theta}_i}(a|s) e^{-\frac{2\eta}{1-\gamma}} \geq \pi_{\tilde{\theta}_i}(a|s)\left(1 - 2\frac{\eta}{1-\gamma}\right),
\end{aligned}$$

where the last inequality comes from the fact that $e^x \geq x + 1$ for all x. We focus now on the Hessian of $\Phi^{\pi_\theta}(\mu)$ with respect to θ, that is $\nabla^2 \Phi^{\pi_\theta}(\mu)$. It suffices to show that the spectral norm of the matrix $\left(D^\theta\right)^{-\frac{1}{2}} \cdot \nabla^2_\theta \Phi^{\pi_{\tilde{\theta}}}(\mu) \cdot \left(D^\theta\right)^{-\frac{1}{2}}$ [2] is bounded by L (notice that the Hessian is computed at $\tilde{\theta}$ and the diagonal matrix D^θ is computed at θ with $\left\|\tilde{\theta} - \theta\right\|_\infty \leq \frac{\eta}{1-\gamma}$).

Fix agent i, scalar $t \geq 0$, and a vector $u = \left(D^{\tilde{\theta}}\right)^{-1} v$ with v a unit vector. Moreover, let $V(t) = V_i^{\pi_{\theta+tu}}(\mu)$.

It holds that

$$V(t) = \sum_{s \in S} \mu(s) \sum_{\mathbf{a} \in A} \pi_{\theta + tu}(\mathbf{a}|s) Q_i^{\pi_{\theta+tu}}(s, \mathbf{a}).$$

---

[2] which has the same spectral norm with $\left(D^\theta\right)^{-1} \cdot \nabla^2_\theta \Phi^{\pi_{\tilde{\theta}}}(\mu)$

Hence taking the second derivative we have

$$
\begin{aligned}
\frac{d^2 V(0)}{dt^2} =& \sum_s \frac{\mu(s)}{d_\mu^{\pi_{\tilde{\theta}}}(s)} \sum_{\mathbf{a} \in A} \pi_\theta(\mathbf{a}|s) \frac{d^2 Q_i^{\pi_{\theta+tu}}(s,\mathbf{a})}{dt^2}\bigg|_{t=0} \\
&+ \sum_s \frac{\mu(s)}{d_\mu^{\pi_{\tilde{\theta}}}(s)} \sum_{\mathbf{a} \in A} \frac{d^2 \pi_{\theta+tu}(\mathbf{a}|s)}{dt^2}\bigg|_{t=0} Q_i^{\pi_\theta}(s,\mathbf{a}) \\
&+ 2 \sum_{s \in S} \frac{\mu(s)}{d_\mu^{\pi_{\tilde{\theta}}}(s)} \sum_{\mathbf{a} \in A} \frac{d\pi_{\theta+tu}(\mathbf{a}|s)}{dt}\bigg|_{t=0} \frac{dQ_i^{\pi_{\theta+tu}}(s,\mathbf{a})}{dt}\bigg|_{t=0}.
\end{aligned}
\tag{15}
$$

The following calculations hold (see Agarwal et al. (2020), section D).

1. Let a_i be the i-th coordinate of $\mathbf{a}$ and u_{i,s,b_i} the coordinate of u corresponding to agent i choosing action b_i at state s. We have
$\frac{d\pi_{\theta+tu}(\mathbf{a}|s)}{dt}\big|_{t=0} = \pi_\theta(\mathbf{a}|s) \sum_{i \in N} \sum_{b_i \in A_i} u_{i,s,b_i} \cdot (\mathbf{1}_{b_i=a_i} - \pi_{\theta_i}(b_i|s))$. We conclude that

$$
\begin{aligned}
\left| \frac{d\pi_{\theta+tu}(\mathbf{a}|s)}{dt}\bigg|_{t=0} \right| &\leq \left| \sum_{i \in N} \sum_{b_i \in A_i} \frac{\pi_\theta(\mathbf{a}|s) \cdot (\mathbf{1}_{b_i=a_i} - \pi_{\theta_i}(b_i|s))}{\pi_{\tilde{\theta}_i}(b_i|s)} \right| \\
&\overset{(14)}{\leq} \sum_{i \in N} \left(1 + \frac{4\eta}{1-\gamma}\right) \pi_{\theta_{-i}}(\mathbf{a}_{-i}|s)(1 - \pi_{\theta_i}(a_i|s)) \\
&+ \left(1 - \frac{2\eta}{1-\gamma}\right) \sum_{i \in N} \sum_{b_i \neq a_i} \pi_\theta(\mathbf{a}|s) \leq \left(1 + \frac{4\eta}{1-\gamma}\right) n A_{\max}.
\end{aligned}
$$

2. We also have $\frac{d^2 \pi_{\theta+tu}(\mathbf{a}|s)}{dt^2}\big|_{t=0} = \pi_\theta(\mathbf{a}|s) \sum_{i,j \in N} \sum_{b_i \in A_i} \sum_{b_j \in A_j} u_{i,s,b_i} u_{j,s,b_j}(\mathbf{1}_{b_i=a_i} - \pi_{\theta_i}(b_i|s)) \cdot (\mathbf{1}_{b_j=a_j} - \pi_{\theta_j}(b_j|s)) - \pi_\theta(\mathbf{a}|s) \sum_{i \in N} \sum_{b_i,c_i \in A_i} u_{i,s,b_i} u_{i,s,c_i} \pi_{\theta_i}(b_i|s))(\mathbf{1}_{b_i=c_i} - \pi_{\theta_i}(c_i|s))$. Similarly, We conclude that

$$
\left| \frac{d^2 \pi_{\theta+tu}(\mathbf{a}|s)}{dt^2}\bigg|_{t=0} \right| \leq \left(1 + \frac{4\eta}{1-\gamma}\right)^2 n^2 A_{\max}^2
$$

To bound the derivative of the Q-function, observe that $Q_i^{\pi_{\theta+tu}}(s,\mathbf{a}) = e_{s,\mathbf{a}}^\top (I - \gamma P(t))^{-1} r$, where $r(s,\mathbf{a})$ is the expected reward of agent i if agent choose action $\mathbf{a}$ at state s and $P(t)$ is state-action transition matrix w.r.t the joint distribution of all agents and the environment.

It was shown in Agarwal et al. (2020) and Leonardos et al. (2021) that

$$
\left| \frac{dQ_i^{\pi_\theta+tu}(s,\mathbf{a})}{dt} \right| = \gamma \left| e_{s,a}^\top (I - \gamma P(0))^{-1} \frac{dP(0)}{dt} (I - \gamma P(0))^{-1} r \right| \leq \frac{\gamma \sqrt{|A_i|}}{(1-\gamma)^2} \leq \frac{\gamma \sqrt{A_{\max}}}{(1-\gamma)^2},
\tag{16}
$$

and also

$$
\begin{aligned}
\left| \frac{d^2 Q_i^{\pi_\theta+tu}(s,\mathbf{a})}{dt^2} \right| &= 2\gamma^2 \left| e_{s_0,a}^\top (I - \gamma P(0))^{-1} \frac{dP(0)}{dt} (I - \gamma P(0))^{-1} \frac{dP(0)}{dt} (I - \gamma P(0))^{-1} r \right| \\
&\leq \frac{2\gamma^2 |A_i|}{(1-\gamma)^3} \leq \frac{2\gamma A_{\max}}{(1-\gamma)^3},
\end{aligned}
\tag{17}
$$

Combining the above and assuming $\eta \leq \frac{1-\gamma}{2}$ we get

$$
\left| \frac{d^2 V(0)}{dt^2} \right| = \sum_s \mu(s) \left(\frac{6\gamma A_{\max}}{(1-\gamma)^3} + 9n^2 A_{\max}^2 + \frac{3n A_{\max} \gamma \sqrt{A_{\max}}}{(1-\gamma)^2} \right) \leq M \frac{27 n^2 A_{\max}^2}{(1-\gamma)^3},
$$

where M is the mismatch coefficient. We choose $L = \frac{27 n^2 A_{\max}^2 M}{(1-\gamma)^3}$. Since u is arbitrary, the proposition is proved.

$\square$