Linear Convergence of Independent Natural Policy Gradient in Games With Entropy Regularization

Youbang Sun, Tao Liu, P. R. Kumar, *Life Fellow, IEEE*, and Shahin Shahrampour, *Senior Member, IEEE*

***Abstract*—This letter focuses on the entropy-regularized independent natural policy gradient (NPG) algorithm in multi-agent reinforcement learning. In this letter, agents are assumed to have access to an oracle with exact policy evaluation and seek to maximize their respective independent rewards. Each individual's reward is assumed to depend on the actions of all agents in the multi-agent system, leading to a game between agents. All agents make decisions under a policy with bounded rationality, which is enforced by the introduction of entropy regularization. In practice, a smaller regularization implies that agents are more rational and behave closer to Nash policies. On the other hand, with larger regularization agents tend to act randomly, which ensures more exploration. We show that, under sufficient entropy regularization, the dynamics of this system converge at a linear rate to the quantal response equilibrium (QRE). Although regularization assumptions prevent the QRE from approximating a Nash equilibrium (NE), our findings apply to a wide range of games, including cooperative, potential, and two-player matrix games. We also provide extensive empirical results on multiple games (including Markov games) as a verification of our theoretical analysis.**

***Index Terms*—Game theory, multi-agent reinforcement learning, natural policy gradient, quantal response equilibrium.**

I. Introduction

IN THE emerging field of reinforcement learning (RL), the topic of multi-agent reinforcement learning (MARL) has been increasingly gaining attention. This surge in interest may be attributed to the fact that many real-world problems are multi-agent in nature, including tasks such as robotics [19], modern production systems [3], economic decision making [25], and autonomous driving [21].

Although applying single-agent RL algorithms, like policy gradient (PG) and natural policy gradient (NPG), to individual agents in MARL may seem straightforward, analyzing multi-agent systems presents numerous challenges. In the single-agent setting, the optimal policy selects the action with the highest cumulative reward and converges to the unique global optimal solution. However, in the multi-agent setting, the global policy is constructed by taking the product of the local policies. Agents have individual rewards in general, but each individual reward depends on the actions of all agents, leading to a game between agents. Even for a game as simple as a two-agent cooperative matrix game, there can be potentially multiple local stationary points. Each stationary point is known as a Nash equilibrium (NE), where no agent can enjoy a larger reward by unilaterally changing its strategy.

For general games, it is known that a system where each agent follows the policy gradient update (i.e., gradient play) can easily fail [22]. For a game to converge to an NE through gradient play, additional assumptions are needed, such as existence of a potential function and isolatedness of the NEs [24]. In addition, in MARL we encounter similar challenges to single-agent RL, including navigating sub-optimal regions characterized by flat curvatures and managing the exploration-exploitation trade-off. One mitigation strategy in practice is to enforce an entropy regularization [12], [13], [16].

Intuitively speaking, the addition of an entropy regularization term penalizes the policies that are not stochastic enough. The entropy regularization introduces rationality, where decisions are selected to be satisfactory rather than optimal, which encourages the exploration of agents and prevents the system from being stuck at local sub-optimal policies caused by pure strategies. The introduction of entropy was also highlighted by Soft Actor Critic [11], which is widely used today in robotics. When entropy is introduced into the problem, the system converges to the quantal response equilibrium (QRE) [15] instead of NE. A QRE refers to an equilibrium with bounded rationality, which we formally define in Definition 2.

In this letter, we consider a general static game, where the system state is assumed to be fixed and no additional assumptions on rewards are imposed (except boundedness).

Manuscript received 8 March 2024; revised 9 May 2024; accepted 28 May 2024. Date of publication 5 June 2024; date of current version 19 June 2024. This work was supported in part by the U.S. Office of Naval Research under Grant N00014-21-1-2385; in part by the U.S. Army Contracting Command under Grant W911NF-22-1-0151; in part by U.S. ARO under Grant W911NF2120064; and in part by the U.S. National Science Foundation under Grant CNS-2328395 and Grant CMMI-2038625. Recommended by Senior Editor S. Olaru. *(Corresponding author: Shahin Shahrampour.)*

Youbang Sun and Shahin Shahrampour are with the Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA 02115 USA (e-mail: sun.youb@northeastern.edu; s.shahrampour@northeastern.edu).

Tao Liu and P. R. Kumar are with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: tliu@tamu.edu; prk@tamu.edu).

Digital Object Identifier 10.1109/LCSYS.2024.3410149

2475-1456 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Our framework subsumes various settings, such as cooperative games, potential games, and two-player matrix games.

A. Contributions

Motivated by the effectiveness of entropy regularization in both single-agent RL and certain multi-agent settings in games, we have adapted the entropy-regularized natural policy gradient algorithm to games. While some existing works like [5] use QRE to approximate NE for some structured games, we consider the regularization as a given factor and study the convergence for general static games. We summarize our contributions as follows.

1) We consider the NPG update with entropy regularization in static games (Section III-A) without structured assumptions such as potential games.
2) We study the convergence properties of the proposed algorithm and establish in Section III-B that the system can reach a QRE with a linear convergence rate when entropy regularization is sufficiently large.
3) In Section IV, we present extensive numerical experiments demonstrating the effectiveness of the algorithm and provide some discussion on its performance across various settings.

Although our theoretical analysis only considers the static game setting, in Section IV-C we conduct experiments for stochastic (Markov) games. We show that similar empirical results also hold for Markov games, the theoretical investigation of which is interesting for future work.

B. Related Literature

This section provides a review of the related literature on the topics of policy gradient-based algorithms in RL and independent learning in games.

a) Policy Gradient: There is a lot of interest in the theoretical understanding of policy gradient methods in recent literature [1], [4]. There are many variations of policy gradient methods under different parameterizations. An important extension of the policy gradient method is the natural policy gradient (NPG) method [1], [14], [17], which introduces the addition of pre-conditioning in the policy update based on the problem geometry. To promote exploration and improve stochasticity within the system, entropy regularization has been introduced. In general, entropy regularization has been shown to accelerate convergence rates for several algorithms. Policy gradient methods with entropy regularization include [16] for PG and [6] for NPG. Additionally, a broad class of convex regularizers has been proposed in [13], [27].

b) Independent Learning in Games: Recent years have witnessed significant progress in understanding the system dynamics of independent learning algorithms in games. It has been shown in game theory that a system where agents use simple gradient play in a game could fail to converge, such as the "cycling problem" shown in [20]. Therefore, additional settings, such as competitive and cooperative settings have been considered. For the competitive setting, zero-sum games have been studied by [10], [26]. A framework more general than the cooperative setting is the potential game setting [5], where agents do not have the same rewards, but there exists a potential function tracking the value changes across all agents.

These settings have also been extended from static games to stochastic games, where the system follows a Markov state transition model. A series of works tackle the system convergence rate in the Markov potential games setting [9], [24], [28], [29].

Entropy regularization is also widely considered in games. The convergence rate has been shown to be linear for two-player zero-sum games [7], [8], and sub-linear for potential games [5]. In particular, [5] studies NPG for potential games with entropy regularization and is of great relevance to our work.

We note that though many works consider entropy regularization in games and study the system convergence to QRE, most of the previous works, including [5], address arbitrarily small regularization factors and view QRE as an approximation of NE. Although these works provide theoretical insights that effectively demonstrate the intended use case, their effectiveness is largely confined to games with structure, such as zero-sum games or potential games. These works do not consider more general games. In contrast, our work considers regularization as a constant penalizing factor and discusses the convergence of the system dynamics for regularized rewards.

II. Problem Formulation

In this section, we introduce the basic setting for a general multi-player game with the consideration of entropy regularization.

Throughout this letter, we use $\|\cdot\|_1$ to denote the ℓ^1 norm and $\|\cdot\|_\infty$ to represent the ℓ^∞ norm, respectively. We denote $[n] \coloneqq \{1, \dots, n\}$. For the time-varying sequence of a set of parameters $\{\theta_i^k\}_{i\in[n],k\in\mathbf{N}}$, we use superscript k to denote the k-th time step and subscript i to denote the i-th agent.

A. Games in the Multi-Agent System

Consider a tabular strategic game $\mathcal{G} = (n, \mathcal{A}, \{r_i\}_{i\in[n]})$ consisting of n agents. The global discrete action space $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ is the product of individual action spaces, where the global action is denoted by $\boldsymbol{a} \coloneqq (a_1, \dots, a_n)$. The reward for each agent $i \in [n]$ is denoted as $r_i : \mathcal{A} \to [0, 1]$.

A mixed strategy for the entire system is a *decentralized* multi-agent policy [28], where all agents make decisions independently. Therefore, the global system policy is denoted by $\pi \in \Delta(\mathcal{A}_1) \times \cdots \times \Delta(\mathcal{A}_n) \subset \Delta(\mathcal{A})$, where Δ denotes the probability simplex operator. We can write $\pi(\boldsymbol{a}) = \prod_{i\in[n]} \pi_i(a_i)$, where $\pi_i \in \Delta(\mathcal{A}_i)$ is the local policy for agent i. We also denote the combined policy of all agents other than i as $\pi_{-i} \coloneqq \prod_{j\in[n]\setminus\{i\}} \pi_j$, so that $\pi = \pi_i \times \pi_{-i}$. Similarly, we denote the combined action as $\boldsymbol{a} = (a_i, a_{-i})$, where $a_{-i} \coloneqq \{a_j\}_{j\neq i}$.

With a slight abuse of notation, we represent the expectation of the reward r_i under policy π as

$$r_i(\pi) \coloneqq \mathbb{E}_{\boldsymbol{a}\sim\pi}[r_i(\boldsymbol{a})] = \langle \pi, r_i \rangle.$$

We also define the *marginalized reward function* of reward r_i with respect to the policy π_{-i} as

$$\bar{r}_i(a_i) \coloneqq \mathbb{E}_{a_{-i}\sim\pi_{-i}}\big[r_i(a_i, a_{-i})\big].$$

We note that the calculation of the marginalized reward requires π_{-i}, the policy of all agents but i. Next, we introduce the notion of Nash equilibrium (NE) [18] in games.

Definition 1 (Nash Equilibrium): A joint policy π^* is a Nash equilibrium if

$$r_i(\pi_i^*, \pi_{-i}^*) \geq r_i(\pi_i', \pi_{-i}^*) \ \forall \pi_i' \in \Delta(\mathcal{A}_i) \ \forall i \in [n].$$

It is known that if mixed strategies (where a player assigns a strictly positive probability to every pure strategy) are allowed, at least one NE exists in any finite game [18].

B. Entropy Regularization in Games

The Shannon entropy of policy π_i, defined as

$$\mathcal{H}(\pi_i) := -\sum_{a_k \in \mathcal{A}_i} \pi_i(a_k) \log \pi_i(a_k)$$

measures the level of randomness in actions of agent i. When entropy is added to the problem with regularization factor τ, the regularized objective for agent i is modified to $\hat{r}_i(\pi) := r_i(\pi) + \tau \mathcal{H}(\pi_i)$. With the consideration of entropy, a new type of equilibrium for the system has been defined in [15], referred to as the quantal response equilibrium (QRE) or logit equilibrium.

Definition 2 (Quantal Response Equilibrium): For any given $\tau > 0$, a joint policy π^* is a corresponding quantal response equilibrium when it holds that,

$$\hat{r}_i(\pi_i^*, \pi_{-i}^*) \geq \hat{r}_i(\pi_i', \pi_{-i}^*), \ \forall \pi_i' \in \Delta(\mathcal{A}_i), \ \forall i \in [n].$$

It can be easily verified that when a QRE has been reached, each agent uses a policy that assigns probability of actions according to the marginalized reward, i.e., $\pi_i^*(\cdot) \propto \exp(\bar{r}_i(\cdot)/\tau)$. This is often referred to by the literature as the policy with bounded rationality [5] with rationality parameter $\frac{1}{\tau}$. Intuitively, an NE refers to a perfectly rational policy with $\tau \to 0$, whereas a fully random policy with $\tau \to \infty$ is considered as completely irrational.

III. Main Results

In this section, we study the dynamics of a multi-agent system where each agent seeks a policy to maximize the individual regularized reward. We first provide the exact algorithm update in Section III-A, and then in Section III-B we show that with a sufficient entropy regularization, the game converges to a QRE with a linear rate.

A. Algorithm Update

We first outline the NPG update applied by agents, which has been studied for single-agent RL in [1]. Since policies are constrained on the probability simplex, in order to relax this constraint, the softmax parameterization has been widely adopted. A set of unconstrained parameters $\theta_i \in \mathbb{R}^{|\mathcal{A}_i|}$ are updated, and the policy is calculated by

$$\pi_i(a_i) = \frac{\exp(\theta_i(a_i))}{\sum_{a_j \in \mathcal{A}_i} \exp(\theta_i(a_j))}.$$

In the static games setting, the NPG algorithm performs gradient updates that are pre-conditioned on the problem geometry [1],

$$\theta_i^{k+1} = \theta_i^k + \eta \mathcal{F}_{\theta_i^k}^{\dagger} \frac{\partial}{\partial \theta_i^k} \hat{r}_i\left(\pi^k\right), \tag{1}$$

where η denotes the step-size, and $\mathcal{F}_{\theta_i}^{\dagger}$ is the Moore-Penrose pseudo-inverse of the Fisher information matrix [1], defined as,

$$\mathcal{F}_{\theta_i} := \mathbb{E}_{a_i \sim \pi_i}\Big[\big(\nabla_{\theta_i} \log \pi_i(a_i)\big)\big(\nabla_{\theta_i} \log \pi_i(a_i)\big)^{\top}\Big].$$

Based on the definition of the regularized reward $\hat{r}_i(\pi)$, the policy gradient for agent i can be calculated as

$$\frac{\partial \hat{r}_i(\pi)}{\partial \theta_i(a_j)} = \pi_i(a_j)\big(\bar{r}_i(a_j) - \tau \log(\pi_i(a_j)) - \hat{r}_i(\pi)\big).$$

Detailed calculation of this derivative can be found in [23].

Using step-size η, the corresponding update for agent i under softmax parameterization [5] becomes

$$\pi_i^{k+1}(a_i) \propto \pi_i^k(a_i)^{1-\eta\tau} \exp\left(\eta \bar{r}_i^k(a_i)\right), \tag{2}$$

where $\bar{r}_i^k(a_i) = \mathbb{E}_{a_{-i} \sim \pi_{-i}^k}[r_i(a_i, a_{-i})]$. Intuitively speaking, if the policy π_{-i} is fixed at iteration k, the game reduces to a single-agent RL problem for agent i, and the updates shown in (2) will converge to a local optimal policy, with

$$\pi_i^{k*}(a_i) \propto \exp\left(\bar{r}_i^k(a_i)/\tau\right), \quad \text{for } 0 < \eta\tau < 1. \tag{3}$$

B. Convergence Analysis

Before presenting our main theorem, we first introduce the notion of *QRE-gap* as

$$\textit{QRE-gap}(\pi) := \max_{i \in [n], \pi_i' \in \Delta(\mathcal{A}_i)} \left[\hat{r}_i(\pi_i', \pi_{-i}) - \hat{r}_i(\pi_i, \pi_{-i})\right],$$

where at iteration k, given policy π_{-i}^k, the maximum for agent i is reached at $\pi_i' = \pi_i^{k*}$ given in (3).

For the case of $\tau = 0$, the agents are purely rational, and the *QRE-gap* recovers *NE-gap*. It is easily verified that a system has reached a QRE if and only if *QRE-gap* $= 0$. We study the convergence of the *QRE-gap* in this section. For the ease of notation, we denote the *QRE-gap* at iteration k (i.e., policy π^k) by *QRE-gap*^{k}.

Given the definition of π_i^{k*} in (3), we have:

$$\begin{aligned}
\textit{QRE-gap}^k &= \max_{i \in [n]} \left[\hat{r}_i\left(\pi_i^{k*}, \pi_{-i}^k\right) - \hat{r}_i\left(\pi_i^k, \pi_{-i}^k\right)\right] \\
&= \max_{i \in [n]} \left[\langle \pi_i^{k*} - \pi_i^k, \bar{r}_i^k \rangle + \tau\left(\mathcal{H}(\pi_i^{k*}) - \mathcal{H}(\pi_i^k)\right)\right] \\
&= \max_{i \in [n]} \Big[\langle \pi_i^{k*} - \pi_i^k, \tau \log \pi_i^{k*} \rangle \\
&\qquad - \tau \langle \pi_i^{k*}, \log \pi_i^{k*} \rangle + \tau \langle \pi_i^k, \log \pi_i^k \rangle\Big] \\
&= \tau \max_{i \in [n]} \Big[\langle \pi_i^k, \log \pi_i^k - \log \pi_i^{k*} \rangle\Big].
\end{aligned} \tag{4}$$

Motivated by [6], we introduce the following auxiliary sequence $\{\xi_i^k \in \mathbb{R}^{|\mathcal{A}_i|}, i \in [n]\}$ to help further with analysis.

$$\begin{aligned}
\xi_i^0(a_i) &= \| \exp\left(\bar{r}_i^0/\tau\right) \|_1 \pi_i^0(a_i) \\
\xi_i^{k+1}(a_i) &= \xi_i^k(a_i)^{1-\eta\tau} \exp\left(\eta \bar{r}_i^k(a_i)\right).
\end{aligned} \tag{5}$$

By the definition of the auxiliary sequence, two consecutive iterates $\xi_i^{k+1}(a_i)$ and $\xi_i^k(a_i)$ satisfy the following equality

$$\begin{aligned}&\log \xi_i^{k+1}(a_i) - \bar{r}_i^{k+1}(a_i)/\tau \\ &= (1-\eta\tau)\Big(\log \xi_i^k(a_i) - \bar{r}_i^k(a_i)/\tau\Big) \\ &\quad + \Big(\bar{r}_i^k(a_i) - \bar{r}_i^{k+1}(a_i)\Big)/\tau. \end{aligned} \tag{6}$$

It can be observed that $\pi_i^k \propto \xi_i^k$ according to Equations (2) and (5). We now introduce the following lemma to establish direct relationships between π_i^k and ξ_i^k.

Lemma 1 [6]: For any two probability distributions $\pi_1, \pi_2 \in \Delta(\mathcal{A})$ that satisfy

$$\pi_1(a) \propto \exp\left(\theta_1(a)\right), \text{ and } \pi_2(a) \propto \exp\left(\theta_2(a)\right),$$

with $\theta_1, \theta_2 \in \mathbb{R}^{|\mathcal{A}|}$, the following inequality holds

$$\|\log(\pi_1) - \log(\pi_2)\|_\infty \leq 2\|\theta_1 - \theta_2\|_\infty.$$

The proof of this lemma is provided by [6]. Next, we introduce the following lemma regarding decentralized multi-agent policies.

Lemma 2: For two sets of policies $\{\pi_i^1\}, \{\pi_i^2\}, i \in [n]$, where each policy $\pi_i^j \in \Delta(\mathcal{A}_i)$, we have the following inequality:

$$\begin{aligned}&\sum_{a_1,\ldots,a_n} |\pi_1^1(a_1) \times \cdots \times \pi_n^1(a_n) - \pi_1^2(a_1) \times \cdots \times \pi_n^2(a_n)| \\ &\leq \sum_{i\in[n]} \sum_{a_i} |\pi_i^1(a_i) - \pi_i^2(a_i)|,\end{aligned}$$

i.e., for two policies π^1, π^2, we have $\|\pi^1 - \pi^2\|_1 \leq \sum_i \|\pi_i^1 - \pi_i^2\|_1$.

The proof can be found in the full version of this letter [23]. From Lemma 2, we are able to evaluate the difference between the marginalized rewards of two consecutive iterations, which is used to prove our main result below.

Theorem 1: Consider a static game with independent NPG update shown in (2), where the regularization factor $\tau > 2\sum_{i\in[n]} |\mathcal{A}_i|$ and the step-size $0 < \eta < \frac{1}{\tau - 2\sum_{i\in[n]}|\mathcal{A}_i|}$. We then have that

$$\begin{aligned}&\textit{QRE-gap}^k \\ &\leq 2\tau\left(1 - \eta\tau + 2\eta\sum_i |\mathcal{A}_i|\right)^k \max_{i\in[n]} \|\log \pi_i^0 - \log \pi_i^{0*}\|_\infty,\end{aligned}$$

where $\pi_i^{0*} \propto \exp(\bar{r}_i^0/\tau)$.

Proof: For any agent i, we know that the reward is upper-bounded by 1, so with the help of Lemma 2 we have

$$\begin{aligned}\|\bar{r}_i^{k+1} - \bar{r}_i^k\|_\infty &\leq \|\pi_{-i}^{k+1} - \pi_{-i}^k\|_1 \leq \sum_{j\neq i} \|\pi_j^{k+1} - \pi_j^k\|_1 \\ &\leq \sum_j |\mathcal{A}_j| \|\pi_j^{k+1} - \pi_j^k\|_\infty \\ &\leq \sum_j |\mathcal{A}_j| \|\log \pi_j^{k+1} - \log \pi_j^k\|_\infty \\ &\leq 2\sum_j |\mathcal{A}_j| \|\log \xi_j^{k+1} - \log \xi_j^k\|_\infty \\ &\leq \left(2\eta\tau \sum_j |\mathcal{A}_j|\right) \max_{i\in[n]} \|\log \xi_i^k - \bar{r}_i^k/\tau\|_\infty,\end{aligned} \tag{7}$$

where the third line follows by mean-value theorem, the fourth line follows from Lemma 1, and the last inequality follows from the definition of ξ_i^k. We can then provide an upper bound on the following term,

$$\begin{aligned}&\max_{i\in[n]} \|\log \xi_i^{k+1} - \bar{r}_i^{k+1}/\tau\|_\infty \\ &\leq \max_{i\in[n]}\Big[(1-\eta\tau)\|\log \xi_i^k - \bar{r}_i^k/\tau\|_\infty + \|\bar{r}_i^k - \bar{r}_i^{k+1}\|_\infty/\tau\Big] \\ &\leq \left(1 - \eta\tau + 2\eta\sum_j |\mathcal{A}_j|\right) \max_{i\in[n]} \|\log \xi_i^k - \bar{r}_i^k/\tau\|_\infty \\ &\leq \left(1 - \eta\tau + 2\eta\sum_j |\mathcal{A}_j|\right)^{k+1} \max_{i\in[n]} \|\log \xi_i^0 - \bar{r}_i^0/\tau\|_\infty.\end{aligned} \tag{8}$$

Here, the first inequality is provided given the properties of the auxiliary sequence in (6), the second inequality comes directly from (7), and the final bound is obtained by recursion.

With the help of (4), we can now bound the QRE-gap by

$$\begin{aligned}&\textit{QRE-gap}^k \\ &= \max_{i\in[n]}\Big[\tau\langle \pi_i^k, \log \pi_i^k - \log \pi_i^{k*}\rangle\Big] \\ &\leq \tau \max_{i\in[n]} \|\log \pi_i^k - \log \pi_i^{k*}\|_\infty \\ &\leq 2\tau \max_{i\in[n]} \|\log \xi_i^k - \bar{r}_i^k/\tau\|_\infty \\ &\leq 2\tau\left(1 - \eta\tau + 2\eta\sum_i |\mathcal{A}_i|\right)^k \max_{i\in[n]} \|\log \xi_i^0 - \bar{r}_i^0/\tau\|_\infty \\ &\leq 2\tau\left(1 - \eta\tau + 2\eta\sum_i |\mathcal{A}_i|\right)^k \max_{i\in[n]} \|\log \pi_i^0 - \log \pi_i^{0*}\|_\infty.\end{aligned} \tag{9}$$

We applied the Hölder's inequality in the second line, and then Lemma 1 and (8) are used to complete the proof. ■

Theorem 1 presents an interesting perspective on the choice of the regularization factor τ. For small values of τ, the system is not guaranteed to converge. Once τ satisfies the lower bound condition in the theorem, the system is guaranteed convergence to a QRE with a suitable step-size. As τ increases further, the rate of convergence becomes faster, yet the corresponding QRE becomes less rational and more stochastic (and less desirable generally). As shown in the next section, it is crucial to find a suitable τ that provides a fast convergence speed but still retains rationality.

IV. Numerical Results

In the previous section, we established the convergence rate for *QRE-gap* in games. In this section, we verify the analytical results through three sets of experiments.

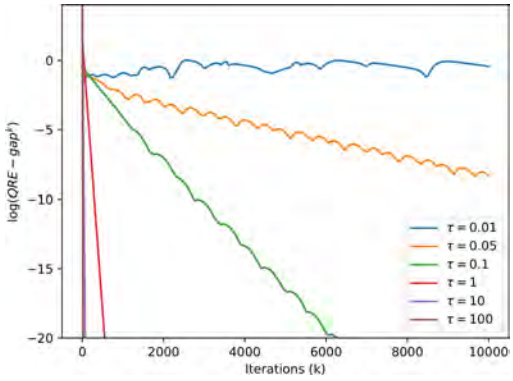


Fig. 1. Convergence of *QRE-gap* in random reward setting for different values of τ.

A. Synthetic Game Setting

We first consider a multi-agent system where the rewards r_i are generated randomly and independently. We set the number of agents to $n = 3$, with each agent having a different discrete action space size, $|\mathcal{A}_1| = 3, |\mathcal{A}_2| = 4, |\mathcal{A}_3| = 5$. At the start of the experiment, all agents are initialized with random policies. This setting is similar to that of [24]; however, the rewards assigned to the agents in [24] satisfy the potential game assumption, but they are set to be independent in our experiment. We use the same initialization across a selection of regularization factors. The step-size is set to $\eta_\tau = \frac{1}{2(\tau+2\sum_i |\mathcal{A}_i|)}$, which is within the range given in Theorem 1 and guarantees convergence.

The results are presented in Fig. 1 in log-scale. It can be seen that when there is no regularization, or the regularization factor is negligible, the system fails to converge. As τ increases, the system converges. When τ is large enough, the *QRE-gap* decreases monotonically and converges to zero at a linear rate, with the decay rate increasing as τ gets larger. The system dynamics in Fig. 1 perfectly verifies our finding on conditions of τ, η and the convergence rate in Theorem 1.

Analytical results and experiments both indicate that the system converges faster when there is a larger weight on regularization, with the system actually failing to converge if the regularization factor is too small. This observation aligns with our analytical results in Theorem 1. In practice, a trade-off needs to be maintained, such that the system convergence is guaranteed, yet the QRE is still meaningful. Furthermore, we find that for the synthetic reward experiment above, a regularization factor of $\tau \approx 0.1$ is enough for the system to converge with a linear rate. Interestingly, this requirement is much more relaxed than the theoretical requirement provided in Theorem 1. This could be due to the random generation of rewards, whereas the theorem provides a bound in the worst-case scenario.

B. Network Zero-Sum Games

Next, we focus on zero-sum games in the network setting with poly-matrix rewards [2]. This problem cannot be solved using the vanilla independent NPG update and requires additional design elements such as using extra-gradient methods [6]. However, most of the methods are restricted to the two-agent setting and cannot be adapted to the network setting.

We consider a 5-agent network with a ring graph. Each edge denotes a randomly generated zero-sum matrix game between the two neighbors. All agents are assumed to have the same action space $|\mathcal{A}_1| = \cdots = |\mathcal{A}_5| = 10$.

We first present the *QRE-gap* in Fig. 2(a) (log-scale). It can be seen that the convergence properties in this setting mirror those of the synthetic game in Section IV-A. We also present the results on the *NE-gap* of the system shown in Fig. 2(b). As previously mentioned, *NE-gap* can be recovered by setting $\tau = 0$ in *QRE-gap*. We note that while *NE-gap* only depends on the current policies and is independent of τ, the update steps to calculate the policies still depend on the regularization factor τ. With a moderate regularization, the system is able to converge to stationarity with relatively small *NE-gap*. When the regularization term is too large, the system does converge but with a somewhat undesirable *NE-gap*.

C. Markov Games

We now extend our experiments to the stochastic setting and use independent NPG to solve general Markov games. The Markov games setting can be seen as a generalization of the Markov decision process used in single-agent RL. Both the policy and reward depend on the current state of the system, which evolves according to a transition probability kernel $P{:}\mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, and the agent value function is a cumulative reward with a discount factor γ.

We refer to [28] for the exact problem formulation and definition for natural gradients. We define the exact update for entropy-regularized NPG in Markov games as

$$\pi_i^{k+1}(a_i|s) \propto \pi_i^k(a_i|s)^{1-\eta\tau} \exp\left(\frac{\eta}{1-\gamma}\bar{A}_i^k(s,a_i)\right),$$

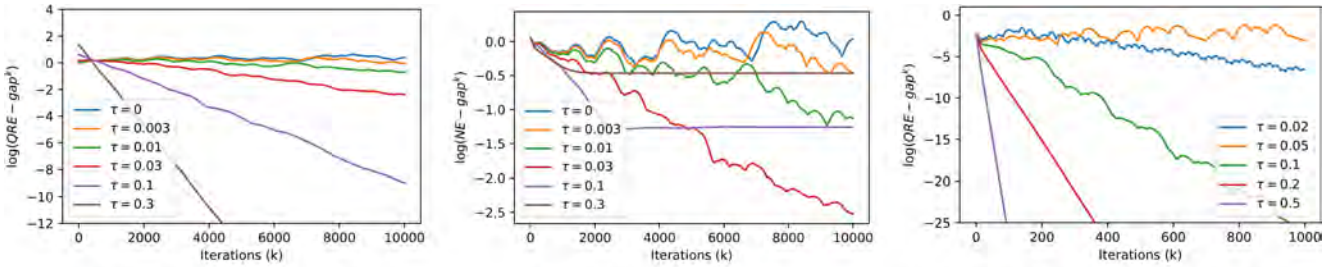
where $\bar{A}_i^k(s,a_i)$ denotes the marginalized advantage function defined therein. We consider a synthetic Markov game with agent number $n = 3$, individual action space $|\mathcal{A}_i| = 5$, and the total number of states is set to $|\mathcal{S}| = 5$.

We plot the log-scale results in Fig. 2(c). The figure shows that the convergence properties of Markov games closely resemble those of static games, suggesting that our theoretical results in Section III could potentially be extended to the stochastic setting.

V. Conclusion and Future Directions

In this letter, we have studied the independent NPG algorithm with entropy regularization under the static game setting. We have shown that the system converges to QRE under independent NPG updates, and the rate of convergence is linear. However, such convergence only occurs with a sufficiently large regularization. On the other hand, a system with inadequate regularization may fail to converge. Experimental results were provided for both cases across various settings.

There are still many open problems for policy gradient-based algorithms in games. A future direction is to extend the analytical results to the stochastic (Markov) game setting. Our preliminary experiments have suggested that stochastic games could enjoy similar convergence to static games. This topic may contribute to multi-agent reinforcement learning, where the system is generally assumed to be Markov. Another potential direction is to consider the scenario where oracle



(a) Convergence of *QRE-gap* for network zero-sum games.

(b) System *NE-gap* for network zero-sum games.

(c) System *QRE-gap* in synthetic Markov games with different regularization.

Fig. 2. Network zero-sum game and Markov game.

information is unavailable, and the policy gradient needs to be estimated via sampling. Lastly, our analysis can be extended to policy gradient-based algorithms such as safe MARL, robust MARL, and multi-objective MARL, following recent literature in single-agent RL [30], [31].

REFERENCES

[1] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "On the theory of policy gradient methods: Optimality, approximation, and distribution shift," *J. Mach. Learn. Res.*, vol. 22, no. 1, pp. 4431–4506, 2021.

[2] J. P. Bailey and G. Piliouras, "Multi-agent learning in network zero-sum games is a Hamiltonian system," *Momentum*, vol. 10, pp. 1–17, Mar. 2019.

[3] J. Bakakeu, S. Baer, H. Klos, J. Peschke, M. Brossog, and J. Franke, "Multi-agent reinforcement learning for the energy optimization of cyber-physical production systems," in *Artificial Intelligence in Industry 4.0: A Collection of Innovative Research Case-studies that are Reworking the Way We Look at Industry 4.0 Thanks to Artificial Intelligence*. Cham, Switzerland: Springer, 2021, pp. 143–163.

[4] J. Bhandari and D. Russo, "Global optimality guarantees for policy gradient methods," *Oper. Res.*, 2024. [Online]. Available: https://doi.org/10.1287/opre.2021.0014

[5] S. Cen, F. Chen, and Y. Chi, "Independent natural policy gradient methods for potential games: Finite-time global convergence with entropy regularization," in *Proc. IEEE 61st Conf. Decis. Control (CDC)*, 2022, pp. 2833–2838.

[6] S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi, "Fast global convergence of natural policy gradient methods with entropy regularization," *Oper. Res.*, vol. 70, no. 4, pp. 2563–2578, 2022.

[7] S. Cen, Y. Chi, S. S. Du, and L. Xiao, "Faster last-iterate convergence of policy optimization in zero-sum Markov games," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–42.

[8] S. Cen, Y. Wei, and Y. Chi, "Fast policy extragradient methods for competitive games with entropy regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 27952–27964.

[9] M. Cheng, R. Zhou, P. R. Kumar, and C. Tian, "Provable policy gradient methods for average-reward markov potential games," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2024, pp. 4699–4707.

[10] C. Daskalakis, D. J. Foster, and N. Golowich, "Independent policy gradient methods for competitive reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 5527–5540.

[11] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.

[12] W. Kim and Y. Sung, "An adaptive entropy-regularization framework for multi-agent reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 16829–16852.

[13] G. Lan, "Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes," *Math. Program.*, vol. 198, no. 1, pp. 1059–1106, 2023.

[14] Y. Liu, K. Zhang, T. Basar, and W. Yin, "An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 7624–7636.

[15] R. D. McKelvey and T. R. Palfrey, "Quantal response equilibria for normal form games," *Games Econ. Behav.*, vol. 10, no. 1, pp. 6–38, 1995.

[16] J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans, "On the global convergence rates of softmax policy gradient methods," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 6820–6829.

[17] J. Müller and G. Montúfar, "Geometry and convergence of natural policy gradient methods," *Inf. Geom.*, vol. 7, no. 1, pp. 485–523, 2024.

[18] J. F. Nash Jr., "Equilibrium points in *n*-person games," *Proc. Nat. Acad. Sci.*, vol. 36, no. 1, pp. 48–49, 1950.

[19] G. Sartoretti, Y. Wu, W. Paivine, T. Kumar, S. Koenig, and H. Choset, "Distributed reinforcement learning for multi-robot decentralized collective construction," in *Proc. 14th Int. Symp. Distrib. Auton. Robotic Syst.*, 2019, pp. 35–49.

[20] F. Schäfer and A. Anandkumar, "Competitive gradient descent," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[21] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "Safe, multi-agent, reinforcement learning for autonomous driving," 2016, *arXiv:1610.03295*.

[22] L. Shapley, "Some topics in two-person games," in *Adventures in Game Theory*, vol. 52. New York, NY, USA: American Elsevier Publ. Co., 1964, pp. 1–29.

[23] Y. Sun, T. Liu, P. R. Kumar, and S. Shahrampour, "Linear convergence of independent natural policy gradient in games with entropy regularization," 2024, *arXiv:2405.02769*.

[24] Y. Sun, T. Liu, R. Zhou, P. R. Kumar, and S. Shahrampour, "Provably fast convergence of independent natural policy gradient for Markov potential games," in *Proc. 37th Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 1–21.

[25] A. Trott, S. Srinivasa, D. van der Wal, S. Haneuse, and S. Zheng. "Building a foundation for data-driven, interpretable, and robust policy design using the AI economist." SSRN.com. 2021. [Online]. Available: http://dx.doi.org/10.2139/ssrn.3900237

[26] C. Wei, C. Lee, M. Zhang, and H. Luo, "Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive Markov games," in *Proc. Conf. Learn. Theory*, 2021, pp. 4259–4299.

[27] W. Zhan, S. Cen, B. Huang, Y. Chen, J. D. Lee, and Y. Chi, "Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence," *SIAM J. Optim.*, vol. 33, no. 2, pp. 1061–1091, 2023.

[28] R. Zhang, J. Mei, B. Dai, D. Schuurmans, and N. Li, "On the global convergence rates of decentralized softmax gradient play in Markov potential games," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 1923–1935.

[29] R. Zhang, Z. Ren, and N. Li, "Gradient play in stochastic games: Stationary points, convergence, and sample complexity," *IEEE Trans. Autom. Control*, early access, Apr. 10, 2024, doi: 10.1109/TAC.2024.3387208.

[30] R. Zhou, T. Liu, M. Cheng, D. Kalathil, P. R. Kumar, and C. Tian, "Natural actor-critic for robust reinforcement learning with function approximation," in *Proc. 37th Adv. Neural Inf. Process. Syst.*, 2023, pp. 1–37.

[31] R. Zhou, T. Liu, D. Kalathil, P. R. Kumar, and C. Tian, "Anchor-changing regularized natural policy gradient for multi-objective reinforcement learning," in *Proc. 36th Adv. Neural Inf. Process. Syst.*, 2022, pp. 1–13.